

A Physical Model-Guided Framework for Underwater Image Enhancement and Depth Estimation

Dazhao Du, Enhao Li, Lingyu Si, Fanjiang Xu, Jianwei Niu, *Senior Member, IEEE*,
and Fuchun Sun, *Fellow, IEEE*

Abstract—Due to the selective absorption and scattering of light by diverse aquatic media, underwater images usually suffer from various visual degradations. Existing underwater image enhancement (UIE) approaches that combine underwater physical imaging models with neural networks often fail to accurately estimate imaging model parameters such as depth and veiling light, resulting in poor performance in certain scenarios. To address this issue, we propose a physical model-guided framework for jointly training a Deep Degradation Model (DDM) with any advanced UIE model. DDM includes three well-designed sub-networks to accurately estimate various imaging parameters: a veiling light estimation sub-network, a factors estimation sub-network, and a depth estimation sub-network. Based on the estimated parameters and the underwater physical imaging model, we impose physical constraints on the enhancement process by modeling the relationship between underwater images and desired clean images, i.e., outputs of the UIE model. Moreover, while our framework is compatible with any UIE model, we design a simple yet effective fully convolutional UIE model, termed UIEConv. UIEConv utilizes both global and local features for image enhancement through a dual-branch structure. UIEConv trained within our framework achieves remarkable enhancement results across diverse underwater scenes. Furthermore, as a byproduct of UIE, the trained depth estimation sub-network enables accurate underwater scene depth estimation. Extensive experiments conducted in various real underwater imaging scenarios, including deep-sea environments with artificial light sources, validate the effectiveness of our framework and the UIEConv model.

Index Terms—Underwater image enhancement, underwater physical imaging model, depth estimation.

I. INTRODUCTION

Underwater images often suffer from color distortion, low contrast, and blurriness caused by light absorption and scattering in water. These issues are exacerbated by varying water conditions such as turbidity and depth which result in uneven lighting and haziness. By improving the visibility and

clarity of underwater images, underwater image enhancement (UIE) facilitates the analysis, monitoring, and exploration of underwater scenes.

Early UIE methods simulate the degradation process by an underwater physical imaging model and estimate its parameters to invert clear images [1]–[4]. With the introduction of paired datasets containing underwater images and reference images [5]–[7], data-driven approaches have gradually gained attention. Numerous UIE models have been proposed to improve performance and efficiency [8]–[12]. However, data-driven methods often suffer from poor generalization and interpretability. To combine the advantages of physical model-based methods and the powerful representational capacity of neural networks, some approaches [13], [14] employ neural networks to estimate the parameters of the physical imaging model and then invert the degradation process based on the estimated parameters to obtain enhanced images. We argue that previous methods face two primary issues: (1) *inaccurate imaging parameter estimation*; and (2) *even with accurate parameter estimation, there remains a discrepancy between the physical imaging model and real underwater conditions*.

To address the first issue, we design a Deep Degradation Model (DDM), which includes three well-designed sub-networks to estimate scene depth, veiling light, attenuation coefficient, and scattering coefficient. We briefly outline the limitations of previous methods in estimating parameters and our solutions. (1) **Scene Depth**: To obtain absolute depth, relative depth is typically scaled based on manually set maximum and minimum depths. For example, in HybrUR [14], the minimum and maximum depth in each image are empirically set to 0.1m and 6m, respectively. These predefined depth ranges are unreasonable. Therefore, we propose a depth estimation sub-network to output accurate relative depth maps and a factors estimation sub-network to adaptively output scales and offsets for scaling relative depth values for each image. (2) **Veiling Light**: Veiling light, considered as the value of backscatter at infinity, is usually represented by a monochrome image where all pixel values are identical across spatial locations [14]–[16]. This assumption holds in shallow-water environments where light primarily originates from natural sources but fails in deep-sea scenes with artificial light sources, where underwater images often suffer from uneven lighting [17], [18]. To adapt to various complex environments, we propose a veiling light

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA 0370604.

Dazhao Du, Enhao Li and Jianwei Niu are with Hangzhou Innovation Institute, Beihang University, Hangzhou, Zhejiang 310051, China (e-mail: dudazhao@buaa.edu.cn; lienhan@buaa.edu.cn; niujianwei@buaa.edu.cn).

Lingyu Si and Fanjiang Xu are with Institute of Software, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lingyu@iscas.ac.cn; fanjiang@iscas.ac.cn).

Fuchun Sun is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: fc-sun@mail.tsinghua.edu.cn).

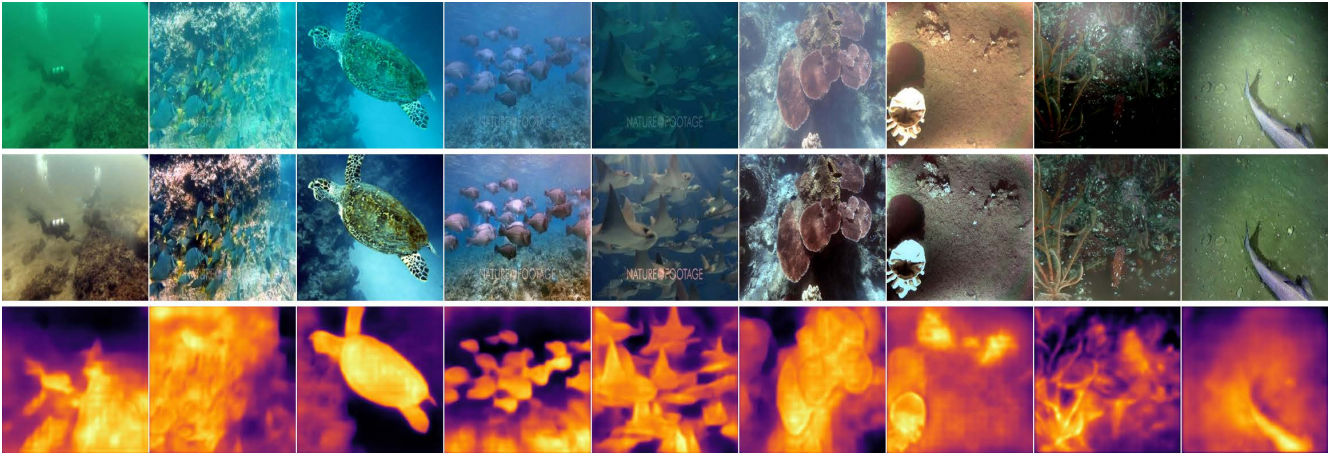


Fig. 1. For the original underwater images with various degradation effects, including bluish, greenish, yellowish hues, turbidity, blur, haziness, and uneven lighting (first row), we visualize the enhancement results (second row) and depth estimation results (third row) obtained by our method. Our method demonstrates impressive performance across various complex underwater environments.

estimation sub-network that outputs a local background light map where pixel values vary across spatial locations instead of a monochrome image. (3) **Attenuation and Scattering Coefficients:** When estimating scattering and attenuation coefficients, some methods directly assume they are identical [19]. Even if some methods estimate the transmission map instead of these two coefficients [13], [15], [20], they still rely on the assumption that both coefficients are identical. Recent work [21] has experimentally shown that they are related but not identical. Therefore, our factors estimation sub-network employs two separate heads to distinguish between them, providing a more accurate modeling approach.

To address the second issue, we do not directly use the estimated imaging parameters to invert clean images, as done by previous methods. Instead, we use the physical imaging model to degrade the enhanced image produced by the UIE model. We enforce that the degraded enhanced image closely resembles the original underwater image, thereby imposing physical constraints on the enhancement process of the UIE model. In our training framework, physical constraints manifest as an additional regularization term that assists in the training of the UIE model. In this way, the UIE model with strong fitting capabilities can further be guided and supplemented by domain knowledge of the imaging process. Furthermore, considering that the scene depth of the original and enhanced underwater images should remain consistent, we introduce a depth consistency loss to further assist in the training process. Our framework can improve the performance of many UIE models.

In UIE models, modeling both local and global features is crucial for image restoration. Global features are essential for correcting overall information such as color and lighting, while local features help restore high-frequency details. Due to the limited receptive field of convolutions, some models adopt frequency domain operations [12] or Transformers [7] to extract global features. In contrast, we design a simple yet effective fully convolutional UIE model, termed UIEConv, which includes both a global branch and a local branch.

The global branch adopts a U-Net-like architecture, where the encoder progressively downsamples to obtain high-level features, and the decoder upsamples to restore the image. The upsampling and downsampling operations provide a large receptive field, sufficient for modeling long-range dependencies. The local branch, with a smaller receptive field, consists of several convolutions that maintain the image resolution throughout. The final enhanced result is obtained by combining the outputs of the two branches.

