SGUIE-Net: Semantic Attention Guided Underwater Image Enhancement With Multi-Scale Perception

Qi Qi, *Student Member, IEEE*, Kunqian Li, *Member, IEEE*, Haiyong Zheng, *Senior Member, IEEE*, Xiang Gao, Guojia Hou, *Member, IEEE*, and Kun Sun, *Member, IEEE*

***Abstract*— Due to the wavelength-dependent light attenuation, refraction and scattering, underwater images usually suffer from color distortion and blurred details. However, due to the limited number of paired underwater images with undistorted images as reference, training deep enhancement models for diverse degradation types is quite difficult. To boost the performance of data-driven approaches, it is essential to establish more effective learning mechanisms that mine richer supervised information from limited training sample resources. In this paper, we propose a novel underwater image enhancement network, called SGUIE-Net, in which we introduce semantic information as high-level guidance via region-wise enhancement feature learning. Accordingly, we propose semantic region-wise enhancement module to better learn local enhancement features for semantic regions with multi-scale perception. After using them as complementary features and feeding them to the main branch, which extracts the global enhancement features on the original image scale, the fused features bring semantically consistent and visually superior enhancements. Extensive experiments on the publicly available datasets and our proposed dataset demonstrate the impressive performance of SGUIE-Net. The code and proposed dataset are available at https://trentqq.github.io/SGUIE-Net.html.**

***Index Terms*— Underwater image enhancement, deep learning, semantic guidance, attention mechanism, SUIM-E dataset.**

Manuscript received 8 January 2022; revised 31 July 2022; accepted 30 September 2022. Date of publication 26 October 2022; date of current version 31 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61906177 and Grant 62176242, in part by the Natural Science Foundation of Shandong Province under Grant ZR2019BF034, and in part by the Fundamental Research Funds for the Central Universities under Grant 201964013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marta Mrak. *(Corresponding author: Kunqian Li.)*

Qi Qi is with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China (e-mail: qiqi2013@stu.ouc.edu.cn).

Kunqian Li is with the College of Engineering, Ocean University of China, Qingdao 266100, China (e-mail: likunqian@ouc.edu.cn).

Haiyong Zheng is with the Intelligent Information Sensing and Processing Laboratory, College of Electronic Engineering, Ocean University of China, Qingdao 266100, China (e-mail: zhenghaiyong@ouc.edu.cn).

Xiang Gao is with the College of Engineering, Ocean University of China, Qingdao 266100, China, and also with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xgao@ouc.edu.cn).

Guojia Hou is with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, China (e-mail: hgjouc@126.com).

Kun Sun is with the School of Computer Science, China University of Geosciences, Wuhan 430078, China (e-mail: sunkun@cug.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2022.3216208, provided by the authors.

Digital Object Identifier 10.1109/TIP.2022.3216208

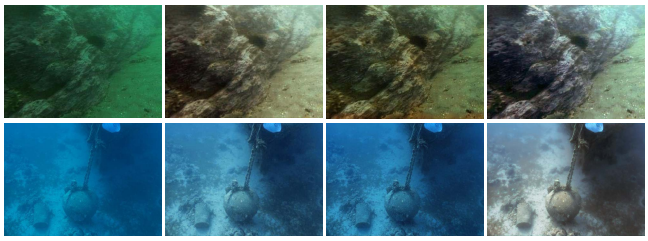
I. Introduction

UNMANNED underwater robots equipped with high-performance vision modules have been increasingly becoming a major tool for ocean exploration. However, the underwater imaging environment is quite complicated and it changes dynamically due to diverse disturbances, such as suspended particles, plankton, light differences, etc. Therefore, underwater images are often characterized by low contrast and blurred details. Moreover, underwater images usually suffer from severe color distortion due to the light wavelength-dependent attenuation. These obstacles pose significant challenges to vision-based underwater operations [1]. Recently, underwater image enhancement has received extensive attention to improve the visual quality of underwater images [2].

With the great success of deep learning in many areas of computer vision, it has also brought new strategies and perspectives to underwater image enhancement task [5]. However, it is almost impossible to simultaneously obtain both degraded and clear underwater images in situ, which makes it difficult to obtain a large number of paired samples to train a comprehensive and complete deep enhancement model suitable for diverse degradation types. As a compromise, researchers proposed to use synthetic underwater datasets [6], [7] or manually selected enhancements with optimal subjective visual quality as model training references [4], [8]. Unfortunately, for the former strategy, there is still a clear gap between synthetic underwater images and real underwater images in terms of verisimilitude and scene diversity. As to the manually selected enhancement references, their labelling are costly in terms of labor and they are not veritable ground-truth, which still carry clear human subjective visual preference and may introduce label contradiction for similar underwater scenarios.

Based on the above analysis, we realize that establishing more effective learning mechanisms is quite important for the mining of reliable supervisory information and robust enhancement. Besides, collecting information from more reliable labelled data as supplement is a promising strategy. The semantic map portrays the boundaries of semantic regions, which can be viewed as high-level associations across different images. In addition, as introduced in [9], explicitly harnessing the scene semantics into the enhancement process can reduce the introduction of artifacts.

In this paper, in order to leverage semantic segmentation maps to assist in underwater image enhancement task,



(a) input (b) Ucolor[3] (c) UICoE-Net[4] (d) SGUIE-Net

Fig. 1. Enhancement examples of deep learning-based methods on real underwater images. The proposed SGUIE-Net gives natural results with few over/under-enhancements or artifacts. Moreover, SGUIE-Net produces visually consistent enhancements for different images with the same semantics.

we propose an underwater image enhancement network with semantic attention guidance and multi-scale perception, which is named as Semantic attention Guided Underwater Image Enhancement Network (SGUIE-Net). Due to the diversity and complexity of underwater imaging degradation, only learning the model from global view for image-to-image enhancement is quite difficult. With the guidance of semantic map, we propose to learn region-wise enhancement features of different semantics separately and demonstrate that these region-wise enhancement features with semantic clues can be an effective complement to global-wise learning.

Figure 1 shows enhancement examples of deep models on real underwater images with severe visual degradation. Unlike other methods that produce obvious over/under-enhancements (Figure 1(b)(c)) or artifacts (first row in Figure 1(c)), our SGUIE-Net generates visually pleasing enhancements. More impressively, SGUIE-Net produces visually consistent enhancements for different images with the same semantics. The contributions of this paper are summarized as follows

(1) We propose a semantic attention guided underwater image enhancement network called SGUIE-Net. It learns enhancement features incorporating semantic clues to help recover those degradations that are uncommon in the training sample distribution but semantically relevant with the well-learned types.
(2) We design SGUIE-Net as a deep enhancement network with multi-scale perception, whose main branch for global-wise learning and semantic branch for region-wise learning are fused to complement each other. The main branch is used to provide end-to-end enhancement while preserving image texture details in original scale, and the semantic branch is used to complement the semantic attention guided features with multi-scale perception.
(3) We establish a new benchmark, namely SUIM-E, by extending the Segmentation of Underwater IMagery (SUIM) dataset [10] with corresponding enhancement reference images. Then, comprehensive experiments, evaluations and analyses conducted on multiple datasets verify the good performance of the proposed SGUIE-Net.

II. Related Works

A. Underwater Image Enhancement

1) Traditional Methods: Traditional underwater image enhancement methods, i.e., methods that do not employ deep learning techniques, can be roughly classified into two categories depending on whether or not a physical model is introduced. Physical model-free methods mainly aim to improve the visual quality of underwater images by directly adjusting the pixel values. The commonly used strategies include pixel distribution adjustment [11], [12], [13], Retinex-based approaches [14], [15], [16] and fusion-based methods [11], [17], [18]. However, without considering the underwater imaging mechanism, physical model-free approaches usually produce unnatural artifacts and over/under-enhancement.

Physical model-based approaches estimate the parameters of underwater imaging model with hand-crafted priors, which include red channel prior [19], [20], dark channel prior [21], [22], light attenuation prior [23], [24], haze-lines prior [25], [26], etc. However, diverse underwater imaging degradations make the accurate estimation of complicated model parameters difficult. Therefore, most of these methods are usually sensitive to the types of underwater imaging degradation.

2) Deep Learning-Based Methods: As deep learning has made great strides in many low-level vision tasks, more and more researchers have started to introduce deep learning into the field of underwater image enhancement [5]. In the last few years, researchers have made considerable effort on new sample generation strategies, more effective learning strategies and network architectures.

Thanks to the great success of image transfer techniques, Fabbri et al. [6] proposed to synthesize underwater images with CycleGAN [27]. Then, GAN-based deep models, such as FGAN [28], DenseGAN [29], MLFcGAN [30] and FUnIE-GAN [31], are widely explored. Later, Li et al. [8] constructed a real-world underwater image enhancement benchmark (UIEB), which contains 890 paired real-world underwater images. The paired enhancement references are manually selected from enhancement candidates according to subject visual preference. UIEB has greatly contributed to the reference-based performance evaluation [32] and the research of deep learning-based methods for underwater image enhancement, especially CNN-based methods [3], [4], [33]. However, despite the above strategies expanding the sources of paired training data, there is still a shortage of high-quality training samples that truly match real-world underwater scenarios and diverse degradation.

