An Underwater Image Enhancement Benchmark Dataset and Beyond

Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, Fellow, IEEE, and Dacheng Tao, Fellow, IEEE

Abstract—Underwater image enhancement has been attracting much attention due to its significance in marine engineering and aquatic robot. Numerous underwater image enhancement algorithms have been proposed in the last few years. However, these algorithms are mainly evaluated using either synthetic datasets or few selected real-world images. It is thus unclear how these algorithms would perform on images acquired in the wild and how we could gauge the progress in the field. To bridge this gap, we present the first comprehensive perceptual study and analysis of underwater image enhancement using large-scale real-world degraded images. In this paper, we construct an Underwater Image Enhancement Benchmark Dataset (UIEBD) including 950 real-world underwater images, 890 of which have the corresponding reference images. We treat the rest 60 underwater images which cannot obtain satisfactory references as challenging data. Using this dataset, we conduct a comprehensive study of the state-of-the-art underwater image enhancement algorithms qualitatively and quantitatively. In addition, we propose an end-to-end Deep Underwater Image Enhancement Network (DUIENet) trained on this benchmark as a baseline, which indicates the generalization of the proposed UIBED for training Convolutional Neural Networks (CNNs). The benchmark evaluations and the proposed DUIENet demonstrate the performance and limitations of state-of-the-art algorithms which shed light on the future research in underwater image enhancement. The code and dataset are available at: https://li-chongyi.github.io/homepage.github.io/proj_benchmark.html

Index Terms—underwater image enhancement, real-world underwater images, comprehensive evaluation, deep learning.

I. INTRODUCTION

During the past few years, underwater image enhancement has drawn considerable attention in both image processing and computer vision [1], [2]. Due to the complicated underwater environment and lighting conditions, enhancing underwater image is a challenging problem. Usually, an underwater image is degraded by wavelength-dependent absorption and scattering including forward scattering and backward scattering [3]–[7]. In addition, the marine snow introduces noise and increases the effects of scattering. These adverse effects reduce visibility, decrease contrast, and even introduce color casts, which limits the practical applications of underwater images and videos in marine biology and archaeology [8], marine ecological [9], to name a few. To solve this problem, earlier methods rely on multiple underwater images or polarization filters, while recent algorithms deal with this problem by using only information from a single image.

Despite the prolific work, both the comprehensive study and insightful analysis of underwater image enhancement algorithms remain largely unsatisfactory due to the lack of a publicly available real-world underwater image dataset. Additionally, lacking sufficient and effective training data, the performance of deep learning-based underwater image enhancement algorithms does not match the success of recent deep learning-based high-level [10], [11] and low-level [12], [13] vision problems. To advance the development of underwater image enhancement, we construct a large-scale real-world Underwater Image Enhancement Benchmark Dataset (UIEBD). Several sampling images and the corresponding reference images from UIBED are presented in Fig. 1. As shown, the raw underwater images in the UIEBD have diverse color ranges and degrees of contrast decrease. In contrast, the corresponding reference results are color casts-free (at least relatively genuine color) and have improved visibility and brightness. With the proposed UIBED, we carry out a comprehensive study for several state-of-the-art single underwater image enhancement algorithms both qualitatively and quantitatively, which enables insights into their performance and sheds light on the future research. In addition, with the constructed UIBED, CNNs can be easily trained to improve the visual quality of an underwater image. To demonstrate this application, we propose an underwater image enhancement CNN model trained by the proposed UIBED.

The main contributions of this paper are summarized as follows:

- We construct a large-scale real-world underwater image enhancement benchmark dataset (i.e., UIBED) which contains 950 degraded underwater images. Moreover, the corresponding reference results for 890 images are provided according to laborious, time-consuming, and well-designed pairwise comparisons. UIBED provides a platform to evaluate, at least to some extent, the performance of different underwater image enhancement...
algorithms. It also makes strongly supervised underwater image enhancement models which are out of the constraints of specific underwater scenes possible.

- With the constructed UIEBD, we conduct a comprehensive study of the state-of-the-art single underwater image enhancement algorithms ranging from qualitative to quantitative evaluations. Our evaluations and analyses provide comprehensive insights into the strengths and limitations of current underwater image enhancement algorithms, and suggest new research directions.
- We propose a CNN model (i.e., DUIENet) trained by the constructed UIEBD for underwater image enhancement, which demonstrates the generalization of the constructed UIEBD and the advantages of our DUIENet, and also motivates the development of deep learning-based underwater image enhancement.

II. EXISTING METHODOLOGY, EVALUATION METRIC, AND DATASET: AN OVERVIEW

A. Underwater Image Enhancement Method

Exploring underwater world has become an active issue in recent years [14]–[16]. Underwater image enhancement as an indispensable step to improve the visual quality of recorded images has drawn much attention. A variety of methods have been proposed and can be organized into four groups: supplementary information-based methods, non-physical model-based methods, physical model-based methods, and data-driven methods.

Supplementary information-based methods: In the earlier stage, supplementary information from multiple images [17] or specialize hardware devices (e.g., polarization filtering [18]–[21], range-gated imaging [22], [23], and fluorescence imaging [24], [25]) were utilized to improve the visibility of underwater images. Compared to supplementary information-based methods, single underwater image enhancement has been proven to be more suitable for challenging situations such as dynamic scenes, and thus, gains extensive attention.

Non-physical model-based methods: Non-physical model-based methods aim to modify image pixel values to improve the visual quality. Iqbal et al. [26] stretched the dynamic pixel range in RGB color space and HSV color space to improve the contrast and saturation of an underwater image. Chani and Isa [27], [28] modified the work of [26] to reduce the over/under-enhanced regions by shaping the stretching process following the Rayleigh distribution. Ancuti et al. [29] proposed an underwater image enhancement method by blending a contrast-enhanced image and a color-corrected image in a multi-scale fusion strategy. In [30], a two-step approach for underwater image enhancement was proposed, which includes a color correction algorithm and a contrast enhancement algorithm.