Our main contributions can be summarized as follows:

- We propose a physical model-guided training framework that jointly performs image enhancement and depth estimation, with the two tasks complementing and enhancing each other. Additionally, any advanced UIE model can be trained within this framework to achieve further performance improvements.
- Within our framework, we meticulously design various sub-networks to accurately estimate crucial parameters for the underwater physical imaging model, including veiling light, scene depth, attenuation coefficient, and scattering coefficient.
- We introduce a simple yet effective UIE model, termed UIEConv. It employs a dual-branch structure to fully exploit both global and local information in the image. Without any bells and whistles, UIEConv outperforms other state-of-the-art UIE models.
- We test our method in various underwater scenarios, including different water quality environments and deep-sea scenes with limited lighting. There are some examples in Fig. 1. Our approach consistently achieves impressive enhancement and depth estimation results. Extensive experiments validate the effectiveness of our approach.

II. RELATED WORK

A. Underwater Physical Imaging Model

According to [22], as shown in Fig. 2, the light entering the camera mainly consists of three components: backscattered light B , direct transmitted light D , and forward scattered

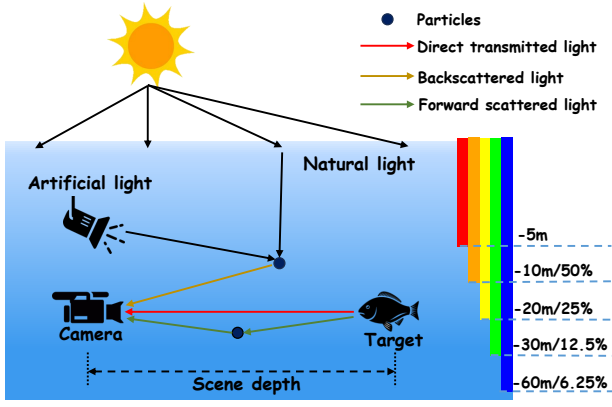


Fig. 2. Schematic diagram of underwater imaging process.

light F . Given that $F \ll D$, forward scattered light does not significantly contribute to the degradation of an image. Therefore, the underwater physical imaging model can be represented as the sum of direct transmitted light D and backscattered light B :

$$I = D + B \quad (1a)$$

$$= J \cdot e^{-\beta^D \cdot d} + B^\infty \cdot (1 - e^{-\beta^B \cdot d}), \quad (1b)$$

where $I \in \mathbb{R}^{3 \times H \times W}$ is the observed underwater image, $J \in \mathbb{R}^{3 \times H \times W}$ is the restored clear image (i.e., scene radiance), and $B^\infty \in \mathbb{R}^{3 \times 1 \times 1}$ is the veiling light (i.e., global background light). d is the scene depth, representing the distance between the camera and the observed object. Due to the different absorption rates of different wavelengths of light in water, the attenuation coefficient β^D and the scattering coefficient β^B are channel-dependent, i.e., $\beta^D, \beta^B \in \mathbb{R}^{3 \times 1 \times 1}$. As shown on the right side of Fig. 2, red light is absorbed the most in water, resulting in underwater images typically having a bluish-green hue. In some previous literature [2], [3], [15], [19], it was assumed that $\beta^D = \beta^B = \beta$. Therefore, the underwater physical imaging model can be rewritten as:

$$I = J \cdot T + B^\infty \cdot (1 - T), \quad (2)$$

where $T = e^{-\beta \cdot d}$ is called the transmission map. In this paper, we consider distinguishing between β^D and β^B to achieve more accurate modeling. Besides, the veiling light B^∞ in Eq. (1b) is assumed to be a monochrome image, i.e., $B^\infty \in \mathbb{R}^{3 \times 1 \times 1}$. This assumption is not reasonable in deep-sea scenarios with artificial light sources where the lighting is typically uneven. In this paper, we address this by assuming that the veiling light may vary at different pixel positions to adapt to scenes with uneven artificial lighting, i.e., $B^\infty \in \mathbb{R}^{3 \times H \times W}$.

B. Underwater Image Enhancement

Underwater image enhancement methods can be categorized into three types: physical model-free, physical model-based, and deep learning-based methods. These methods are mainly applied in shallow-water environments and perform poorly in deep-sea environments. Therefore, some methods designed for deep-sea environments have been proposed recently.

a) *Physical model-free methods*: Physical model-free methods directly adjust image pixel values based on various metrics [23]. To improve color cast and contrast degradation in underwater images, some methods perform weighted summation of multiple enhanced versions. For example, Fusion [24] combines the results of white balance and contrast local adaptive histogram equalization. ACCC [25] integrates the complementary advantages of local and global contrast-enhanced versions through multi-scale fusion [26]. MLLE [27] explores adaptive algorithms for color correction and contrast enhancement. Additionally, some methods based on Retinex theory decompose underwater images into reflection and illumination components and enhance them separately [28]–[30]. However, these methods often suffer from over-enhancement and color distortions in complex underwater environments.

b) *Physical model-based methods*: Physical model-based methods estimate the parameters of the underwater physical imaging model in Eq. (1b) or (2) to invert the degradation process. These parameters can be estimated based on various priors, such as the dark channel prior (DCP) [3], [31], maximum intensity prior [1], image blurriness [4], [32], and minimum information loss [33]. UDCP [2], [34] adapts DCP to underwater environments by using only the blue and green channels to estimate the transmission map. Sea-thru [35] proposes a physically accurate model and restores color based on RGBD images. Berman et al. [36] estimate the attenuation ratios of the blue-red and blue-green color channels by evaluating every possible water type. Zhou et al. [37] estimate depth maps using a channel intensity prior and eliminate backscatter through adaptive dark pixels. However, these methods are sensitive to the assumptions made during parameter estimation, making them less robust in complex and varied underwater environments.

c) *Deep learning-based methods*: Deep learning-based methods have gained popularity recently due to their remarkable performance. Li et al. [5] introduce the UIEB dataset, which consists of many underwater images and corresponding reference images, for supervised training. This dataset has facilitated the development of various advanced UIE models [11], [12], [38]. Peng et al. [7] construct a larger dataset LSUI and design a Transformer-based UIE model. In addition, UGAN [8] and FUnIEGAN [9] utilize adversarial training to generate clean images. Similar to our work, some methods integrate physical imaging models with neural networks. For instance, some researchers incorporate transmission maps and depth maps as attention or additional inputs to guide neural networks [10], [39]. PUGAN [19] estimates attenuation coefficients and depth maps to derive transmission maps, which then guide the decoding process of recovered images. To circumvent the reliance on paired training data, several self-supervised methods [13]–[16] employ neural networks to estimate transmission maps, global background light, and even depth maps, scattering coefficients. These parameters are subsequently used to construct inversion models for image recovery. In comparison to these methods, our approach enables more accurate parameter estimation and better adaptation to various complex underwater environments.

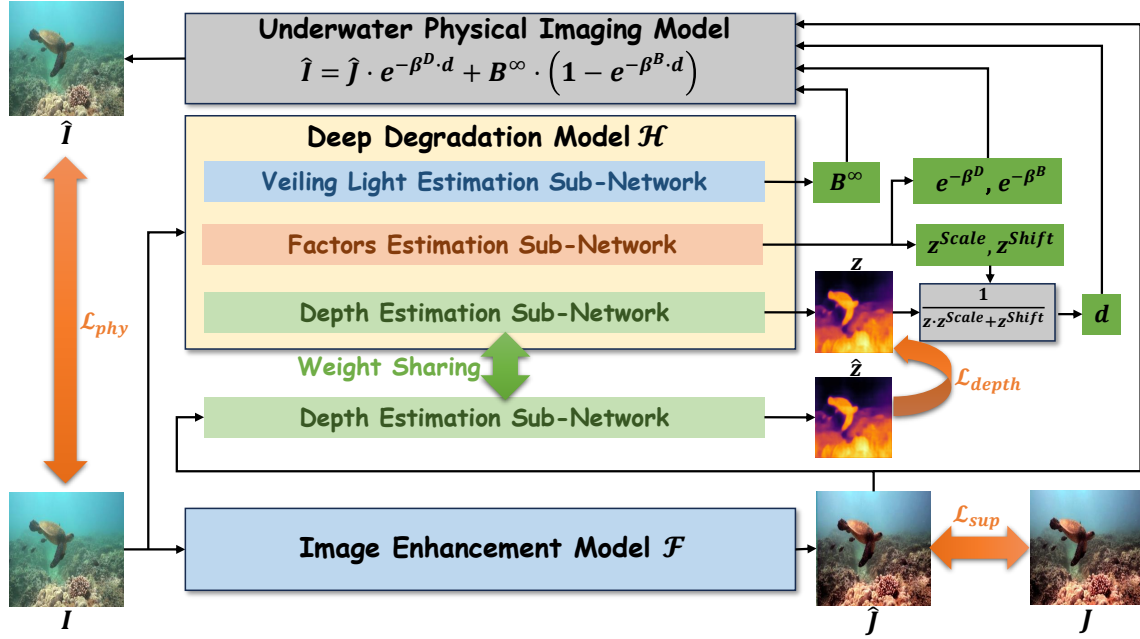


Fig. 3. Schematic diagram of our framework. It mainly consists of a deep degradation model (DDM) \mathcal{H} and an underwater image enhancement (UIE) model \mathcal{F} . The UIE Model is employed to obtain the enhanced image \hat{J} . The DDM is composed of three sub-networks that estimate various parameters of the physical imaging model. Using the outputs from these models, the degraded image \hat{I} can be generated.