To make full use of the limited high-quality underwater training images, researchers also explored new learning strategies and network structures. With UIEB dataset, Li et al. [8] proposed a gated fusion network called Water-Net, which fuses the inputs with three predicted confidence maps to achieve the enhanced result. Qi et al. [4] proposed an Underwater Image Co-Enhancement Network (UICoE-Net) by introducing correlation feature matching units to communicate the mutual correlation of the two input branches. More recently, Li et al. [3] proposed to learn rich feature representations from diverse color spaces and attention weights by medium transmission map, and accordingly designed an encode-decoder enhancement network called Ucolor. While, the shortage of high-quality paired training samples has always

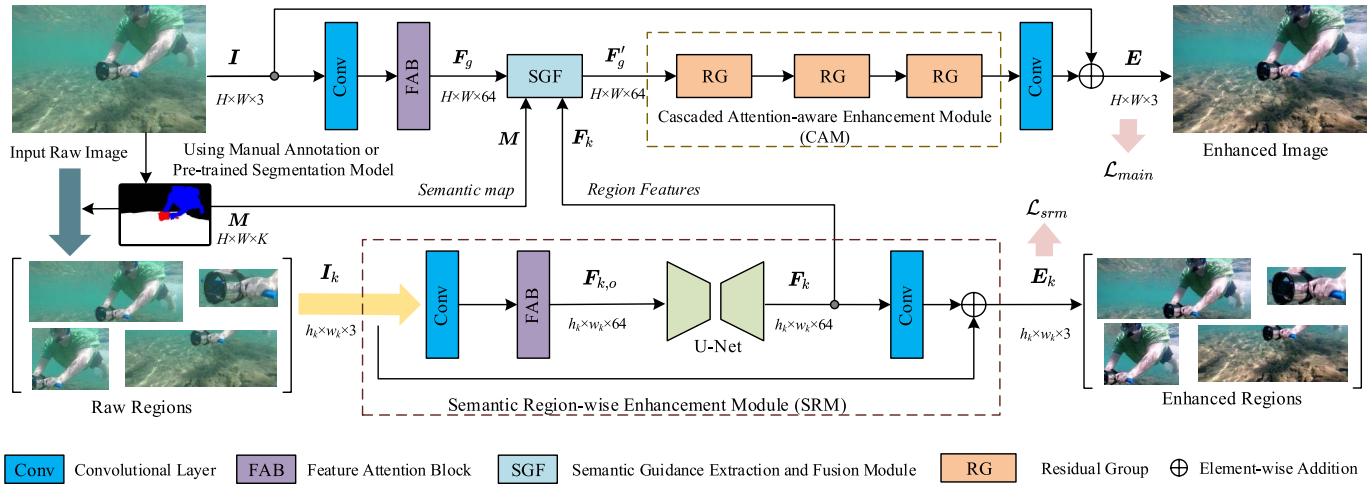


Fig. 2. Architecture of SGUIE-Net. Our SGUIE-Net contains two enhancement branches for multi-scale perception. The main (top) branch works on the original input scale through the cascaded attention-aware enhancement module (CAM) which contains three residual groups (RGs), and additionally embeds a semantic guided feature extraction and fusion module (SGF) to receive enhanced multi-scale features. The bottom branch learns multi-scale enhancement features with semantic attention using an encoder-decoder structure. It builds semantic enhancement guidance through the semantic region-wise enhancement module (SRM) and then feeds them back to the main branch. The details of the feature attention blocks (FAB), RG and SGF modules are shown in Figure 3.

constrained the performance of the above models. And the inherent ambiguity and non-consistency of the training labels with various quality may even further hinder the learning of robust enhancement models. Although some researchers have started to explore shared features among images for richer constraints, such as UICoE-Net, there is still a gap in mining high-level constraints for scene-consistent and robust enhancement.

B. Semantic Guidance in Computer Vision

In previous studies, semantic guidance has been widely explored in various computer vision tasks, such as visual navigation [34], [35], image generation [36], [37], image-image matching [38], [39], image-sentence matching [40], [41], Re-IDentification [42], [43], etc. Most of these tasks are high-level or mid-level tasks, where semantic information plays a guiding and error-correcting role. A common form is that the result of the associated task is constrained by semantic information, and the result violating the constraint will be penalized by an additional cost [37], [38], [39], [40], [41]. In another form, semantic information will be used as prior knowledge to provide guidance for subsequent tasks, thus providing additional constraints while reducing the search space [34], [35], [36], [42], [43].

In addition to the above tasks, we also notice that semantic information has played important roles in low-level image processing tasks, such as low-light image enhancement [9], [44], [45], image dehazing [46], [47], image inpainting [48] and underwater image enhancement [49]. As an early attempt in semantic image enhancement area, Lindner et al. [44] established a non-parametric statistical framework to link the image characteristics and semantic concepts. Then the semantic context is used to guide the image processing tasks, such as gray-level tone-mapping, emphasizing semantically relevant colors, and performing a defocus magnification for the given image. Xie et al. [9] directly utilized semantic mask as differentiated weights to refine rough low-light image enhancement from retinex model. When it comes to deep learning-based image processing frameworks, semantic features are usually represented as the high-level features from pre-trained models [48], [49], such as pre-trained VGG16 with ImageNet. In [48], Liu et al. proposed a deep generative model for image inpainting with a coherent semantic attention layer, which can help to preserve contextual structure and model the VGG feature-based semantic relevance between the content-missing holes. Within an encoder-decoder UIE framework, Shi et al. [49] introduced the multi-scale VGG16 features into different layers of the encoder to provide semantic clues for better enhancement. Another common form of semantic guidance is to directly introduce the semantic segmentation maps as features into the main feature branch [46], [47]. However, the above two strategies to provide semantic guidance for deep models have the similar concern, i.e., the features used for fusion are all only semantic features rather than task-oriented features with semantic clues. In order to provide more specific auxiliary information, a task-oriented feature fusion strategy with semantic guidance is urgently needed.

III. Proposed Method

To address the concerns mentioned in Section II, we propose a deep semantic attention-guided underwater image enhancement framework, i.e., SGUIE-Net. Specifically, we designed a semantic guided region-wise enhancement module to gather enhancement-oriented features with multi-scale semantic attention for the main enhancement branch. Additionally, the main branch works on the original input scale through the cascaded attention-aware enhancement module to recover rich local details. In the following subsections, overall network structure and key modules will be introduced in detail.

A. Overall Network With Multi-Scale Perception

The architecture of the proposed SGUIE-Net is shown in Figure 2. With semantic segmentation maps as high-level

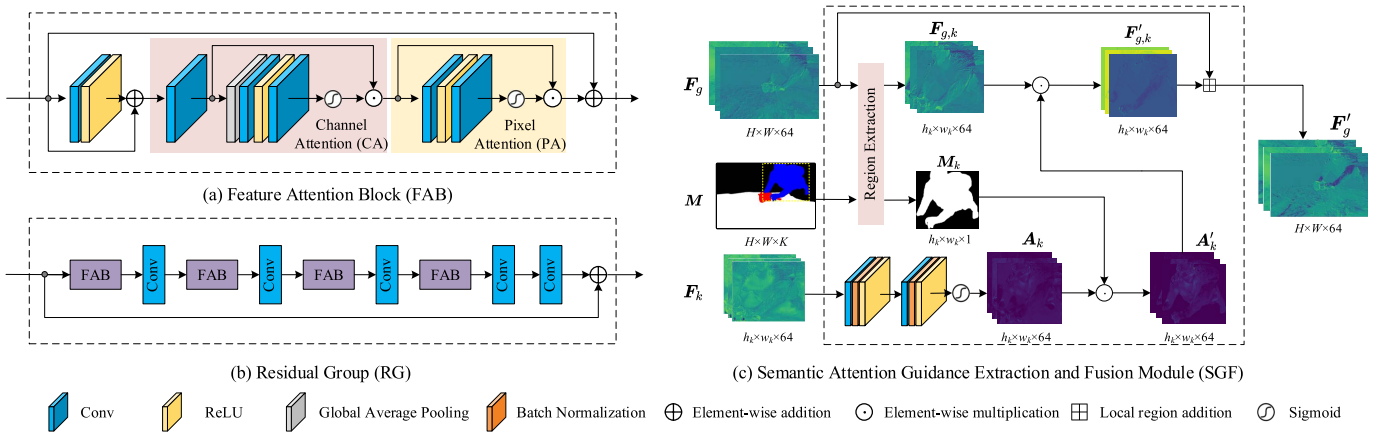


Fig. 3. Architectures of the basic blocks of our SGUIE-Net. (a) Feature attention block (FAB) consists of cascaded channel attention (CA) and pixel attention (PA) module. (b) illustrates the RG module formed by multiple FABs with shorted connections. (c) The SFG module takes the intermediate global enhancement residual feature $\boldsymbol{F}_g$, semantic segmentation mask $\boldsymbol{M}$ and the region enhancement residual feature $\boldsymbol{F}_k$ as the inputs. Then, the SFG module extracts the local parts of $\boldsymbol{F}_g$ according to the guidance of $\boldsymbol{M}$ for the later fusion with $\boldsymbol{F}_k$.

complementary information, it provides a new perspective to more effectively mine limited supervisory information for the training of UIE models. By manually performing mask annotation or using pre-trained underwater semantic segmentation models, such as SUIM-Net [10], we partition the input image into semantic regions. Then, we deliver them to the semantic region-wise enhancement module (SRM), which uses an encoder-decoder structure to extract multi-scale enhancement features with semantic attention. Subsequently, the semantic attention guidance extraction and fusion module (SGF) fuses these semantic-aware local features to the global features in original resolution according to the guidance of the semantic masks. Since the encoder-decoder structure prone to lose spatial detail features, the main branch (top branch in Figure 2) uses a cascaded attention-aware enhancement module (CAM), which operates at the original resolution of input image to provide differential enhancement for uneven degradation and maintain detailed texture. This dual-branch complementary structure provides multi-scale perception for the underwater image enhancement.

B. Cascaded Attention-Aware Enhancement Module

As discussed above, although the existing single-branch enhancement methods based on encoder-decoder architecture can well capture multi-scale features with multiple down-sampling operations, they also inevitably lose pixel-wise spatial details. To well preserve the local details of the input image for better recovery, we design a cascaded attention-aware enhancement module (CAM) as the basic structure of the main branch. The structure of CAM is shown in Figure 2. It is worth noting that CAM is designed to perceive the image features with the original resolution of the input image. Therefore, CAM does not contain any down-sampling or up-sampling operations, and the convolutional layers used in this module are all set with $kernel\_size = 3$ and $padding = 1$ to keep the feature resolution fixed during the process in the main branch and to maximally preserve spatial detail features.

CAM consists of three sets of cascaded residual groups (RG) with skip connections as shown in Figure 3(b). In each RG module, we stack four feature attention blocks (FAB) as shown in Figure 3(a) to learn higher-level features, and use long skip connections to solve the problem of training difficulties caused by the network being too deep. Inspired by [50], to further learn the differentiated attention that describe the uneven degradation of different pixels, FAB consists of a channel attention module (CA) and a pixel attention module (PA), which are placed in sequence for block residual learning.

