Underwater Image Quality Assessment: Benchmark Database and Objective Method

Yutao Liu, Baochao Zhang, Runze Hu, Ke Gu, *Senior Member, IEEE*, Guangtao Zhai, *Senior Member, IEEE*, and Junyu Dong, *Member, IEEE*

***Abstract*—Underwater image quality assessment (UIQA) plays a crucial role in monitoring and detecting the quality of acquired underwater images in underwater imaging systems. Currently, the investigation of UIQA encounters two major challenges. First, a lack of large-scale UIQA databases for benchmarking UIQA algorithms remains, which greatly restricts the development of UIQA research. The other limitation is that there is a shortage of effective UIQA methods that can faithfully predict underwater image quality. To alleviate these two challenges, in this paper, we first construct a large-scale UIQA database (UIQD). Specifically, UIQD contains a total of 5369 authentic underwater images that span abundant underwater scenes and typical quality degradation conditions. Extensive subjective experiments are executed to annotate the perceived quality of the underwater images in UIQD. Based on an in-depth analysis of underwater image characteristics, we further establish a novel baseline UIQA metric that integrates channel and spatial attention mechanisms and a transformer. Channel- and spatial attention modules are used to capture the image channel and local quality degradations, while the transformer module characterizes the image quality from a global perspective. Multilayer perception is employed to fuse the local and global feature representations and yield the image quality score. Extensive experiments conducted on UIQD demonstrate that the proposed UIQA model achieves superior prediction performance compared with the state-of-the-art UIQA and IQA methods.**

***Index Terms*—Attention mechanism, image database, image quality assessment (IQA), transformer, underwater image.**

I. Introduction

UNDERWATER imaging provides intuitive and vivid underwater information for a variety of underwater operations, such as ocean monitoring, resource exploitation, animal and plant protection, and engineering construction. However, unlike imaging in the atmosphere, underwater imaging involves many severe challenges, such as light scattering and absorption, exposure to low light or dark environments, low contrast, and marine snow. Hence, acquired underwater images often suffer from various undesirable effects, such as color casting, low contrast, blurring, and fog, which heavily degrade image quality. Therefore, effectively evaluating underwater image quality is an important premise for monitoring and detecting underwater image quality in underwater imaging systems.

Subjective quality assessment or visual inspection still occupies the dominant position in underwater image quality assessment (UIQA). Although subjective assessment by observers provides the most credible results, this approach is taxing and uneconomical and cannot satisfy real-time requirements. It is more promising to develop objective UIQA algorithms to evaluate underwater image quality automatically. In fact, objective IQA has been studied for a long time, and numerous superior IQA measures have been established [1], [2], [3], [4]. According to the availability of original pristine images, existing IQA measurements can be classified into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. Specifically, FR methods employ full information of the original image to predict the image quality [5], [6], [7], RR methods employ partial information of the original image for quality evaluation [2], [8], [9], and NR methods measure the image quality without referring to the original image [10], [11], [12]. Although these IQA models predict image quality effectively, most of them are designed for natural images, namely, images captured in the air. However, their ability to predict underwater images is still largely limited [13]. In underwater photography, underwater images are characterized by color casting, fog, and low contrast due to light absorption and scattering, which are divergent from natural images photographed in the air. Such divergence between the underwater image and the natural image leads to a decrease in the performance of the natural IQA methods in evaluating the underwater image quality. Therefore, specialized UIQA methods need to be developed.

Compared with natural image quality evaluation, the investigation of UIQA lags behind severely. First, large-scale underwater image quality databases (UIQDs) for fully examining the prediction performance of UIQA algorithms are lacking. The existing UIQDs are mostly small scale and contain only several hundred or even fewer than one hundred images [14], [15],

Manuscript received 1 November 2023; revised 8 January 2024 and 15 February 2024; accepted 22 February 2024. Date of publication 28 February 2024; date of current version 24 April 2024. This work was supported in part by the National Science Foundation of China under Grant 62201538, Grant 62301041, Grant 62076013, Grant 62273011, and Grant 62322302, and in part by Shandong Natural Science Foundation under Grant ZR2022QF006. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. B. Adsumilli. *(Corresponding author: Runze Hu.)*

Yutao Liu, Baochao Zhang, and Junyu Dong are with the School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: liuyutao@ouc.edu.cn; zhangbaochao@stu.ouc.edu.cn; dongjunyu@ouc.edu.cn).

Runze Hu is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100080, China (e-mail: hrzlpk2015@gmail.com).

Ke Gu is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke@bjut.edu.cn).

Guangtao Zhai is with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhaiguangtao@gmail.com).

The proposed UIQD and UIQA models will be released at https://github.com/YT2015?tab=repositories.

Digital Object Identifier 10.1109/TMM.2024.3371218

[16]. These databases cannot provide a fundamental test bed for evaluating UIQA methods comprehensively. Second, compared with imaging in air, underwater imaging is much more challenging, which leads to more complicated quality degradation of underwater images. However, existing UIQA methods mostly characterize several image properties by handcrafted features for quality evaluation [13], [17], which is preliminary and less effective. More advanced UIQA models are expected to be developed.

To facilitate UIQA research, we construct a large-scale UIQD through extensive subjective experiments and propose a novel baseline UIQA model in this paper. Specifically, we collected a total of 5369 underwater images to constitute our UIQD, which includes rich image content of different water types, depths, lighting conditions and underwater scenes. Then, subjective user studies are conducted to annotate the quality of these images. The obtained quality scores go through strict processing, e.g., outlier elimination, which results in the formation of an extensive and standard UIQD.

To predict underwater image quality more precisely, we design a novel neural network to learn more effective feature representations and deliver more accurate quality predictions. On the one hand, due to the different attenuation rates of different wavelengths of light, underwater images usually exhibit heavy color casts, e.g., greenish and bluish, which significantly affect image quality. Moreover, an imbalance in quality degradation can also occur across different image regions. For instance, underwater images often contain areas that are too bright or too dark due to the introduction of artificial light during acquisition. Inspired by the attention mechanisms in computer vision [18], we employ channel and spatial attention modules to emphasize quality-degraded channels and regions to deal with such quality imbalances across different channels and regions. On the other hand, image quality is also perceived from a global or nonlocal perspective [19], [20], [21], [22]. Specifically, the quality dependencies across different regions of an image have an impact on image quality [21]. Therefore, we introduce transformer module to characterize the global image quality, which has a strong capability to model the nonlocal relationship in an image [23], [24]. Finally, we employ a multilayer perception module to integrate the quality-related features extracted from the attention modules and transformer module and obtain the image quality score. For convenience, we name the proposed UIQA model attention and transformer-driven underwater image quality predictor, abbreviated as ATUIQP. The experimental results performed on our UIQD dataset demonstrate that ATUIQP yields more accurate prediction results than the state-of-the-art IQA and UIQA methods.

We summarize the contributions of this paper as follows:

- To the best of our knowledge, we constructed the largest authentic UIQD in the field of UIQA research to date. A total of 5369 underwater images were collected online; these images encompassed a wide range of underwater image contents and typical quality degradation conditions. We then carry out rigorous subjective experiments to label the quality of each image. The proposed UIQD will be made publicly available to serve as a comprehensive test platform for benchmarking UIQA algorithms.
- Different from most existing UIQA approaches that resort to handcrafted features to characterize image quality, we make the first attempt to employ channel-, spatial-attention mechanisms and transformers to characterize underwater image quality in UIQA research and establish a novel baseline UIQA model, i.e., ATUIQP. Specifically, ATUIQP learns more effectively feature representations end-to-end so that the model can represent the image quality precisely.
- Thorough experiments conducted on UIQD verify that the proposed ATUIQP achieves better prediction performance than the existing state-of-the-art IQA/UIQA models.

The remainder of this paper is arranged as follows. In Section II, we review the related works of this paper. In Section III, we introduce the proposed UIQD in detail. Section IV introduces the proposed UIQA method. Section V provides the experimental results and analysis. In Section VI, we present the conclusions of this paper.

II. Related Works

In this section, we overview related works about this paper, including existing UIQDs and objective UIQA methods.