Another line of research tries to enhance underwater images based on the Retinex model. Fu et al. [31] proposed a Retinex-based method for underwater image enhancement, which consists of color correction, layer decomposition, and enhancement. Zhang et al. [32] proposed an extended multi-scale Retinex based underwater image enhancement method. In this work, it is interesting that the underwater turbidity conditions are simulated by using a mixture of whole milk and grape juice in the water.

Physical model-based methods: Physical model-based methods regard the enhancement of an underwater image as an inverse problem, where the latent parameters of an underwater image formation model are estimated from input image. These methods usually follow the same pipeline: 1) building a physical model of the degradation; 2) estimating the unknown model parameters; and 3) addressing this inverse problem.

One line of research is to modify the Dark Channel Prior (DCP) [33] for underwater image enhancement. In [34], DCP was combined with the wavelength-dependent compensation algorithm to restore the underwater image. In [35], an Underwater Dark Channel Prior (UDCP) was proposed based on the fact that the information of the red channel in an underwater image is undependable. Based on the observation that the dark channel of the underwater image tends to be a zero map, Liu and Chau [36] formulated a cost function and minimized it so as to find the optimal transmission map which is able to maximize the image contrast. Instead of the DCP, Li et al. [37] employed the random forest regression model to estimate the medium transmission of the underwater scenes. Recently, Peng et al. [38] proposed a Generalized Dark Channel Prior (GDCP) for image restoration, which incorporates adaptive color correction into an image formation model.

Another line of research is to employ the optical properties of underwater imaging. Carlevaris-Bianca et al. [39] proposed
a prior that exploits the difference in attenuation among three color channels in RGB color space to predict the medium transmission of an underwater scene. The idea behind this prior is that the red light usually attenuates faster than the green light and the blue light in an underwater scenario. Galdran et al. [40] proposed a Red Channel method, which recovers the lost contrast of an underwater image by restoring the colors associated with short wavelengths. According to the findings that the background color of underwater images has relations with the inherent optical properties of water medium, Zhao et al. [41] enhanced the degraded underwater images by deriving inherent optical properties of water from the background color.

Li et al. [42] proposed an underwater image enhancement method based on the minimum information loss principle and histogram distribution prior. Peng et al. [43] proposed a depth estimation method for underwater scenes based on image blurriness and light absorption, which is employed in an underwater image formation model to enhance underwater images. Berman et al. [44] took multiple spectral profiles of different water types into account and reduced the problem of underwater image restoration to single image dehazing by estimating two additional global parameters. Wang et al. [45] combined the adaptive attenuation-curve prior with the characteristics of underwater light propagation for underwater image restoration.

Data-driven methods: Recent years have witnessed significant advance of deep learning in low-level vision problems, including image super-resolution [46]–[48], image denoising [49], [50], image deblurring [51], [52], image dehazing [53], [54], etc. These methods can be trained using synthetic pairs of degraded images and high-quality counterparts. However, underwater image formation models depend on specific scenes and lighting conditions, and even are related to temperature and turbidity. Thus, it is difficult to synthesize realistic underwater images for CNN training. Further, the learned distribution by CNN trained on synthetic underwater images does not always generalize to the real-world underwater images. Therefore, the performance and the amount of deep learning-based underwater image enhancement methods do not match the success of recent deep learning-based low-level vision problems due to the lack of sufficient and effective training data.

Recently, Li et al. [55] proposed a deep learning-based underwater image enhancement model, called WaterGAN. WaterGAN first simulates realistic underwater images from the in-air image and depth pairings in an unsupervised pipeline. With the synthetic training data, the authors use a two-stage network for underwater image restoration, especially for color casts removal. However, the testing data are similar to the training data since the generalization of the WaterGAN is limited due to the training data simulated in specific scenes. More recently, a weakly supervised underwater color transfer model [56] was proposed based on Cycle-Consistent Adversarial Networks [57]. Benefiting from the adversarial network architecture and multi-term loss function, this network model relaxes the need for paired underwater images for training and allows the underwater images being taken in unknown locations. However, it tends to produce inauthentic results in some cases due to the nature of multiple possible outputs. In [58], Hou et al. synthesized underwater images by setting random parameters of an underwater image formation using the prior of wavelength-dependent attenuation and developed an underwater residual CNN model to learn the transmission. However, the employed underwater image formation cannot hold well in real cases. Therefore, the robustness and generalization of deep learning-based underwater enhancement methods still fall behind conventional state-of-the-art algorithms.

B. Underwater Image Quality Evaluation

Different from other low-level vision problems where the ground truth can be easily obtained (e.g., image super-resolution), it is challenging to achieve a large amount of paired degraded underwater images and the corresponding ground truth. In the following, we will give a brief introduction of the image quality evaluation metrics which are usually used for the performance evaluation of underwater image enhancement methods.

Full-reference metrics: For an underwater image with ground truth, the full-reference image quality evaluation metrics (e.g., MSE, PSNR, and SSIM [59]) were employed for evaluation. Such underwater images usually are a few color checker images or color image patches taken in the simulated or real underwater environment. For example, Zhao et al. [41] treated a plastic color disk as ground truth and captured its underwater image in a water pool as the testing image.

