UID2021: An Underwater Image Dataset for Evaluation of No-Reference Quality Assessment Metrics

GUOJIA HOU, YUXUAN LI, and HUAN YANG, Qingdao University
KUNQIAN LI, Ocean University of China
ZHENKUAN PAN, Qingdao University

Achieving subjective and objective quality assessment of underwater images is of high significance in underwater visual perception and image/video processing. However, the development of underwater image quality assessment (UIQA) is limited for the lack of publicly available underwater image datasets with human subjective scores and reliable objective UIQA metrics. To address this issue, we establish a large-scale underwater image dataset, dubbed UID2021, for evaluating no-reference (NR) UIQA metrics. The constructed dataset contains 60 multiply degraded underwater images collected from various sources, covering six common underwater scenes (i.e., bluish scene, blue-green scene, greenish scene, hazy scene, low-light scene, and turbid scene), and their corresponding 900 quality improved versions are generated by employing 15 state-of-the-art underwater image enhancement and restoration algorithms. Mean opinion scores with 52 observers for each image of UID2021 are also obtained by using the pairwise comparison sorting method. Both in-air and underwater-specific NR IQA algorithms are tested on our constructed dataset to fairly compare their performance and analyze their strengths and weaknesses. Our proposed UID2021 dataset enables ones to evaluate NR UIQA algorithms comprehensively and paves the way for further research on UIQA. The dataset is available at https://github.com/Hou-Guojia/UID2021.

CCS Concepts: • Computing methodologies → Computer graphics; Image processing; Image manipulation;

Additional Key Words and Phrases: Underwater image, image quality assessment, benchmark dataset, image enhancement and restoration, mean opinion score

ACM Reference format:
Guojia Hou, Yuxuan Li, Huan Yang, Kunqian Li, and Zhenkuan Pan. 2023. UID2021: An Underwater Image Dataset for Evaluation of No-Reference Quality Assessment Metrics. ACM Trans. Multimedia Comput. Commun. Appl. 19, 4, Article 151 (February 2023), 24 pages.
https://doi.org/10.1145/3578584

G. Hou and Y. Li contributed equally to this work.

The research work is partially supported by the National Natural Science Foundation of China (No. 61901240, 61906177), the National Science Foundation of Shandong Province, China (No. ZR2019BF042, ZR2019BF034), China Scholarship Council, China (No. 201908370002), and the China Postdoctoral Science Foundation (No. 2017M612204).

Authors’ addresses: G. Hou, Y. Li, H. Yang, and Z. Pan, College of Computer Science & Technology, Qingdao University, No. 308 Ningxia Road, Qingdao, China, 266071; emails: hgj2015@qdu.edu.cn, lyy776239423@gmail.com, cathy_huanyang@hotmail.com, zkpan@126.com; K. Li (corresponding author), College of Engineering, Ocean University of China, No. 238 Songling Road, 266100, Qingdao, China; email: likunqian@ouc.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/03-ART151 $15.00
https://doi.org/10.1145/3578584
1 INTRODUCTION

The quality of underwater images or videos plays a critical role in marine scientific research and ocean engineering, such as sea life monitoring, object detection, and tracking. However, images captured underwater often suffer from low contrast, blurring, distorted color, noise, and other degradations due to the absorption and scattering effects when light travels through the water. To address these problems, many underwater image enhancement and restoration (UIER) algorithms [1–3] have been developed. In these methods, subjective observation is always used to evaluate their performance. Although subjective evaluation can give a reliable result, it is expensive, laborious, and non-automatic. Therefore, it is important to achieve an effective objective metric for image quality assessment (IQA). Generally, existing objective IQA metrics can be classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR) algorithms, depending on the availability of a reference image. For underwater scenarios where a reference image is usually not available, the NR IQA metrics are highly desired. But unfortunately, not all existing natural image quality metrics can be effectively applied to evaluate the underwater image because they fail to correlate well with subjective perception. So far, there are only a few efficient NR metrics specifically designed for underwater image quality assessment (UIQA), such as underwater color image quality evaluation (UCIQE) [4], underwater image quality measure (UIQM) [5], an imaging-inspired underwater image quality evaluation metric dubbed CCF [6], and a frequency domain based underwater image quality metric dubbed FDUM [7], which constrains the development of underwater image/video enhancement and restoration. To figure out the performance of the state-of-the-art (SOTA) NR IQA metrics for underwater image evaluation, it is necessary to conduct a comprehensive evaluation of the modern NR metrics on the available dataset. However, all the available natural image datasets for IQA research cover a diverse set of in-air images, and they are not applicable to judge and optimize UIER algorithms.

Recently, some researchers developed several underwater benchmark datasets for evaluating UIER methods. In the work of Li et al. [8], Liu et al. [9], and Berman et al. [10], three real-world underwater image datasets, namely the underwater image enhancement benchmark (UIEB), real-world underwater image enhancement (RUIE), and stereo quantitative underwater image dataset (SQUID), are respectively constructed for studying the SOTA UIER algorithms qualitatively and quantitatively. Additionally, several synthetic underwater image datasets are accordingly developed, such as the 3D TURBID dataset [11], the underwater image database of laboratory of embedded systems and integrated circuits applications (UID-LEIA) dataset [12], an underwater images degradation dataset [13], and a synthetic underwater image dataset (SUID) [14]. Since these proposed synthetic datasets contain reference images, which creates a possibility for FR evaluation of existing UIER technologies. More recently, several real-world underwater image datasets for evaluating NR UIQA algorithms are also successively proposed, such as underwater image quality assessment (UWIQA) [7], an underwater optical image quality database (UOQ) [15], and an IQA database for underwater image enhancement (UEIQA) [16]. However, some of these datasets contain a very limited number of images and cover fewer challenging scenes, which limits their applications.

In this article, we construct a new underwater image dataset, namely UID2021, and further provide a comprehensive evaluation of NR quality assessment metrics for underwater images. The main contributions of this work include the following:

(i) A large-scale underwater image enhancement dataset is established containing 60 multiply degraded underwater images and 900 corresponding quality improved versions. In our UID2021, the selected 60 raw underwater images cover six common underwater scenes with different types of degradation. Moreover, 900 images are produced by employing 15 SOTA algorithms, well designed for UIER.
Table 1. Comparisons ofExisting 14 In-Air IQA Datasets

| Dataset    | Ref. | Dataset Type | No. of Source Images | No. of Distorted Images | Distortion Type | Distortion Level | No. of Subjects | Subjective Test Method | Subjective Data Type | Score Range |
|------------|------|--------------|----------------------|-------------------------|-----------------|------------------|-------------------|------------------------|----------------------|-------------|
| IVC        | [29] | STD          | 10                   | 185                     | 4               | 5                | 15                | DSIS                   | DMOS                 | 1–5         |
| LIVE       | [17] | STD          | 29                   | 779                     | 5               | 5–9              | 20–29             | SS                     | DMOS                 | 0–100       |
| MICT       | [18] | STD          | 14                   | 196                     | 2               | 6                | 16                | SS                     | RAW                  | 1–5         |
| TID2008    | [19] | STD          | 25                   | 1,700                   | 17              | 4                | 838               | PC                     | MOS                  | 0–9         |
| CSIQ       | [30] | STD          | 30                   | 866                     | 6               | 4–5              | 25                | N/A                   | DMOS                 | 0–1         |
| MDIQ       | [23] | MTD          | 15                   | 405                     | 3               | 4                | 37                | SS-HR                  | DMOS                 | 0–100       |
| TID2013    | [24] | MTD          | 25                   | 3,000                   | 24              | 5                | 971               | PC                     | MOS                  | 0–9         |
| CID2013    | [20] | MTD          | 8                    | 480                     | 12–14           | N/A              | 188               | ACR-DR                 | RAW                  | 0–100       |
| MDID       | [21] | MTD          | 20                   | 1,600                   | 5               | 4                | 192               | PCS                    | MOS                  | 0–8         |
| MDIVL      | [25] | MTD          | 10                   | 750                     | 2               | 12–24            | 12                | SS                     | MOS                  | 1–100       |
| WED        | [22] | MTD          | 4,744                | 94,880                  | 4               | 5                | N/A               | N/A                   | N/A                  | N/A         |
| LIVEC      | [27] | RLD          | N/A                  | 1,162                   | N/A             | N/A              | N/A               | >18,100                | SS-CP                | 3.42–92.43 |
| BID        | [26] | MTD          | 120                  | 6,000                   | 2               | 25               | 180               | SS                     | MOS                  | 1–5         |
| KonIQ-10k  | [28] | RLD          | 10,073               | N/A                     | N/A             | N/A              | 1,459             | ACR                    | MOS                  | 3.91–88.39 |

N/A means the information is not available or unknown.
RAW stands for the raw data.

In the Subjective Test Method column, DSIS represents the double-stimulus impairments scale, ACR-DR represents absolute category rating-dynamic reference, PCS stands for pair comparison sorting, ACR represents absolute category rating, and SS-CP stands for single stimulus continuous procedure.

(ii) The proposed dataset also contains the corresponding mean opinion scores (MOS) from 52 observers. With the constructed UID2021, we further conduct a comprehensive study of 11 NR quality assessment metrics.