A. Underwater Image Quality Assessment Databases

To facilitate research on underwater image quality evaluation, several representative UIQDs have been constructed. For instance, in [15], Wang et al. proposed an underwater image quality database by photographing a color checker in a water tank, which contains 87 authentic underwater images. Twenty students were invited to rate the subjective image quality. Tang et al. [16] collected 30 underwater images of different quality levels to benchmark underwater IQA algorithms. However, this database lacks subjective quality annotations. In [14], a relatively large-scale underwater image quality database was built by collecting or self-capturing 950 authentic underwater images. Various underwater image enhancement methods were applied to those images, and the best images were manually selected as reference images. However, this database also lacks subjective mean opinion score (MOS) values. To better assist in detecting fish in underwater images, Lin et al. [25] collected 145 high-quality underwater images and then applied different types of distortions to each image artificially, which led to a total of 2670 distorted underwater images. Subjective tests were also conducted to assess image quality. Due to the complexity of underwater imaging, there is still a large gap between artificially generated underwater images and authentic underwater images. In [26], Yang et al. constructed an underwater image quality assessment (UWIQA) database that contains 890 underwater images taken from [14]. Twenty-one observers participated in the subjective quality ratings. Zheng et al. [27] constructed an underwater image enhancement database (UIED) that contains 100 authentic underwater images and 1000 underwater images enhanced by 10 representative UIE methods. A subjective test was also conducted to evaluate the quality of these images. Liu et al. [28] set up a multiview

underwater imaging system to capture authentic underwater images and videos. Four thousand underwater images were selected to construct three underwater image databases: the underwater image quality set (UIQS), the underwater color cast set (UCCS) and the underwater higher-level task-driven set (UHTS). For the UIQS dataset, the underwater image quality metric UCIQE [13] was adopted to annotate the image quality. However, UCIQE scores are not consistent with subjective quality evaluations [14].

B. Underwater Image Quality Assessment Methods

Specialized UIQA approaches have been proposed by investigating the characteristics of underwater images. For example, in [13], Yang et al. proposed the widely adopted underwater IQA metric, named UCIQE, which combines the measurements of image chroma, contrast and saturation to estimate underwater image quality. Another well-known underwater IQA metric, the UIQM, characterizes image colorfulness, sharpness and contrast for quality evaluation [17]. The CCF metric measures image quality by characterizing colorfulness, contrast and fog density [15]. In [32], Li et al. designed an underwater IQA measure by evaluating the degree of color cast and visibility degradation. To improve underwater image visibility, Lu et al. [33] proposed a contrast enhancement method involving descattering and color correction. Then, they established a more comprehensive quality assessment scheme to benchmark the performance of underwater image restoration. Considering the degradation and color bias of underwater images, Tang et al. predicted underwater image quality by characterizing the sharpness, contrast and chroma, which are closely related to the human visual system (HVS) [34]. Yang et al. [35] proposed a reference-free underwater IQA metric by properly fusing colorfulness, contrast, and sharpness cues. Liu et al. [36] established an underwater image quality index to comprehensively evaluate underwater image quality by characterizing multiple underwater image properties, such as contrast, sharpness, color cast, fog, and noisiness. Zheng et al. [27] proposed an effective metric, namely, UIF, to evaluate the quality of enhanced underwater images; UIF exploits the statistical features of images across three key aspects, i.e., naturalness, sharpness, and structure. These features are subsequently combined with a saliency map to derive the image quality score. Wang et al. [37] proposed a powerful UIQA network named GLCQE that considers the degradation factors from both underwater optical imaging and enhancement algorithms. GLCQE generates two reference images, i.e., the unenhanced and the optimal enhanced versions of the input-enhanced images, which are then employed to characterize the image chrominance and luminance distortions. A parallel SA module was also designed to characterize the image sharpness. In [30], Guo et al. constructed a quality assessment method to evaluate the quality of enhanced underwater images; this method extracts quality-related features to characterize the colorfulness, sharpness and contrast of the image and subsequently uses support vector regression (SVR) to predict the image quality.

In summary, most existing UIQA methods extract quality-aware features to characterize limited image properties, such as contrast, color cast, and sharpness; then, they predict image quality by assigning different weights to the features. This prediction approach is restricted because handcrafted features cannot characterize image quality comprehensively. With the rapid development of deep neural networks, developing end-to-end UIQA approaches that can automatically learn more useful features from the data itself and thus predict image quality more accurately is highly desirable.

III. UIQD: Benchmark Underwater Image Quality Database

In this section, we elaborate upon the construction process of the proposed UIQD, which includes underwater image material collection, subjective tests and subjective score processing.

A. Underwater Image Material Collection

To construct a large-scale authentic underwater IQA database, We collected a total of 5369 underwater images from existing underwater computer vision works, such as underwater object detection, marine animal detection, and live fish recognition [38], [39], [40], [41], which encompass rich underwater scenes, e.g., plants, animals, wrecks, and rocks, and a variety of quality degradation conditions, e.g., color casting, low contrast, and insufficient luminance. Here, we show some sample images from UIQD in Fig. 1. It is clear that underwater scenes in UIQD are very abundant, which ensures the completeness of the image quality database.

Here, we also compare the proposed UIQD database with existing representative underwater image quality databases, as listed in Table I clearly. In general, existing underwater image databases can be categorized into two types: UIQA databases, i.e., UIQDs, and UIE assessment databases, i.e., UIEA databases. Among them, the UIEA database includes enhanced underwater images generated by different enhancement algorithms, which are used to benchmark underwater image enhancement methods. The UIQD includes underwater images acquired directly from imaging devices without enhancement; these images can be used to evaluate the quality of underwater images. Our proposed UIQD database belongs to the UIQA database.

By observing the Table I, we can see that the proposed UIQD is much larger than the existing UIQA databases, such as the Wang database [15], Tang database [16] and UWIQA [26]. Specifically, the number of annotated images in UIQD is approximately six times that in the UWIQA database, which guarantees more reliable UIQA benchmarking. The UIFD database [25] is composed of synthetic underwater images, which have a certain difference from authentically acquired underwater images. Therefore, the proposed UIQD is highly promising for use as a specialized UIQA database.

B. Subjective Tests

After image collection, a subjective test was performed to annotate the image quality. We set the configurations of the subjective experiments strictly in line with those of ITU-R

Fig. 1. Sample underwater images in the proposed UIQD. Images are resized for better visualization.

TABLE I
COMPARISON OF EXISTING UNDERWATER IMAGE QUALITY DATABASES

Database	Year	Image Amount	Image Types	Subjective Tests	Annotations	Usage
SUID [29]	2020	7495	Synthetic	N.A.	N.A.	UIEA
UIEB [14]	2019	950	Authentic	N.A.	N.A.	UIEA
RUIE [28]	2020	4230	Authentic	N.A.	N.A.	UIEA
Guo et al. [30]	2021	240	Authentic	SS	MOS	UIEA
SAUID [31]	2022	1100	Authentic	N.A.	N.A.	UIEA
UIED [27]	2022	1100	Authentic	SS	MOS	UIEA
Wang et al. [15]	2018	87	Authentic	SS	MOS	UIQA
Tang et al. [16]	2020	30	Authentic	N.A.	N.A.	UIQA
UWIQA [26]	2021	890	Authentic	SS	MOS	UIQA
UIFD [25]	2021	2670	Synthetic	SS	MOS	UIQA
UIQD (Pro.)	2023	5369	Authentic	SS	MOS	UIQA

'Usage' refers to underwater image enhancement (UIE) and UIQA, 'N.A.' indicates not available, 'SS' refers to single-stimulus methodology.

TABLE II
CONFIGURATIONS OF SUBJECTIVE TEST

Attribute	Setting
Method	Single-stimulus (SS)
Evaluation scales	Discrete scale from 1 to 5
Image number	5369
Subjects	18 inexperienced viewers
Image resolution	416×360 to 3840×2160
Viewing distance	Three times image height
Room illuminance	Dark

BT.500-12 [42]. For clarity, we list the main configurations of the subjective test in Table II. Since pristine images are unavailable, we chose the single-stimulus (SS) methodology to carry out a subjective quality test. Specifically, the SS method involves displaying an image for a short and fixed duration, and an observer is asked to rate its quality using a predefined quality scale, such as the commonly used Likert-type scale. As recommended by [42], the image quality is scaled into 5 levels, bad, poor, fair, good and excellent, corresponding to scores of 1, 2, 3, 4 and 5, respectively. We invited 18 inexperienced observers to rate the quality of each image according to their own quality perceptions. All the observers were college students who were naive to image quality assessment. The ages ranged from 20 to 30 years, and all the participants had normal vision. The room illuminance was set to dark to eliminate ambient lighting effects and minimize irrelevant visual stimuli from the surrounding environment, which allows subjects to be fully engaged in evaluating underwater image quality.