d) Deep-sea image enhancement methods: Underwater images in deep-sea environments often suffer from uneven lighting and low light conditions. L²UWE [17] proposes two contrast-guided atmospheric illumination models that enhance details and reduce dark regions. Cao et al. [40] simulate point light sources by adding light spots to the ground truth, creating a dataset that includes both synthetic and real data for training neural networks. Hou et al. [18] observe that the illumination channel of a uniform-light underwater image in the HSI color space contains few pixels close to zero, and present an effective illumination channel sparsity prior (ICSP) based on this observation. IACC [41] unifies the luminance features of underwater artificial and natural light and guides consistent enhancement across similar luminance regions. Other methods use CNNs to enhance non-uniform illumination images from the HSV [42] and LAB [43] color spaces.

C. Monocular Depth Estimation

Recently, various monocular depth estimation (MDE) algorithms [44], [45] have been proposed to estimate scene depth from a single image. These methods typically require training on datasets with detailed depth annotations [46]. However, collecting labeled datasets in underwater scenes is more challenging. Therefore, some researchers synthesize underwater-styled images by processing in-air images with depth data and physical imaging models to construct labeled datasets [6], [47]. Nonetheless, a domain gap exists between synthetic and real underwater images. To address this issue, Atlantis [48] proposes a novel pipeline for generating photorealistic underwater images based on terrestrial depth data and diffusion models. Additionally, some works attempt to estimate depth in a self-supervised manner using GANs [49] or the relationships

between consecutive frames in underwater videos [16], [50]. Many physical model-based UIE algorithms indirectly obtain depth maps while estimating the transmission maps [2], [3], [32], [34], [51]. However, due to unknown scattering and attenuation coefficients, the estimated depth maps are most likely wrong. A recent work, Depth Anything [45], demonstrates strong performance across various complex scenarios due to its pre-training on large-scale datasets. Although it is not specifically designed for underwater scenes, it serves as a good starting point for estimating depth in underwater environments.

III. METHODOLOGY

A. Overview of Framework

Our proposed framework is illustrated in Fig. 3. Given an underwater image I , the UIE model \mathcal{F} produces an enhanced image \hat{J} . The supervised loss \mathcal{L}_{sup} ensures that \hat{J} closely approximates the reference image J . Additionally, I is input into a deep degradation model \mathcal{H} , which comprises three sub-networks: veiling light estimation sub-network (VLEN), depth estimation network (DEN), and factors estimation sub-network (FEN). VLEN and DEN estimate the veiling light B^∞ and relative inverse depth map z of the underwater image I , respectively. FEN estimates the attenuation coefficient, scattering coefficient, and the scale factor z^{Scale} and shift factor z^{Shift} for the relative inverse depth. The two factors z^{Scale} , z^{Shift} , and relative inverse depth map z together derive the scene depth map d . Finally, all variables are substituted into the underwater physical imaging model in Eq. (1b), resulting in a re-degraded image \hat{I} . We enforce consistency between I and \hat{I} to impose a physical constraint loss \mathcal{L}_{phy} . Considering that scene depth should be consistent before and after enhancement, we also utilize DEN to estimate the inverse

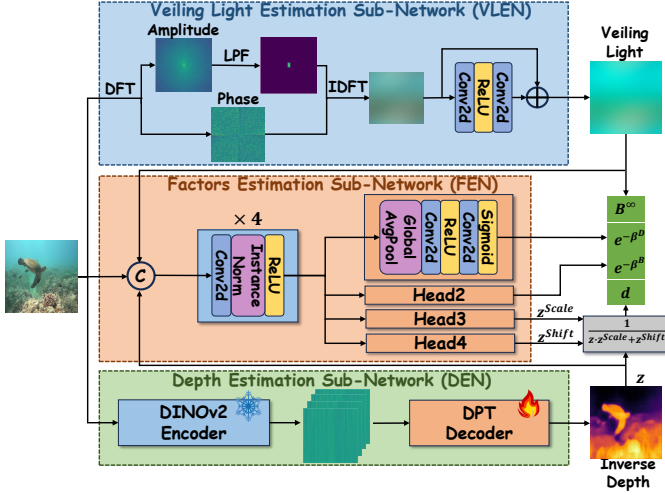


Fig. 4. Detailed structure of the three sub-networks in the deep degradation model.

depth map \hat{z} of the enhanced image \hat{J} and introduce an additional depth consistency loss \mathcal{L}_{depth} . Ultimately, three losses jointly train the learnable components \mathcal{H} and \mathcal{F} within our framework. We will detail the structures of \mathcal{H} and \mathcal{F} in the following subsections.

B. Deep Degradation Model

As shown in Fig. 4, the deep degradation model (DDM) \mathcal{H} consists of three well-designed sub-networks that accurately estimate various parameters of the underwater physical imaging model in Eq. (1b). Below, we provide a detailed explanation of the design principles and structures.

a) *Veiling Light Estimation Sub-Network*: Most previous works assume that the veiling light B^∞ can be represented by a monochrome image with identical pixel values across spatial locations. However, this assumption fails in deep-sea scenarios where the light primarily comes from artificial point light sources. As shown in the two examples on the far right of Fig. 1, deep-sea images suffer from non-uniform illumination phenomena. Therefore, it is necessary to design a more reasonable veiling light estimation method to adapt to various lighting conditions from shallow to deep-sea environments. Fourmer [52] demonstrated that brightness, as a global feature, is primarily preserved in the center of the amplitude component of the image. Inspired by this idea, we first perform a Discrete Fourier Transform (DFT) on the underwater image to obtain the amplitude and phase components. Then we obtain the initial veiling light estimation by performing an Inverse DFT (IDFT) on the phase component and the amplitude component after applying a low-pass filter (LPF). To introduce flexibility and adjustability, the initial estimation is fed into convolutional modules with a residual connection to obtain the final veiling light B^∞ . Fig 5 compares our method with IBLA [32] and GDCP [51], both of which represent veiling light as monochrome images. For the first underwater image, the estimated veiling light maps of the three methods are similar. However, for non-uniform illumination images

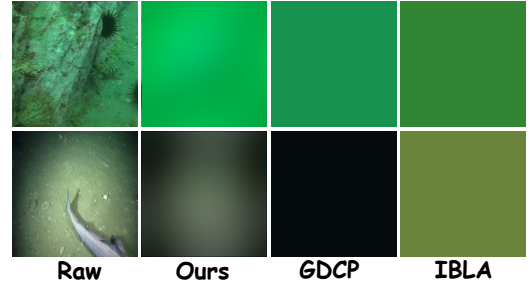


Fig. 5. Two underwater images under different lighting conditions and the veiling light maps estimated by three methods.

in deep-sea scenarios, the other two methods fail while our method achieves relatively accurate veiling light estimation.

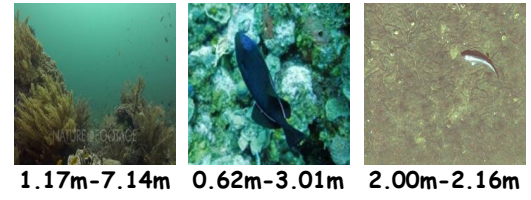


Fig. 6. Three underwater images with varying ranges of scene depth. The depth range below each image is derived from the output z^{Scale} and z^{Shift} of FEN. The first image features a large maximum depth in the water region; the second image has smaller maximum and minimum depths; in the last image, the minimum and maximum depths are nearly identical.

b) *Depth Estimation Sub-Network*: Scene depth is crucial for the physical imaging model, but accurately estimating it from a single underwater image is very challenging. A recent work, Depth Anything [45], demonstrates strong performance across various complex scenarios, including underwater environments, due to its pre-training on large-scale datasets. Depth Anything includes a DINOv2 encoder [53] to extract image features and a dense prediction Transformer (DPT) decoder [54] to predict relative depth, i.e., inverse depth. We aim to transfer the powerful zero-shot depth estimation capability of Depth Anything to our depth estimation sub-network (DEN). Specifically, we adopt the same encoder-decoder architecture and set the Dinov2 encoder to the small version of ViT [55] for efficiency. We initialize DEN with the weights from Depth Anything. During training, we freeze the DINOv2 encoder weights and only fine-tune the DPT decoder. Ultimately, DEN can accurately estimate the normalized inverse depth map z .