(iii) The evaluation results can provide a better understanding of the strengths and limitations of current NR IQA metrics and point to the top-performing NR metrics for UIQA. Additionally, it can suggest a new research direction, so as to facilitate researchers to select appropriate NR metrics to evaluate their newly developed UIER algorithms. (iv) Our constructed UID2021 can also be regarded as a benchmark to guide the optimization process of developing new NR IQA algorithms for specific underwater images.

The rest of the article is organized as follows. The existing IQA datasets including the natural in-air image datasets and underwater image datasets are reviewed in Section 2. Section 3 presents a review of some popular NR IQA metrics. Section 4 describes the generation of the proposed UID2021 dataset in detail. In Section 5, we present the evaluation results, as well as the associated discussions. Finally, Section 6 concludes the article.

2 REVIEW OF EXISTING IQA DATASETS

2.1 In-Air Image Dataset

Over the past couple of decades, many publicly available image databases have been reported to evaluate the FR and NR IQA algorithms, such as laboratory image and video engineering (LIVE) [17], multimedia information and communication technology (MICT) [18], tampere image database 2008 (TID2008) [19], camera image database (CID2013) [20], multiply distorted image database (MDID) [21], and Waterloo exploration database (WED) [22]. We pick out some detailed information from 14 publicly available in-air IQA databases, as presented in Table 1. As shown the table, IQA databases can be categorized into three types: single type of distortion (STD), multiple types of distortions (MTD), and real-life distortion (RLD). Different from STD, MTD contains multiple
types of distortions of each distorted image. Among them, the STD image databases have originally attracted the greatest attention. Recently, several multiply distorted image databases have been developed, such as multiply distorted image quality (MDIQ) [23], tampere image database 2013 (TID2013) [24], CID2013, MDID, the multiple distorted IVL (MDIVL) database [25], and the blur image database (BID) [26]. However, both STD and MTD databases obtain distorted images by synthetically introducing single or multiple distortions on a pristine image, which ignores the important and frequently occurring mixtures of real-life distortions. More recently, some real-life distortion image databases have been introduced, such as LIVE challenge (LIVEC) [27] and KonIQ-10k [28]. For collecting the subjective data, single stimulus (SS) and pair comparison (PC) are two of the most widely used schemes. The benchmark evaluation rating scores are in the form of MOS, difference mean opinion score (DMOS), and raw data. MOS is the average value of the subjective ratings of an image, and a higher MOS indicates a better image quality. On the contrary, DMOS is defined as the difference between the subjective ratings of the distorted and the corresponding reference images, meaning that it can only be used when the reference image exists. A lower value of DMOS represents a better image quality. The raw data is the original subjective ratings before being calculated into MOS or DMOS from each image, providing more opportunities to those researchers who would like to utilize the data to its full potential. In the subjective evaluation, image and video communication (IVC) [29], LIVE, categorical subjective image quality (CSIQ) [30], and MDIQ databases provide subjective data in the form of DMOS. Differently, in TID2008, TID2013, MDID, and MDIVL databases, the MOS was computed for subjective testing. It is necessary to note that WED is the current largest one for in-air IQA, which contains 4,744 clear reference images and 94,880 distorted images. Because of the large scale of WED, it is extremely difficult to collect the MOS via subjective testing. Instead, they utilized three criteria to perform a systematic evaluation of 20 well-known IQA algorithms. Although these existing datasets cover a wide variety of in-air images, they contain few underwater images, which is limited for the development of UIER algorithms. To fit this gap, some researchers transfer their attention to constructing underwater image datasets.

### 2.2 Underwater Image Dataset

Currently, to the best of our knowledge, there are two kinds of underwater image datasets used to test the performance of UIER algorithms. The details of these two kinds of underwater image datasets can be found in Table 2. With regard to the performance validation of UIER algorithms, Codevilla et al. [31] proposed an underwater turbidity images dataset called TURBID. All images in TURBID are captured in a 1,000-L water tank, and the level of turbidity is controlled by successively adding milk into the water tank. Afterward, Ma et al. [32] used real turbidity lake water to simulate the real underwater situation and proposed an underwater turbidity image dataset, namely NWPU. NWPU contains 6,240 underwater images of 40 objects, which are captured under six different levels of turbidity, four light conditions, and six different distances. More recently, Li et al. [8] proposed a large-scale underwater image enhancement benchmark, which includes 890 real underwater images taken under natural light, artificial light, and a mixture of natural and artificial light. In the same year, Liu et al. [9] constructed a real-world underwater enhancement dataset, RUIE, with more than 4,000 underwater images captured by a multi-view imaging system under seawater. RUIE consists of three subsets: underwater image quality set (UIQS), underwater color cast set (UCCS), and underwater higher-level task-driven set (UHTS), which are used to validate the capability of UIER algorithms to improve image visibility, correct color cast, and examine their effectiveness from the aspect of high-level underwater tasks.

The aforementioned underwater image datasets focus on testifying the performance of UIER algorithms but do not provide the corresponding MOS values, which are not suitable for evaluating
Table 2. Comparisons of Different Underwater Image Datasets

| Datasets  | No. of Images | Resolution | Image Generation                                                                 | Usage  | No. of Subjects | Subjective Test Method | Subjective Data Type |
|-----------|---------------|------------|---------------------------------------------------------------------------------|--------|----------------|------------------------|----------------------|
| TURBID [31] | 570           | 4000 × 3000 | Adding milk to simulate different levels of turbidity                           | UIER   | N/A            | N/A                    | N/A                  |
| UIEB [8]   | 950           | 225 × 225 to 2180 × 1447 | Real underwater images taken under natural light, artificial light            | UIER   | N/A            | N/A                    | N/A                  |
| RUIE [9]   | 4,230         | 400 × 300   | Real underwater images captured by a multi-view imaging system under seawater    | UIER   | N/A            | N/A                    | N/A                  |
| NWPU [32]  | 6,240         | 1600 × 1200 | Using different level of turbidity water, light conditions, view distances      | UIER   | N/A            | N/A                    | N/A                  |
| UOQ [15]   | 216           | 512 × 512   | Using five image enhancement algorithms                                          | UIQA   | N/A            | SS                     | MOS                  |
| UWIQA [7]  | 890           | 225 × 225 to 2180 × 1447 | Images collected from the UIEB                                                  | UIQA   | 21             | SS                     | MOS                  |
| UEIQA [16] | 240           | 1280 × 720  | Using five image enhancement algorithms                                          | UIQA   | 18             | SS                     | MOS                  |
| UID2021    | 960           | 512 × 384   | Using 15 UIER algorithms                                                        | UIQA   | 52             | PCS                    | MOS                  |

UIQA algorithms. Due to the outstanding performance of FR IQA algorithms, some researchers attempted to adopt FR IQA metrics in evaluating the quality of underwater images to remedy the shortage of inconsistent results between subjective evaluation and the results of NR IQA metrics. Nevertheless, the existing classical FR IQA algorithms are not available on account of lacking reference images in the underwater environment. To overcome this issue, several synthetic underwater image datasets have been developed. Li et al. [13] proposed an underwater image synthesis algorithm based on the underwater imaging physical model to generate an underwater image degradation dataset. In the same year, Hou et al. [14] constructed a large-scale synthetic underwater image dataset by utilizing hierarchical searching and a red channel prior algorithm to acquire the underwater background light and transmission map from the real-world underwater image. SUID contains 900 degraded images, covering a diverse set of different turbidity types and degradation levels.

However, there still exists a gap between synthetic and real-world underwater images. Additionally, since the reference images are not available in underwater scenes, NR algorithms are still the first choice of objective quality assessment for underwater images. Nevertheless, all the aforementioned underwater image datasets are not developed for evaluating NR algorithms. Until recently, several underwater image datasets specifically designed for evaluating underwater NR algorithms were emerging. Wu et al. [15] built an underwater optical image quality database. It consists of 36 typical underwater images with the size of 512 × 512, and 180 enhanced images, covering 10 categories of degradation types. Additionally, Yang et al. [7] proposed a large-scale underwater image quality evaluation benchmark dataset by directly utilizing 890 images from UIEB as the source images and conducting a subjective quality evaluation to obtain MOS, but the MOS values of the UWIQA dataset are concentrated in 10 discrete scales (from 0.1 to 1), which may be not accurate enough. Moreover, Guo et al. [16] constructed a UEIQA dataset, consisting of 40 source images captured by a remotely operated underwater vehicle and corresponding 200 enhanced images by five image enhancement algorithms. However, all 40 source images in UEIQA are captured on a sea cucumber farm, indicating that it only covers a limited set of underwater scenes.

The current underwater datasets pay more attention to validating the performance of UIER algorithms rather than UIQA algorithms. Even though several datasets for testing UIQA algorithms...
are emerging, they usually contain monotonous content, limited scenes, few degradation characteristics, and insufficient data, which makes it difficult or even impractical to fairly evaluate underwater NR algorithms, and hinders the development of new NR algorithms as well. To address these limitations, a new real-world underwater image dataset is highly desired to provide the benchmark for UIQA algorithms.