For convenience, we developed a graphical user interface in MATLAB to display the images and collect subjective ratings, as shown in Fig. 2. As observed, the subjective test for each person included two stages: the training stage and the testing

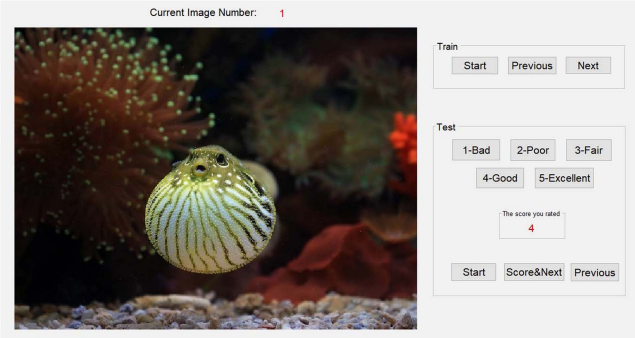


Fig. 2. Graphical user interface of subjective test.

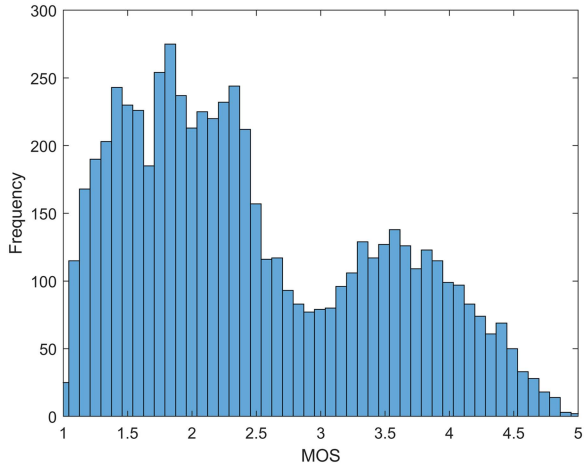


Fig. 3. Histogram of MOS values in UIQD.

stage. In the training stage, we show some examples of underwater images to the viewers, which calibrate the viewers to the UIQA task. Training images will not appear in the testing phase. In the formal testing stage, the presentation time of each image is set to 5 seconds, which is enough for viewers to judge the image quality correctly. Considering accidental clicking, we implemented the 'previous' button so that observers can modify the rating scores of previous images. Furthermore, each observer was asked to score continuously for no more than 30 minutes to prevent inaccurate ratings caused by visual fatigue. The viewing distance was set to three times the image height, and the room illuminance was set to dark. The evaluation criterion for image quality involves three main aspects: clarity, content discrimination, and color deviation.

C. Raw Subjective Quality Score Processing

After the subjective experiments of all the viewers, we obtained the raw subjective quality scores of the underwater images. However, outliers in these scores can occur due to various factors, such as participant bias, technical glitches, or extreme ratings. To ensure the reliability and quality of UIQD, we first performed outlier removal on the collected raw data. Guided by ITU-R BT.500-12 [42], we leveraged the 95% confidence interval to detect and remove outliers. Specifically, for the j-th underwater image, the 95% confidence interval is defined as $[\mu_j - \epsilon_j, \mu_j + \epsilon_j]$, where μ_j refers to the mean score of the j-th image for M subjective scores and ϵ_j is calculated as follows:

$$\epsilon_j = 1.96 \times \frac{\sigma_j}{\sqrt{M}}, \quad \sigma_j = \sqrt{\sum_{i=1}^{M} \frac{(r_{ij} - \mu_j)^2}{M-1}}, \tag{1}$$

where $i = 1, \ldots, M$ represents the number of observers and r_{ij} represents the subjective score for the j-th image from the i-th observer. μ_j is the mean of all the ratings for the j-th image. Scores outside the 95% confidence interval were considered outliers and were removed. In addition, if an observer gives a certain number of out-of-bounds scores, this observer is deemed not eligible, and all his or her ratings will be discarded. The final quality score of the j-th image was the arithmetic average value of the remaining scores and was computed as follows:

$$s_j = \frac{1}{\hat{M}} \sum_{i=1}^{\hat{M}} r_{ij}, \tag{2}$$

where $\hat{M}$ is the remaining score of the j-th image. s_j is termed the mean opinion score (MOS) of the jth image. Here, we give the MOS distribution of the entire UIQD in Fig. 3. By observing this figure, we can see that the MOS values span the entire abscissa axis, which intuitively demonstrates the favorable quality diversities of the underwater images in UIQD.

IV. The Proposed Attention and Transformer-Driven Underwater Image Quality Predictor

A. Overview of the Proposed ATUIQP

Most UIQA models still depend on handcrafted features to characterize several image properties for quality evaluation; however, these methods have relatively low efficiency and cannot adequately represent image quality. Recently, deep neural networks (DNNs) have achieved great success in computer vision, such as in image recognition [43], [44], classification [45], and segmentation [46], due to their outstanding nonlinear fitting ability. Therefore, in this paper, we build a novel UIQA model by means of DNN technologies to predict underwater image quality more effectively. The architecture of the proposed ATUIQP is shown in Fig. 4. ATUIQP contains two essential branches for extracting quality features in parallel. The upper branch contains multiscale feature extraction and the CSAM group. The multiscale feature extraction module includes three-scale convolutional layers to extract quality-aware features at different scales. The CSAM group is composed of five CSAM blocks, which integrate CA and SA modules to emphasize more informative channels and regions for quality evaluation, as shown in Fig. 4. The lower branch contains the transformer encoder and decoder layers for extracting quality features from a global perspective. The MLP module is introduced to fuse the features extracted from these two branches and output the quality score.

B. Characterization of Channel and Spatial Distortions

1) Multiscale Feature Representation: To comprehensively capture the distortions at different scales, we introduce three parallel convolution layers with different kernel sizes to yield a

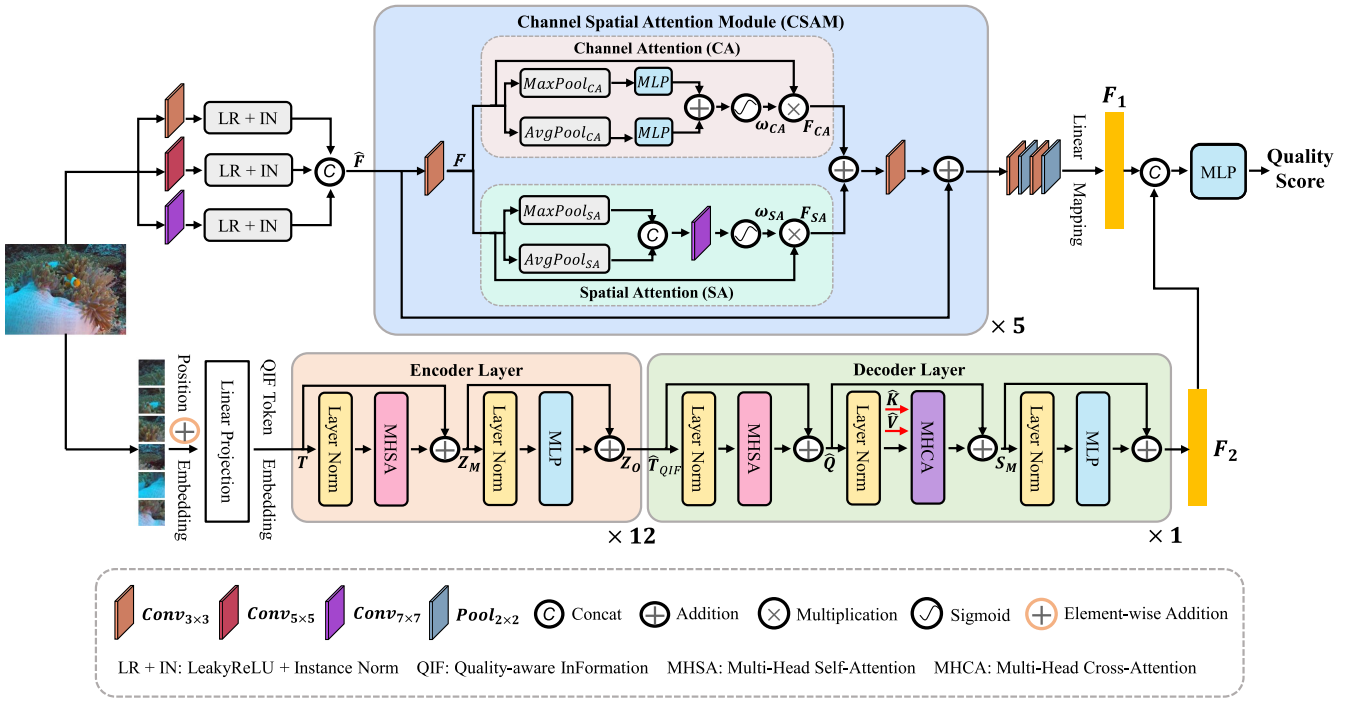


Fig. 4. Framework of the proposed ATUIQP.