c) *Factors Estimation Sub-Network*: DEN estimates a relative inverse depth map, but what we need is the absolute depth map. To obtain the absolute depth, previous methods used predefined maximum and minimum depth values, such as 0.1-6m, to scale the relative depth of all images to the same range. However, the depth ranges of different underwater images can vary significantly, as illustrated in Fig. 6. Therefore, we introduce a factors estimation sub-network (FEN) to estimate the scaling factor z^{Scale} and shift factor z^{Shift} for the relative inverse depth. The absolute depth d is then computed as $d = \frac{1}{z \cdot z^{Scale} + z^{Shift}}$. Additionally, FEN is responsible for estimating the scattering and

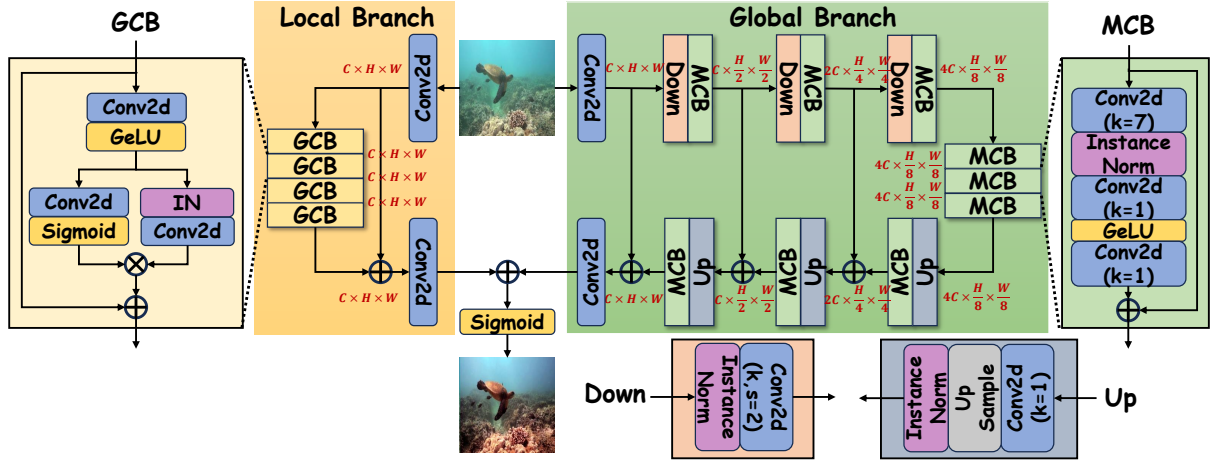


Fig. 7. The structural diagram of UIEConv. UIEConv includes the local branch on the left and the global branch on the right. We annotate the shapes of the intermediate features, where C , H , and W represent the number of channels, height, and width, respectively. k and s respectively represent the kernel size and stride in Conv2d.

attenuation coefficients. Specifically, the underwater image I , the estimated veiling light map B^∞ , and the inverse depth map z are concatenated along the channel dimension and then input into FEN, which consists of a backbone and four structurally identical yet independent heads. The backbone includes four simple convolutional layers, each comprising convolution, instance normalization, and a ReLU activation function. The four independent heads are designed to estimate z^{Scale} , z^{Shift} , $e^{-\beta^D}$, and $e^{-\beta^B}$, respectively. Given that the attenuation coefficient β^D and scattering coefficient β^B are positive values, a Sigmoid activation function is applied to produce values between 0 and 1 as the terms $e^{-\beta^D}$ and $e^{-\beta^B}$. To more reasonably scale the relative depth, we add 0.1 to the Sigmoid output to generate values greater than 0.1 as z^{Shift} and multiply the Sigmoid output by 2 to generate values between 0 and 2 as z^{Scale} . The approach allows the maximum and minimum absolute depths to vary significantly within a range of 0-10m.

C. UIEConv Model

Although the enhancement model \mathcal{F} in our framework can be any advanced UIE model, such as UShape [7] or UIEC²Net [11], we have designed a simple yet more effective fully convolutional model called UIEConv. As shown in Fig. 7, UIEConv includes two branches: a global branch and a local branch.

a) Global Branch: The global branch follows the common U-Net [56] encoder-decoder architecture. In the encoding part, downsampling reduces the image resolution, while in the decoding part, upsampling gradually restores the image to its original resolution. We use convolutions with a stride and kernel size of 2 for downsampling and bilinear interpolation for upsampling. The core module of the global branch is the Modern Convolutional Block (MCB) derived from the advanced vision convolutional backbone ConvNeXt [57], which employs large kernel convolutions (kernel size of 7) and a transformer-like [55] design. We replace Layer Normalization [58] with Instance Normalization [59], which has been

proven to be more suitable for image restoration tasks [60]. The combination of the U-Net-like architecture and large kernel convolutions results in a very large receptive field, allowing the capture of global features in the image without the need for complex attention mechanisms.

b) Local Branch: During the downsampling process in the global branch, the image resolution is reduced, leading to a loss of details. To focus on local areas and supplement detailed information, the local branch stacks four Gated Convolutional Blocks (GCB), maintaining resolution throughout the forward process. GCB primarily employs convolutions with a kernel size of 3 and GeLU nonlinear activation functions to extract local features. Considering that underwater images may have varying degrees of degradation across different spatial locations and channels, we also introduce gating mechanism [41] in GCB. Specifically, we use convolutions and sigmoid activation functions to generate dynamic weights, which adaptively control the retention of features in different spatial locations and channels. The local branch, with its smaller receptive field, focuses on local features and image details.

The outputs of the global and local branches are added and passed through a Sigmoid activation function to obtain the final enhanced image \hat{J} .

D. Loss Function

The total loss function comprises three components: supervised loss \mathcal{L}_{sup} , physical constraint loss \mathcal{L}_{phy} , and depth consistency loss \mathcal{L}_{depth} .

a) Supervised Loss: The supervised loss calculates the L1 distance between the enhanced image \hat{J} and the reference image J :

$$\mathcal{L}_{sup} = \frac{1}{H * W} \sum_i \sum_j |J_{ij} - \hat{J}_{ij}|, \quad (3)$$

where H represents the height and W represents the width of the image, and ij represents the pixel position.

b) Physical Constraint Loss: The physical constraint loss computes the L1 distance between the degraded image \hat{I} , which is obtained by applying the underwater physical imaging model to the enhanced image \hat{J} , and the original underwater image I :

$$\mathcal{L}_{phy} = \frac{1}{H * W} \sum_i^H \sum_j^W |I_{ij} - \hat{I}_{ij}|. \quad (4)$$

The physical constraint loss is a primary optimization objective for UIE models in some self-supervised UIE methods [15], [16], while it serves as an additional regularization term to assist the training of the UIE model \mathcal{F} in our framework.

c) Depth Consistency Loss: Enhancing an image should not alter the scene depth. A robust depth estimation sub-network should consistently output the same depth map regardless of image quality changes. Therefore, we compute the scale- and shift-invariant loss [44] between the inverse depth map z of the underwater image and the inverse depth map \hat{z} of the enhanced image as the depth consistency loss:

$$\mathcal{L}_{depth} = \frac{1}{H * W} \sum_i^H \sum_j^W \rho(z_{ij}, \hat{z}_{ij}), \quad (5)$$

where ρ is the affine-invariant mean absolute error loss defined in [45].

The overall loss function is a weighted sum of the above three losses:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{phy} + \lambda_2 \mathcal{L}_{depth}, \quad (6)$$

where λ_1 and λ_2 are 0.2 and 1, respectively.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We conduct experiments on three shallow-water datasets (where the light source for most images comes from sunlight) and two deep-sea datasets (where almost all images are illuminated by artificial lights). Additionally, we perform quantitative evaluations of depth estimation on an underwater dataset with depth annotations.

The shallow water datasets include LSUI [7], UIEB [5] and U45 [61]. LSUI contains 4279 pairs of real-world underwater images and clear reference images. We use 3879 pairs of these images as the training set, and the remaining 400 pairs of images as the test set. To evaluate generalization, we randomly select 90 images with reference images in UIEB as the second test set, named **T90**. In addition, UIEB also includes a set of 60 challenge images (**C60**) that do not have corresponding reference images. And **U45** contains 45 carefully selected underwater images, serving as an important benchmark for UIE. For all the supervised methods, we train the models on the training set of LSUI and test them on the LUSI test set, T90, C60, and U45. The deep-sea datasets are UIID [40] and OceanDark [62]. UIID contains 3486 pairs of non-uniform illumination images with reference images. This dataset consists primarily of synthetic images, with a small portion of real images. We randomly select 3136 pairs for the training set and reserve the remaining 350 pairs for

the test set. OceanDark comprises 183 underwater images captured by video cameras located in profound depths using artificial lighting. For all supervised methods, we train the models on the UIID training set and evaluate them on both the UIID test set and OceanDark. To quantitatively evaluate depth estimation, we conduct evaluations using the Sea-thru's D3 and D5 subsets [35], which includes underwater images and corresponding depth maps obtained via the Structure-from-Motion algorithm.