C. Semantic Region-Wise Enhancement Module

Like most existing UIE networks, the main branch learns global feature transformation for whole image enhancement. However, an unavoidable concern is that only leaning image-to-image enhancement from global view is difficult due to the complexity and diversity of underwater imaging degradation. The scarceness of high-quality training samples brings more difficulty for learning comprehensive and robust models. To alleviate this dilemma and learn multi-scale feature representation, we raise a semantic region-wise enhancement module (SRM) for local enhancement feature learning, which can be an effective complement and guidance for the global feature of main branch. We use the semantic segmentation map to divide the whole image into semantically consistent regions for local enhancement model learning. In this way, we would like the region-wise enhancement module to learn robust local enhancement features with semantic attention.

Benefiting from the fact that CAM maintains feature transformation and learning on the original resolution of the input image, we can more comfortably propose a complementary enhancement module that mainly focus on local feature representation and multi-scale perception. With the following SGF module, these multi-scale and region-wise enhancement features with semantic consistency can be fused into the global features with original resolution. The above design enables the CAM to perform independent learning for different semantic regions within their own local context. After getting

rid of global learning for image-level degradation, the local degradation oriented enhancement feature learning is easier and is an effective supplement for global features. Moreover, SRM establishes a soft guidance for those degradation types that are uncommon in the training sample distribution but semantically relevant with the well-learned types.

The input of SRM is a collection of sub-images containing different semantic regions, which are obtained by splitting the original input image according to the semantic mask. The mask can be either manually annotated or generated with a pre-trained semantic segmentation model for underwater image, such as SUIM-Net [10]. For convenience, we denote the raw input image as $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$. Then, $\boldsymbol{I}_k \in \mathbb{R}^{h_k\times w_k\times 3}, h_k \in [0, H], w_k \in [0, W], k = 1, 2, \ldots, K$ represents the k-th semantic region of $\boldsymbol{I}$. K is the number of semantic categories contained in $\boldsymbol{I}$. In SRM, each input semantic region $\boldsymbol{I}_k$ is first passed through the convolutional layer and the FAB module to learn the residuals. This preliminary region enhancement residual feature $\boldsymbol{F}_{k,o} \in \mathbb{R}^{h_k\times w_k\times 64}$ can be formulated as follows

$$\boldsymbol{F}_{k,o} = \mathcal{C}_{FAB}\left(\mathcal{C}_{conv}\left(\boldsymbol{I}_k\right)\right), \tag{1}$$

where $\mathcal{C}_{FAB}$ represents the processing of the FAB module and $\mathcal{C}_{conv}$ denotes the convolutional operation. Since the U-Net-like structure has been widely applied in underwater enhancement missions due to its good performance in extracting multi-scale information [3], [4], we use the standard U-Net structure [51] for multi-scale feature transformation on the preliminary region residual features. Then, the region enhancement residual feature $\boldsymbol{F}_k \in \mathbb{R}^{h_k\times w_k\times 64}$ can be represented as

$$\boldsymbol{F}_k = \mathcal{C}_{U-Net}\left(\boldsymbol{F}_{k,o}\right), \tag{2}$$

where $\mathcal{C}_{U-Net}$ represents the U-Net-based feature transformation. With a convolutional layer for further channel squeeze, we get the enhancement residual. By adding the enhancement residual back to raw semantic region, the region-wise enhancement $\boldsymbol{E}_k$ can be acquired as follows

$$\boldsymbol{E}_k = \boldsymbol{I}_k \oplus C_{conv}\left(\boldsymbol{F}_k\right), \tag{3}$$

where $\oplus$ is element-wise addition. During the residual learning of semantic-guided region enhancement, $\boldsymbol{F}_k$ generated by U-Net module have contained rich residual features for enhancement with semantic attention and multi-scale perception. Therefore, we next feed these features into the main branch to inject semantic-guided enhancement residual information for high-level and cross-image enhancement alignment.

D. Semantic Attention Guidance Extraction and Fusion

In the proposed network, the CAM misses multi-scale information since it is designed to preserve pixel-wise spatial details and dependency at the original resolution, while the SRM acquires rich multi-scale semantic attention information for different semantic regions. To introduce the semantic attention guidance with region-wise enhancement features into the global enhancement branch, we design the semantic attention guidance extraction and fusion module (SGF). As shown in Figure 2, the SGF module has three inputs, i.e.,

- $\boldsymbol{F}_g \in \mathbb{R}^{H\times W\times 64}$, the intermediate global enhancement residual feature from main branch;
- $\boldsymbol{F}_k \in \mathbb{R}^{h_k\times w_k\times 64}$, the region enhancement residual feature of $\boldsymbol{I}_k$, which is extracted from SRM;
- $\boldsymbol{M} \in \mathbb{R}^{H\times W\times K}$, the semantic segmentation map for $\boldsymbol{I}$.

A processing example of SGF with visualized intermediate features is presented in Figure 3(c). For the regional residual enhancement feature $\boldsymbol{F}_k$ extracted from the SRM module, it first passes through two stacked Conv-BN-ReLU units and a sigmoid function. Then, we can receive regional semantic attention feature for enhancement task, i.e. $\boldsymbol{A}_k$, which is formulated as

$$\boldsymbol{A}_k = \sigma\left(\mathcal{C}_{2-conv}\left(\boldsymbol{F}_k\right)\right), \tag{4}$$

where $\mathcal{C}_{2-conv}$ denotes the two stacked Conv-BN-ReLU units and σ represents the sigmoid function. Note that, $\boldsymbol{A}_k$ covers the enclosing rectangle of the semantic object. To further extract pixel-wise semantic attention $\boldsymbol{A}'_k$ for the fusion on the main branch, we weight $\boldsymbol{A}_k$ with regional segmentation mask $\boldsymbol{M}_k \in \mathbb{R}^{h_k\times w_k\times 1}$, which is a clipped slice of $\boldsymbol{M}$, i.e.,

$$\boldsymbol{A}'_k = \boldsymbol{A}_k \odot \boldsymbol{M}_k, \tag{5}$$

where $\odot$ indicates element-wise multiplication. To smoothly incorporate the semantic attention into the global features $\boldsymbol{F}_g$ from main branch, we extract its corresponding part $\boldsymbol{F}_{g,k}$ in advance according to the location of the semantic region $\boldsymbol{I}_k$. Then, this cropped feature highlighted by pixel-wise semantic attention can be acquired as

$$\boldsymbol{F}'_{g,k} = \boldsymbol{F}_{g,k} \odot \boldsymbol{A}'_k. \tag{6}$$

Finally, the intermediate global enhancement residual features that incorporated with semantic attention guidance can be generated as

$$\boldsymbol{F}'_g = \boxplus\left(\boldsymbol{F}_g, \boldsymbol{F}'_{g,k}, x_k, y_k\right), \quad \forall k \in \{1, 2, \ldots, K\}, \tag{7}$$

where $\boxplus$ is defined as a local region addition between $\boldsymbol{F}'_{g,k}$ and $\boldsymbol{F}_g$, the coordinates x_k and y_k indicate the location on $\boldsymbol{F}_g$ where the top-left corner of $\boldsymbol{F}'_{g,k}$ should be placed. After all $\boldsymbol{F}'_{g,k}, k \in \{1, 2, \ldots, K\}$ have been incorporated, $\boldsymbol{F}'_g$ will be sent to the following CAM for further residual perception and final enhancement.

E. Loss Function

Inspired by [7] and [4], to better preserve the sharpness of edges and details in enhanced images, we use the ℓ_2 loss to simultaneously minimize the pixel-wise error of the enhancements from the main branch and SRM. Accordingly, the full training loss $\mathcal{L}$ for SGUIE-Net can be expressed as

$$\mathcal{L} = \mathcal{L}_{main} + \mathcal{L}_{srm}, \tag{8}$$

where $\mathcal{L}_{main}$ represents the pixel-wise ℓ_2 distance between the predicted enhancement $\boldsymbol{E}$ from the main branch and the enhancement reference $\boldsymbol{E}_{ref}$, $\mathcal{L}_{srm}$ denotes the pixel-wise ℓ_2 distance between the enhancement $\boldsymbol{E}_k$ from SRM for

TABLE I
EXPERIMENT SETTINGS FOR ALL COMPARED METHODS

Methods	Supervised/Unsupervised	Train Dataset	Test Dataset
ULAP [23]	Unsupervised / Physical model-based	n/a	SUIM-E / UIEB / UIEB Challenging / RUIE / EUVP / Color-Checker7
CBF [11]	Unsupervised / Physical model-free	n/a	SUIM-E / UIEB / UIEB Challenging / RUIE / EUVP / Color-Checker7
HUE [18]	Unsupervised / Physical model-free	n/a	SUIM-E / UIEB / UIEB Challenging / RUIE / EUVP / Color-Checker7
Water-Net [8]	Supervised	SUIM-E	SUIM-E / RUIE / EUVP / SQUID / Color-Checker7
		UIEB	UIEB / UIEB Challenging
Ucolor [3]	Supervised	SUIM-E	SUIM-E / RUIE / EUVP / SQUID / Color-Checker7
		UIEB	UIEB / UIEB Challenging
UICoE-Net [4]	Supervised	SUIM-E	SUIM-E / RUIE / EUVP / SQUID / Color-Checker7
		UIEB	UIEB / UIEB Challenging
SGUIE-Net	Supervised	SUIM-E	SUIM-E / RUIE / EUVP / SQUID / Color-Checker7
		UIEB	UIEB / UIEB Challenging

semantic region $\boldsymbol{I}_k$ and the region-wise enhancement reference $\boldsymbol{E}_{k,ref}$ clipped from $\boldsymbol{E}_{ref}$. $\mathcal{L}_{main}$ and $\mathcal{L}_{srm}$ are formulated as

$$\mathcal{L}_{main} = \sum_{i=1}^{H}\sum_{j=1}^{W}\left(\boldsymbol{E}(i,j) - \boldsymbol{E}_{ref}(i,j)\right)^2, \tag{9}$$

and

$$\mathcal{L}_{srm} = \sum_{k=1}^{K}\sum_{i=1}^{h_k}\sum_{j=1}^{w_k}\left(\boldsymbol{E}_k(i,j) - \boldsymbol{E}_{k,ref}(i,j)\right)^2, \tag{10}$$

respectively, where i and j are the pixel location indexes. SGUIE-Net is trained in an end-to-end manner, and the prediction of its main branch will be taken as the final enhancement.