Non-reference metrics: For a real-world underwater image where the ground truth was unavailable, non-reference image quality evaluation metrics, such as image entropy, visible edges [60], dynamic range independent image quality assessment [61], were utilized. In addition, some authors employed specific applications, like feature point matching, edge detection, and image segmentation, to evaluate their results. Recently, several specific non-reference metrics were proposed for underwater images evaluation. Yang and Sowmya [62] proposed an underwater color image quality evaluation metric (i.e., UCIQE). UCIQE first quantifies the non-uniform color casts, blurring, and low contrast, and then combines these three components in a linear manner, which has been widely used. In [63], the authors proposed a non-reference underwater image quality measure, called UIQM, which comprises three attribute measures: a colorfulness measure, a sharpness measure, and a contrast measure. Each presented attribute measure is inspired by the properties of the human visual system.

C. Underwater Image Dataset

There are several real-world underwater image datasets such as Fish4Knowlege dataset for underwater target detection and recognition1, underwater images in SUN dataset for scene recognition and object detection2, and MARIS dataset for marine autonomous robotics3. However, existing datasets are not suitable for the enhancement task due to the monotonous

---

1. http://groups.inf.ed.ac.uk/f4k/
2. http://groups.csail.mit.edu/vision/SUN/
3. http://rimlab.ce.unipr.it/Maris.html
content and limited scenes, few degradation characteristics, and insufficient data. Besides, to synthesize underwater images, Li et al. [55] proposed a GAN based method while Duarte et al. [65] simulated underwater image degradation using milk, chlorophyll, or green tea in a tank. Blasinski et al. [66] provided an open-source underwater image simulation tool and a three parameter underwater image formation model [67]. However, there still exists a gap between the synthetic and real-world underwater images. Therefore, it is difficult to evaluate the state-of-the-art underwater image enhancement methods fairly and comprehensively, and is hard to develop effective deep learning-based underwater image enhancement models without reference images.

III. PROPOSED BENCHMARK DATASET

After systematically reviewing previous work, we found the main issue existing in the community of underwater image enhancement is lacking a large-scale real-world underwater image dataset with reference images. In what follows, we introduce the constructed dataset in detail, including data collection and reference image generation.

A. Data Collection

There are three objectives for underwater image collection:

1) a diversity of underwater scenes, different characteristics of quality degradation, and a broad range of image content should be covered;
2) the amount of degraded underwater images should be large; and
3) the corresponding high-quality reference images should be provided so that pairs of images enable fair image quality evaluation and end-to-end learning.

To achieve the first two objectives, we first collect a large number of underwater images, and then refine them. We mainly retain the underwater images which meet the first objective. After data refinement, most of the collected images are weeded out, and about 950 candidate images are remaining. The resolution of the remaining images ranges from $183 \times 275$ to $1350 \times 1800$. Fig. 2 gives some examples of the images in the conducted UIEBD. These images have obvious characteristics of underwater image quality degradation (e.g., color casts, decreased contrast, and blurring details) and are taken in a diversity of underwater scenes.

![Fig. 2. Examples of the images in the conducted UIEBD. These images have obvious characteristics of underwater image quality degradation (e.g., color casts, decreased contrast, and blurring details) and are taken in a diversity of underwater scenes.](image)

Fig. 2. Examples of the images in the conducted UIEBD. These images have obvious characteristics of underwater image quality degradation (e.g., color casts, decreased contrast, and blurring details) and are taken in a diversity of underwater scenes.

real-world underwater images. All the employed methods except the fusion-based [29] method are realized by the source codes provided by their authors. We reimplement the fusion-based [29] method since the source code is unavailable. For the dive+, we tune its parameter settings to generate satisfactory results. At last, we totally generate $12 \times 950$ enhanced results.

![Fig. 3. An example of pair comparisons. Treating the raw underwater image as reference, a volunteer needs to independently decide which one is better between the results of method A and method B.](image)

Fig. 3. An example of pair comparisons. Treating the raw underwater image as reference, a volunteer needs to independently decide which one is better between the results of method A and method B.

With raw underwater images and the enhanced results, we invite 50 volunteers (25 volunteers with image processing experience; 25 volunteers without related experience) to perform pairwise comparisons among the 12 enhanced results of each raw underwater image under the same monitor. The pairwise comparisons have been proven to be more robust and consistent in subjective perception than individual rating [68], [69], though it is laborious and time-consuming.

Specifically, each volunteer is shown a raw underwater image and a set of enhanced result pairs. The enhanced image pairs are drawn from all the competitive methods randomly, and the result winning the pairwise comparisons will be compared again in the next round, until the best one is selected. There is no time constraint for volunteers and zoom-in operation is allowed. An example of pair comparisons is shown in Fig. 3. For each pair of enhanced results, taking the raw underwater image as reference, a volunteer first needs to independently decide which one is better than the other. Therefore, for each volunteer, the best result of a raw underwater image
will be selected after 11 pair comparisons. In addition, the volunteer needs to inspect the best result again and then label the best result as being satisfactory or dissatisfactory. At last, the reference image for a raw underwater image is first selected by majority voting after pair comparisons. After that, if the selected reference image has greater than half the number of votes labeled dissatisfaction, its corresponding raw underwater image is treated as a challenging image and the reference image is discarded. We totally achieve 890 available reference images which have higher quality than any individual methods and a challenging set including 60 underwater images. To visualize the process of reference image generation, we present some cases that the results of some methods are shown and indicate which one is the final reference image in Fig. 4. Furthermore, the percentage of the reference images from the results of different methods is present in Table I. In this paper, we highlight the top 1 performance in red, whereas the second top one is in blue.

| Method               | Percentage (%) |
|----------------------|-----------------|
| Fusion-based [29]    | 24.72           |
| Two-step-based [30]  | 7.30            |
| Retinex-based [31]   | 0.22            |
| DCP [33]             | 2.58            |
| UDCP [35]            | 0.00            |
| Regression-based [37]| 1.80            |
| GDCP [38]            | 0.34            |
| Red Channel [40]     | 0.90            |
| Histogram Prior [42] | 13.37           |
| Blurriness-based [43]| 3.48            |
| MSCNN [53]           | 0.90            |
| dive+                | 43.93           |

In summary, the results with improved contrast and genuine color are most favored by observers while the over/under enhancement, artifacts, and color casts lead to visually unpleasing results. Finally, the constructed UIEBD includes two subsets: 890 raw underwater images with the corresponding high-quality reference images; 60 challenging underwater images. To the best of our knowledge, it is the first real-world underwater image dataset with reference images so far. We will periodically update the results for noticeable underwater image enhancement methods. The UIEBD has various potential applications, such as performance evaluation and CNN training. We will introduce these two applications in the next sections.