We use the commonly employed PSNR and SSIM as full-reference image quality evaluation metrics. For datasets without reference images, we use UIQM [63] and UCIQE [64], which are designed for underwater image quality assessment, as no-reference metrics. In deep-sea scenarios, we also use the PIQE [65] metric. To compare the efficiency of different methods, we report the runtime for all methods and the parameter count and FLOPs only for the deep learning-based models. For depth estimation, we use six metrics in [45]: root mean square error (RMSE), absolute mean relative error (Abs.Rel), absolute error in log-scale (\log_{10}), and the percentage of inlier pixels (δ_i) with threshold 1.25^i .

B. Implementation Details

We replicated all the compared methods based on their official codes and hyper-parameters. All models were trained for 160 epochs with a batch size of 8, using the ADAM optimizer. The initial learning rate was set to 5×10^{-5} and reduced by half after 128 epochs. In our proposed training framework, the learning rate for the depth estimation sub-network was set to 0.3 times that of the other modules. To ensure a fair comparison, all images were resized to 256×256 during both training and testing. For runtime evaluation, we ran all deep learning-based models on an NVIDIA TITAN V GPU and traditional methods on an Intel(R) Core(TM) i7-13700K CPU. We inferenced on 1000 images to report the average runtime per image.

C. Performance Evaluation

We comprehensively compare our method with other UIE methods in various scenarios. Besides, we qualitatively and quantitatively evaluate the performance of depth estimation.

a) Shallow-water Scene: We use the training set of LSUI for training and the LSUI test set, T90, C60, and U45 for testing. Almost all the images in these datasets come from shallow-sea scenes, where the light mainly come from the sun. In Table I, we report not only the results of training UIEConv integrated into our framework (last row) but also the results of training UIEConv alone (second to last row). For datasets with reference images like LSUI and T90, our method achieves the best PSNR and SSIM metrics. For images without reference images in C60 and U45, Fusion achieves the best UCIQE, while FUnIEGAN and UIEC2Net perform best on UIQM. However, as shown in Figs. 8, 9, 10, their enhancement effects are not satisfactory from a human visual perspective. This observation suggests that these two non-reference quality assessment metrics sometimes do not align with human visual perception, as noted in many previous works [5], [19], [60].

TABLE I
QUANTITATIVE COMPARISON OF VARIOUS UIE METHODS ACROSS FOUR DATASETS PRIMARILY CONSISTING OF SHALLOW-WATER IMAGES. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Method	LSUI		T90		C60		U45		FLOPs↓	#Param.↓	Time↓
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	UIQM↑	UCIQE↑	UIQM↑	UCIQE↑			
GDCP [51]	13.4928	0.6852	14.3821	0.7360	2.1487	0.5873	2.2481	0.5937	-	-	0.076s
Fusion [24]	18.6835	0.7923	23.0684	0.9194	2.6077	0.6123	2.9691	0.6393	-	-	0.056s
IBLA [32]	17.2079	0.7516	18.6182	0.7715	1.9790	0.5874	2.3578	0.5836	-	-	2.934s
UGAN [8]	21.1094	0.8103	19.5773	0.8280	2.8587	0.5485	3.1022	0.5670	18.15G	54.40M	0.006s
FUnIEGAN [9]	21.7107	0.7967	19.7440	0.8381	2.8725	0.5457	2.9190	0.5594	10.24G	7.02M	0.002s
MLLE [27]	17.6313	0.7127	19.8410	0.8215	2.2075	0.5689	2.4845	0.5947	-	-	0.031s
Ucolor [10]	21.2919	0.8324	23.2240	0.9039	2.6610	0.5520	3.2023	0.5796	443.85G	157.42M	0.832s
PUGAN [19]	20.5350	0.8075	22.6060	0.8774	2.8460	0.5998	3.1774	0.6124	72.05G	95.66M	0.021s
UIEC ² Net [11]	25.0757	0.8708	23.3631	0.9008	2.8481	0.5810	3.2157	0.5820	26.06G	0.53M	0.041s
Ushape [7]	25.7630	0.8296	20.3947	0.7763	2.4828	0.5475	2.8956	0.5725	2.98G	22.82M	0.046s
UIEConv	28.9155	0.9187	24.1197	0.9283	2.3540	0.5709	3.0953	0.5907	121.53G	3.31M	0.020s
Ours	29.9253	0.9248	24.9395	0.9429	2.5486	0.5767	3.1542	0.5966	146.91G	29.55M	0.049s

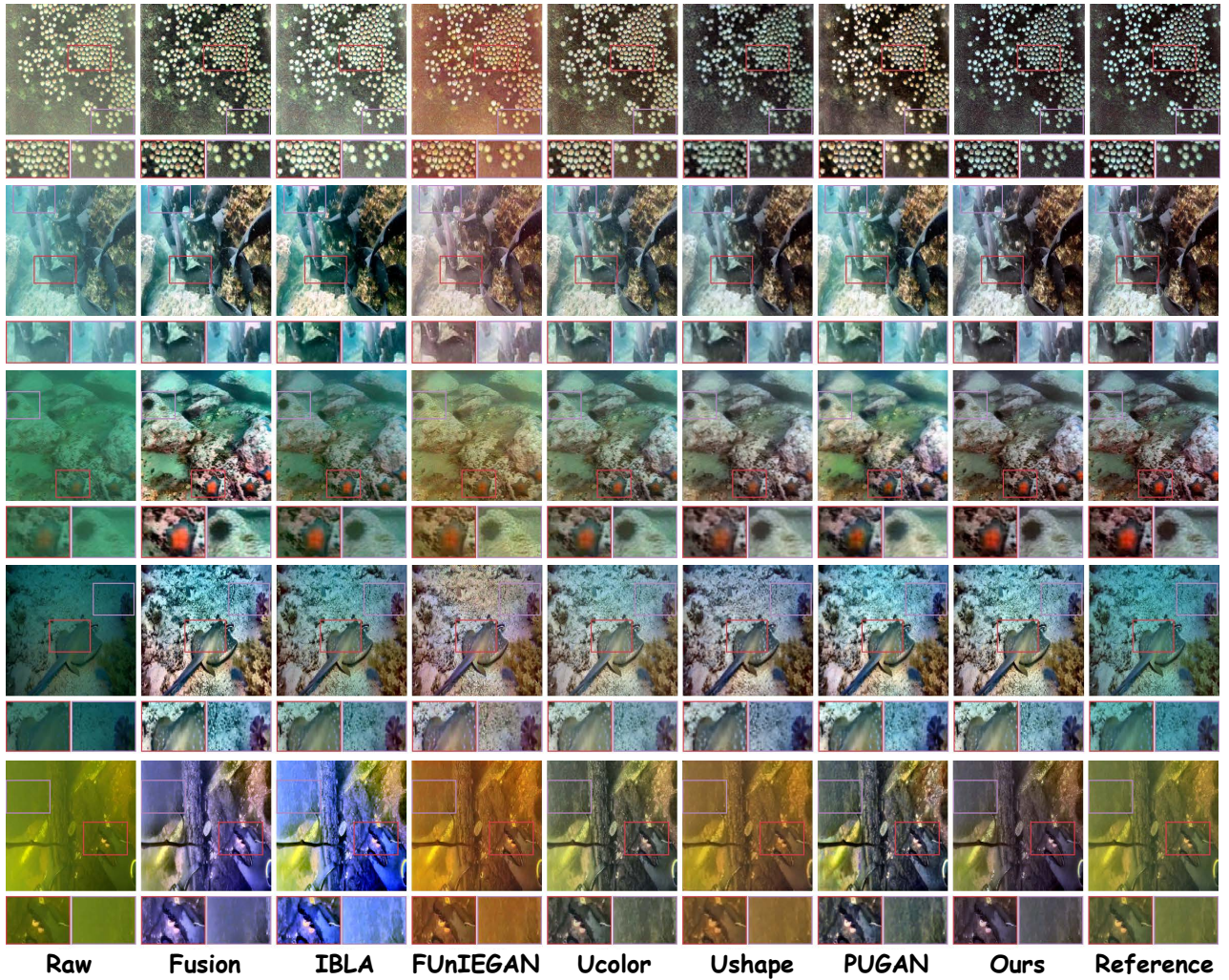


Fig. 8. Enhancement results of five images from the LSUI and UIEB dataset. The first three images come from LSUI, while the last two come from UIEB. We enlarge the local areas of the enhanced images to compare the details.