IV. EXPERIMENTS

In this part, we first introduce the implementation details about the proposed SGUIE-Net in Section IV-A. Then, in Section IV-B, experiment settings including compared methods, datasets and evaluation metrics are introduced. In Sections IV-C and IV-D, comprehensive comparisons from the perspectives of visual quality improvement and color restoration performance are given, respectively. In Section IV-E, a series of ablation studies are conducted. Finally, we give a discussion on the limitation and prospect for future research. For more visual comparisons, please refer to the supplement materials.

A. Implementation Details

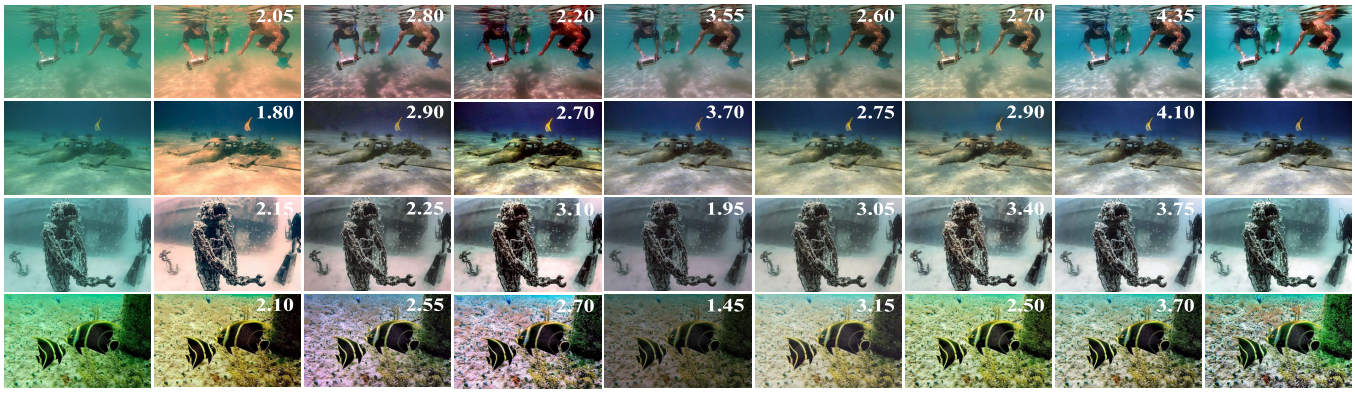
We implemented the proposed SGUIE-Net on PyTorch platform. During the training, the filter weights of each layer were initialized with Kaiming initialization. We used the Adam for network optimization and the initial learning rate was set to 1e-4 with the scheduler of linearly decaying to zero. Besides, batch size is set to 1. We make the input image size to 256×256 and perform data augmentation by random cropping and flipping. The experiments are conducted on a PC with an NVIDIA RTX 3090 GPU, a 2.3GHz Intel Xeon processor, 128GB RAM and Ubuntu 18.04 operation system.

B. Experiment Settings

1) Datasets: To the best of our knowledge, there is no underwater dataset that contains both corresponding enhancement reference and semantic segmentation map as ground truth. To verify the performance of our method and to facilitate the future research, we constructed the **SUIM-E** dataset by supplementing the underwater images of SUIM dataset [10] with corresponding enhancement references in a hand-picked manner similar to [8]. In detail, the SUIM-E dataset contains a total of 1635 real underwater images with pixel annotations for eight categories: background, fish/vertebrates, reefs/invertebrates, aquatic plant/seagrass, wrecks/ruins, human divers, robots and sea-floor/rocks. We used 1525 of these images for training/validation and the remaining 110 images for testing. Due to the limited space, more details about the construction methodology, statistical information and quality validation of SUIM-E dataset can be found in the supplementary material. SUIM-E dataset has been made publicly available at https://github.com/trentqq/SUIM-E.

To verify the performance of training SGUIE-Net without manually labelled segmentation maps, we also conducted model training and evaluation on the **UIEB** [8] dataset. UIEB includes 890 underwater images, which only have enhancement references. Pre-trained SUIM-Net [10] was adopted to predict semantic segmentation maps for both training and test. We randomly selected 800 images for training/validation. The rest 90 images were used for testing and evaluation.

In addition, we verified the generalizability and color correction accuracy of our method on the **UIEB Challenging** [8], **RUIE** [52], **EUVP** [31], **SQUID** [26] and **Color-Checker7** [11] datasets. UIEB Challenging set contains 60 real-world underwater images without enhancement references. RUIE contains a total of 4230 real-world underwater images, consisting of three subsets, namely UIQS, UCCS and UHTS, all of which were used for the test in our experiment. EUVP contains more than 12K paired and 8K unpaired underwater images with diverse scenes. In our experiment, we used all the officially specified 2300 underwater images



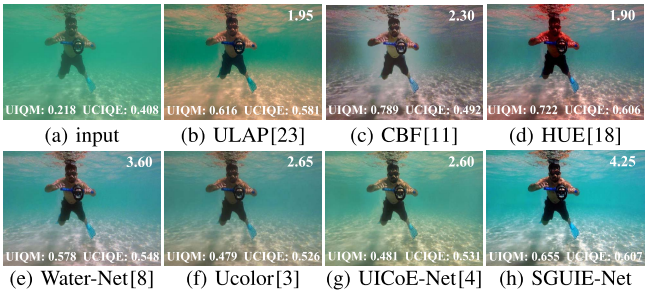
(a) input (b) ULAP[23] (c) CBF[11] (d) HUE[18] (e) Water-Net[8] (f) Ucolor[3] (g) UICoE-Net[4] (h) SGUIE-Net (i) Reference

Fig. 4. Visual comparisons on underwater images from SUIM-E test set. The perceptual score is marked on the upper right corner of each enhancement.

TABLE II

QUANTITATIVE COMPARISON ON SUIM-E TEST SET IN TERMS OF AVERAGE PSNR, SSIM, UIQM, UCIQE AND PERCEPTUAL SCORES (PS). TRADITIONAL AND DEEP LEARNING-BASED METHODS ARE SEPARATED WITH A LINE. TOP THREE SCORES ARE IN RED, GREEN AND BLUE, RESPECTIVELY. (BEST VIEWED IN COLOR)

Method	PSNR↑	SSIM↑	UIQM↑	UCIQE↑	PS↑
ULAP [23]	16.561	0.769	0.616	0.614	2.142
CBF [11]	16.465	0.828	0.836	0.571	2.397
HUE [18]	17.970	0.822	1.022	0.651	2.434
Water-Net [8]	17.019	0.820	0.415	0.548	2.251
Ucolor [3]	21.063	0.844	0.500	0.577	2.892
UICoE-Net [4]	21.752	0.910	0.668	0.588	3.219
SGUIE-Net	24.820	0.928	0.703	0.615	3.665



(a) input (b) ULAP[23] (c) CBF[11] (d) HUE[18] (e) Water-Net[8] (f) Ucolor[3] (g) UICoE-Net[4] (h) SGUIE-Net

Fig. 5. Example of no-reference metric-based and subjective quality evaluation. The UIQM, UCIQE and perceptual scores are marked on the bottom and upper right corner of each enhancement, respectively.

for evaluation. The SQUID dataset contains 57 stereo image pairs taken from four different sites. In these scenes, multiple color charts are placed as the reference for color correction. We used 57 images from the right camera for testing, which contain the position information of the color charts. Color-Checker7 contains 7 underwater Color Checker images taken with 7 different professional cameras, which is used to evaluate the robustness and accuracy of underwater color correction.

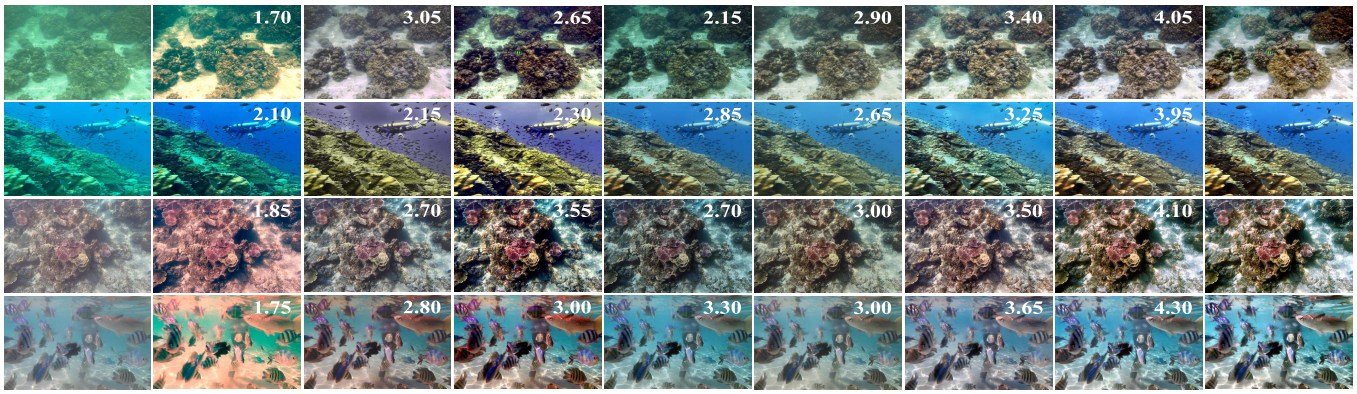
2) Compared Methods: We compared our SGUIE-Net with six typical methods proposed recently, including one physical model-based method (ULAP [23]), two physical model-free methods (CBF [11] and HUE [18]) and three deep learning-based methods (Water-Net [8], Ucolor [3], UICoE-Net [4]). Since the source code of the CBF [11] is not publicly available, as a compromise, we used the code[1] implemented by other researchers. For the other compared methods, we reproduced the results using the code published by their authors.