IV. EVALUATION AND DISCUSSION

A comprehensive and fair evaluation of underwater image enhancement methods has long been missing from the literature. Using the constructed UIEBD, we evaluate the state-of-the-art underwater image enhancement methods (i.e., Fusion-based [29], Two-step-based [30], Retinex-based [31], UDCP [35], Regression-based [37], GDCP [38], Red Channel [40], Histogram Prior [42], Blurriness-based [43]) both qualitatively and quantitatively.

A. Qualitative Evaluation

We first extract several underwater images from the UIEBD, and then divide these images into four categories: greenish underwater images, bluish underwater images, shallow water images, and underwater image with limited illumination. The results of different methods and the corresponding reference images are shown in Figs. 5-8. Note that these underwater images cannot cover the entire UIEBD.

In general, the red light first disappears in water because of its longest wavelength, followed by the green light and then the blue light. Such selective attenuation in water results in the greenish or bluish underwater images, such as the raw underwater images in Figs. 5 and 6. Color deviation seriously affects the visual quality of underwater images and is difficult to be removed. As shown in Figs. 5 and 6, Fusion-based [29], Retinex-based [31], and the histogram distribution prior [42], Retinex-based [31] well deals with the color deviation, while UDCP [35] and GDCP [38] tend to aggravate the effect of the color casts. Two-step-based [30] can effectively increase the contrast of underwater images. Red Channel [40] and Blurriness-based [43] have the less positive effect on the greenish and bluish underwater images on account of the limitations of the priors used in these methods. In fact, it is...
Fig. 5. Subjective comparisons on greenish underwater images. From left to right are raw underwater images, and the results of Fusion-based [29], Retinex-based [31], Two-step-based [30], UDCP [35], Red Channel [40], Histogram Prior [42], Regression-based [37], Blurriness-based [43], GDCP [38], and reference images. Best viewed with zoom-in on a digital display.

Fig. 6. Subjective comparisons on bluish underwater images. From left to right are raw underwater images, and the results of Fusion-based [29], Retinex-based [31], Two-step-based [30], UDCP [35], Red Channel [40], Histogram Prior [42], Regression-based [37], Blurriness-based [43], GDCP [38], and reference images. Best viewed with zoom-in on a digital display.

Fig. 7. Subjective comparisons on shallow water images. From left to right are raw underwater images, and the results of Fusion-based [29], Retinex-based [31], Two-step-based [30], UDCP [35], Red Channel [40], Histogram Prior [42], Regression-based [37], Blurriness-based [43], GDCP [38], and reference images. Best viewed with zoom-in on a digital display.

Fig. 8. Subjective comparisons on underwater images with limited illumination. From left to right are raw underwater images, and the results of Fusion-based [29], Retinex-based [31], Two-step-based [30], UDCP [35], Red Channel [40], Histogram Prior [42], Regression-based [37], Blurriness-based [43], GDCP [38], and reference images. Best viewed with zoom-in on a digital display.
almost impossible for a color correction algorithm or a kind of prior effective for all types of underwater images.

Images taken in shallow water usually have sufficient illumination as shown in Fig. 7. The methods such as Fusion-based [29], Retinex-based [31], Histogram Prior [42], Regression-based [37], and Blurriness-based [43] significantly remove the effect of haze on the underwater images while UDCP [35] and GDCP [38] bring in obvious color deviation. Two-step-based [30] and Red Channel [40] leave the haze on the results. Compared to other results, the enhanced images by the Fusion-based [29] are visually pleasing.

Different from shallow water images, the images with limited illumination look dark. Generally, such underwater images are hard to be enhanced or recovered. In Fig. 8, Fusion-based [29] and Histogram Prior [42] significantly increase the brightness and contrast of underwater images, which benefits from the histogram modification strategy used in these two methods.

In summary, Fusion-based method [29] has relatively good performance on a variety of underwater images. UDCP [35] tends to produce artifacts on enhanced results in some cases. The other compared methods are effective to some extent.

### B. Quantitative Evaluation

To quantitatively evaluate the performance of different methods, we perform the full-reference evaluation, non-reference evaluation, and running time evaluation.