In contrast, our model consistently produces enhanced images with no color cast and clear details across various underwater conditions and degradation effects. In the last two examples, our enhanced images even appear more satisfactory than the reference images, removing excessive blue and yellow tones.

Additionally, the UIEConv model has a small number of parameters and runs quickly. While integrating UIEConv into our framework reduces efficiency, it significantly improves performance. Future work will focus on designing more efficient sub-network structures.

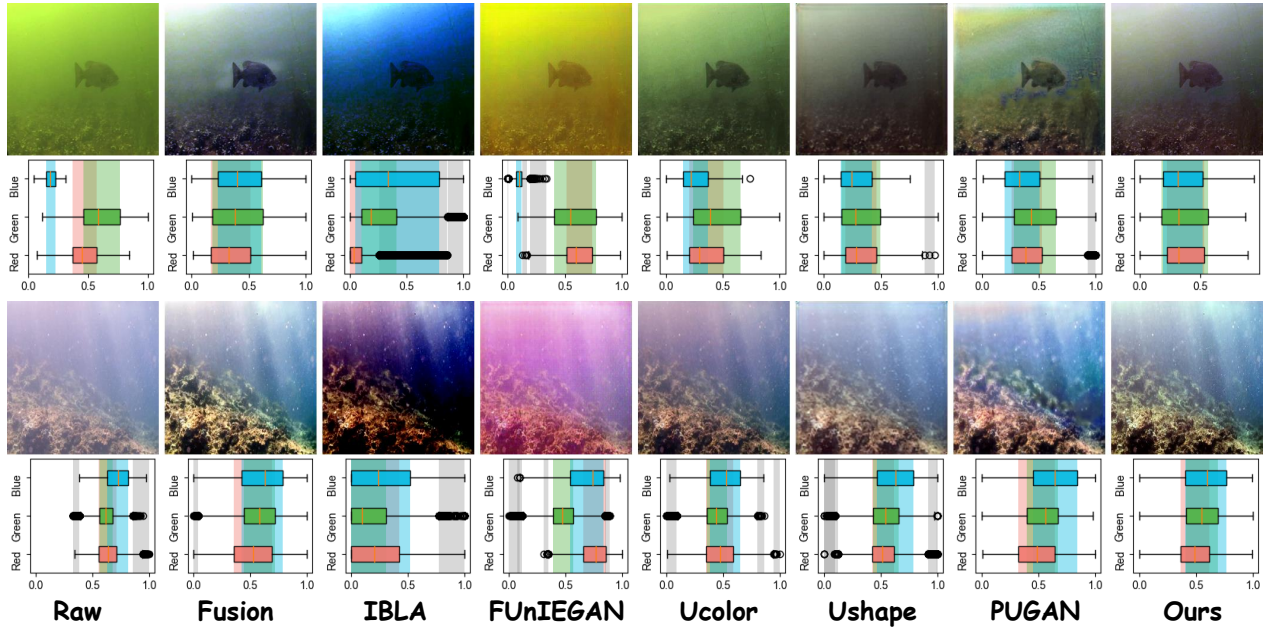


Fig. 9. Enhancement results of two images from the C60 dataset. The intensity distribution of the RGB channels is presented in boxplots. The black points represent outliers in each channel, and the gray rectangles indicate the areas where outliers are clustered.

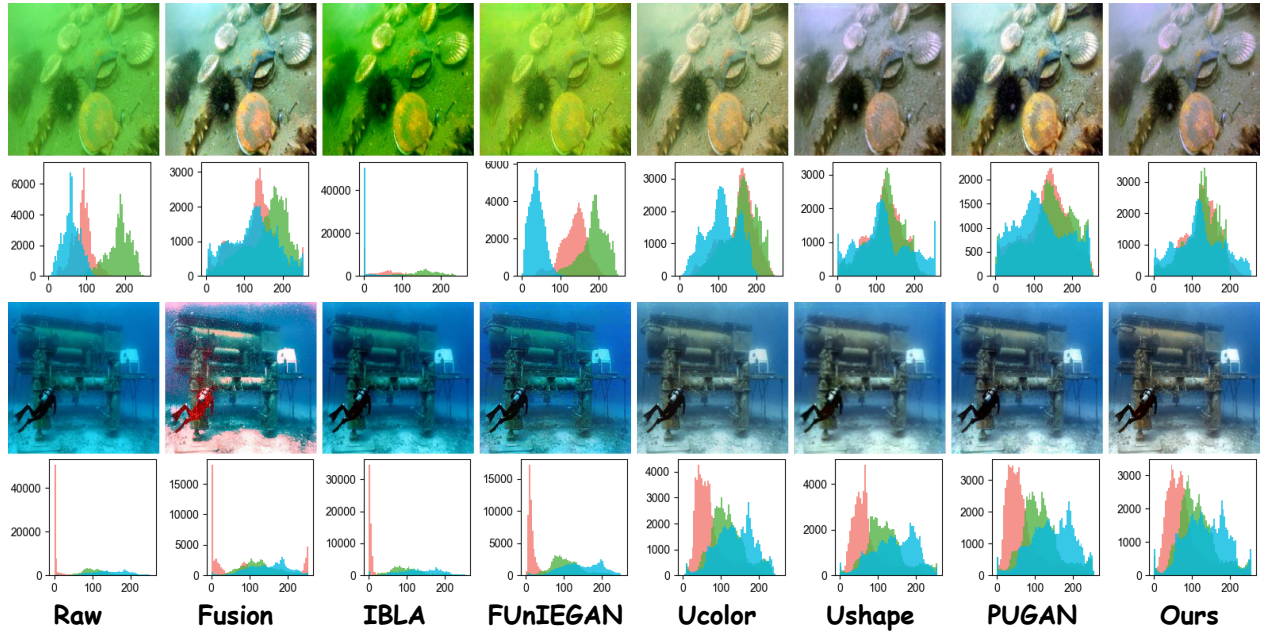


Fig. 10. Enhancement results of two images from the U45 dataset. We display the histogram distribution of the RGB channels.

b) Challenging Scene: The C60 dataset consists of 60 challenging underwater images of very poor quality. As shown in Fig. 9, two underwater images suffer from severe blurriness and abnormal colors, and other UIE methods perform poorly on these two images. For instance, the enhanced visual effect of IBLA is even worse than the original underwater image, and PUGAN’s first enhanced image exhibits significant distortion. In contrast, our method achieves the best visual results. Moreover, the intensity distribution of the RGB channels in our enhanced images is more even and reasonable, with almost no outlier pixels.

c) Severe Color Distortion Scene: There are 45 images suffering from severe color casts in the U45 dataset. As shown in Fig. 10, we present two typical examples: one image exhibits a pronounced green hue, while the other has a strong blue tint. Analyzing the RGB channel histograms reveals that the green channel histogram of the first image is skewed to the right, while the red channel in the second image is notably attenuated. Fusion overcompensates for the red channel, and IBLA fails to correct the color in both images. Our model demonstrates superior capability in suppressing dominant channels and compensating for weaker channels,

TABLE II
QUANTITATIVE COMPARISON OF VARIOUS UIE METHODS ACROSS TWO DATASETS FEATURING DEEP-SEA IMAGES WITH UNEVEN LIGHTING. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Method	UIID					OceanDark			FLOPs↓	#Param.↓	Time↓
	PSNR↑	SSIM↑	UIQM↑	UCIQE↑	PIQE↓	UIQM↑	UCIQE↑	PIQE↓			
ICSP [18]	8.6463	0.3544	1.1620	0.6251	20.4159	2.3674	0.5761	8.9793	-	-	0.034s
L ² UWE [17]	11.1388	0.4964	1.9282	0.6121	11.3622	3.2258	0.5514	6.1632	-	-	2.302s
NUICNet [40]	23.5712	0.8159	2.9312	0.5086	7.4141	2.6533	0.5081	5.6451	49.95G	15.70M	0.014s
IACC [41]	25.0095	0.8895	3.0979	0.5137	8.3629	2.8983	0.5148	5.1617	132.44G	2.10M	0.054s
UIEConv	27.4934	0.9129	3.1545	0.5231	8.5412	2.9005	0.5094	5.5921	121.53G	3.31M	0.020s
Ours	28.6446	0.9203	3.1226	0.5240	8.6548	2.8771	0.5059	5.2309	146.91G	29.55M	0.049s