For clarity, the experiment settings of all compared methods are summarized in Table I. All deep learning-based methods use the same training and test sets as our proposed method. We trained two models for each of them on the SUIM-E dataset and UIEB dataset, respectively. In particular, tests on the SUIM-E, RUIE, EUVP, SQUID and Color-Checker7 datasets use the models trained on SUIM-E dataset. The models trained on the UIEB dataset are used on the test set of UIEB and UIEB Challenging set. The semantic segmentation maps for all test datasets were generated using the pre-trained SUIM-Net, except for the SUIM-E dataset.

3) Evaluation Metrics: For image quality comparison, we adopted multiple widely used metrics for comprehensive evaluation, such as full-reference quality metrics including peak signal-to-noise ratio (PSNR) and structural similarity index metric (SSIM), no-reference quality metrics including underwater image quality measure (UIQM) [53] and underwater color image quality evaluation (UCIQE) [54] metric. On SQUID dataset, to quantitatively evaluate the performance of color restoration, we calculated the average reproduction angular error $\overline{\psi}$ between the enhanced gray-scale patches and their corresponding pure gray colors in RGB space. The evaluation code provided by Berman et al. [26] was adopted. For Color-Checker7 dataset, we followed Ancuti et al. [11] to employ CIEDE2000 [55] to measure the relative differences between the corresponding color patches of ground-truth Macbeth Color Checker and the enhancement results.

Besides, we invited 20 volunteers with basic knowledge of image processing to independently evaluate the perceptual quality scores (PS) of the enhanced images. We followed the setting of [3] that the perceptual quality is scored on a scale from 1 to 5 (from the worst to the best quality) and

[1]https://github.com/fergaletto/Color-Balance-and-fusion-for-underwater-image-enhancement



(a) input (b) ULAP[23] (c) CBF[11] (d) HUE[18] (e) Water-Net[8] (f) Ucolor[3] (g) UICoE-Net[4] (h) SGUIE-Net (i) Reference

Fig. 6. Visual comparisons on underwater images from UIEB test set. The perceptual score is marked on the upper right corner of each enhancement.

volunteers scored the results based on the following evaluation criteria: 1) whether the enhanced image introduces other color deviation; 2) whether the enhanced image contains artifacts; 3) whether the enhanced image looks like natural image; and 4) whether the enhanced image has improved contrast and better visibility. The enhancements on the same image of different methods are anonymous to volunteers, which are displayed in a random order on the same screen. There is no time limit on the scoring process for each image so that volunteers can give their evaluation after sufficient observation and comparison. We performed above subjective evaluation on all the test images of SUIM-E, UIEB, UIEB Challenging and SQUID datasets. As for RUIE and EUVP, considering the huge labor costs and the fact that there are thousands of test images, we only selected a representative fraction of images for the evaluation of perception scores. Specifically, we sampled 10 images proportionally from each subfolder of the RUIE dataset, resulting in 130 images. In addition, we took the first 30 images from each subfolder of the EUVP dataset separately, resulting in 120 images.

C. Comparison of Visual Quality Improvement

To give a comprehensive comparison of visual enhancement performance, we first conducted qualitative and quantitative comparison on the SUIM-E, UIEB, UIEB Challenging, RUIE and EUVP datasets.

Figure 4 shows the results of different methods on the SUIM-E dataset. The traditional methods, such as ULAP [23], CBF [11] and HUE [18], usually cause color deviation due to the introduction of excessive red components. Since the Water-Net method introduces a white balance channel during the enhancement process, which is not always reliable for underwater images, this may lead to color bias as well. The enhancement results of the Ucolor [3] achieve a better visual quality from the perspective of color correction. However, as shown in the first three rows of the examples, Ucolor does not improve the low contrast and blurred details very well. UICoE-Net [4] requires paired images matching similar scenes for correlation feature learning. Therefore, given that the SUIM-E dataset does not have any scene classification prior, it may lead to unstable performance of UICoE-Net.

TABLE III

QUANTITATIVE COMPARISON ON UIEB TEST SET IN TERMS OF AVERAGE PSNR, SSIM, UIQM, UCIQE AND PS SCORES. TRADITIONAL AND DEEP LEARNING-BASED METHODS ARE SEPARATED WITH A LINE. TOP THREE SCORES ARE IN RED, GREEN AND BLUE, RESPECTIVELY. (BEST VIEWED IN COLOR)

Method	PSNR↑	SSIM↑	UIQM↑	UCIQE↑	PS↑
ULAP [23]	16.786	0.776	0.742	0.598	2.157
CBF [11]	17.678	0.849	1.028	0.551	2.753
HUE [18]	18.630	0.834	1.260	0.648	2.794
Water-Net [8]	19.298	0.873	0.501	0.563	2.547
Ucolor [3]	20.742	0.847	0.802	0.548	2.696
UICoE-Net [4]	20.070	0.875	0.838	0.585	3.106
SGUIE-Net	24.074	0.908	0.851	0.601	3.571

On the contrary, our SGUIE-Net provides stable color correction for severe color deviation and does not introduce other artifacts. Our method is able to naturally improve the contrast while enhancing it in low-light underwater scenes. In particular, as shown in the third row of Figure 4, the enhancement of the white sands by our method is more complete compared to other compared methods, and the visual quality even outperforms the given enhancement reference. The comparison shows that our method achieves a better balance between visual quality and color correction. Thanks to the cascade attention-aware enhancement module, our method can effectively preserve and recover the texture details of the degraded images. Moreover, with the high-level semantic guidance from SRM, our method can apply more consistent and robust enhancements to the same semantic regions, such as the white-sand sea-floor in the top three rows.

The quantitative numerical results for the SUIM-E dataset are shown in Table II. By comparing the results of all metrics in combination, we can observe that our method achieves better results overall. Specifically, compared to the second best method, our method improves the PSNR by 14.1% (from 21.752 to 24.820). For the SSIM metric, UICoE-Net scores 0.910 since it introduces correlation feature matching units for better local detail recovery. While, our SGUIE-Net achieved a better SSIM score of 0.928, which further demonstrates the superiority of our method in maintaining and recovering

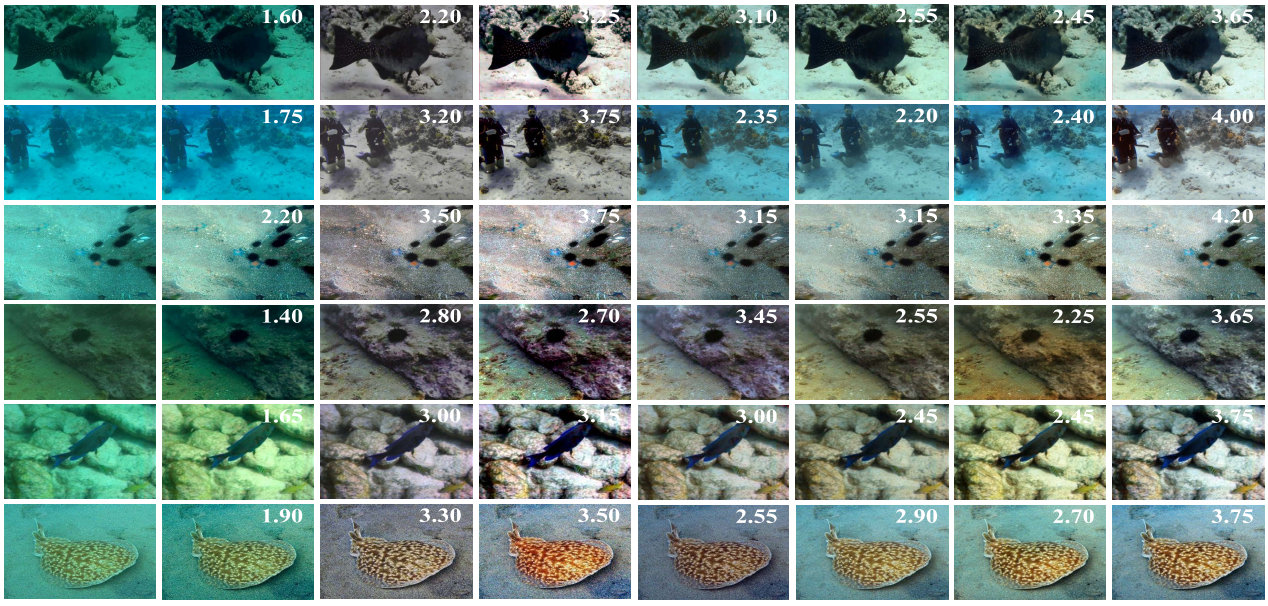


Fig. 7. Visual comparisons on underwater images from UIEB Challenging set (the top two rows), RUIE (the middle two rows) and EUVP (the bottom two rows) datasets. The perceptual score is marked on the upper right corner of each enhancement.

image local details. In addition, our method scores in the top three on both UIQM and UCIQE metrics, and ranks first among all deep learning-based methods in the comparison. Since the HUE method contains operations for color saturation and contrast enhancement, and the CBF method contains processes for white balance as well as edge sharpening, their enhancements are naturally preferred by UIQM and UCIQE. Thus, it also leads to the fact that these two methods can achieve relatively higher scores on these two metrics. However, as shown in Figure 5, despite their high UIQM and UCIQE scores, the visual quality of the enhancements generated by HUE and CBF is not satisfactory enough since more severe and unnatural color distortion was introduced. The perceptual scores given by 20 volunteers also show that the enhancements of SGUIE-Net were more highly rated and generally accepted.

To further verify the effectiveness and robustness of our method when only incomplete segmentation maps are available for training and test, we conducted experiments on UIEB dataset. The visual comparisons are shown in Figure 6. Unlike the compared methods which usually introduce over-saturation and artifacts or fail to correct the deviated colors, our method can handle enhancement properly in scenes with severe backscatter and is able to perform color correction while maintaining the details of the input image. As to quantitative comparison presented in Table III, our SGUIE-Net achieves the best results on full reference metric-based and subjective evaluations, which is consistent with the qualitative analysis given above. Besides, SGUIE-Net also performs the best among all the deep learning-based approaches in terms of no-reference quality metrics. Note that we directly use the pre-trained SUIM-Net without any fine-tuning, which results in some semantic information obtained by our network may not be completely correct. However, even with the incomplete semantic information, SGUIE-Net achieves the most obvious improvement in image visual quality. It proves that the training process of SGUIE-Net is able to learn effective guidance from incomplete semantic segmentation maps. More ablation experiments and discussion on the role of semantic guidance can be found in Section IV-E.