3 REVIEW OF NR IQA METRICS

In this section, we present a brief review of some well-known NR IQA algorithms that are usually used for evaluating the performance of UIER methods. NR IQA methods aim to evaluate the quality of a digital image without depending on any reference information, and they are also known as blind IQA (BIQA) methods. In the past decade, natural scene statistics (NSS)-based algorithms have been widely used for NR IQA. Based on the NSS, Moorothy and Bovik [33] designed a new two-step framework for NR IQA, namely BIQI, which can assess image quality across no distortion-specific categories. Mittal et al. [34] proposed a blind/referenceless image spatial quality evaluator (BRISQUE) based on the spatial domain NSS because of its higher computational efficiency. In BRISQUE, they extracted the natural scene statistics by using local divisive normalization. Afterward, Mittal et al. [35] further developed a natural image quality evaluator (NIQE) without training with any human opinion scores. Unlike BRISQUE, the NIQE algorithm only considered some image patches with high contrast for fitting NSS features. By using the NSS model of discrete cosine transform coefficients, Saad et al. [36] proposed a novel NR IQA method called BLIIDNS-II, which utilized a lower-dimensional feature space and a simpler single-stage framework, achieving appealing computational complexity.

With the fast development of deep neural networks, designing data-driven algorithms to tackle this challenge becomes possible. Ma et al. [37] presented a multi-task end-to-end optimized network (MEON) for NR IQA. MEON is performed by training two sub-networks: a distortion type identification network and a quality prediction network relying on pre-trained early layers. Motivated by meta-learning, Zhu et al. [38] proposed a new NR IQA metric called MetaIQA by employing bi-level gradient optimization to learn the shard prior knowledge model of various distortions from plenty of NR IQA tasks, and then fine-tune the prior model to obtain the target quality model. Wu et al. [39] proposed an IQA-oriented convolutional neural network (CNN) method, namely CaHDC, inspired by the hierarchical perception mechanism in the human visual system. Benefiting from the hierarchical degradation concatenation and the end-to-end optimization, CaHDC can better learn the nature of quality degradation and accurately predict image quality. Instead of evaluating the quality by accumulating features over the entire image, Kang et al. [40] estimated image quality on small image patches by using a modified CNN since local image quality is crucial for image denoising and reconstruction. Su et al. [41] proposed a self-adaptive hyper network for assessing authentically distorted images, which consists of three stages including content understanding, perception rule learning, and quality predicting. Zhang et al. [42] developed a CNN-based two-stream network architecture for evaluating synthetically and authentically distorted images, respectively. By adopting the Siamese networks, Bosse et al. [43] proposed a data-driven deep neural network based approach for IQA, namely weighted average deep image quality measure for NR IQA (WaDIQaM-NR). The network comprises 10 convolutional layers and five pooling layers for feature extraction, and two fully connected layers for regression and trained on artificially augmented datasets. It would be desirable to combine multiple IQA databases to enrich the training samples for the sake of data diversity. However, the annotation shift usually causes unreliable data fusion. Zhang et al. [44] proposed a database combination strategy and the pairwise learning-to-rank model to account for this problem. Basically, by using a continual learning strategy, a BIQA model can learn continually from a stream of IQA datasets, integrating
new knowledge from the current dataset while preventing the forgetting of acquired knowledge from previously seen datasets. Inspired by it, Zhang et al. [45] also developed a simple yet effective continual learning method for BIQA. Wang et al. [46] proposed a novel NR IQA method based on the Swin Transformer, which aggregates information from both local and global features by fusing features from multiple stages to accurately predict the quality.

Since haze and blur are two special types of problems of underwater images, some specially designed NR metrics [47–52] are employed to assess the performance of defogging and deblurring algorithms. Among them, Ferzli and Karam [47] first proposed an NR perceptual-based blur metric depending on the notion of just noticeable blur (JNB). JNB can accurately predict the relative amount of blurriness with different content. Afterward, they presented an improved blurriness metric [48] to estimate the probability of detecting blur based on the cumulative probability of blur detection (CPBD) and JNB. Li et al. [49] developed a blind image blur evaluation (BIBLE) metric using discrete orthogonal moments. They divided the gradient image into equal-size blocks and computed their Tchebichef moments to characterize the image shape. The proposed BIBLE algorithm not only can generate accurate image blur scores but also achieves highly consistent with subjective evaluations. In 2015, Choi et al. [50] designed an NR fog density prediction model called fog aware density evaluator (FADE). FADE can accurately predict the fog density benefiting from its dependence on NSS and fog-aware statistical features. Liu et al. [51] proposed a new NR metric called PSQA, which is based on the analysis of pre-attention and spatial dependency’s influence on the perception of distortion. Yan et al. [52] first trained an image distortion classifier by employing the SOTA Inception-ResNet-v2 neural networks. Then, based on the distortion classification, they developed a novel NR metric, namely DIQM, to characterize the image quality in a targeted manner.

Unlike in-air images, underwater images encounter the effect of serious absorption and scattering. Some IQA metrics for in-air images are not applicable to evaluate underwater image quality. To address this challenge, some authors focus on developing the specific non-reference metrics for UIQA. In 2015, Yang and Sowmya [4] designed a UCIQE algorithm to qualify the non-uniform color cast, blurring, and low contrast by a linear combination of these three components. The following year, Panetta et al. [5] presented an NR UIQM inspired by the properties of the human visual system. The UIQM algorithm comprises three attribute measures in terms of colorfulness measure, sharpness measure, and contrast measure. Afterward, Wang et al. [6] proposed an imaging-inspired metric for underwater color IQA, called CCF, which is weighted with a linear combination of colorfulness, contrast, and fog density. CCF can quantify the color loss, blurring, and fogging, which are caused by absorption, forward scattering, and backward scattering, respectively. Likewise, Yang et al. [7] presented a reference-free UIQA metric, namely FDUM, in the frequency domain by combining colorfulness, contrast, and sharpness. Recently, based on the naturalness, sharpness, and structure indexes and saliency-based pooling strategy, Zheng et al. [53] established an underwater image fidelity (UIF) metric for objective evaluation of the UIER-enhanced underwater images. In the same year, Jiang et al. [54] proposed an effective NR underwater image quality metric (NUIQ) that can fairly rank enhanced underwater images according to the visual quality by using a set of quality-aware features from chromatic and luminance components. Most recently, some ranking-based methods have been proposed. Motivated by the observation that a mid-quality image can be generated by mixing a high-quality image and its low-quality version, Fu et al. [55] proposed a novel rank learning guided NR quality assessment method to evaluate different UIER algorithms, termed Twice Mixing. Guo et al. [56], based on the efficient conv-attentional image Transformer, proposed a ranking-based UIQA method, namely URanker, by specifically devising the histogram prior and the dynamic cross-scale correspondence to model the underwater image degradations. In addition, some authors perform specific applications such as edge detection...
local feature point matching \([59, 60]\), and image segmentation \([61, 62]\) to assess their enhanced and restored results.

Because of the fact that most existing NR metrics are not specially designed for UIQA. The in-air IQA metrics mentioned previously are also often employed for UIQA. However, to our best knowledge, there lacks a systematic study of their performance in evaluating the UIER algorithms. Actually, some in-air NR objective metrics show a poor correlation with subjective evaluations, which cannot accurately predict underwater image quality. A detailed description of the 11 well-known and widely used NR IQA metrics being evaluated on the proposed dataset will be given in Section 5.

4 UID2021 DATASET

After systematically summarizing previous work, we find that the lack of comprehensive human subjective user study with publicly available datasets and reliable objective UIQA metrics makes it difficult to better understand the real performance of UIER algorithms. Generally, an in-air image dataset is constructed by dozens of reference images and lots of generated distorted images. In what follows, we will introduce the constructed dataset UID2021 in detail, including source image collection and enhanced or restored image generation.

4.1 Source Images Selection

To our knowledge, the existing underwater datasets do not classify the underwater images based on different underwater scenes, which makes it difficult to evaluate the performance of UIQA algorithms under different underwater scenes. Aiming to provide a platform for more comprehensively evaluating the performance of the UIQA algorithms, the selected source images in our UID2021 cover six common types of underwater scenes (i.e., bluish scene, blue-green scene, greenish scene, hazy scene, low-light scene, and turbid scene). Additionally, the source images we selected in our UID2021 are likely taken under natural light, artificial light, or a mixture of natural light and artificial light, as well as cover a variety of underwater scenes including archaeological relics, divers, marine life, underwater pipeline, and seabed, among others. The details of these six scenes are given in Table 3. The 60 source images are selected from several popular underwater datasets \([8–10]\) and a public website \([63]\). The selected underwater images are in a different size because they are picked from various public real-world image datasets. We crop them into the uniform size of \(512 \times 384\) to make sure that they can be completely and clearly displayed on the monitor. All source underwater images are presented in Figure 1.