multiscale feature representation of the input image. The convolution kernel sizes are set to 3×3, 5×5, and 7×7. The obtained feature maps after convolution are concatenated to form a multiscale feature representation, which can be formulated as follows:

$$\begin{aligned} \boldsymbol{F}_i &= \mathrm{IN}(\mathrm{ReLU}(\mathrm{conv}_i(I))), \text{for } i = 1 \sim 3, \\ \hat{\boldsymbol{F}} &= \mathrm{Concat}(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3), \end{aligned} \tag{3}$$

where IN refers to the instance normalization. The leaky rectified linear unit (ReLU) is adopted as the activation function. Conv is the convolutional layer with kernel sizes of 3×3, 5×5, and 7×7. *The zero padding* strategy is adopted to make the obtained feature maps the same size for direct concatenation. $\hat{\boldsymbol{F}}$ will be sent to the CSAM module to characterize the channel and spatial information of the image. As illustrated in Fig. 4, prior to the computation of channel and SA, $\hat{\boldsymbol{F}}$ is passed through a 3×3 convolutional layer to integrate these concatenated features, leading to $\boldsymbol{F}$, as in $\boldsymbol{F} = \mathrm{conv}(\hat{\boldsymbol{F}})$.

2) Channel Attention for Channel Context Characterization: Unlike natural images captured in the air, underwater images frequently suffer from heavy color casting due to differences in attenuation rates for light of different wavelengths. Specifically, red light has the highest attenuation rate, resulting in the acquired underwater images exhibiting green or bluish coloration. To address this effect, we introduce the channel-attention (CA) mechanism [18] to emphasize those quality-informative channels and suppress the effect of the less relevant or noisy channels in characterizing image quality.

As shown in Fig. 4, given the feature map $\boldsymbol{F} \in \mathbb{R}^{C \times H \times W}$ (Input), we first squeeze its spatial information through average pooling and max pooling. As a result, we can obtain two spatial context descriptors, as shown in $\boldsymbol{F}_{\text{avg}}^{(\text{CA})} \in \mathbb{R}^{C \times 1 \times 1}$ and $\boldsymbol{F}_{\text{max}}^{(\text{CA})} \in \mathbb{R}^{C \times 1 \times 1}$. Both average pooling and max pooling can be beneficial for computing the CA. Average pooling serves as the global context descriptor of each channel, whereas max pooling describes the channelwise salience information. Hence, they are complementary and enable the network to generate a more complete CA map. $\boldsymbol{F}_{\text{avg}}^{(\text{CA})}$ and $\boldsymbol{F}_{\text{max}}^{(\text{CA})}$ are then passed through a multilayer perceptron (MLP) block of one hidden layer to establish the CA map, denoted by $\boldsymbol{W}_{\text{CA}}$, written as:

$$\boldsymbol{W}_{\text{CA}} = \sigma\left(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{F}))\right), \tag{4}$$

where σ is the sigmoid activation function. We can obtain the channel-information enhanced feature map as follows:

$$\boldsymbol{F}_{\text{CA}} = \boldsymbol{F} \odot_{\text{c}} \boldsymbol{W}_{\text{CA}}, \tag{5}$$

where $\odot_{\text{c}}$ is the channelwise multiplication.

3) Spatial Attention for Local Distortion Characterization: Underwater images are prone to local quality degradation, e.g., local overbrightness caused by introducing artificial light sources in underwater imaging. Local distortion has an inevitable effect on underwater image quality and tends to draw increased amounts of attention when determining underwater image quality [47], [48]. To address this issue, we leverage the spatial attention (SA) mechanism to emphasize those regions that significantly affect image quality.

As Fig. 4 shows, we compute the SA map by exploring the interspatial relationships among features. The average-pooling and max-pooling operations are conducted on the feature map $\boldsymbol{F}$ in the channel dimension. We thus obtain two efficient feature significance descriptors, as in $\boldsymbol{F}_{\text{avg}}^{(\text{SA})} \in \mathbb{R}^{1 \times H \times W}$ and $\boldsymbol{F}_{\text{max}}^{(\text{SA})} \in$

$\mathbb{R}^{1\times H\times W}$. $\boldsymbol{F}_{\text{avg}}^{(\text{SA})}$ and $\boldsymbol{F}_{\text{max}}^{(\text{SA})}$ are then concatenated and sent to a convolutional layer to produce the final SA map, denoted by $\boldsymbol{W}_{\text{SA}}$, formulated as follows:

$$\boldsymbol{W}_{\text{SA}} = \sigma\left(\text{conv}(\text{Concat}(\text{AvgPool}(\boldsymbol{F}), \text{MaxPool}(\boldsymbol{F})))\right). \tag{6}$$

The spatially refined feature map can be obtained through spatialwise multiplication, written as follows:

$$\boldsymbol{F}_{\text{SA}} = \boldsymbol{F} \odot_{\text{s}} \boldsymbol{W}_{\text{SA}}. \tag{7}$$

where $\odot_{\text{s}}$ denotes the spatialwise multiplication. The CA and SA modules can be applied in either a parallel or a sequential paradigm. Here, we adopt the parallel architecture based on the experimental analysis (a detailed discussion can be found in Section V-D). We obtain the feature vector that characterizes the channel and spatial distortions through:

$$\boldsymbol{F}_1 = \text{MLP}(\text{MaxPool}(\text{conv}(\hat{\boldsymbol{F}} + \text{conv}(\boldsymbol{F}_{\text{CA}} + \boldsymbol{F}_{\text{SA}})))), \tag{8}$$

where $\boldsymbol{F}_1 \in \mathbb{R}^U$ refers to the feature vector whose dimension is determined by the hidden-layer size of the MLP block. U is set to 384 in this work.

C. Global Feature Representation via a Transformer

Image quality perception is actually a synthesis of local and global quality perception of the HVS [19], [20], [21], [22]. In the previous sections, we focused on characterizing the channel and local quality of the underwater image. In this section, we turn to characterize the image quality from the global perspective. Specifically, the HVS scans the entire image and perceives the qualities of different regions. All the results are subsequently compared and fused to obtain the global quality representation [21]. From this quality perception process, we know that the quality dependencies among different regions of the image play an important role in characterizing the global image quality. Recently, the transformer method was proposed to model nonlocal dependencies in a sentence or in an image; this approach has been proven to be very powerful in dealing with natural language processing [49] and computer vision [23] tasks. Inspired by this, we propose employing a transformer to characterize the quality dependencies among the nonadjacent regions to represent the global image quality.

As shown in Fig. 4, given an underwater image I, we first divide I into N patches as in $t_n \in \mathbb{R}^{p^2\times C_{\text{I}}}$, where p indicates the patch size and $N = \frac{H_{\text{I}}W_{\text{I}}}{p^2}$. Each patch is transformed into a D-dimensional embedding through a linear projection layer. Then, we employ a learnable embedding, namely, the quality-aware information (QIF) token $\boldsymbol{T}_{\text{QIF}} \in \mathbb{R}^{1\times D}$, to aggregate the perceptual information of underwater images. $\boldsymbol{T}_{\text{QIF}}$ is added to the N patch embeddings, yielding a total number of $N+1$ embeddings. Position embedding is also introduced into these $N+1$ embeddings to preserve the positional information.