More visual comparisons on UIEB Challenging, RUIE and EUVP datasets are given in Figure 7. The presented real-world underwater images carry diverse visual degradation. In addition to different degrees of color distortion and blur appearance, the illumination in different images also varies. However, we can notice that these images share a lot of common scene elements, such as the sea-floor (sand and rocks). For deep learning-based approaches, such as Water-Net, Ucolor and UICoE-Net, the enhancements of the same method on the same scene elements usually show obvious inconsistency. This inconsistency manifests itself in varying degrees of color correction, contrast enhancement and detail improvement. In contrast, our proposed SGUIE-Net produces more consistent enhancements for the same scene elements across different images while keeping their own visual characteristics. With the enhancement of SGUIE-Net, even the illumination of different images is appropriately adjusted to a relatively uniform range for visual pleasure. The results of quantitative evaluation are presented in Table IV. Benefiting from the high-level guidance of semantic information, our method is able to maintain a stable enhancement performance for challenging underwater images and thus achieves the highest perceptual scores. In addition, our method achieves the highest UIQM and UCIQE scores compared to other deep learning-based methods.

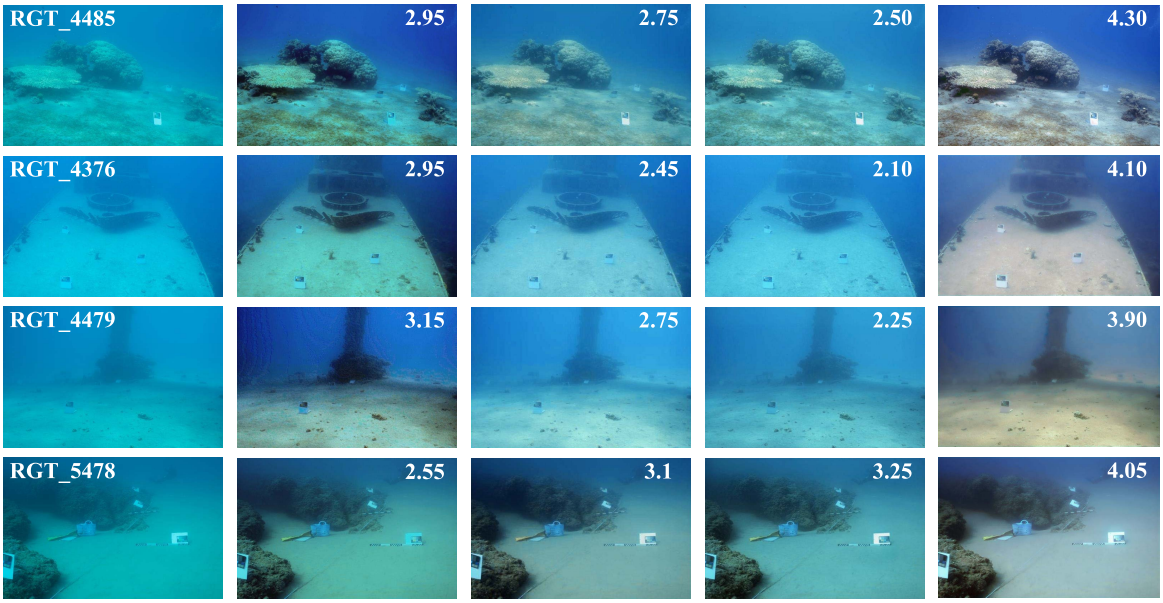
D. Comparison of Color Restoration Performance

To analyse the robustness and accuracy of color restoration, we further conducted the comparisons on SQUID and Color-

TABLE IV

NO-REFERENCE IMAGE QUALITY ASSESSMENT IN TERMS OF UIQM, UCIQE AND PS ON UIEB CHALLENGING SET, RUIE DATASET AND EUVP DATASET. TRADITIONAL AND DEEP LEARNING-BASED METHODS ARE SEPARATED WITH A LINE. TOP THREE SCORES ARE IN RED, GREEN AND BLUE, RESPECTIVELY. (BEST VIEWED IN COLOR)

Method	UIEB Challenging			RUIE			EUVP		
	PS↑	UIQM↑	UCIQE↑	PS↑	UIQM↑	UCIQE↑	PS↑	UIQM↑	UCIQE↑
ULAP [23]	1.904	0.368	0.540	1.703	0.162	0.490	2.191	0.954	0.600
CBF [11]	2.753	0.802	0.508	2.984	0.773	0.489	2.887	1.156	0.562
HUE [18]	2.792	0.873	0.600	2.995	1.078	0.608	2.893	1.443	0.638
Water-Net [8]	2.826	0.248	0.549	2.841	0.590	0.526	2.228	0.709	0.534
Ucolor [3]	2.790	0.265	0.520	2.789	0.621	0.526	2.769	0.814	0.561
UICoE-Net [4]	2.822	0.312	0.539	2.436	0.517	0.518	2.995	0.910	0.586
SGUIE-Net	3.340	0.524	0.578	3.553	0.688	0.556	3.483	1.007	0.603



(a) input (b) Water-Net [8] (c) Ucolor [3] (d) UICoE-Net [4] (e) SGUIE-Net

Fig. 8. Visual comparisons on challenging underwater images from SQUID. The perceptual score is marked on the upper right corner of each enhancement.

Checker7 datasets. For all the deep learning-based approaches, we used their models trained on SUIM-E dataset for testing.

In Figure 8, we present the enhancement examples of deep learning-based methods on images from SQUID dataset. The input underwater images photographed from different water depths (from 3-6 meters to 20-30 meters) present diverse tough problems for all compared methods. The Water-Net and UICoE-Net remove the haze better, but fail to recover the color. Ucolor slightly improves the color appearance in some cases but fails to deblur and recover the details. Our SGUIE-Net achieves impressive visual enhancement with the guidance of region-wise enhancement features from the proposed SRM module for the recovery of tough degradations. The presented enhancements show that our SGUIE-Net not only successfully performs color restoration without introducing artifacts but also significantly improves the local details and global contrast of the input underwater images.

To quantitatively measure the accuracy of color restoration, we evaluate the performance of deep learning-based methods in terms of average reproduction angular error $\overline{\psi}$. Figure 9 shows the $\overline{\psi}$ scores for each color card in Figure 8,

TABLE V

AVERAGE ANGULAR REPRODUCTION ERROR ($\overline{\psi}$), PS, UIQM AND UCIQE ON SQUID DATASET. THE BEST SCORES ARE IN RED. (BEST VIEWED IN COLOR)

Method	$\overline{\psi}\downarrow$	PS↑	UIQM↑	UCIQE↑
input	34.414	—	0.052	0.399
Water-Net [8]	20.105	2.680	0.124	0.528
Ucolor [3]	20.771	2.883	0.153	0.498
UICoE-Net [4]	20.788	2.784	0.254	0.482
SGUIE-Net	14.675	3.836	0.271	0.553

demonstrating that our results not only provide a pleasing visual perception, but also receive quite good color restoration accuracy. Full average evaluation results on all the 57 test images are shown in Table V. Our SGUIE-Net achieves the lowest mean $\overline{\psi}$ on all test images. In addition, our method improves by more than 1 point in terms of PS score compared to the second best performance, which is consistent with qualitative analysis of the visual comparison.

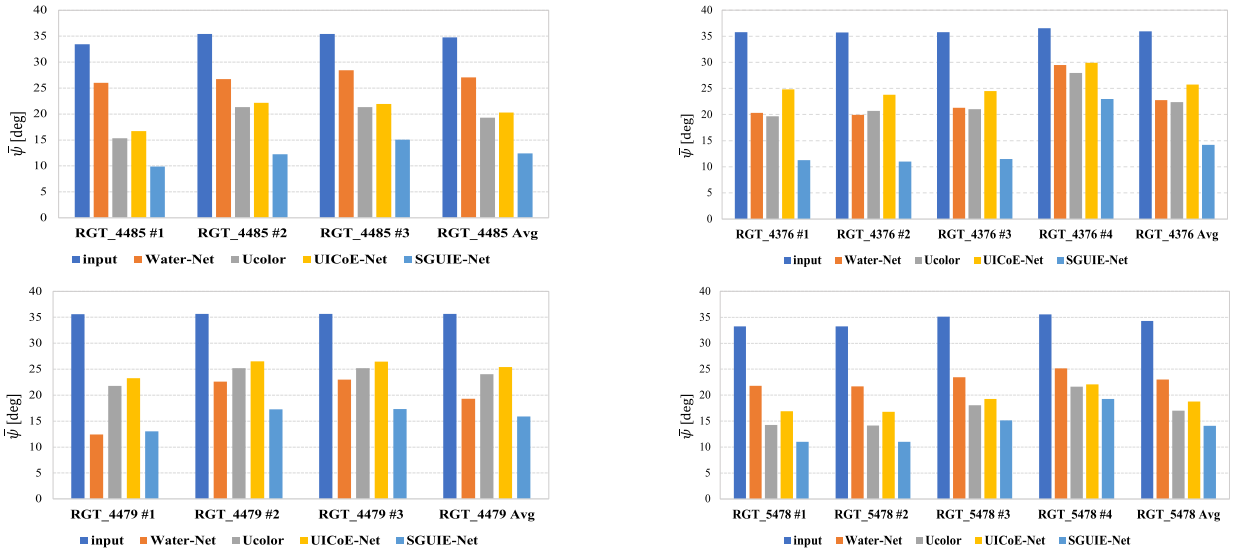


Fig. 9. Comparison of average reproduction angular error $\overline{\psi}$ between the gray-scale patches and the pure gray colors. The sub-bargraph reports the average performance of each comparison deep learning method on all color charts in test image. The average performance of each approach is reported on the right.