4.2 Dataset Creation

In our UID2021, based on the 60 source underwater images under different challenging scenes, 900 enhanced or restored images are generated by using 15 SOTA UIER algorithms (Table 4 presents more details). These existing SOTA methods are specially designed for improving the quality of underwater images from different aspects of contrast enhancement, dehazing, deblurring, color correction, and so on. It is worth noting that these 15 UIER algorithms are not randomly selected, because images enhanced by different UIER algorithms may be visually similar, marking it difficult for human observers to distinguish the quality of the enhanced images. In this article, the 15 UIER algorithms are carefully selected to ensure that the enhanced images have large visual gaps, covering the image quality from the best to the worst. Due to the limited space, we present one image of each scene accompanied by their 15 enhanced or restored versions for demonstration, as shown in Figure 2. As we can see from the figure, the enhanced or restored versions of the source images cover a broad range of image qualities, assuring that they can be regarded as a benchmark to evaluate the NR algorithms.
Table 3. Details of the Six Subsets with Different Challenge Scenes

| Subsets | Scenes      | Source Images | Total Images | Descriptions                                                                 |
|---------|-------------|---------------|--------------|-----------------------------------------------------------------------------|
| 1       | Blush       | 10            | 160          | In underwater environment, the decay of light is related to the wavelength of the color. When traveling through water, the red light decays fastest than green and blue wavelengths because of its largest wavelength. Therefore, underwater captured images always appear to have bluish, blue-green or greenish tones. |
| 2       | Blue-Green  | 10            | 160          | The particles suspended underwater lead to images suffer from haze.          |
| 3       | Greemish    | 10            | 160          | In deep water, there's not enough light, and artificial light is needed.     |
| 4       | Hazy        | 10            | 160          | Underwater mud makes images turbid.                                         |

Fig. 1. Source images with different challenging scenes: bluish scene (a), blue-green scene (b), greenish scene (c), hazy scene (d), low-light scene (e), and turbid scene (f).

4.3 Subjective IQA

Using reliable subjective evaluation to represent the ground truth is the most significant component for obtaining MOS values for a database. Various subjective methods are employed to provide the reliability of perceptual quality evaluations, such as the double stimulus continuous quality scale (DSCQS) scheme, SS scheme, and PC scheme. The DSCQS method is successfully introduced in the IVC database because it is appropriate for evaluating a small number of images. The SS method is a widely used scheme that has been exploited in LIVE, MICT, MDIQ, and MDIVL databases. In this method, the observers are asked to absolutely rate only one image each
Table 4. Details of the 15 Underwater Enhancement or Restoration Methods

| Category       | Methods | Public Implementation Codes |
|----------------|---------|------------------------------|
| Enhancement    | Bayesian-retinex [64] | https://github.com/zhuangpeixian/Bayesian-Retinex-Underwater-Enhancement |
|                | CBF [62] | https://github.com/bilityniu/underimage-fusion-enhancement |
|                | CHS [65] | Implemented by ourselves |
|                | HP [67] | https://github.com/Hou-Guojia/HP |
|                | L₂UWE [69] | https://github.com/tunai/l2uwe |
|                | TS [70] | https://xueyangfu.github.io/paper/2017/ISPACS/code.zip |
|                | UWB-VCSE [60] | https://github.com/Hou-Guojia/UWB-VCSE |
|                | VR [74] | https://xueyangfu.github.io/projects/icip2014.html |
| Restoration    | IBLA [68] | https://github.com/ytpeng-aimlab/Underwater-Image-Restoration-based-on-Image-Blurriness-and-Light-Absorption |
|                | RCP [57] | https://github.com/agaldran/UnderWater |
|                | UNTV [72] | https://github.com/Hou-Guojia/UNTV |
|                | UTV [73] | https://github.com/Hou-Guojia/UTV |
| Deep learning  | GLN-CHE [66] | https://xueyangfu.github.io/projects/spic2020.html |
|                | Ucolor [71] | https://github.com/Li-Chongyi/Ucolor |
|                | UWCNN [13] | https://github.com/saeed-anwar/UWCNN |

Fig. 2. Examples of source images and their enhanced or restored versions. From left to right: source images (a), and the enhanced or restored results by using Bayesian-retinex (b), CBF (c), CHS (d), GLN-CHE (e), HP (f), IBLA (g), L₂UWE (h), RCP (i), TS (j), Ucolor (k), UNTV (l), UTV (m), UWB-VCSE (n), UWCNN (o), and VR (p), respectively.

time. In fact, it is often difficult for observers to assign a score to an image. In many cases, the observers find that they give an improper score when they later evaluate another image, but many subjects are reluctant to change their previous scores. Unlike SS, the PC scheme does not require the observers to assign an absolute score of an image. Instead, they are only asked to choose the better one of two images, which helps observers easily make a decision. This method is successfully utilized in TID2008 and TID2013 databases. Nevertheless, we find that it is also difficult for observers to compare two quality-indistinctive images. Fortunately, Sun et al. [21] proposed a novel subjective scheme dubbed pairwise comparison sorting (PCS). PCS allows observers to make an “equal” decision besides “greater” or “less.” Actually, these subjective schemes are not viable when the number of images is tremendous, such as in the WED database [22] which contains 4,744 source images and 94,880 distorted images. Ma et al. [22] presented three criteria, referred to as the pristine/distorted image discriminability test, the listwise ranking consistency test, and the pairwise preference consistency test, to perform a subjective evaluation of image quality.
Fig. 3. The qualities of underwater images generated by different methods appear visually equal.

Table 5. Five Aspects of Subjective Quality Assessment

| Protocol              | Description                                                                 |
|-----------------------|-----------------------------------------------------------------------------|
| Color distortion      | The underwater image usually suffers from severe color distortion.          |
| Contrast distortion   | The backward scattering effect reduces the contrast of underwater images.   |
| Texture distortion    | The forward scattering results in blurry underwater images, and the texture information of underwater images are seriously lost. |
| Visibility            | Visibility in underwater images is often poor due to reduced contrast, poor lighting conditions, and other problems. |
| Recognizable foreground| Recognizable foreground is important for high-level underwater tasks.       |

It is common that the enhanced or restored images by some different methods are visually equal, and it is hard to decide which one is better. To further demonstrate the concept of visually equal images, we adopt the underwater image sharpness measure (UISM) and underwater image color measure (UICM) to objectively evaluate the visually equal images, as shown in Figure 3. It is clear to see that compared with the right image, the left image got a higher UISM and a lower UICM, indicating that it discloses more texture and rich details while its color is quite unsaturated. On the contrary, the right image got a higher UICM and a lower UISM, presenting more pleasurable color correction results but revealing fewer details. The preceding phenomena make it difficult for human subjects to judge which image is better. We call these images visually equal images. To make the subjective evaluation less annoying for the participants, we adopt the PCS scheme to conduct the quality comparison. We perform the subjective evaluation in an isolated room with natural illumination. In all sessions, a 21.5-inch LED Lenovo monitor at the solution of 1920 × 1080 pixels displays the two compared images. Observers are seated in front of the monitor at a distance of three times the screen height.

4.4 Methodology

In our study, a total of 52 undergraduate and graduate students participate in the subjective evaluation experiment. For improving the efficiency and convenience of rating and collecting, we design a subjective evaluation software, as shown in Figure 4. Before the experiment, the participants are trained on using this software. Underwater images usually have practical applications; therefore, to make sure that subjects are not affected by any aesthetic factors in the process of subjective quality assessment, subjects are told to evaluate image quality from five aspects, including color distortion, contrast distortion, texture distortion, visibility, and recognizable foreground (Table 5 presents more details). In the experiment, all 960 images are divided into six image scenes according to their degradation types. In each subset, observers can load the images with one scene
or more than one scene at a time. Moreover, the software is programmed to ensure that the observers evaluate all images of the same scene before evaluating images of the other scenes. Additionally, observers are instructed to rate each two compared images within no more than 10 seconds. To improve the shortcomings of the original PCS algorithm, it should be pointed out that when there are more than two pairs of images with equal quality, the system will automatically ask observers to rate them again.

4.5 Data Processing
After using PCS in subjective evaluation, we obtain an integer sequence that presents the quality ranking of the images. To convert the ranking integers obtained from PCS into scores, we use the following formula to normalize these integers into $[0, 9]$:

$$\text{Score}_N (i) = 9 \times \frac{\max (r) - r (i)}{\max (r) - \min (r)},$$

where $r$ is the vector of ranking integers within each image group obtained from PCS and an image group including 1 source image and its 15 corresponding enhanced images. Additionally, $i$ is the
number of images contained in one scene \( i = 1, 2, \ldots, 16 \), and \( \text{Score}_N (i) \) is the subjective elevation score of \( i^{th} \) image rated by \( N^{th} \) observer. Evaluation results of image quality by different subjects usually contain abnormal results, which may be caused by wrong clicks, visual fatigue, and even depressed mood. These abnormal results will seriously contaminate the accuracy of our dataset. To prevent this, outlier detection and subject rejection are conducted with the same method in LIVE [17]. Specifically, the mean and standard deviation of the raw data for each image are calculated as follows:

\[
\bar{S}_i = \frac{1}{N} \sum_{n=1}^{N} S_{i,n},
\]

\[
\sigma_i = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (S_{i,n} - \bar{S}_i)^2},
\]

where \( N \) stands for the number of subjects and \( S_{i,n} \) represents the score of \( i^{th} \) images rated by the \( n^{th} \) subject. A raw data for an image is considered to be an outlier if \( S_{i,n} \) is outside the interval \( [\bar{S}_i - F_1 \times \sigma_i, \bar{S}_i + F_1 \times \sigma_i] \). Additionally, if more than \( F_2 \) elements of subject \( i \) are considered to be outliers, all evaluation results of subject \( i \) are rejected. The \( F_1 \) and \( F_2 \) are obtained by minimizing the average width of the 95% confidence interval of each distorted image. Approximately 3.77% of abnormal results and one subject have been removed from further consideration, and MOS values are calculated by averaging the remaining reliable individual scores.