1) **Full-reference Evaluation:** We first conduct a full-reference evaluation using three commonly-used metrics (i.e., MSE, PSNR, and SSIM). We treat the reference images as “ground truth”. The results of full-reference image quality evaluation can provide realistic feedback of the performance of different methods to some extent, although the real ground truth might be different from the reference images. A higher PSNR score and a lower MSE score denote the result is closer to the reference image in terms of image content, while a higher SSIM score means the result is more similar to the reference image in terms of image structure and texture. We present the average scores of different methods on the 890 images with reference images in the UIEBD. As shown in Table II, Fusion-based [29] stands out as the best performer across all metrics. In addition, Two-step-based [30] rank the second best in terms of the full-reference metrics.

| Method              | MSE ($\times 10^3$) | PSNR (dB) | SSIM   |
|---------------------|---------------------|-----------|--------|
| Fusion-based [29]   | 0.8679              | 18.7461   | 0.8162 |
| Two-step-based [30] | 1.1146              | 17.6596   | 0.7199 |
| Retinex-based [31]  | 1.3351              | 16.8757   | 0.6233 |
| UDCP [35]           | 5.1300              | 11.0296   | 0.4999 |
| Regression-based [37]| 1.1365              | 17.5751   | 0.6543 |
| GDCP [38]           | 3.6345              | 12.5264   | 0.5503 |
| Red Channel [40]    | 2.1073              | 14.8935   | 0.5973 |
| Histogram Prior [42]| 1.6282              | 16.0137   | 0.5888 |
| Blurriness-based [43]| 1.5826             | 16.1371   | 0.6582 |

2) **Non-reference Evaluation:** We employ two non-reference metrics (i.e., UCIQE [62] and UIQM [63]) which are usually used for underwater image quality evaluation [37], [38], [42], [43]. A higher UCIQE score indicates the result has better balance among the chroma, saturation, and contrast, while a higher UIQM score indicates the result is more consistent with the human visual perception. The average scores are shown in Table III.

| Method              | UCIQE [62] | UIQM [63] |
|---------------------|------------|-----------|
| Fusion-based [29]   | 0.6414     | 1.3310    |
| Two-step-based [30] | 0.5776     | 1.4002    |
| Retinex-based [31]  | 0.6026     | 1.4338    |
| UDCP [35]           | 0.5852     | 1.6297    |
| Regression-based [37]| 0.5971    | 1.2996    |
| GDCP [38]           | 0.5993     | 1.4301    |
| Red Channel [40]    | 0.5421     | 1.2147    |
| Histogram Prior [42]| 0.6778     | 1.5440    |
| Blurriness-based [43]| 0.6001    | 1.3757    |

In Table III, Histogram Prior [42] and UDCP [35] obtain the highest scores of UCIQE and UIQM, respectively. It is interesting that the good performers in terms of UCIQE and UIQM metrics are not consistent with the subjective pair comparisons, though both UCIQE and UIQM claim that they take the human visual perception into account. In addition, Figs. 5-8 show the results generated by Histogram Prior [42] and UDCP [35] still suffer from color casts and over-enhancement, and are not as good as those of Fusion-based [29]. Through further analyzing, we found these two non-reference metrics (i.e., UCIQE and UIQM) might be biased to some characteristics (not entire image) and did not take the color shift and artifacts into account. For example, the results with high contrast (e.g., the results of Histogram Prior [42]) are usually favored by the UCIQE metric. To illustrate this phenomenon, we present an example in Fig. 9.

![Visual comparisons in terms of UCIQE and UIQM metrics.](a) Raws (b) Histogram Prior [42] (c) dive+)

Fig. 9. Visual comparisons in terms of UCIQE and UIQM metrics. From (a) to (c) are the raw underwater images, the results of Histogram Prior [42] and dive+. Higher scores are in red. It is obvious that higher quantitative scores do not lead to better subjectively quality.

In Fig. 9, the results generated by Histogram Prior [42] have obvious reddish color shift and artifacts; however, they...
obtain better quantitative scores in terms of UCIQE and UIQM metrics than the results of dive+. Thus, we believe there is a gap between the quantitative scores of non-reference metrics and the subjectively visual quality. In other words, the current image quality evaluation metrics designed for the underwater image have limitations in some cases.

3) Running Time Evaluation: We compare the average running time of different methods for the images of different sizes. Experiments are conducted using MATLAB R2014b on a PC with an Intel(R) i7-6700 CPU, 32GB RAM. The average running time is shown in Table IV. Two-step-based [30] is the fastest across different image sizes, while Retinex-based [31] ranks the second fastest. Regression-based [37] is the slowest method due to the time-consuming random forest-based transmission prediction, especially for images with large sizes, which limits its promising applications.

| Method                  | 500 × 500 | 640 × 480 | 1280 × 720 |
|-------------------------|-----------|-----------|------------|
| Fusion-based [29]       | 0.6904    | 0.6798    | 1.8431     |
| Two-step-based [30]     | 0.6975    | 0.8829    | 2.1089     |
| Retinex-based [31]      | 2.6288    | 3.3165    | 9.9019     |
| UDCP [35]               | 138.6138  | 167.1711  | 415.4935   |
| Regression-based [37]   | 3.2676    | 3.8974    | 9.5934     |
| GDCP [38]               | 2.7523    | 3.2503    | 9.7447     |
| Red Channel [40]        | 4.6284    | 5.8289    | 16.9229    |
| Histogram Prior [42]    | 37.0018   | 47.2538   | 146.0253   |
| Blurriness-based [43]   | 0.6044    | 0.6798    | 1.8431     |

After reviewing and evaluating the state-of-the-art underwater image enhancement methods, we found that Fusion-based [29] is the relatively best performer in most cases, while other compared methods have obvious disadvantages. However, there is no method which always wins when facing a large-scale real-world underwater image dataset (i.e., UIEBD). All in all, due to neglecting the underwater imaging physical models, the non-physical model-based methods, such as Two-step-based [30] and Retinex-based [31], produce over-enhanced or under-enhanced results. Physical model-based methods such as UDCP [35] employ an outdoor haze formation based model to predict the medium transmission which is not well-suited for the underwater scenario. Inaccurate physical models and assumptions result in color casts and remaining haze in the results of physical model-based methods such as Regression-based [37], GDCP [38], Red Channel [40], Histogram Prior [42], and Blurriness-based [43]. Some methods such as Retinex-based [31] and Histogram Prior [42] tend to introduce noise and artifacts, which leads to visually unpleasing results. The running time of some methods seriously limits their practical applications.