(a) input (b) Water-Net [8] (c) Ucolor [3] (d) UICoE-Net [4] (e) SGUIE-Net

Fig. 10. Visual comparisons on images of Color-Checker7. The two images are taken by the Fuji Z33 (first row) and Olympus T6000 (second row), respectively.

The Color-Checker7 dataset contains 7 underwater images taken from a shallow swimming pool with different cameras. Color checker is also photographed in each image. It provides a good path to demonstrate the robustness of our method to different imaging devices and the accuracy of color restoration. As shown in Figure 10, the professional underwater cameras (Fuji Z33 and Olympus T6000) also inevitably introduce various color casts. Although Water-Net removes the bluish distortion of the input images, it still leaves the problems of low contrast and excessive reddish compensation. Our SGUIE-Net, UICoE-Net and Ucolor perform better than Water-Net and do not introduce obvious artificial colors. In Table VI, we report the CIEDE2000 scores of different methods in each image and give the average score subsequently. Our SGUIE-Net achieves the best performance on four cameras, and receives very competitive scores on the other three cameras, resulting in the best average score in the end. It demonstrates that our method can effectively perform color correction for underwater images taken by different cameras with state-of-the-art restoration accuracy and good robustness.

E. Ablation Study

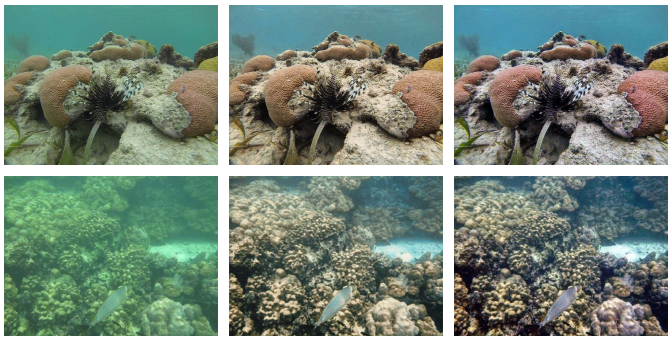
To demonstrate the effectiveness of the core components in our network, we conduct a series of ablation studies involving the following experiments:

- **-w/o-CAM**: without cascaded attention-aware enhancement module;
- **-w/o-SRM**: without semantic region-wise enhancement module;
- **-w/o-UNet**: using SRM without U-Net block;
- **-w/o-SS-train**: using SRM without semantic-based region split, which is replaced with random region split in the training stage;
- **-w/o-SS-test**: using SRM with semantic-based region split in the training stage, but with random region split in the test stage;

TABLE VI

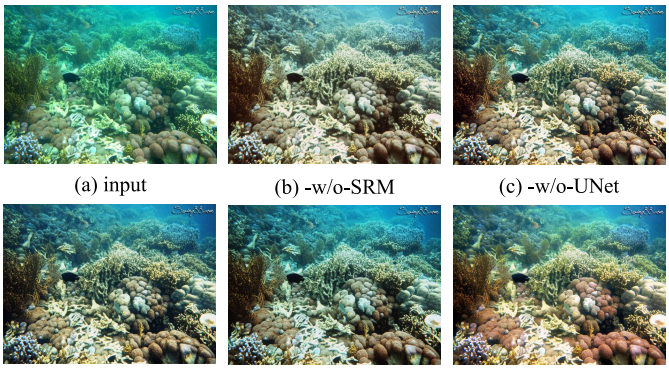
COLOR DISSIMILARITY COMPARISONS OF DIFFERENT METHODS ON COLOR-CHECK7 IN TERMS OF THE CIEDE2000. TRADITIONAL METHODS AND DEEP LEARNING-BASED METHODS TRAINED WITH PAIRED REFERENCE IMAGES ARE SEPARATED WITH A LINE. THE BEST SCORES ARE IN RED. (BEST VIEWED IN COLOR)

Method	Can D10	Fuji Z33	Oly T6000	Oly T8000	Pan TS1	Pen W60	Pen W80	Avg
input	12.910	16.648	14.990	19.301	16.152	11.966	14.123	15.156
ULAP [23]	16.788	11.671	11.508	17.176	23.469	17.729	18.547	16.698
CBF [11]	9.447	10.249	9.656	12.746	9.815	9.664	11.458	10.434
HUE [18]	14.214	12.863	12.943	13.785	12.496	14.287	14.003	13.513
Water-Net [8]	13.920	18.779	12.132	16.889	18.971	11.851	17.630	15.739
Ucolor [3]	10.424	10.706	8.758	11.668	10.587	10.110	10.678	10.419
UICoE-Net [4]	10.023	13.085	13.291	11.846	11.438	10.416	10.159	11.465
SGUIE-Net	10.526	10.227	8.654	11.292	9.572	10.999	11.504	10.396



(a) input (b) -w/o-CAM (c) SGUIE-Net

Fig. 11. Ablation study on the contribution of the cascaded attention-aware enhancement module. With CAM, SGUIE-Net recovers more local details.



(a) input (b) -w/o-SRM (c) -w/o-UNet

(d) -w/o-SS-train (e) -w/o-SS-test (f) SGUIE-Net

Fig. 12. Ablation study on the contribution of the semantic region-wise enhancement module. Full SGUIE-Net model achieves better performance in color and detail recovery.

- **-w/o-FAB**: without feature attention block;
- **-w/o-CA**: using FAB without channel attention;
- **-w/o-PA**: using FAB without pixel attention.

The PSNR and SSIM scores on SUIM-E and UIEB datasets are given in Table VII. On both two datasets, our full model achieves the best performance compared to all the ablated models, which also proves the effectiveness of the CAM, SRM and FAB modules.

1) Ablation Study on CAM: The CAM module is designed without down-sampling or up-sampling operations to preserve the detailed texture appearance of the input image and thus obtain a better visual quality. Figure 11 shows the visual comparison between the ablated model without CAM and the full SGUIE-Net, demonstrating the important role of the CAM module in recovering detailed information.

2) Ablation Study on SRM: The SRM module is the key component of our network, through which our network can capture region-wise enhancement features with multi-scale perception to build high-level semantic guidance. Compared with SGUIE-Net-w/o-SRM, our full model improves SSIM and PSNR scores with nearly 4% and 13%, respectively. Besides, by removing the U-Net structure from the SRM, SGUIE-Net-w/o-UNet will loss the multi-scale perception capability, which also degrades the enhancement performance according to the evaluation in Table VII and the example in Figure 12(c).

TABLE VII

QUANTITATIVE RESULTS OF ABLATION STUDY ON NETWORK STRUCTURE

Modules	Ablations	SUIM-E		UIEB	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑
CAM	-w/o-CAM	22.000	0.893	21.200	0.863
SRM	-w/o-SRM	21.940	0.897	19.910	0.852
	-w/o-UNet	22.587	0.908	21.357	0.877
	-w/o-SS-train	22.841	0.907	21.175	0.874
	-w/o-SS-test	21.026	0.879	20.496	0.871
FAB	-w/o-FAB	22.887	0.902	20.524	0.865
	-w/o-CA	23.175	0.902	21.824	0.871
	-w/o-PA	23.356	0.914	22.164	0.885
full	—	24.820	0.928	24.074	0.908

Moreover, to fully verify the effectiveness of using semantic segmentation map as enhancement guidance, we test the performance of two ablated models by replacing the semantic-based region split with random region split in the training and test stages, respectively, i.e., SGUIE-Net-w/o-SS-train and SGUIE-Net-w/o-SS-test. Compared with the full model, SGUIE-Net-w/o-SS-train suffers from performance degradation on both SUIM-E and UIEB datasets. It demonstrates that introducing semantic information as high-level

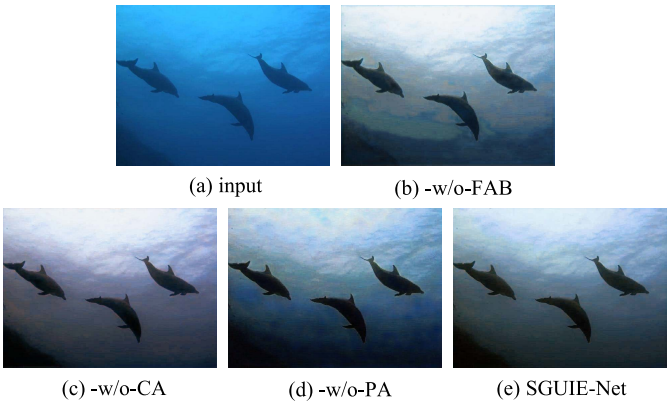


Fig. 13. Ablation study on the effectiveness of feature attention block. With FABs, SGUIE-Net achieves more consistent and robust enhancement performance.

TABLE VIII
QUANTITATIVE RESULTS OF ABLATION STUDY ON LOSS TERMS

Ablations	SUIM-E		UIEB	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
$\mathcal{L}_{main}$	22.658	0.908	21.587	0.883
$\mathcal{L}_{main} + \mathcal{L}_{srm}$	24.820	0.928	24.074	0.908

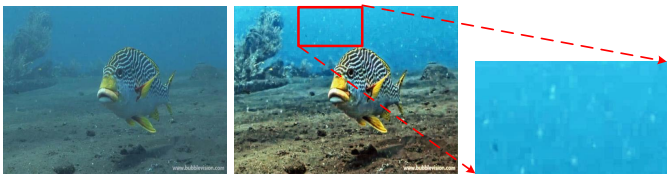


Fig. 14. Failure case. Although our SGUIE-Net effectively corrects the distorted colors and improves the contrast, the background noise is magnified after the enhancement.

guidance for the learning of underwater image enhancement is effective. More importantly, compared with SGUIE-Net-w/o-SRM, which removes SRM branch and only keeps the main branch, SGUIE-Net-w/o-SS-train still provides relatively better performance despite the fact that the semantic information being provided is incorrect. It demonstrates that employing SRM for region-wise enhancement feature learning is a necessary supplement for global-wise enhancement. And feeding SRM with meaningful regions according to the semantic segmentation maps will help to learn better supplementary features than that with random region-based learning. Additionally, compared with the full SGUIE-Net model, the performance of SGUIE-Net-w/o-SS-test is also relatively lower, which demonstrates that the better segmentation maps lead to better enhancements. But even with poor segmentation maps, compared with other deep learning-based methods, SGUIE-Net-w/o-SS-test still achieves competitive results. As shown in Figure 12, the full SGUIE-Net model provides better color correction, detail recovery and visually pleasing enhancement, with the effective guidance of region-wise semantic attention features.