1) Encoder Layer: We obtained the embedding sequence $\boldsymbol{T} = \{\boldsymbol{T}_{\text{QIF}}, \boldsymbol{T}_1, \ldots, \boldsymbol{T}_N\} \in \mathbb{R}^{N+1\times D}$. $\boldsymbol{T}$ is normalized and subsequently sent to the multihead self-attention (MHSA) block to compute the nonlocal dependencies. The MHSA block contains h heads each with dimensions $d = \frac{D}{h}$. $\boldsymbol{T}$ is transformed into three groups of matrices as in the query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and value $\boldsymbol{V}$ using three different linear projection layers as in $\boldsymbol{Q} = \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_h\} \in \mathbb{R}^{(N+1)\times D}$, $\boldsymbol{K} = \{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_h\} \in \mathbb{R}^{(N+1)\times D}$, and $\boldsymbol{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_h\} \in \mathbb{R}^{(N+1)\times D}$ for $\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h \in \mathbb{R}^{(N+1)\times d}$. For each head, the self-attention map that characterizes the nonlocal dependencies is computed by:

$$\text{Attention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h) = \text{softmax}\left(\frac{\boldsymbol{Q}_h {\boldsymbol{K}_h}^T}{\sqrt{d}}\right)\boldsymbol{V}_h. \tag{9}$$

We compute the SAs of all the heads and concatenate them to construct the final SA map. The output of the transformer encoder $\boldsymbol{Z}_O$ can be formulated as:

$$\begin{aligned}
\text{MHSA}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) =\ & \text{Concat}(\text{Attention}(\boldsymbol{Q}_1, \boldsymbol{K}_1, \boldsymbol{V}_1), \ldots \\
& \text{Attention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h))\boldsymbol{W}_L, \\
\boldsymbol{Z}_M =\ & \text{MHSA}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \boldsymbol{T}, \\
\boldsymbol{Z}_O =\ & \text{MLP}\,(\text{Norm}(\boldsymbol{Z}_M)) + \boldsymbol{Z}_M,
\end{aligned} \tag{10}$$

where $\boldsymbol{W}_L$ refers to the weight of the linear projection layer and norm$(\cdot)$ indicates layer normalization. $\boldsymbol{Z}_O$ is denoted as in $\boldsymbol{Z}_O = \{\boldsymbol{Z}_O[0], \ldots, \boldsymbol{Z}_O[N]\} \in \mathbb{R}^{(N+1)\times d}$.

2) Decoder Layer: The transformer encoder module effectively captures the nonlocal dependencies for various vision tasks, e.g., object detection and segmentation. However, the features extracted from the encoder may lack the necessary precision for downstream vision tasks, such as UIQA, which typically relies on a pretrained classification model. To mitigate this, a feature refinement operation is needed to further decouple more relevant features to the image quality. To this end, we introduce the transformer decoder, which incorporates a cross-attention mechanism to decipher the perceptual information provided by the encoder module.

Let $\hat{\boldsymbol{T}}_{\text{QIF}} = \boldsymbol{Z}_O[0] \in \mathbb{R}^{1\times D}$ be the QIF token obtained from the output of the encoder. $\hat{\boldsymbol{T}}_{\text{QIF}}$ is first sent to an MHSA block to model the dependencies between each element and the remaining elements of the QIF token. The output of the MHSA is followed by the residual connection to generate queries for the transformer decoder, written as follows:

$$\hat{\boldsymbol{Q}} = \text{MHSA}\left(\text{Norm}\left(\hat{\boldsymbol{T}}_{\text{QIF}}\right)\right) + \hat{\boldsymbol{T}}_{\text{QIF}}. \tag{11}$$

The role of the MHSA block is to decode the QIF token such that the produced query is more sensitive to quality-aware features. Following this, we utilize $\hat{\boldsymbol{Z}}_O = \{\boldsymbol{Z}_O[1], \ldots, \boldsymbol{Z}_O[N]\} \in \mathbb{R}^{N\times d}$ as the key and value of the decoder, denoted by $\hat{\boldsymbol{K}} = \hat{\boldsymbol{V}} = \hat{\boldsymbol{Z}}_O$, where $\hat{\boldsymbol{Z}}_O \cap \boldsymbol{Z}_O = \hat{\boldsymbol{T}}_{\text{QIF}}$. Then, $\hat{\boldsymbol{Q}}, \hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{V}}$ are sent to a multihead cross-attention (MHCA) block to perform the cross-attention. During this process, we utilize $\hat{\boldsymbol{Q}}$ to reinteract with the features of the image patches preserved in the encoder outputs, ensuring that the attentional features are more significant to the image quality. The output of the cross-attention mechanism is

written as follows:

$$\begin{aligned} \boldsymbol{S}_M &= \text{MHCA}(\text{Norm}(\hat{\boldsymbol{Q}}), \hat{\boldsymbol{K}}, \hat{\boldsymbol{V}}) + \hat{\boldsymbol{Q}}, \\ \boldsymbol{F}_2 &= \text{MLP}\left((\text{Norm}(\boldsymbol{S}_M)\right) + \boldsymbol{S}_M, \end{aligned} \tag{12}$$

where $\boldsymbol{F}_2$ indicates the refined quality-aware features from the encoder outputs, which are more comprehensive and accurate in describing the image quality.

D. Underwater Image Quality Prediction

To evaluate the underwater image quality, we adopt an MLP head of one hidden layer to map the extracted features to the image quality score. Specifically, we build the feature vector by concatenating $\boldsymbol{F}_1$ in (8) and $\boldsymbol{F}_2$ in (12). The quality score of an underwater image is calculated as follows:

$$q = \text{Concat}(\boldsymbol{F}_1, \boldsymbol{F}_2)\boldsymbol{W}_O, \tag{13}$$

where q refers to the predicted quality score and $\boldsymbol{W}_O$ denotes the weight of the MLP head. We minimize $\mathcal{L}_2$ loss to train the entire neural network, which is defined as:

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}(q_i - s_i)^2, \tag{14}$$

where N refers to the total number of training images, and q_i and s_i are the predicted quality score and ground-truth MOS value of the ith training image, respectively.

V. Experimental Results and Analysis

In this section, we fully evaluate the prediction performance of the proposed UIQA model on UIQD.

A. Experimental Protocol

To quantify the prediction performance of the objective UIQA methods, we use four widely adopted statistical measures, i.e., the Spearman rank order correlation coefficient (SRCC), Kendall's rank correlation coefficient (KRCC), Pearsons linear correlation coefficient (PLCC) and root mean square error (RMSE) are computed between the predicted quality scores and the subjective MOS values [59]. Specifically, the SRCC and KRCC values quantify the prediction consistency of the objective IQA methods, and the PLCC and RMSE indicate the prediction accuracy. Higher values of the SRCC, KRCC and PLCC and lower values of the RMSE demonstrate the preferable performance of an objective IQA model. Prior to calculating these values, a nonlinear logistic function is suggested to map the predicted quality scores to the subjective MOS values [59], defined as follows:

$$m(q) = \theta_1\left(\frac{1}{2} - \frac{1}{1+\exp\left(\theta_2\cdot(q-\theta_3)\right)}\right) + \theta_4\cdot q + \theta_5 \tag{15}$$

where $m(q)$ refers to the mapped quality score, q indicates the objective quality score, and $\theta_1 \sim \theta_5$ are the model parameters to be fitted in the curve fitting process.

B. Implementation Configurations

Some parameters in ATUIQP should be noted explicitly. For example, the number of CSAM modules was set to 5. A detailed discussion about the number of CSAM modules can be found in Section V-D. The numbers of input and output nodes of the fully connected layer were 3136 and 384, respectively. The number of transformer encoder layers was set to 12, and the number of heads in each encoder was set to 6. The number of transformer decoder layers was set to 1. We implemented the proposed ATUIQP on the PyTorch framework. The entire neural network was trained by the Adam optimizer on one NVIDIA RTX3090 card for 10 epochs. During training, the learning rate and batch size were set to 1e-4 and 32, respectively.

C. Prediction Performance Evaluation of UIQDs

We compare the proposed ATUIQP with nineteen state-of-the-art IQA and UIQA metrics, including CNN-IQA [50], WaDIQaM-NR [51], DBCNN [52], HyperIQA [53], TReS [54], BRISQUE [55], UNIQUE [56], DipIQ [57], NIQE [58], SNP-NIQE [3], NPQI [10], CCF [15], UCIQE [13], URQ [32], FDUM [35], UIQM [17], UIQI [36], UEIQM [30], and GLCQE [37]. Among these methods, the first eleven methods are general-purpose IQA methods, which are designed for natural images (photographed in the air). The remaining eight methods, along with the proposed ATUIQP, are UIQA methods. In addition, DipIQ, NIQE, SNP-NIQE, NPQI, CCF, UCIQE, URQ, FDUM, and UIQM fall into the category of unsupervised methods and do not require MOS values to construct the quality prediction model. In contrast, the other methods are supervised and need MOS values for training the quality prediction model. For the evaluation experiments, we employed the common 80%-20% split of UIQD, where 80% of the images were randomly selected to form the training set and the remaining 20% images formed the testing set. Correspondingly, the supervised IQA models were trained on the training set and then tested on the testing set. These unsupervised methods were tested only on the testing set. Furthermore, to mitigate performance bias, we repeated these training experiments 100 times. The average performance values across these 100 experiments on the testing sets of all the quality models are reported in Table III, where we use "(S)" and "(U)" to indicate supervised or unsupervised methods.