In the future, a comprehensive method that can robustly deal with a variety of underwater image degradation is expected. The non-reference metrics which are more effective and consistent with human visual perception are desired in the community of underwater image enhancement.

V. PROPOSED MODEL

Despite the remarkable progress of underwater image enhancement methods, the generalization of deep learning-based underwater image enhancement models still falls behind the conventional state-of-the-art methods due to the lack of effective training data and well-designed network architectures for underwater image enhancement. With the constructed UIEBD, we propose a CNN model for underwater image enhancement, called DUIENet. The purpose of the proposed DUIENet as a baseline is to call for the development of deep learning-based underwater image enhancement, and demonstrate the generalization of the constructed UIEBD for training CNN. Note that the proposed DUIENet is only a baseline model which can be further improved by well-designed network architectures, task-related loss functions, and the like.

In this section, we first present input generation from an underwater image and the architecture of the proposed DUIENet. Then we present the training and implementation details. At last, we perform experiments to demonstrate the advantage of the proposed DUIENet. All the source code of the proposed deep model will be made available to the public.

A. Input Generation

As discussed in Section IV, there is no algorithm generalized to all types of underwater images due to the complicated underwater environment and lighting conditions. In general, the Fusion-based method [29] achieves decent results, which benefits from the inputs derived by multiple pre-processing operations and a fusion strategy. In the proposed DUIENet, we also employ such a manner. Based on the characteristics of underwater image degradation, we generate three inputs by respectively applying White Balance (WB), Histogram Equalization (HE) and Gamma Correction (GC) algorithms to an underwater image. Specifically, WB algorithm is used to correct the color casts of an underwater image, while HE and GC algorithms aim to improve the contrast and light up dark regions, respectively. We directly employ the WB algorithm in [29], whose effectiveness has been turned out. For the HE algorithm, we apply the adapthisteq function [70] provided by MATLAB to the L component in Lab color space, and then transform back into RGB color space. We set the gamma value of GC algorithm as 0.7 empirically.

B. Network Architecture

DUIENet employs a gated fusion network architecture to learn three confidence maps which will be used to combine the three input images into an enhanced result. The learned confidence maps determine the most significant features of inputs remaining in the final result. Gated fusion network architecture has not been explored in the context of underwater image enhancement. In addition, the impressive performance of the fusion-based underwater image enhancement method [29] also encourages us to explore the fusion-based networks.

The architecture of the proposed DUIENet and parameter settings are shown in Fig. 10. As a baseline model, the proposed DUIENet is a plain fully CNN. We believe that the
widely used backbones such as the U-Net architecture [71] and the residual network architecture [72] can be incorporated to improve the performance. We leave it as the future work. We feed the three derived inputs and original input to the proposed DUIENet with aforementioned parameter settings took roughly 3 hours. With an NVIDIA GeForce GTX 1080 Ti GPU, the proposed DUIENet can process an image with a size of 640 × 1080 in less than one second. During training, a batch-mode learning method with a batch size of 16 was applied. The filter weights of each layer were initialized by “Xavier” [76]. Bias was initialized as a constant. We used  ADAM [77] with default parameters for our network optimization. We initialized the learning rate to 1e−3 and decreased the learning rate by 0.1 every 10,000 iterations until the proposed DUIENet converges.

We implemented the proposed DUIENet with TensorFlow on a PC with an Intel(R) i7 6700 CPU, 32GB RAM, and an NVIDIA GeForce GTX 1080 Ti GPU. During training, a batch-mode learning method with a batch size of 16 was applied. The filter weights of each layer were initialized by “Xavier” [76]. Bias was initialized as a constant. We used ADAM [77] with default parameters for our network optimization. We initialized the learning rate to 1e−3 and decreased the learning rate by 0.1 every 10,000 iterations until the proposed DUIENet converges.

The training for the proposed DUIENet with aforementioned parameter settings took roughly 6 hours. With an NVIDIA GTX 1080 Ti GPU, the proposed DUIENet can process an image with a size of 640 × 480 within 0.128s (8FPS), which is very fast in practical applications.

D. Experiments

To demonstrate the advantages achieved by the proposed DUIENet, we compare it against five state-of-the-art underwater image enhancement methods. The experiments are conducted on the testing set which includes 90 underwater images and the challenging set including 60 underwater images. We show several results in Figs. 11 and 12.
Fig. 11. Subjective comparisons on underwater images from testing set. From left to right are raw underwater images, and the results of Fusion-based [29], Retinex-based [31], Histogram Prior [42], Blurriness-based [43], GDCP [38], the proposed DuwieNet, and reference images.

As shown in Fig. 11, the proposed DuwieNet effectively removes the haze on the underwater images and remits color casts, while the compared methods introduce unexpected colors (e.g., Fusion-based [29], GDCP [38], and Histogram Prior [42]) and artifacts (e.g., Fusion-based [29], Retinex-based [31], and Histogram Prior [42]) or have little effect on inputs (e.g., Blurriness-based [43]). In addition, it is interesting that our results even achieve better visual quality than the corresponding reference images (e.g., less noise and better details). This is because that the perceptual loss-based DuwieNet can learn the potential attributes of good visual quality from the constructed large-scale real-world underwater image dataset. For the results of different methods on challenging set shown in Fig. 12, the proposed DuwieNet produces visually pleasing results. By contrast, other methods tend to introduce artifacts, over-enhancement (e.g., the backgrounds and the foregrounds), and color casts (e.g., the reddish or greenish color).