3) Ablation Study on FAB: The FAB module, as the basic component of our network, contains the cascaded channel attention block and pixel attention block. To demonstrate the effectiveness of these two attention blocks, we also performed ablation experiments. Numerically, the PSNR score of SGUIE-Net-w/o-PA is higher than that of SGUIE-Net-w/o-CA by about 0.18, and both of these two ablated models are better than SGUIE-Net-w/o-FAB. However, their best scores are still lower than our full model in terms of both SSIM and PSNR metrics. Figure 13 shows the results of different ablations of the FAB module. Due to the lack of weighted perception of the channel attention block for different feature components, only using the pixel attention module may introduce artificial colors into the enhancement results (see Figure 13(c)). While, using the channel attention block alone ignores the spatially uneven degradation of different pixels, resulting in unnatural local color distortion (see Figure 13(d)). Moreover, the visual quality and quantitative evaluation of SGUIE-Net-w/o-FAB are disappointing. Both qualitative and quantitative results can demonstrate that FAB effectively combines the advantages of the two attention blocks to achieve better performance.

4) Ablation Study on Loss Terms: In Table VIII, we further present the quantitative results of ablation study on loss function. If we remove the $\mathcal{L}_{srm}$, the SRM module will loss the region-wise enhancement supervision and thus will not learn effective complementary features with semantic guidance. The evaluation demonstrates that without such effective complementary features, the performance of final enhancement will be degraded, which also illustrates the effectiveness of the proposed learning strategy.

F. Limitation and Prospect for Future Research

As described in Section III-B, the main branch of SGUIE-Net is designed to keep working at the original resolution to maximally recover local details. While, as a common dilemma for all image enhancement approaches, recovering more details usually means that more noise will be activated during the enhancement. Figure 14 shows an enhancement example of our SGUIE-Net with the background noise amplified. Actually, some reference images contain magnified noise, which also prevents the deep models from learning to denoise. In the future, exploring noise suppression modules embedded within the network is a promising research direction, such as adaptive high-frequency suppression module, which pays differentiated attention to different kinds of high-frequency components in the perspective of frequency domain. Besides, exploring the role of semantic attention in transformer-based enhancement frameworks [56], [57], [58] is another promising topic.

V. CONCLUSION

In this paper, we propose a novel underwater image enhancement network with semantic attention guidance and multi-scale perception. By introducing semantic information as the high-level guidance, we design a semantic region-wise enhancement module to bridge the gap between uncommon

degradation types and the learned distribution of underwater degradation. The complementary dual-branch, multi-scale feature perception architecture allows the model to obtain good global enhancement while recovering clear local details. We conducted extensive experiments on real-world underwater benchmarks to verify the effectiveness and the color restoration accuracy of our network. Ablation studies on the key modules and loss function demonstrate the effectiveness of our method.

REFERENCES

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

[2] 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.

[3] 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.

[4] Q. Qi et al., "Underwater image co-enhancement with correlation feature matching and joint learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1133–1147, Mar. 2022.

[5] S. Anwar and C. Li, "Diving deeper into underwater image enhancement: A survey," *Signal Process., Image Commun.*, vol. 89, Nov. 2020, Art. no. 115978.

[6] 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.

[7] 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.

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

[9] J. Xie, H. Bian, Y. Wu, Y. Zhao, L. Shan, and S. Hao, "Semantically-guided low-light image enhancement," *Pattern Recognit. Lett.*, vol. 138, pp. 308–314, Oct. 2020.

[10] M. J. Islam et al., "Semantic segmentation of underwater imagery: Dataset and benchmark," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 1769–1776.

[11] 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.

[12] D. Huang, Y. Wang, W. Song, J. Sequeira, and S. Mavromatis, "Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2018, pp. 453–465, doi: 10.1007/978-3-319-73603-7_37.

[13] X. Li, G. Hou, K. Li, and Z. Pan, "Enhancing underwater image via adaptive color and contrast enhancement, and denoising," *Eng. Appl. Artif. Intell.*, vol. 111, May 2022, Art. no. 104759.

[14] P. Zhuang and X. Ding, "Underwater image enhancement using an edge-preserving filtering retinex algorithm," *Multimedia Tools Appl.*, vol. 79, nos. 25–26, pp. 17257–17277, Jul. 2020.

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

[16] P. Zhuang, "Retinex underwater image enhancement with multiorder gradient priors," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1709–1713.

[17] 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.

[18] X. Li, G. Hou, L. Tan, and W. Liu, "A hybrid framework for underwater image enhancement," *IEEE Access*, vol. 8, pp. 197448–197462, 2020.

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

[20] C. Li, J. Quo, Y. Pang, S. Chen, and J. Wang, "Single underwater image restoration by blue-green channels dehazing and red channel correction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1731–1735.

[21] P. Drews-Jr, 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./Apr. 2016.

[22] G. Hou, J. Li, G. Wang, H. Yang, B. Huang, and Z. Pan, "A novel dark channel prior guided variational framework for underwater image restoration," *J. Vis. Commun. Image Represent.*, vol. 66, Jan. 2020, Art. no. 102732.

[23] 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 *Proc. Pacific Rim Conf. Multimedia*, 2018, pp. 678–688.

[24] 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.

[25] D. Berman, T. Treibitz, and S. Avidan, "Diving into haze-lines: Color restoration of underwater images," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[26] 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.

[27] 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. 2242–2251.

[28] H. Li and P. Zhuang, "DewaterNet: A fusion adversarial real underwater image enhancement network," *Signal Process., Image Commun.*, vol. 95, Jul. 2021, Art. no. 116248.

[29] Y. Guo, H. Li, and P. Zhuang, "Underwater image enhancement using a multiscale dense generative adversarial network," *IEEE J. Ocean. Eng.*, vol. 45, no. 3, pp. 862–870, Jul. 2020.

[30] 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.

[31] 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.

[32] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater image enhancement quality evaluation: Benchmark dataset and objective metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5959–5974, Sep. 2022.

[33] H. Yang, F. Tian, Q. Qi, Q. J. Wu, and K. Li, "Underwater image enhancement with latent consistency learning-based color transfer," *IET Image Process.*, vol. 16, no. 6, pp. 1594–1612, 2022.

[34] A. Mousavian, A. Toshev, M. Fiser, J. Kosecka, A. Wahid, and J. Davidson, "Visual representations for semantic target driven navigation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8846–8852.

[35] Y. Wu, Y. Wu, A. Tamar, S. Russell, G. Gkioxari, and Y. Tian, "Bayesian relational memory for semantic visual navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2769–2779.

[36] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7986–7994.

[37] H. Tang, D. Xu, Y. Yan, P. H. S. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7867–7876.

[38] C. B. Choy, J. Y. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2414–2422.

[39] P. Truong, M. Danelljan, and R. Timofte, "GLU-Net: Global-local universal network for dense flow and correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6257–6267.

[40] Y. Huang, Q. Wu, W. Wang, and L. Wang, "Image and sentence matching via semantic concepts and order learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 636–650, Mar. 2020.

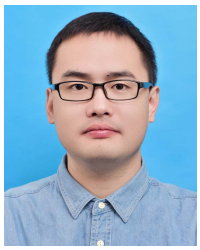
[41] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image-text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, Dec. 2020.

[42] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.
[43] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.
[44] A. Lindner, A. Shaji, N. Bonnier, and S. Süsstrunk, "Joint statistical analysis of images and keywords with applications in semantic image enhancement," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, 2012, pp. 489–498.
[45] D. Liang et al., "Semantically contrastive learning for low-light image enhancement," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1555–1563.
[46] W. Ren et al., "Deep video dehazing with semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1895–1908, Apr. 2019.
[47] S. Zhang et al., "Semantic-aware dehazing network with adaptive feature fusion," *IEEE Trans. Cybern.*, early access, Nov. 19, 2021, doi: 10.1109/TCYB.2021.3124231.
[48] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4169–4178.
[49] D. Shi, L. Ma, R. Liu, X. Fan, and Z. Luo, "Semantic-driven context aggregation network for underwater image enhancement," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, 2021, pp. 29–40.
[50] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11908–11915.
[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.*, 2015, pp. 234–241.
[52] 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.
[53] 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. 2015.
[54] 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.
[55] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, Feb. 2005.
[56] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[57] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12299–12310.
[58] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3D position embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 5812–5820.

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 Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China. His research interests include image and video enhancement, underwater vision, image processing, and visual recognition.

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 Lecture with the College of Engineering, Ocean University of China, Qingdao. His research interests include image processing, visual recognition, and deep learning. His has published over 20 peer-reviewed articles in these fields.

Haiyong Zheng (Senior Member, IEEE) received the B.E. degree in electronic information engineering and the Ph.D. degree in ocean information sensing and processing from the Ocean University of China, Qingdao, China, in 2004 and 2009, respectively. In 2009, he joined the College of Electronic Engineering, Ocean University of China, where he is currently a Professor. His research interests include computer vision, underwater vision, and deep learning.



Xiang Gao received the B.S. and M.S. degrees from the Ocean University of China, Qingdao, China, in 2012 and 2015, respectively, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2019. From 2019 to 2022, he was a Lecture at the College of Engineering, Ocean University of China. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include 3D computer vision, with an emphasis on image-based large-scale 3D reconstruction.

Guojia Hou (Member, IEEE) received the Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, China, in 2015. He is currently an Associate Professor with the College of Computer Science and Technology, Qingdao University. His current research interests include underwater vision, image/video processing, and image quality assessment.

Kun Sun (Member, IEEE) received the Ph.D. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan, China. He is currently working as an Associate Professor at the School of Computer Science, China University of Geosciences (CUG), Wuhan. His research focuses on 3D related computer vision algorithms, such as multi-view image matching, large scale structure from motion (SfM), and 3D point cloud processing.