From Table III, we can make several meaningful observations. First, for the IQA methods, most supervised methods perform much better than unsupervised methods. This is because underwater images have different characteristics than natural images. Therefore, unsupervised methods that are designed for natural images, such as DipIQ and SNP-NIQE, fail to evaluate underwater image quality accurately. However, deep neural network-driven IQA methods, such as WaDIQaM-NR and DBCNN, which are designed for natural images, also perform superiorly; hence, DNNs for evaluating natural image quality have great potential for use in underwater image quality evaluation. This is mainly attributed to the strong abilities of DNNs to perform feature extraction and nonlinear fitting. Moreover, for the compared UIQA methods, i.e., CCF, UCIQE, URQ, FDUM,

TABLE III
PREDICTION PERFORMANCE COMPARISON IN TERMS OF SRCC, KRCC, PLCC, AND RMSE VALUES ON UIQD

Method	Type	SRCC	KRCC	PLCC	RMSE
CNN-IQA [50]	IQA (S)	0.7939	0.6005	0.7611	0.6297
WaDIQaM-NR [51]	IQA (S)	0.8910	0.7161	0.8937	0.4354
DBCNN [52]	IQA (S)	0.9023	0.7319	0.9010	0.4211
HyperIQA [53]	IQA (S)	0.8893	0.7153	0.8854	0.4511
TReS [54]	IQA (S)	0.8856	0.7099	0.8810	0.4586
BRISQUE [55]	IQA (S)	0.6528	0.4680	0.6143	0.7674
UNIQUE [56]	IQA (S)	0.8269	0.6303	0.8315	0.5399
DipIQ [57]	IQA (U)	0.4747	0.3216	0.5190	0.8313
NIQE [58]	IQA (U)	0.6225	0.4486	0.4551	0.8580
SNP-NIQE [3]	IQA (U)	0.4886	0.3324	0.3739	0.8998
NPQI [10]	IQA (U)	0.7078	0.5334	0.6841	0.7091
CCF [15]	UIQA (U)	0.5855	0.3984	0.5401	0.8181
UCIQE [13]	UIQA (U)	0.7545	0.5625	0.7743	0.6151
URQ [32]	UIQA (U)	0.5939	0.4397	0.5665	0.8009
FDUM [35]	UIQA (U)	0.8464	0.6580	0.8372	0.5318
UIQM [17]	UIQA (U)	0.7937	0.6079	0.7867	0.6003
UIQI [36]	UIQA (S)	0.8358	0.6409	0.8387	0.5292
UEIQM [30]	UIQA (S)	0.8121	0.6216	0.8232	0.5508
GLCQE [37]	UIQA (S)	0.8988	0.7287	0.8972	0.4283
ATUIQP (Pro.)	UIQA (S)	**0.9192**	**0.7570**	**0.9186**	**0.3837**

(S) refers to supervised method, (U) refers to unsupervised methods. The best results are emphasized with boldface.

UIQM, UIQI, and UEIQM, although they are designed specifically for evaluating underwater image quality, they can achieve only moderate prediction performance. These observations can be attributed to the fact that they all employ handcrafted features and assign different weights to different features artificially for quality evaluation. Nevertheless, such quality prediction methods are less effective because handcrafted features have difficulty characterizing underwater image quality comprehensively. In contrast, GLCQE and the proposed ATUIQP, which learn quality feature representations end-to-end through DNNs, perform much better than the other handcrafted UIQA models.

To examine the statistical significance of the obtained results, we performed a t test with 95% confidence on the differences between the mapped objective scores and the subjective MOS values, which are assumed to follow a Gaussian distribution. Specifically, the t test results of -1, 0, and 1 indicate that the proposed ATUIQP model is inferior, equal, and superior to the other IQA models. The experimental results are listed in Table IV. From this table, we observe that all the results are 1, which proves that ATUIQP is significantly superior to the other IQA/UIQA competitors.

In Fig. 5, we also present the scatter plots of the MOS values against the predicted scores of all IQA metrics. Each data point in each subfigure represents one underwater image, and the red curve was obtained through nonlinear regression. These experimental results are taken from one testing set in the 100-time experiments. The fitted curve of ATUIQP is a straight line, and the data points cluster more tightly around the red line, which intuitively demonstrates the superior prediction performance of the proposed ATUIQP. By comparison, the data points of some competing methods, such as CNN-IQA, DipIQ, and CCF, are distributed relatively loosely and irregularly around the fitted curve, implying a poor correlation between the MOS values and the predicted quality scores.

To intuitively demonstrate the correlation between the subjective MOSs and the predicted scores of ATUIQP, we select 8 underwater images of different quality levels from the UIQD database, whose MOSs range from 1.25 to 4.75 with an interval of 0.5, associated with their predicted scores of ATUIQP, as shown in Fig. 6. As the MOS increases, there is an obvious improvement in the visual quality of the images. Images with lower MOS values (less than 3) exhibit pronounced color casting and blurring. However, images of higher quality (MOS value greater than 4) reveal a more vibrant spectrum of colors and well-defined details. Moreover, the predicted quality scores of our proposed ATUIQP closely align with the subjective MOS values, which clearly demonstrates the efficacy of ATUIQP in predicting underwater image quality.

D. Ablation Study

To determine the importance of the proposed ATUIQP, we performed necessary ablation studies on UIQD. Generally, we design two key components in ATUIQP, i.e., the CSAM group and transformer, to characterize the channel, local and global quality of the underwater image. Therefore, we examined the individual contribution of each component to the prediction performance. Specifically, we utilized each component to construct the ATUIQP and tested its prediction accuracy. The train-test configurations are the same as those described in Section V-C. The experimental results in terms of the average values of the SRCC, KRCC, PLCC, and RMSE are reported in Table V.

From this table, we observe that each component (CSAM group or transformer) can lead to relatively high prediction accuracy, which verifies the effectiveness of these two components in characterizing image quality. In addition, it is noteworthy that the CSAM group demonstrated a greater level of significance in influencing the model's performance than the transformer component. These observations demonstrate that channel and spatial quality representations can play more important roles in the process of quality evaluation. Through the integration of these two components, the proposed ATUIQP achieves the best performance, which verifies that the two components work in a complementary manner to represent image quality.

Furthermore, in the CSAM component, we employ channel attention and SA mechanisms to characterize the channel and local quality degradations, respectively. We further investigated the effects of the CA and SA modules on ATUIQP. Similarly, we removed CA or SA from ATUIQP and tested it against UIQD. The experimental results measured by the SRCC, KRCC, PLCC and RMSE are listed in Table VI. The prediction performance of "CA+Transformer" is better than that of "SA+Transformer", which can be attributed to the fact that channel distortions are more common than spatial distortions in underwater images. ATUIQP, which integrates both CA and SA, achieves the best performance, which also indicates that CA and SA are complementary in capturing image distortions.

The arrangement of CA and SA modules in the CSAM component is also an influential factor of ATUIQP. Specifically, CA

TABLE IV
T-TEST RESULTS (95% CONFIDENCE) ON UIQD DATABASE BASED UPON THE GAUSSIANITY OF THE DIFFERENCES BETWEEN THE PREDICTED SCORES (AFTER NONLINEAR FITTING) AND THE MOS VALUES

IQA Methods											UIQA Methods							
CNN-IQA	WaDIQaM-NR	DBCNN	HyperIQA	TReS	BRISQUE	UNIQUE	DipIQ	NIQE	SNP-NIQE	NPQI	CCF	UCIQE	URQ	FDUM	UIQM	UIQI	UEIQM	GLCQE
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