Table V reports the quantitative results of different methods in terms of MSE, PSNR, and SSIM on the testing set. The quantitative results are obtained by comparing the results of each method with the corresponding reference image. We discard the non-reference metrics designed for underwater image enhancement based on the conclusion presented in Sec. IV. Our method achieves the best performance in terms of full-reference image quality assessment. In addition, instead of pairwise comparison, we conduct a user study to score the visual quality of the results on challenging set. This is because some images in the challenging set are too difficult to
obtain satisfactory results following the procedure of reference image generation. Thus, we invited 50 participants (the same volunteers with the reference image generation) to score results. The scores have five scales ranging from 5 to 1 which represent “Excellent”, “Good”, “Fair”, “Poor” and “Bad”, respectively. The average scores of the results by each method on challenging set are shown in Table VI. Besides, we also provide the standard deviation of the results by each method on challenging set. Our method receives the highest average score and lowest standard deviation, which indicates our method produces better results from a subjective perspective and has more robust performance than several state-of-the-art methods.

**TABLE VI**

| Method            | Average Score | Standard Deviation |
|-------------------|---------------|--------------------|
| Fusion-based [29] | 2.28          | 0.8475             |
| Retinex-based [31]| 2.23          | 0.8720             |
| GDCP [38]         | 1.90          | 0.8099             |
| Histogram Prior   | 2.08          | 0.7897             |
| Blurriness-based  | 2.02          | 0.7762             |
| DuwieNet          | 2.57          | 0.7280             |

Qualitative and quantitative experiments demonstrate the effectiveness of the proposed DUIENet and also indicate the constructed dataset can be used for training CNN. However, there is still room for the improvement of underwater image enhancement. In addition, underwater images in the challenging set still cannot be enhanced well. We look forward to more effective and efficient methods.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have constructed an underwater image enhancement benchmark dataset which offers large-scale degraded underwater images and the corresponding reference images. This benchmark dataset enables us to comprehensively study the existing underwater image enhancement methods, and easily train CNN for underwater image enhancement. As analyzed in qualitative and quantitative evaluations, there is no method which always wins in terms of full- and no-reference metrics. In other words, there is still much room for the improvement of underwater image enhancement. In addition, effective non-reference underwater image quality evaluation metrics are highly desirable. To promote the development of deep learning-based underwater image enhancement methods, we proposed an underwater image enhancement CNN model trained by the constructed dataset. Experimental results demonstrate the proposed CNN model performs favorably against the state-of-the-art methods, and also verify the generalization of the constructed dataset for training CNN. In the future work, we will extend the constructed dataset towards more challenging underwater images and underwater videos. Moreover, we tend to investigate more effective underwater image enhancement CNN models which take advantages of the prior information of underwater imaging.

REFERENCES

[1] J. Jaffe, “Underwater optical imaging: The past, the present, and the prospects,” *IEEE J. Oceanic Eng.*, vol. 40, no. 3, pp. 683-700, 2015.

[2] M. Sheinin and Y. Schechner, “The next best underwater view,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2016, pp. 3764-3773.

[3] B. McGlamery, “A computer model for underwater camera systems,” in *Ocean Optics VI*, pp. 221-231, 1980.

[4] J. Jaffe, “Computer modeling and the design of optimal underwater imaging systems,” *IEEE J. Oceanic Eng.*, vol. 15, no. 2, pp. 101-111, 1990.

[5] W. Hou, S. Woods, E. Jarosz, et al., “Optical turbulence on underwater image degradation in natural environments,” *Appl. Opt.*, vol. 15, no. 14, pp. 2678-2686, 2012.

[6] D. Akkaynak and T. Treibitz, “A revised underwater image formation model,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2017, pp. 6723-6732.

[7] D. Akkaynak, T. Treibitz, T. Shlesinger, et al., “What is the space of attenuation coefficients in underwater computer vision,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2017, pp. 568-577.

[8] M. Ludvigsen, B. Sortland, G. Johnsen, et al., “Applications of georeferenced underwater photo mosaics in marine biology and archaeology,” *J. Oceanography*, vol. 20, no. 4, pp. 140-149, 2007.

[9] N. Strachan, “Recognition of fish species by colour and shape,” *J. Image Vis. Comput.*, vol. 11, no. 1, pp. 2-10, 1993.

[10] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[11] T. He, X. Zhang, S. Ren, et al., “Deep residual learning for image recognition,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2016, pp. 770-778.

[12] R. Li, K. Li, Y. C, et al., “Referring image segmentation via recurrent refinement networks,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2018, pp. 5745-5753.

[13] K. Zhang, W. Zuo, and S. Gu, “Learning deep CNN denoiser prior for image restoration,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2017, pp. 2808-2817.

[14] R. Schettini and S. Corchs, “Underwater image processing: State of the art of restoration and image enhancement methods,” *J. Advan. Signal Process.*, vol. 2010, pp. 1-14, 2010.

[15] M. Han, Z. Lyu, and T. Qiu, “A review on intelligence dehazing and color restoration for underwater images,” *IEEE Trans. Syst., Man, Cybern., Syst.*, pp. 1-13, 2018 (Early Access).

[16] R. Cui, L. Chen, C. Yang, et al., “Extended state observer-based integral sliding mode control for an underwater robot with unknown disturbances and uncertain nonlinearities,” *IEEE Trans. Ind. Electron.*, vol. 64, no. 8, pp. 6785-6795, 2017.

[17] S. Narasimhan and S. Nayar, “Contrast restoration of weather degraded images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25 no. 6, pp. 713-724, 2003.

[18] N. Shashar, R. Hanlon, and A. deM. Petz, “Polarization vision helps detect transparent prey,” *Nature*, vol. 393 no. 6682, pp. 222-223, 1998.