### 4.6 Analysis of the Subjective Evaluation Results

As suggested in the work of Winkler [75], the distribution of MOS and standard deviation of the subjective ratings are commonly adopted to verify the accuracy and suitability of the proposed dataset.

The distribution of MOS is indicative of the image quality range of a dataset. Generally, a uniform distribution of MOS is highly desired because it demonstrates that the range of subjective rating scales is completely utilized. In Figure 5, we present the scatter plots of MOS distributions of our proposed dataset and the UWIQA dataset, respectively. Figure 6 shows the MOS distribution of the six subsets in our UID2021. As shown in Figure 5, we can observe that the MOS values of the UWIQA dataset are mostly located between 0.3 and 0.7, indicating that UWIQA covers a narrow range of image quality. On the contrary, the MOS distribution of our constructed UID2021 is very close to the uniform distribution, visually indicating that our UID2021 contains equal frequencies of images varying in different levels of quality. Moreover, in Figure 6, the MOS distributions of the
The standard deviation is another important benchmark. Here, to compare the MOS values with single individual scores, we further carry out a specific analysis of standard deviation by calculating the standard deviation of subjective scores on every single image in our UID2021, and obtained an average standard deviation of 1.2498. To further analyze the standard deviation, we divide all 960 images into three groups according to their corresponding MOS. The first group is called Bad with MOS from 0 to 3, followed by the second group called Middle with MOS from 3 to 6. Likewise, the last group is called Good with MOS larger than 6. Then, we calculate the mean of standard deviations separately for each group, and their results are presented in Table 6. As we can see from the table, the standard deviation is typically higher in the Middle group than in the Bad and Good groups. In our opinion, this phenomenon means that subjects’ opinions are quite unified when evaluating images with bad or good quality, but they differ greatly when evaluating images with middle quality. All in all, the uniform distribution of MOS and a small standard deviation demonstrate that subjective evaluation in our proposed dataset is reliable.

## 5 Evaluations

Since there is no available ground truth for real underwater images, UIQA datasets are specially developed for testing the performance of different NR IQA algorithms. In this section, to fully

---

**Table 6. Mean of Standard Deviations for Bad, Middle, and Good Groups**

| Groups | Bad     | Middle  | Good    |
|--------|---------|---------|---------|
| Mean of Standard Deviation | 1.1230  | 1.5009  | 1.1256  |

six subsets are also uniform, demonstrating that our proposed dataset fully utilizes the entire MOS range without emphasizing any part of that range.
understand the performance of the current NR metrics for IQA research, we perform a comprehensive evaluation of various widely used NR IQA metrics on our proposed UID2021 dataset.

5.1 Criteria to Evaluate NR IQA Metrics

The four most widely used criteria for evaluating algorithms include the Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), Kendall rank order correlation coefficient (KROCC), and root mean squared error (RMSE). Generally, the higher PLCC, SROCC, and KROCC as well as smaller RMSE indicate a better performance of the NR IQA algorithm. Specifically, before calculating PLCC and RMSE, the results of NR IQA algorithms are fitted to MOS through a five-parameter logistic function, which is defined by

\[ f(s) = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + \exp \left( \beta_2 \cdot (s - \beta_3) \right)} \right) + \beta_4 \cdot s + \beta_5, \]  

(4)

where \( s \) is the results of NR IQA algorithms, \( f(s) \) is the fitted scores after nonlinear regression, and \( \beta_i (i = 1, 2, \ldots, 5) \) are the regression function parameters that are estimated through the nonlinear fitting.

5.2 Comparison Results

Without reference images, researchers focus on exploring NR IQA metrics to evaluate underwater image quality. To the best of our knowledge, many NR IQA metrics have been used for testing the performance of UIER algorithms [2, 62, 66, 68, 71–73, 76–83], including four underwater-specific NR IQA algorithms in terms of UCIQE [4], UIQM [5], CCF [6], FDUM [7], and five in-air NR IQA algorithms with regard to BIQI [33], CPBD [48], BLIINDS-II [36], NIQE [35], and FADE [50]. In addition to these traditional IQA algorithms, two of the latest deep learning algorithms, including HyperIQA [41] and TReS [84], are also introduced in our experiment. (See Table 7 for details.) However, there is no evidence to demonstrate whether they are suitable for UIQA or not. Here,
Table 8. Performance Comparison of 11 Popular NR IQA Algorithms on Six Subsets

| Subset  | Criterion | Traditional In-Air IQA Algorithms | Underwater IQA Algorithms | Deep Learning IQA Algorithms |
|---------|-----------|-----------------------------------|----------------------------|----------------------------|
|         |           | BIQI | CPBD | BLIINDS-II | NIQE | FADE | UCiQE | UIQM | CCF | FDUM | HyperIQA | TReS |
| Bluish  | SROCC     | 0.1714 | 0.0032 | 0.2212 | 0.2828 | 0.5766 | 0.6182 | 0.3593 | 0.4570 | 0.6767 | 0.8614 | 0.8704 |
|         | PLCC      | 0.2730 | 0.2343 | 0.3277 | 0.3741 | 0.4109 | 0.6409 | 0.5908 | 0.5416 | 0.6862 | 0.8695 | 0.8787 |
|         | KROCC     | 0.1211 | 0.0042 | 0.1532 | 0.1953 | 0.2558 | 0.4575 | 0.3753 | 0.3281 | 0.5047 | 0.6794 | 0.7003 |
|         | RMSE      | 2.0238 | 2.0451 | 1.9875 | 1.9509 | 1.9179 | 1.6149 | 1.6972 | 1.7685 | 1.5302 | 1.1035 | 1.0089 |
| Blue-green | SROCC   | 0.3980 | 0.1320 | 0.1611 | 0.3224 | 0.2890 | 0.6668 | 0.5864 | 0.4043 | 0.6588 | 0.8850 | 0.8692 |
|         | PLCC      | 0.4030 | 0.2224 | 0.2127 | 0.2399 | 0.2562 | 0.5767 | 0.5792 | 0.2699 | 0.5851 | 0.9128 | 0.8935 |
|         | KROCC     | 0.2686 | 0.0877 | 0.1126 | 0.2188 | 0.1015 | 0.3820 | 0.3531 | 0.1510 | 0.3889 | 0.7282 | 0.7077 |
|         | RMSE      | 1.9902 | 2.1495 | 2.1543 | 2.1403 | 2.1311 | 1.8012 | 1.7973 | 2.1229 | 1.7880 | 1.0367 | 1.2992 |
| Greenish | SROCC     | 0.3345 | 0.0859 | 0.1028 | 0.2507 | 0.1547 | 0.6237 | 0.4291 | 0.4899 | 0.5261 | 0.8841 | 0.8236 |
|         | PLCC      | 0.3533 | 0.2598 | 0.1363 | 0.2485 | 0.2804 | 0.6905 | 0.6220 | 0.3097 | 0.6716 | 0.9057 | 0.8925 |
|         | KROCC     | 0.2297 | 0.0467 | 0.0711 | 0.1692 | 0.1898 | 0.4892 | 0.4164 | 0.2945 | 0.4783 | 0.7256 | 0.7154 |
|         | RMSE      | 2.1080 | 2.1748 | 2.2311 | 2.1815 | 2.1618 | 1.6290 | 1.7634 | 2.1414 | 1.6686 | 1.0367 | 1.2992 |
| Hazy    | SROCC     | 0.2821 | 0.027 | 0.0231 | 0.3442 | 0.364 | 0.6237 | 0.4291 | 0.4899 | 0.5261 | 0.8841 | 0.8236 |
|         | PLCC      | 0.3078 | 0.1504 | 0.1541 | 0.4036 | 0.4433 | 0.6688 | 0.5761 | 0.5686 | 0.5795 | 0.8577 | 0.8136 |
|         | KROCC     | 0.1933 | 0.0208 | 0.017 | 0.2352 | 0.2507 | 0.4558 | 0.3103 | 0.3496 | 0.379 | 0.7077 | 0.6410 |
|         | RMSE      | 2.0242 | 2.1034 | 2.1021 | 1.9466 | 1.9071 | 1.5817 | 1.739 | 1.7528 | 1.7339 | 1.2322 | 1.3842 |
| Low light | SROCC   | 0.183 | 0.1655 | 0.1767 | 0.2618 | 0.5254 | 0.665 | 0.4928 | 0.434 | 0.62 | 0.7336 | 0.7289 |
|         | PLCC      | 0.2878 | 0.162 | 0.191 | 0.2947 | 0.5278 | 0.6727 | 0.5297 | 0.451 | 0.6261 | 0.7200 | 0.7114 |
|         | KROCC     | 0.1223 | 0.1127 | 0.1132 | 0.1775 | 0.3421 | 0.4755 | 0.3356 | 0.2978 | 0.439 | 0.5513 | 0.5385 |
|         | RMSE      | 1.9163 | 1.9468 | 1.9366 | 1.8853 | 1.7121 | 1.4597 | 1.6734 | 1.7609 | 1.5383 | 1.3514 | 1.6384 |
| Turbid  | SROCC     | 0.1982 | 0.0804 | 0.2826 | 0.5473 | 0.3897 | 0.5599 | 0.7156 | 0.6089 | 0.6725 | 0.8956 | 0.9283 |
|         | PLCC      | 0.3104 | 0.0881 | 0.4101 | 0.4093 | 0.3926 | 0.5844 | 0.7145 | 0.6383 | 0.6882 | 0.8906 | 0.9124 |
|         | KROCC     | 0.1366 | 0.0536 | 0.1962 | 0.3763 | 0.2693 | 0.4060 | 0.3537 | 0.4508 | 0.4987 | 0.7538 | 0.7872 |
|         | RMSE      | 2.0400 | 2.1377 | 1.9573 | 1.9580 | 1.9737 | 1.7414 | 1.5015 | 1.6520 | 1.5571 | 1.3309 | 1.1621 |