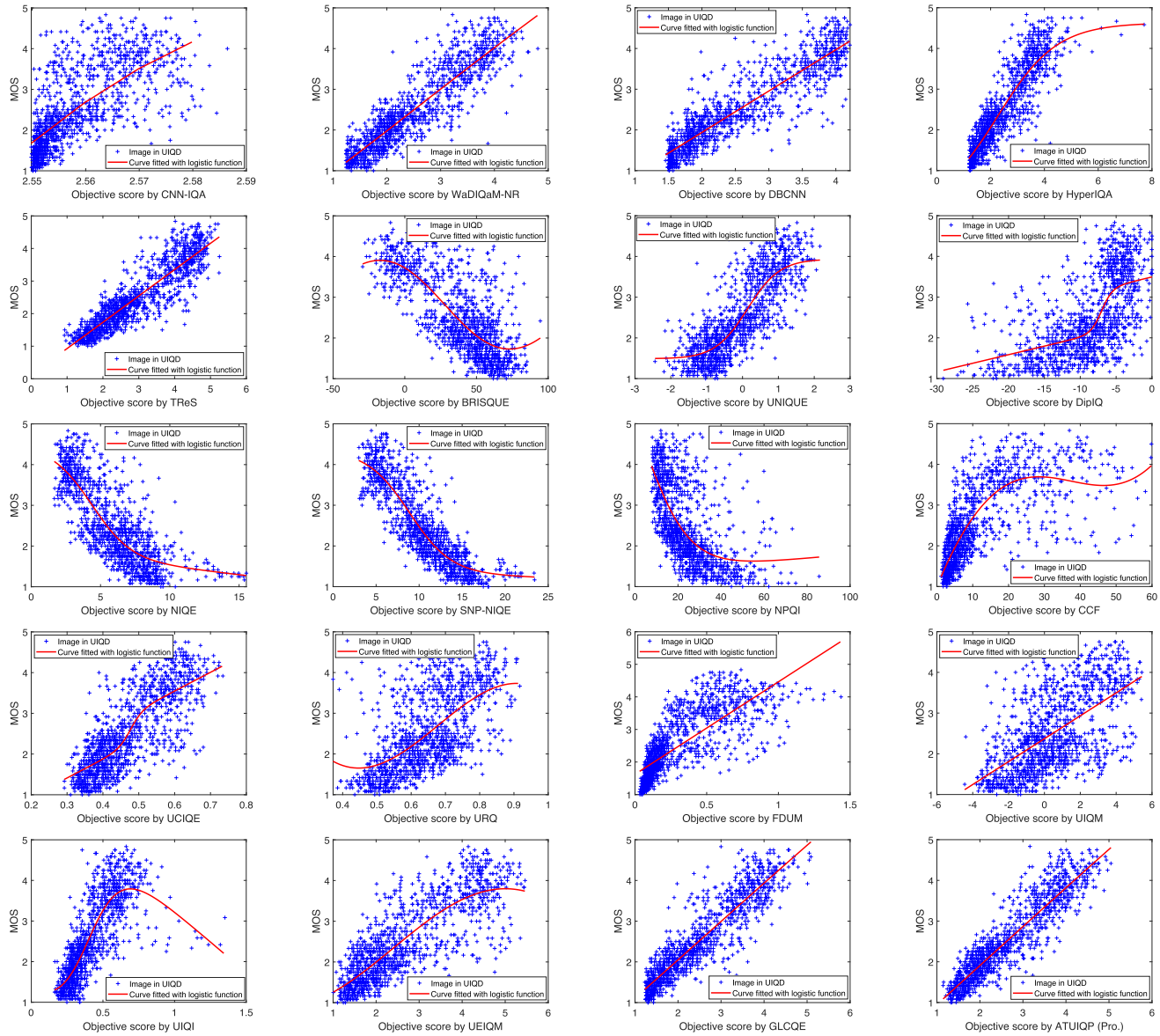


Fig. 5. Scatter plots of the objective scores against the subjective MOS values on one testing set. It's clearly observed that objective scores predicted by ATUIQP are more consistent with the MOS values.

Fig. 6. Sample underwater images of different quality levels associated with their MOS values and quality scores predicted by ATUIQP (MOS value / ATUIQP score). The MOS values range from 1.25 to 4.75 at the increment of 0.5.

TABLE V
PREDICTION PERFORMANCE EVALUATION OF EACH COMPONENT IN ATUIQP

Module	SRCC	KRCC	PLCC	RMSE
CSAM Group	0.8941	0.7199	0.8891	0.4591
Transformer	0.8491	0.6628	0.8501	0.5368
ATUIQP	0.9192	0.7570	0.9186	0.3837

TABLE VI
PREDICTION PERFORMANCE EVALUATION OF CA AND SA IN CSAM MODULE

Module	SRCC	KRCC	PLCC	RMSE
CA+Transformer	0.8834	0.7068	0.8860	0.4608
SA+Transformer	0.8742	0.6952	0.8733	0.4963
ATUIQP	0.9192	0.7570	0.9186	0.3837

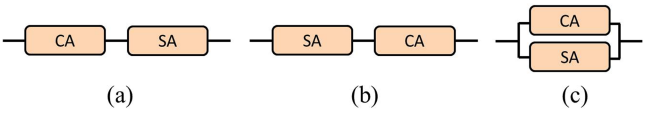


Fig. 7. Three types of arrangement manners of CA and SA in CSAM module. (a) CA-SA; (b) SA-CA; (c) Parallel.

TABLE VII
PREDICTION PERFORMANCE EVALUATION W.R.T. THE ARRANGEMENT MANNER OF CA AND SA

Module	SRCC	KRCC	PLCC	RMSE
CA-SA	0.9069	0.7356	0.9057	0.4366
SA-CA	0.9028	0.7300	0.9061	0.4330
Parallel (ATUIQP)	0.9192	0.7570	0.9186	0.3837

TABLE VIII
PREDICTION PERFORMANCE EVALUATION OF TRANSFORMER DECODER

Module	SRCC	KRCC	PLCC	RMSE
w/o Decoder	0.9011	0.7289	0.9046	0.4367
w Decoder (ATUIQP)	0.9192	0.7570	0.9186	0.3837

TABLE IX
PREDICTION PERFORMANCE EVALUATION OF TRANSFORMER ENCODERS. PARA. REFERS TO THE NUMBER OF TRAINABLE PARAMETERS

Module	SRCC	KRCC	PLCC	RMSE	Para.
Light-weight Encoder [24]	0.8616	0.6720	0.8502	0.5244	1.15Mb
Standard Encoder (ATUIQP)	0.9192	0.7570	0.9186	0.3837	1.77Mb

Para. refers to the number of trainable parameters.

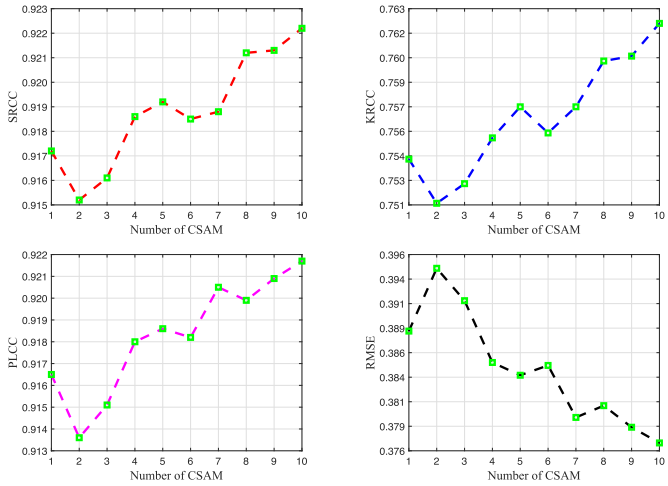


Fig. 8. Prediction performance variation w.r.t. the number of CSAM.

and SA can be arranged in a sequential or parallel manner, as illustrated in Fig. 7. We used these three kinds of arrangement manners to construct ATUIQPs. Their prediction performance values are listed in Table VII. From this table, we find that sequential arrangement manners, i.e., "CA-SA" and "SA-CA", deliver comparative prediction performance, which implies that the order of CA and SA in a sequential manner has little effect on the prediction performance. As observed, the parallel approach of CA and SA outperforms the sequential arrangement approach. Therefore, we arranged CA and SA in parallel in ATUIQP.

The transformer decoder module within the ATUIQP framework is introduced to refine the features extracted by the encoder layers. It enhances the discriminative capacity of features toward image quality attributes as opposed to object description. To verify the effect of the decoder, we removed the decoder from the ATUIQP dataset and tested its prediction performance. The results measured by the SRCC, KRCC, PLCC and RMSE are reported in Table VIII. It is clearly observed that the exclusion of the decoder module results in an obvious decrease in performance, which indicates that the decoder can further decode more accurate quality features for quality prediction.

We also investigate the lightweight architecture of the transformer encoder to assess its potential for enhancing model efficiency. To achieve this, we replaced the standard transformer encoder in ATUIQP with a lightweight variant, namely, LightViT [24]. The experimental results in terms of the SRCC, KRCC, PLCC, RMSE and parameter quantity are given in Table IX. As observed, the model's performance declines notably upon the replacement of the transformer encoder with the LightViT encoder, despite a substantial reduction in the number of parameters of the encoder (from 1.77 Mb to 1.15 Mb). This observation is not surprising since a higher-performing model requires a more complex and deeper network architecture. Therefore, the choice of the Transformer encoder can be tailored to meet specific practical application requirements.