[19] Y. Schechner and N. Karpel, “Clear underwater vision,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2004, pp. 536-543.

[20] Y. Schechner and N. Karpel, “Recovery of underwater visibility and structure by polarization analysis,” *IEEE J. Oceanic Eng.*, vol. 30, no. 3, pp. 570-587, 2005.

[21] T. Treibitz and Y. Schechner, “Active polarization descattering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 385-399, 2009.

[22] C. Tan, G. Sluzek, and D. He, “A novel applications of range-gated underwater laser imaging system in near target turbid medium,” *Opticas and Lasers Eng.*, vol. 43, no. 9, pp. 995-1009, 2005.

[23] H. Li, X. Wang, T. Bai, et al., “Speckle noise suppression of range gated underwater imaging system,” *Applid Optics*, vol. 38, no. 18, pp. 3937-3944, 2009.

[24] T. Treibitz and Y. Schechner, “Turbid scene enhancement using multi-directional illumination fusion,” *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4662-4667, 2012.

[25] Z. Murez, T. Treibitz, R. Ramamoorthi, et al., “Photometric stereo in a scattering medium,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2015, pp. 3415-3423.

[26] K. Iqbal, M. Odetayo, and A. James, “Enhancing the low quality images using unsupervised colour correction method,” in *Proc. of IEEE Int. Conf. Syst., Man, Cybern.*, 2010, pp. 1703-1709.

[27] A. Ghani and N. Isa, “Underwater image quality enhancement through integrated color model with Rayleigh distribution,” *Appl. Soft Comput.*, vol. 27, pp. 219-230, 2015.

[28] A. Ghani and N. Isa, “Enhancement of low quality underwater image through integrated global and local contrast correction,” *Appl. Soft Comput.*, vol. 37, pp. 332-344, 2015.
C. Ancuti, C. O. Ancuti, and P. Bekaert, “Enhancing underwater images and videos by fusion,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2012, pp. 81-88.

X. Fu, Z. Fan, and M. Ling, “Two-step approach for single underwater image enhancement,” in *Symposium of IEEE Intel. Signal Process. (SISP)*, 2017, pp. 789-794.

X. Fu, P. Zhang, Y. Huang, et al., “A retina-rebased enhancing approach for single underwater image,” in *Proc. of IEEE Int. Conf. Image Process. (ICIP)*, 2014, pp. 4572-4576.

S. Zhang, T. Wang, J. Dong, et al., “Underwater image enhancement via extended multi-scale Retinex,” *Neurocomputu.*, vol. 245, no. 5, pp. 1-9, 2017.

K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341-2353, 2011.

J. Chiang and Y. Chen, “Underwater image enhancement by wavelength compensation and dehazing,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1756-1769, 2012.

P. Drews-Jr, E. Nascimento, S. Botelho, et al., “Underwater depth estimation and image restoration based on single images,” *IEEE Comput. Graph. Appl.*, vol. 36, no. 2, pp. 24-35, 2016.

H. Liu and L. Chau, “Underwater image restoration based on contrast enhancement,” in *Proc. of IEEE Int. Digit. Signal Process.*, 2016, pp. 584-588.

C. Li, J. Guo, C. Guo, et al., “A hybrid method for underwater image correction,” *Pattern Rec. Lett.*, vol. 94, pp. 64-67, 2017.

X. Fu, T. Cao, and P. Cosman, “Generalization of the dark channel prior for single image restoration,” *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856-2868, 2018.

N. Carlevatis-Bianco, A. Mohan, and R. Eustice, “Initial results in underwater single image dehazing,” in *Proc. of IEEE Int. Conf. Oceans*, 2010, pp. 1-8.

A. Galdran, D. Pardo, and A. Picc, “Automatic Red-Channel underwater image restoration,” *J. Vis. Commu. and Image Repre.*, vol. 26, pp. 132-145, 2015.

X. Zhao, T. Jin, and S. Qu, “Deriving inherent optical properties from background color and underwater image enhancement,” *Ocean Eng.*, vol. 94, pp. 163-172, 2015.

C. Li, J. Guo, R. Cong, et al., “Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior,” *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664-5677, 2016.

Y. Peng and P. Cosman, “Underwater image restoration based on image blurriness and light absorption,” *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1579-1594, 2017.

D. Berman, T. Treibitz, and S. Avidan, “Diving into Haze-lines: Color restoration of underwater images,” in *Proc. of Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1-12.

Y. Wang, H. Liu, and L. Chau, “Single underwater image restoration using adaptive attenuation-curve prior,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 3, pp. 992-1002, 2018.

C. Dong, C. Loy, K. He, et al., “Learning a deep convolutional network for image super-resolution,” in *Proc. of Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 184-199.

C. Dong, C. Loy, K. He, et al., “Image super-resolution using deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38 no. 2, pp. 295-307, 2016.

W. Lai, J. Huang, N. Ahuja, et al., “Deep laplacian pyramid network for fast and accurate superresolution,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2017, pp. 624-632.

X. Mao, C. Shen, and Y. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in *Proc. of Advances Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 2802-2810.

K. Zhang, W. Zuo, Y. Chen, et al., “Beyond a gaussian denoiser: Residual learning of deep cnns for image denoising,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142-3155, 2017.

L. Li, J. Pan, W. Lai, et al., “Learning a discriminative prior for blind image blurring,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2018, pp. 6616-6625.

J. Zhang, J. Pan, J. Ren, et al., “Dynamic scene deblurring using spatially variant recurrent neural networks,” in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Rec. (CVPR)*, 2018, pp. 2521-2529.

W. Ren, S. Liu, H. Zhang, et al., “Single image dehazing via multi-scale convolutional neural networks,” in *Proc. of Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 154-169.