we further test these popular NR algorithms on our UID2021 dataset. Unlike in-air image datasets, to provide valuable guidance, we first perform a more reasonable evaluation on the six subsets with different degradation scenes. As some of the tested NR IQA methods need to train a quality prediction model to adjust the relationship between the extracted features and the MOS values, we randomly partitioned the dataset into two non-overlapping subsets: the training subset and the testing subset. Specifically, the training subset consists of 80% of the images for training the prediction model, whereas the testing subset includes the remaining 20% of the images for testing. Since the proposed UID2021 dataset contains six subsets, to make the data split strategy more reasonable, we randomly selected 960 × 80% ÷ 6 = 128 images from each subset, a total of 128 × 6 = 768 images as the training subset, and the remaining 192 images as the test subset. This training-testing procedure was conducted 100 times to ensure that there was no bias for the specific training-testing split. For those methods that do not require training, we only examine them on the testing subset, which was also repeated 100 times. Finally, the average of PLCC, SROCC, KROCC, and RMSE values over 100 training-testing iterations are used as the final performance indices. The performance of these 11 NR IQA algorithms on the six common degradation scenes is successively presented in Table 8. The boldface values indicate the best performance in each scene.

From Table 8, in terms of the traditional in-air and underwater-specific IQA algorithms, we can see that compared with the underwater-specific NR IQA algorithms, the performance of these
tested in-air NR IQA algorithms is much poorer. To be specific, for the six scenes, UCIQE achieves the best predictive performance with respect to four criteria, closely followed by FDUM and UIQM. As for in-air NR IQA algorithms, FADE shows the best overall performance, followed by NIQE. However, we see that although these two algorithms perform well for turbid and low-light scenes, their performance is much worse for the blue-green and green scenes. For example, in the low-light scene, the SROCC value of FADE reaches up to 0.5254, whereas in the blue-green scene, the value is only about 0.16. Moreover, even the best algorithm, FADE, performs far from satisfactory, not to mention BIQI, CPBD, and BLIINDS-II. For the turbid scene, all four underwater-specific NR IQA algorithms achieve rather good performance with the highest SROCC being 0.7156. However, the SROCC of CCF drops considerably in the blue-green subset. Even BIQI, an in-air NR algorithm, performs better than CCF. To conclude, UCIQE and FDUM are quite similar in performance, whereas FDUM performs better in the bluish scene and UCIQE performs better in the hazy scene. Moreover, UIQM is more suitable for the turbid scene, and CCF performs worst among the four underwater NR algorithms. More unfortunately, all five in-air NR IQA algorithms cannot get satisfactory results. It can be concluded that UIQA is quite challenging for in-air NR IQA algorithms, and developing efficient NR IQA algorithms specific to underwater images is highly desired. In terms of the two latest deep learning algorithms, their performance is far better than the traditional IQA methods, showing that deep learning methods have great potential in the field of IQA.

Additionally, we conduct a comparison on the entire UID2021 dataset. Detailed results can also be found in Table 9. The conclusions are similar to those in Table 8, with deep learning methods demonstrating the best performance, followed by underwater-specific IQA algorithms and in-air IQA algorithms. To visually demonstrate the performance of these NR algorithms, we draw their distribution plots of MOS values with respect to the predicted scores on our UID2021, as presented in Figure 7. Here, we normalize the MOS values and IQA algorithms predicted scores into [0, 1] for better comparison. The blue lines were obtained through the curve fitting process. In Figure 7, it is clear to see that HyperIQA and TReS gather much closer to the fitted curve than the other IQA methods. Based on Figure 7 and Table 9, we can also conclude that the deep learning methods show the best performance, followed by underwater-specific IQA algorithms and in-air IQA algorithms.

To further confirm our judgments, we conduct the statistical significance of the performance of these NR IQA algorithms. The experimental results are tabulated in Table 10, where the values of “1,” “0,” and “−” indicate that the algorithm in the row is statistically superior, comparative, and inferior to the one in the column with 95% confidence. With the data in Table 10, we can observe that HyperIQA and TReS are statistically superior to all the other competitors on our dataset, demonstrating their superiority in quantifying image quality.

To comprehensively evaluate the performance of various IQA methods, computational complexity analysis is also conducted. Specifically, we carry out all IQA methods on the entire UID2021 database and the average running time of each method is recorded. It should be pointed out that this experiment is performed on a standard PC with an Intel Core i7-9700T and 16 GB of RAM. We list the average running time measured in seconds, as shown in Table 11.
Fig. 7. Distribution plots of MOS values with respect to the predicted scores of BIQI (a), CPBD (b), BLIINDS-II (c), NIQE (d), FADE (e), UIQE (f), UIQM (g), CCF (h), FDUM (i), HyperIQA (j), and TReS (k).

### 5.3 Evaluation of UIER Methods

Since the corresponding 900 enhanced images are generated from 60 source images by employing 15 underwater enhancement and restoration algorithms, it is also worthwhile to quantitatively analyze the performance of these algorithms. In the following, we calculate the average MOS values of each method on six subsets and the full UID2021, respectively, and their results are given in Figure 8. By comparing the average MOS values of different UIER algorithms in Figure 8, we can conclude that UWB-VCSE, CBF, and GLN-CHE are the relatively best performers, which proves
Table 10. Statistical Significant Analysis of NR IQA Algorithms

| BIQI | CPBD | BLIINDS-II | NIQE | FADE | UCIQE | UIQM | CCF | FDUM | HyperIQA | TReS |
|------|------|------------|------|------|-------|------|-----|------|-----------|------|
| -    | 1    | 1          | 0    | 0    | 0     | 0    | 0   | 0    | 0         | 0    |
| 0    | -    | 0          | 0    | 0    | 0     | 0    | 0   | 0    | 0         | 0    |
| 0    | 1    | -          | 0    | 0    | 0     | 0    | 0   | 0    | 0         | 0    |
| 1    | 1    | 1          | -    | 0    | 0     | 0    | 0   | 0    | 0         | 0    |
| 1    | 1    | 1          | 1    | -    | 0     | 0    | 0   | 0    | 0         | 0    |
| 1    | 1    | 1          | 1    | 1    | -     | 1    | 1   | -    | 0         | 0    |
| 1    | 1    | 1          | 1    | 1    | 1     | 0    | -   | 0    | 0         | 0    |
| 1    | 1    | 1          | 1    | 1    | 1     | 0    | 0   | -    | 0         | 0    |
| 1    | 1    | 1          | 1    | 1    | 1     | 1    | 1   | 1    | -         | -    |
| 1    | 1    | 1          | 1    | 1    | 1     | 1    | 1   | 1    | -         | -    |

Table 11. Computational Complexity Analysis of Various IQA Methods

| BIQI | CPBD | BLIINDS-II | NIQE | FADE | UCIQE | UIQM | CCF | FDUM | HyperIQA | TReS |
|------|------|------------|------|------|-------|------|-----|------|-----------|------|
| Time (s) | 1.0179 | 0.0672 | 2.4121 | 0.2483 | 0.3979 | 0.0616 | 1.2752 | 0.7694 | 1.1673 | 2.3245 | 2.5976 |

Fig. 8. Comparisons of the average of MOS of different underwater enhancement and restoration methods.

that the enhanced results of these algorithms are more in line with the human visual perception. Furthermore, the limitation of lacking training data of UWCNN leads to producing some unsatisfactory results. Moreover, the average MOS values of the images enhanced by underwater image restoration algorithms are relatively lower than the images enhanced by underwater image enhancement algorithms. The underwater image restoration algorithm builds an appropriate physical model and priors by studying the physical mechanism of underwater image degradation. However, the physical mechanism and the prior knowledge may not be applicable to various underwater scenes, which leads to the poor generalization ability of the underwater image restoration methods. Additionally, to further analyze the stability of these 15 UIER methods, the standard deviations of MOS for each method are also calculated and their results are tabulated in Table 12. By jointly analyzing Figure 8 and Table 12, it is interesting to note that the performance of several UIER algorithms fluctuated under different scenes, meaning that the current UIER algorithms are unable to generalize to all types of underwater images due to the complicated underwater environment and lighting conditions.
Table 12. Comparisons of the Standard Deviation of MOS of Different Underwater Enhancement and Restoration Methods

| Method       | Standard Deviation of MOS |
|--------------|---------------------------|
| Bayesian-retinex | 1.6715                    |
| CBF          | 1.8091                    |
| CHS          | 1.8934                    |
| GLN-CHE      | 1.8866                    |
| HP           | 1.8650                    |
| IBLA         | 1.5434                    |
| L^2UWE       | 1.8201                    |
| RCP          | 1.2489                    |
| TS           | 1.8237                    |
| Ucolor       | 1.4038                    |
| UNTV         | 1.7780                    |
| UTV          | 1.8239                    |
| UWB-VCSE     | 1.5483                    |
| UWCNN        | 0.9657                    |
| VR           | 1.6032                    |