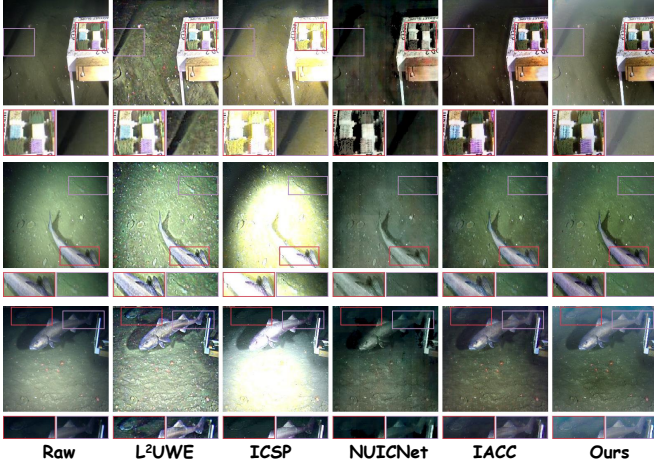


Fig. 11. Enhancement results of three deep-sea images with uneven lighting from the OceanDark dataset. We enlarge the local areas of the enhanced images to compare the details.

thereby effectively correcting color casts. The color distribution of our enhanced images closely approximates that of in-air images, aligning well with the gray world assumption.

d) Deep-sea Scene: For deep-sea scenarios, we train on the UIID training set and test on the UIID test set and OceanDark dataset. Table II compares our method with several UIE methods tailored for deep-sea environments with uneven lighting conditions. Our method achieves superior performance on two full-reference metrics, PSNR and SSIM. While some methods outperform ours on non-reference metrics, a visual comparison in Fig. 11 reveals the opposite. For instance, ICSP excels in the UCIQE metric but produces overexposed images. In contrast, our method enhances brightness in dark regions, mitigating uneven lighting without introducing distortion. This highlights the challenge of accurately evaluating deep-sea image quality using existing non-reference metrics. Developing appropriate non-reference metrics for deep-sea images remains an important research direction.

e) Depth Estimation: We visualize inverse depth maps estimated by several methods in Fig. 12. Methods like GDGP and IBLA, which use visual priors to estimate transmission and depth maps, perform the worst. PUGAN struggles in texture-rich scenes. Depth Anything exhibits strong generalization ability and adapts well to underwater scenes, but its depth maps are overly smooth. In comparison, our depth estimation sub-network produces more detailed depth maps,

TABLE III
QUANTITATIVE COMPARISONS OF DEPTH ESTIMATION ON REAL UNDERWATER IMAGES FROM TWO SUBSETS OF SEA-THRU DATASET. DA IS AN ABBREVIATION FOR DEPTH ANYTHING.

Method	RMSE↓	Abs.Rel↓	log ₁₀ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DA	1.5969	0.4039	0.2444	0.1593	0.3791	0.6628
Ours	1.3811	0.3969	0.2304	0.0938	0.3660	0.7482
DA	4.3064	0.5757	0.4083	0.0676	0.1421	0.2746
Ours	3.8157	0.5361	0.3490	0.0938	0.1922	0.3069

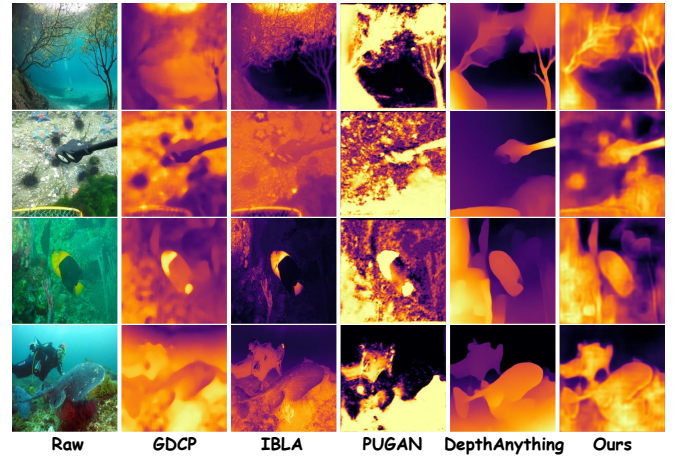


Fig. 12. Visual comparison of depth maps estimated by five different methods.

such as the branches in the first example and the sea urchins in the second example. Additionally, we quantitatively compare the performance of Depth Anything and our method on a real underwater image dataset with depth annotations in Table III. Our method outperforms Depth Anything. Although our depth estimation sub-network is initialized from Depth Anything, through joint training with underwater image enhancement and physical imaging model parameter estimation, our method is better suited for underwater images.

D. Ablation Studies

a) Ablations about the Framework: To further demonstrate the effectiveness of our framework, we use it to train other advanced UIE models. As shown in Table IV, we train UIEC²Net and Ushape on the LSUI dataset, and NUICNet and IACC on the UIID dataset. We find that integrating these models into our framework yields significantly superior results compared to training them alone. For instance, PSNR



Fig. 13. The left panel shows keypoint matching results on raw underwater images and our enhanced images, while the right panel presents the number of matched keypoints for images enhanced using different methods.

TABLE IV
COMPARISON OF SEVERAL ADVANCED UIE MODELS BEFORE AND AFTER INTEGRATION INTO OUR FRAMEWORK. RED NUMBERS INDICATE THE PERFORMANCE IMPROVEMENTS ACHIEVED BY OUR FRAMEWORK.

Dataset	Method	PSNR \uparrow	SSIM \uparrow
LSUI	UIEC ² Net	25.0757	0.8708
	+Ours	27.0192 (+1.9435)	0.9002 (+0.0294)
	UShape	25.7630	0.8296
	+Ours	27.8548 (+2.0918)	0.9068 (+0.0772)
	UIEConv	28.9155	0.9187
	+Ours	29.9253 (+1.0098)	0.9248 (+0.0061)
UIID	NUICNet	23.5712	0.8159
	+Ours	25.2495 (+1.6783)	0.8688 (+0.0529)
	IACC	25.0095	0.8895
	+Ours	27.4576 (+2.4481)	0.9167 (+0.0272)
	UIEConv	27.4934	0.9129
	+Ours	28.6446 (+1.1512)	0.9203 (+0.0074)

TABLE V
ABLATION STUDY OF THE UIECONV MODEL STRUCTURE.

Method	LSUI		T90	
	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
Local Branch	27.9657	0.9080	23.7615	0.9044
Global Branch	25.9986	0.8777	23.8900	0.8977
UIEConv	28.9155	0.9187	24.1197	0.9283

increases by up to 2.4481 and SSIM increases by up to 0.0772. These experimental results indicate that our framework can be integrated with any advanced UIE model, consistently achieving performance enhancements.

b) Ablations about UIEConv: We perform an ablation study on the structure of UIEConv, as presented in Table V. The performance significantly declines when either the global branch or the local branch of UIEConv is removed. This highlights the importance of both global features and local features for image enhancement. Consequently, the dual-branch structure of UIEConv is key to its superiority over other advanced UIE models.

c) Ablations about DDM and Loss Function: We conduct a comprehensive ablation study on the design details of each sub-network within the DDM, as presented in Table VI. "Full Framework" represents the complete version of our proposed framework. We remove each unique design element in the three sub-networks one by one to observe performance changes. For VLEN, we remove its low-pass filter (w/o lowpass) or the subsequent convolutional transformation (w/o

TABLE VI
ABLATION STUDY OF THE DEPTH DEGRADATION MODEL (DDM) STRUCTURE AND LOSS FUNCTION ON THE LUSI DATASET.

Method		PSNR \uparrow	SSIM \uparrow
Full Framework		29.9253	0.9248
VLEN	w/o lowpass	29.1634	0.9212
	w/o transform	28.9733	0.9208
DEN	finetune all	29.5960	0.9221
	freeze all	29.4437	0.9220
	scratch	nan	nan
FEN	$\beta^D = \beta^B$	29.0706	0.9213
	w/o z^{Scale}, z^{Shift}	29.1588	0.9211
	w/o additional inputs	29.4085	0.9218
Loss	w/o \mathcal{L}_{phy}	28.7452	0.9191
	w/o \mathcal{L}_{depth}	29.6402	0.9227

transform). For DEN, we explore three different settings during training: fine-tuning its encoder and decoder (finetune all), freezing its encoder and decoder (freeze all), and training it from scratch without initializing with Depth Anything weights (scratch). For FEN, we experiment with several different structures. First, we retain one head to simultaneously estimate the attenuation coefficient β^D and the scattering coefficient β^B ($\beta^D = \beta^B$). Second, we remove the heads that estimate the scale z^{Scale} and shift z^{Shift} of the relative depth (w/o z^{Scale}, z^{Shift}), and instead use a predefined depth range (0.1m-10m) to scale the relative depth. Finally, we exclude the estimated veiling light B^∞ and inverse depth z as inputs to FEN (w/o additional inputs), using only the underwater image as the input. We find that removing any module or changing the training settings in the sub-networks leads to less accurate estimation performance, thereby reducing enhancement performance. Notably, training DEN from scratch causes the training process to fail to converge (resulting in NaNs), which indicates that the knowledge transferred from Pre-trained Depth Anything is crucial for the training of DEN and the entire framework. Additionally, we study the role of two additional loss functions. Training without the physical constraint loss \mathcal{L}_{phy} (w/o \mathcal{L}_{phy}) significantly decreased performance, highlighting the importance of physical guidance in our framework. Training without the depth consistency loss \mathcal{L}_{depth} (w/o \mathcal{L}_{depth}) also slightly reduced performance.