6 CONCLUSION

In this article, we constructed a new large-scale underwater image dataset, which is highly desired in the field of UIQA. Our proposed dataset contains 900 quality improved images that are derived from 60 real underwater images by utilizing 15 SOTA underwater enhancement and restoration algorithms, generating 960 underwater images with different quality levels in total. The proposed dataset covers a diverse set of common underwater scenes, including a bluish scene, greenish scene, blue-green scene, hazy scene, turbid scene, and low-light scene. Traditional in-air NR-IQA methods, underwater-specific methods, and deep learning based methods are tested on our dataset, and their experimental results suggest that evaluating the qualities of underwater images is quite challenging for in-air IQA algorithms, and indicate the necessity of developing IQA algorithms specific to underwater images. Based on the study and analysis of the existing SOTA UIQA methods, we believe that the future UIQA methods need to be improved in the following two aspects. On the one hand, most of the existing UIQA methods tend to give high scores to the images with higher saturation, which leads to misjudgments of underwater image quality since underwater images usually contain heavy bluish and greenish color cast. On the other hand, UIQA methods tend to give high ratings to images with clear textures and details, even if these textures are too intense and become unnatural. It is recommended that future UIQA methods need to find a balance between blurred details and unnatural sharpening of details. To summarize, our constructed UID2021 not only complements the existing underwater image datasets as a tool for testing the performance of objective NR IQA metrics but also creates a platform for evaluating and optimizing the accuracy of newly upcoming UIQA metrics.

REFERENCES

[1] Miao Yang, Jintong Hu, Chongyi Li, Gustavo Rohde, Yixiang Du, and Ke Hu. 2019. An in-depth survey of underwater image enhancement and restoration. *IEEE Access* 7 (Aug. 2019), 123638–123657.
[2] Min Han, Zhiyu Lyu, Tie Qiu, and Meiling Xu. 2020. A review on intelligence dehazing and color restoration for underwater images. *IEEE Trans. Syst. Man Cybern. Syst.* 50, 5 (May 2020), 1820–1832.
[3] Saeed Anwar and Chongyi Li. 2020. Diving deeper into underwater image enhancement: A survey. *Signal Process. Image Commun.* 89 (Nov. 2020), 115978.
[4] Miao Yang and Arcot Sowmya. 2015. An underwater color image quality evaluation metric. *IEEE Trans. Image Process.* 24, 12 (Dec. 2015), 6062–6071.

[5] Karen Panetta, Chen Gao, and Sos Again. 2016. Human-visual-system-inspired underwater image quality measures. *IEEE J. Ocean. Eng.* 41, 3 (July 2016), 541–551.

[6] Yan Wang, Na Li, Zongying Li, Zhaoru Gu, Haiyong Zheng, Bing Zheng, and Mengnan Sun. 2017. An imaging-inspired no-reference underwater color image quality assessment metric. *Comput. Electron. Eng.* 70 (Dec. 2017), 904–913.

[7] Ning Yang, Qihang Zhong, Kun Li, Runmin Cong, Yao Zhao, and Sam Kwong. 2021. A reference-free underwater image quality assessment metric in frequency domain. *Signal Process. Image Commun.* 94 (March 2021), 116218.

[8] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. 2019. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* 29 (Nov. 2019), 4376–4389.

[9] Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. 2020. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE Trans. Circuits Syst. Video Technol.* 30, 12 (Jan. 2020), 4861–4875.

[10] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. 2021. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 8 (Aug. 2021), 2822–2837.

[11] Amanda Duarte, Felipe Codevilla, Joel De O. Gaya, and Silvia S. C. Botelho. 2016. A dataset to evaluate underwater image restoration methods. In *Proceedings of the 2016 MTS/IEEE OCEANS Conference (OCEANS’16).* 1–6.

[12] C. Sánchez-Ferreira, L. S. Coelho, H. V. H. Ayala, M. C. Q. Farias, and C. H. Llanos. 2019. Bio-inspired optimization algorithms for real underwater image restoration. *Signal Process. Image Commun.* 77 (Sept. 2019), 49–65.

[13] Chongyi Li, Saeed Anwar, and Fatih Porikli. 2020. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* 98 (Feb. 2020), 107038.

[14] Guojia Hou, Xin Zhao, Zhenkuan Pan, Huan Yang, Lu Tan, and Jingming Li. 2020. Benchmarking underwater image enhancement and restoration, and beyond. *IEEE Access* 8 (July 2020), 122078–122091.

[15] Di Wu, Fei Yuan, and En Cheng. 2020. Underwater no-reference image quality assessment for display module of ROV. *Sci. Program.* 2 (Aug. 2020), 1–15.

[16] Pengfei Guo, Lang He, Shuangyin Liu, Delu Zeng, and Hantao Liu. 2021. Underwater image quality assessment: Subjective and objective methods. *IEEE Trans. Multimedia* 24 (April 2021), 1980–1989.

[17] Alan Conrad Bovik, Muhammad Farooq Sabir, and Hamid R. Sheikh. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* 15, 11 (Nov. 2006), 3440–3451.

[18] Y. Horita, K. Shibata, and Y. Kawayoke. 2023. MICT Image Quality Evaluation Database. Retrieved January 10, 2023 from [https://computervisiononline.com/dataset/1105138668](https://computervisiononline.com/dataset/1105138668).

[19] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Jaakko Astola, Marco Carli, and Federica Battisti. 2009. TID2008: A database for evaluation of full-reference visual quality assessment metrics. *Adv. Modern Radioelectron.* 10, 4 (2009), 30–45.

[20] Toni Virtanen, Mikko Nuutinen, Mikko Vaahteronkoska, Pirkko Oittinen, and Jukka Häkkinen. 2015. CID2013: A database for evaluating image no-reference quality assessment algorithms. *IEEE Trans. Image Process.* 24, 1 (Jan. 2015), 390–402.

[21] Wen Sun, Fei Zhou, and Qingmin Liao. 2017. MDID: A multiply distorted image database for image quality assessment. *Pattern Recognit.* 61 (Jan. 2017), 153–168.

[22] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. 2017. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Trans. Image Process.* 26, 2 (Feb. 2017), 1004–1016.

[23] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. 2012. Objective quality assessment of multiply distorted images. In *Proceedings of the 2012 Conference Record of the 46th Asilomar Conference on Signals, Systems, and Computers (ASILOMAR’12).* 1693–1697.

[24] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.* 30 (Jan. 2015), 57–77.

[25] Silvia Corchs and Francesca Gasparini. 2017. A multidistortion database for image quality. In *Proceedings of the International Workshop on Computational Color Imaging (CCIW’17).* 95–104. [http://www.mmsp.unimib.it/image-quality/](http://www.mmsp.unimib.it/image-quality/).

[26] Alexandre Ciancio, André Luiz N. Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. 2011. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Trans. Image Process.* 20, 1 (Jan. 2011), 64–75.

[27] Deepti Ghadiyaram and Alan C. Bovik. 2016. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.* 25, 1 (Jan. 2016), 372–387.
Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2021. Uncertainty-aware blind image quality assessment. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’21), 14131–14140.

Jing Wang, Haotian Fan, Xiaoxia Hou, Yitian Xu, Tao Li, Xuechao Lu, and Lean Fu. 2022. MSTRIQ: No reference image quality assessment based on swin transformer with multi-stage fusion. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR’22). 1268–1277.

Rony Ferzli and Lina J. Karam. 2019. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). IEEE Trans. Image Process. 18, 4 (April 2009), 717–728.

Niranjan D. Narvekar and Lina J. Karam. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). IEEE Trans. Image Process. 20, 9 (Sept. 2011), 2678–2683.

Leida Li, Weisi Lin, Xuesong Wang, Gaobo Yang, Khosro Bahrami, and Alex C. Kot. 2016. No-reference image blur assessment based on discrete orthogonal moments. IEEE Trans. Image Cybern. 46, 1 (Jan. 2016), 39–50.

Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. 2015. Referenceless prediction of perceptual fog density and perceptual image defogging. IEEE Trans. Image Process. 24, 11 (Nov. 2015), 3888–3901.

Lixiong Liu, Tianshu Wang, and Hua Huang. 2019. Pre-attention and spatial dependency driven no-reference image quality assessment. IEEE Trans. Multimedia 21, 9 (Sept. 2019), 2305–2318.

Chenggang Yan, Tong Teng, Yutao Liu, Yongbing Zhang, Haoqian Wang, and Xiangyang Li. 2021. Precise no-reference image quality evaluation based on distortion identification. ACM Trans. Multimed. Comput. Commun. Appl. 17, 110 (Oct. 2021), 1–21.