E. Application Tests

To demonstrate the practicality of our method, we conduct two application tests. We use various UIE methods to enhance



Fig. 14. Results of object detection on raw underwater images and images enhanced using different methods.

underwater images and perform keypoint matching and object detection on both the original and enhanced images. In Fig. 13, we visualize the matched keypoints based on SIFT features. Compared to the raw images and the enhanced images from other methods, our enhanced images obtain the most keypoints. Additionally, we apply the YOLO-World model [66] for underwater object detection, setting the text prompts as *person* and *fish* to enable the detection of these two classes. Our method has the fewest missed detections in Fig. 14.

V. CONCLUSION

In this paper, we introduce a novel physical model-guided framework for underwater image enhancement and depth estimation. This framework leverages the UIEConv model to enhance raw underwater images and employs the Deep Degradation Model (DDM) to estimate various parameters of the physical imaging model, including scene depth. By accurately simulating the degradation process, the physical imaging model bridges the relationship between the enhanced and raw underwater images, guiding the training of both UIEConv and DDM. Extensive experiments across diverse underwater environments validate the effectiveness of our framework, demonstrating significant improvements in both image quality and depth estimation. This robust framework offers a comprehensive solution to underwater imaging challenges, paving the way for further advancements in the field.

REFERENCES

- [1] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice, "Initial results in underwater single image dehazing," in *Oceans 2010 Mts/IEEE Seattle*. IEEE, 2010, pp. 1–8.
- [2] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 825–830.
- [3] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *Journal of Visual Communication and Image Representation*, vol. 26, pp. 132–145, 2015.
- [4] Y.-T. Peng, X. Zhao, and P. C. Cosman, "Single underwater image enhancement using depth estimation based on blurriness," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4952–4956.
- [5] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2019.
- [6] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognition*, vol. 98, p. 107038, 2020.
- [7] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Transactions on Image Processing*, 2023.
- [8] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7159–7165.
- [9] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.
- [10] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Transactions on Image Processing*, vol. 30, pp. 4985–5000, 2021.
- [11] Y. Wang, J. Guo, H. Gao, and H. Yue, "Uiec2-net: Cnn-based underwater image enhancement using two color space," *Signal Processing: Image Communication*, vol. 96, p. 116250, 2021.
- [12] J. Jiang, T. Ye, J. Bai, S. Chen, W. Chai, S. Jun, Y. Liu, and E. Chen, "Five a+ network: You only need 9k parameters for underwater image enhancement," *arXiv preprint arXiv:2305.08824*, 2023.
- [13] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of koschmieder's model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 205–16 215.
- [14] S. Yan, X. Chen, Z. Wu, M. Tan, and J. Yu, "Hybrur: A hybrid physical-neural solution for unsupervised underwater image restoration," *IEEE Transactions on Image Processing*, 2023.
- [15] Z. Fu, H. Lin, Y. Yang, S. Chai, L. Sun, Y. Huang, and X. Ding, "Unsupervised underwater image restoration: From a homology perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 643–651.
- [16] N. Varghese, A. Kumar, and A. Rajagopalan, "Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 248–12 258.
- [17] T. P. Marques and A. B. Albu, "L2uwe: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 538–539.
- [18] G. Hou, N. Li, P. Zhuang, K. Li, H. Sun, and C. Li, "Non-uniform illumination underwater image restoration via illumination channel sparsity prior," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [19] R. Cong, W. Yang, W. Zhang, C. Li, C.-L. Guo, Q. Huang, and S. Kwong, "Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators," *IEEE Transactions on Image Processing*, 2023.
- [20] P. Mu, H. Xu, Z. Liu, Z. Wang, S. Chan, and C. Bai, "A generalized physical-knowledge-guided dynamic model for underwater image enhancement," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7111–7120.
- [21] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6723–6732.
- [22] D. Akkaynak, T. Treibitz, T. Shlesinger, Y. Loya, R. Tamir, and D. Iluz, "What is the space of attenuation coefficients in underwater computer vision?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4931–4940.

- [23] M. S. Hitam, E. A. Awalludin, W. N. J. H. W. Yussof, and Z. Bachok, "Mixture contrast limited adaptive histogram equalization for underwater image enhancement," in *2013 International conference on computer applications technology (ICCAT)*. IEEE, 2013, pp. 1–5.
- [24] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 81–88.
- [25] W. Zhang, Y. Wang, and C. Li, "Underwater image enhancement by attenuated color channel correction and detail preserved contrast enhancement," *IEEE Journal of Oceanic Engineering*, vol. 47, no. 3, pp. 718–735, 2022.
- [26] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 379–393, 2017.
- [27] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Transactions on Image Processing*, vol. 31, pp. 3997–4010, 2022.
- [28] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 4572–4576.
- [29] S. Zhang, T. Wang, J. Dong, and H. Yu, "Underwater image enhancement via extended multi-scale retinex," *Neurocomputing*, vol. 245, pp. 1–9, 2017.
- [30] P. Zhuang, C. Li, and J. Wu, "Bayesian retinex underwater image enhancement," *Engineering Applications of Artificial Intelligence*, vol. 101, p. 104171, 2021.
- [31] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.
- [32] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE transactions on image processing*, vol. 26, no. 4, pp. 1579–1594, 2017.
- [33] 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 Transactions on Image Processing*, vol. 25, no. 12, pp. 5664–5677, 2016.
- [34] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE computer graphics and applications*, vol. 36, no. 2, pp. 24–35, 2016.
- [35] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1682–1691.
- [36] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2822–2837, 2020.
- [37] J. Zhou, Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang, "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *International Journal of Computer Vision*, pp. 1–19, 2023.
- [38] C. Guo, R. Wu, X. Jin, L. Han, W. Zhang, Z. Chai, and C. Li, "Underwater ranker: Learn which is better and how to be better," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 702–709.
- [39] K. A. Skinner, J. Zhang, E. A. Olson, and M. Johnson-Roberson, "UwstereoNet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7947–7954.
- [40] X. Cao, S. Rong, Y. Liu, T. Li, Q. Wang, and B. He, "Nuicnet: Non-uniform illumination correction for underwater image using fully convolutional network," *IEEE Access*, vol. 8, pp. 109 989–110 002, 2020.
- [41] J. Zhou, Q. Gai, D. Zhang, K.-M. Lam, W. Zhang, and X. Fu, "Iacc: Cross-illumination awareness and color correction for underwater images under mixed natural and artificial lighting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.
- [42] W. Zhang, W. Liu, L. Li, H. Jiao, Y. Li, L. Guo, and J. Xu, "A framework for the efficient enhancement of non-uniform illumination underwater image using convolution neural network," *Computers & Graphics*, vol. 112, pp. 60–71, 2023.
- [43] M. Li, K. Wang, L. Shen, Y. Lin, Z. Wang, and Q. Zhao, "Uialn: Enhancement for underwater image with artificial light," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [44] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.
- [45] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.
- [46] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [47] P. Hambarde, S. Murala, and A. Dhall, "Uw-gan: Single-image depth estimation and image enhancement for underwater images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.
- [48] F. Zhang, S. You, Y. Li, and Y. Fu, "Atlantis: Enabling underwater depth estimation with stable diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 852–11 861.
- [49] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 624–628.
- [50] J. Wang, X. Ye, Y. Liu, X. Mei, and J. Hou, "Underwater self-supervised monocular depth estimation and its application in image enhancement," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105846, 2023.
- [51] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2856–2868, 2018.
- [52] M. Zhou, J. Huang, C.-L. Guo, and C. Li, "Fourmer: An efficient global modeling paradigm for image restoration," in *International Conference on Machine Learning*. PMLR, 2023, pp. 42 589–42 601.
- [53] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khaidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby *et al.*, "DinoV2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2023.
- [54] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.
- [55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.
- [56] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [57] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
- [58] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [59] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.
- [60] D. Du, E. Li, L. Si, F. Xu, and J. Niu, "End-to-end underwater video enhancement: Dataset and model," *arXiv preprint arXiv:2403.11506*, 2024.
- [61] H. Li, J. Li, and W. Wang, "A fusion adversarial underwater image enhancement network with a public test dataset," *arXiv preprint arXiv:1906.06819*, 2019.
- [62] T. Porto Marques, A. Branzan Albu, and M. Hoeberechts, "A contrast-guided approach for the enhancement of low-lighting underwater images," *Journal of Imaging*, vol. 5, no. 10, p. 79, 2019.
- [63] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2015.
- [64] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.
- [65] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 twenty first national conference on communications (NCC)*. IEEE, 2015, pp. 1–6.
- [66] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.