Yannan Zheng, Weiling Chen, Rongfu Lin, Tiesong Zhao, and Patrick Le Callet. 2022. UIF: An objective quality assessment for underwater image enhancement. IEEE Trans. Image Process. 31 (2022), 5456–5468.
[54] Qiuping Jiang, Yueze Gu, Chongyang Li, Runmin Cong, and Feng Shao. 2022. Underwater image enhancement quality evaluation: Benchmark dataset and objective metric. *IEEE Trans. Circuits Syst. Video Technol.* 32, 9 (Sept. 2022), 5959–5974.

[55] Zhenqi Fu, Xueyang Fu, Yue Huang, and Xinghao Ding. 2022. Twicewing: A rank learning based quality assessment approach for underwater image enhancement. *Signal Process. Image Commun.* 102 (March 2022), 116622.

[56] Chunjie Guo, Ruiqi Wu, Xin Jin, Linghao Han, Zhi Chai, Weidong Zhang, and Chongyang Li. 2022. Underwater Ranker: Learn which is better and how to be better. arXiv:2208.06857.

[57] Adrian Galdran, David Pardo, Artzai Picón, and Aitor Alvarez-Gila. 2015. Automatic red channel underwater image restoration. *J. Vis. Commun. Image Represent.* 26 (Jan. 2015), 132–145.

[58] Guojia Hou, Zhenkuan Pan, Guodong Wang, Huan Yang, and Jinming Duan. 2019. An efficient nonlocal variational method with application to underwater image restoration. *Neurocomputing* 369 (Dec. 2019), 106–121.

[59] Chongyang Li, Jichang Guo, Runmin Cong, Yanwei Pang, and Bo Wang. 2016. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Trans. Image Process.* 25, 12 (Dec. 2016), 5664–5677.

[60] Xinjie Li, Guojia Hou, Lu Tan, and Wanquan Liu. 2020. A hybrid framework for underwater image enhancement. *IEEE Access* 8 (Oct. 2020), 2169–3536.

[61] Weidong Zhang, Peixian Zhuang, Haihan Sun, Guohou Li, Sam Kwong, and Chongyi Li. 2020. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Trans. Image Process.* 31 (June 2022), 3997–4010.

[62] Codruta O. Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, and Philippe Bekaert. 2018. Color balance and fusion for underwater image enhancement. *IEEE Trans. Image Process.* 27, 1 (Jan. 2018), 379–393.

[63] Nicholas Hope. n.d. Bubble Vision Underwater Imaging. Retrieved January 10, 2023 from https://bubblevision.com.

[64] Peixian Zhuang, Chongyang Li, and Jiamin Wu. 2021. Bayesian retinex underwater image enhancement. *Eng. Appl. Artif. Intell.* 101 (May 2021), 104171.

[65] Kashif Iqbal, Rosalina Abdul Salam, Azam Osman, and Abdullah Zawawi Talib. 2007. Underwater image enhancement using an integrated colour model. *IAENG Int. J. Comput.* 34, 2 (March 2007), 239–244.

[66] Xueyang Fu and Xiangyong Cao. 2020. Underwater image enhancement with global–local networks and compressed-histogram equalization. *Signal Process. Image Commun.* 86 (Aug. 2020), 115892.

[67] Guojia Hou, Zhenkuan Pan, Baoxiang Huang, Guodong Wang, and Xin Luan. 2018. Hue preserving-based approach for underwater colour image enhancement. *IET Image Process.* 12, 2 (Feb. 2018), 292–298.

[68] Yan-Tsung Peng and Pamela C. Cosman. 2017. Underwater retinex-based image restoration based on image bluriness and light absorption. *IEEE Trans. Image Process.* 26, 4 (April 2017), 1579–1594.

[69] Tunai Porto Marques and Alexandra Branzan Alb. 2020. L’UWE: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR’20)*, 538–539.

[70] Xueyang Fu, Zhiwen Fan, Mei Ling, Yue Huang, and Xinghao Ding. 2017. Two-step approach for single underwater image enhancement. In *Proceedings of the 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS’17)*, 789–794.

[71] Chongyang Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. 2021. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Trans. Image Process.* 30 (May 2021), 4985–5000.

[72] Jun Xie, Guojia Hou, Guodong Wang, and Zhenkuan Pan. 2022. A variational framework for underwater image dehazing and deblurring. *IEEE Trans. Circuits Syst. Video Technol.* 32, 6 (June 2022), 3514–3526.

[73] Guojia Hou, Jingming Li, Guodong Wang, Huan Yang, Baoxiang Huang, and Zhenkuan Pan. 2019. An efficient nonlocal variational method with application to underwater image restoration. *Neurocomputing* 369 (Dec. 2019), 106–121.

[74] Kashif Iqbal, Rosalina Abdul Salam, Azam Osman, and Abdullah Zawawi Talib. 2007. Underwater image enhancement using an integrated colour model. *IAENG Int. J. Comput.* 34, 2 (March 2007), 239–244.

[75] Xueyang Fu and Xiangyong Cao. 2020. Underwater image enhancement with global–local networks and compressed-histogram equalization. *Signal Process. Image Commun.* 86 (Aug. 2020), 115892.

[76] Guojia Hou, Zhenkuan Pan, Baoxiang Huang, Guodong Wang, and Xin Luan. 2018. Hue preserving-based approach for underwater colour image enhancement. *IET Image Process.* 12, 2 (Feb. 2018), 292–298.

[77] Yan-Tsung Peng and Pamela C. Cosman. 2017. Underwater retinex-based image restoration based on image bluriness and light absorption. *IEEE Trans. Image Process.* 26, 4 (April 2017), 1579–1594.

[78] Tunai Porto Marques and Alexandra Branzan Alb. 2020. L’UWE: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR’20)*, 538–539.

[79] Xueyang Fu, Zhiwen Fan, Mei Ling, Yue Huang, and Xinghao Ding. 2017. Two-step approach for single underwater image enhancement. In *Proceedings of the 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS’17)*, 789–794.

[80] Chongyang Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. 2021. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Trans. Image Process.* 30 (May 2021), 4985–5000.

[81] Jun Xie, Guojia Hou, Guodong Wang, and Zhenkuan Pan. 2022. A variational framework for underwater image dehazing and deblurring. *IEEE Trans. Circuits Syst. Video Technol.* 32, 6 (June 2022), 3514–3526.

[82] Guojia Hou, Jingming Li, Guodong Wang, Huan Yang, Baoxiang Huang, and Zhenkuan Pan. 2020. A novel dark channel prior guided variational framework for underwater image restoration. *J. Vis. Commun. Image Represent.* 66 (Jan. 2020), 102732.

[83] Xueyang Fu, Peixian Zhuang, Yue Huang, Yinghao Liao, Xiao-Ping Zhang, and Xinghao Ding. 2014. A retinex-based enhancing approach for single underwater image. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP’14)*, 4572–4576.

[84] Stefan Winkler. 2012. Analysis of public image and video databases for quality assessment. *IEEE J. Sel. Top. Signal. Process.* 6, 6 (Aug. 2012), 616–625.

[85] Yang Wang, Yang Cao, Jing Zhang, Feng Wu, and Zhengjun Zha. 2021. Leveraging deep statistics for underwater image enhancement. *ACM Trans. Multimed. Comput. Commun. Appl.* 17, 116 (Oct. 2021), 1–20.

[86] Xinjie Li, Guojia Hou, Kunqian Li, and Zhenkuan Pan. 2022. Enhancing underwater image via adaptive color and contrast enhancement, and denoising. *Eng. Appl. Artif. Intell.* 111 (May 2022), 104759.

[87] Qi Qi, Kunqian Li, Haiyong Zheng, Xiang Gao, Guojia Hou, and Kun Sun. 2022. SGUIE-Net: Semantic attention guided underwater image enhancement with multi-scale perception. *IEEE Trans. Image Process.* 31 (Oct. 2022), 6816–6830.
[79] Peixian Zhuang, Jiamin Wu, Fatih Porikli, and Chongyi Li. 2022. Underwater image enhancement with hyper-Laplacian reflectance priors. *IEEE Trans. Image Process.* 31 (Aug. 2022), 5442–5455.

[80] Jingchun Zhou, Tongyu Yang, Weishen Chu, and Weishi Zhang. 2022. Underwater image restoration via backscatter pixel prior and color compensation. *Eng. Appl. Artif. Intell.* 111 (May 2022), 104785.

[81] Jieyu Yuan, Zhanchuan Cai, and Wei Cao. 2022. TEBCF: Real-world underwater image texture enhancement model based on blurriness and color fusion. *IEEE Trans. Geosci. Remote Sens.* 60 (Oct. 2021), 4204315.

[82] Kuanqin Li, Li Wu, Qi Qi, Wenjie Liu, Xiang Gao, Liqin Zhou, and Dalei Song. 2022. Beyond single reference for training: Underwater image enhancement via comparative learning. *IEEE Trans. Circuits Syst. Video Technol.* Early access, November 28, 2022.

[83] Nan Li, Guojia Hou, Yuhai Liu, Zhenkuan Pan, and Lu Tan. 2022. Single underwater image enhancement using integrated variational model. *Digit. Signal Process.* 129 (2022), 103660.

[84] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV’22)*. 3989–3999.

Received 19 April 2022; revised 19 December 2022; accepted 23 December 2022