Finally, we investigate the impact of the number of CSAM modules on the prediction performance. We varied the number of CSAMs from 1 to 10 and examined the variation in the prediction performance. The experimental results measured by the SRCC, KRCC, PLCC, and RMSE are shown in Fig. 8. As observed, the variations in these four metrics are all nonsignificant as the number of CSAMs increases, i.e., [0.915, 0.923] in the SRCC, [0.751, 0.763] in the KRCC, [0.913, 0.922] in the PLCC, and [0.376, 0.396] in the RMSE, which means that the number of CSAM modules has little impact on the prediction performance. In addition, the performance values reach a local optimum when the CSAM number reaches 5. In addition, additional CSAM

TABLE X
TIME COST OF THE IQA AND UIQA METHODS (MEASURED IN SECONDS)

CNN-IQA	WaDIQaM-NR	DBCNN	HyperIQA	TReS	BRISQUE	UNIQUE	DipIQ	NIQE	SNP-NIQE
0.0012	0.0115	0.0043	0.0130	0.0336	0.5400	0.0071	2.3600	0.1100	1.2998
NPQI	CCF	UCIQE	URQ	FDUM	UIQM	UIQI	UEIQM	GLCQE	ATUIQP
9.7000	4.3000	0.7300	0.3700	9.7700	1.1700	2.5725	0.5833	0.0760	0.0122

Fig. 9. Some failure cases of ATUIQP (MOS value / ATUIQP score).

modules significantly increase the model size. Therefore, by properly balancing the prediction performance and model size, we set the CSAM number to 5 in ATUIQP by default.

E. Computational Time Evaluation

We finally evaluate the computational efficiency of the proposed ATUIQP, which is an important indicator of an IQA model for practical deployment. Specifically, we chose a standard image of size 1920×1080 from our UIQD database and employed all the IQA and UIQA methods to compute its quality score. These experiments were conducted on a computer with a 2.1-GHz Intel Core i7-12700 CPU, 64 GB RAM and a graphics card NVIDIA RTX3090 with 24 G graphics RAM. The software platforms used were MATLAB R2019b and PyTorch. The individual processing times are recorded and listed in Table X. From this table, we observe that our ATUIQP model processes a single image in 0.0122 seconds, which is very promising for real applications.

F. Limitations

We demonstrated the effectiveness of ATUIQP in predicting underwater image quality. However, images cannot be used to evaluate ATUIQP accurately. We present some typical images in Fig. 9. All the MOSs of these three images are higher than the predicted quality scores. This discrepancy suggests that images have clear, discernible semantic information in the foregrounds despite the presence of color casts or fog in these images, which may cause subjects to give relatively higher quality scores. Such a phenomenon aligns with the perceptual characteristics of human visual perception that naturally incline to prioritize meaningful content over certain quality imperfections. For example, in the context of underwater photography, elements such as marine life and underwater structures play a crucial role in how subjects engage with and interpret an image. Consequently, even if an image exhibits some quality degradations, the interpretability of the semantic content can lead subjects to perceive it as relatively high quality. However, the proposed ATUIQP, which integrates local and global information for quality evaluation, does not account for semantics specifically and cannot deliver quality scores as high as subjective MOS values. These observations demonstrate that the useful semantic information could also influence the image quality. In our future work, mechanisms that account for the perceptual significance of semantic content and contextual information can be explored, and thus further enhancing the consistency between model's predictions and human perceptual judgments.

VI. CONCLUSION

To address these challenges in UIQA research, in this paper, we constructed a large-scale authentic UIQA database, named UIQD, containing 5369 authentic underwater images; this dataset is much larger than the existing authentic UIQA database. UIQD covers plentiful underwater scenes and typical quality degradation conditions. Strict subjective experiments were conducted to annotate the quality of the underwater images in UIQD, making it a standard database for UIQA investigations. Then, in contrast to most existing handcrafted UIQA methods, we established an end-to-end UIQA network, named ATUIQP, which incorporates channel and spatial attention mechanisms and a transformer for the first time to characterize the image channel and spatial and global quality degradation precisely. The experimental results fully demonstrate the superiority of ATUIQP over other state-of-the-art IQA and UIQA models in predicting underwater image quality. From this work, we expect more superior UIQA models to be developed in the future.

REFERENCES

[1] B. Yan, B. Bare, and W. Tan, "Naturalness-aware deep no-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2603–2615, Oct. 2019.
[2] Y. Liu et al., "Reduced-reference image quality assessment in free-energy principle and sparse representation," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 379–391, Feb. 2018.
[3] Y. Liu et al., "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, Apr. 2020.
[4] G. Zhai, Y. Zhu, and X. Min, "Comparative perceptual assessment of visual signals using free energy features," *IEEE Trans. Multimedia*, vol. 23, pp. 3700–3713, 2021.
[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[6] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro-and macro-structures," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3903–3912, May 2017.
[7] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[8] F. Qi, D. Zhao, and W. Gao, "Reduced reference stereoscopic image quality assessment based on binocular perceptual information," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2338–2344, Dec. 2015.

[9] L. Ma, S. Li, F. Zhang, and K. N. Ngan, "Reduced-reference image quality assessment using reorganized DCT-based image representation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 824–829, Aug. 2011.

[10] Y. Liu, K. Gu, X. Li, and Y. Zhang, "Blind image quality assessment by natural scene statistics and perceptual characteristics," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 3, pp. 1–91, 2020.

[11] R. Hu, Y. Liu, K. Gu, X. Min, and G. Zhai, "Toward a no-reference quality metric for camera-captured images," *IEEE Trans. Cybern.*, vol. 53, no. 6, pp. 3651–3664, Jun. 2023.

[12] Y. Liu, K. Gu, S. Wang, D. Zhao, and W. Gao, "Blind quality assessment of camera images based on low-level and high-level statistical features," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 135–146, Jan. 2019.

[13] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062–6071, Dec. 2015.

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

[15] Y. Wang et al., "An imaging-inspired no-reference underwater color image quality assessment metric," *Comput. Elect. Eng.*, vol. 70, pp. 904–913, 2018.

[16] S. Tang, C. Li, and Q. Tian, "Underwater image quality assessment based on human visual system," in *Proc. 13th Int. Congress Image Signal Process. BioMed. Eng. Informat.*, 2020, pp. 378–382.

[17] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541–551, Jul. 2016.

[18] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[19] A. Shnayderman, A. Gusev, and A. M. Eskicioglu, "An SVD-based grayscale image quality measure for local and global assessment," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 422–429, Feb. 2006.

[20] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4923–4936, Oct. 2017.

[21] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1778–1791, Apr. 2022.

[22] W. Zhou et al., "Local and global feature learning for blind quality evaluation of screen content and natural scene images," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2086–2095, May 2018.

[23] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] T. Huang et al., "LightViT: Towards light-weight convolution-free vision transformers," 2022, *arXiv:2207.05557*.

[25] R. Lin, T. Zhao, W. Chen, Y. Zheng, and H. Wei, "Underwater image quality database towards fish detection," in *Proc. IEEE/CIC Int. Conf. Commun. China*, 2021, pp. 205–210.

[26] N. Yang et al., "A reference-free underwater image quality assessment metric in frequency domain," *Signal Process.: Image Commun.*, vol. 94, 2021, Art. no. 116218.

[27] Y. Zheng, W. Chen, R. Lin, T. Zhao, and P. L. Callet, "UIF: An objective quality assessment for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 31, pp. 5456–5468, 2022.

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

[29] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, 2020, Art. no. 107038.

[30] P. Guo, L. He, S. Liu, D. Zeng, and H. Liu, "Underwater image quality assessment: Subjective and objective methods," *IEEE Trans. Multimedia*, vol. 24, pp. 1980–1989, 2022.

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

[32] L. Li, "Underwater color image quality assessment," 2020. [Online]. Available: https://github.com/LangtaoLi/Underwater-color-image-quality-assessment

[33] H. Lu et al., "Underwater image descattering and quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1998–2002.

[34] S. Tang, C. Li, and Q. Tian, "Underwater image quality assessment based on human visual system," in *Proc. IEEE Proc. 13th Int. Congr. Image Signal Process. BioMed. Eng. Informat.*, 2020, pp. 378–382.

[35] N. Yang et al., "A reference-free underwater image quality assessment metric in frequency domain," *Signal Process. Image Commun.*, vol. 94, 2021, Art. no. 116218.

[36] Y. Liu et al., "UIQI: A comprehensive quality evaluation index for underwater images," *IEEE Trans. Multimedia*, vol. 26, pp. 2560–2573, 2023.

[37] Z. Wang, L. Shen, Z. Wang, Y. Lin, and Y. Jin, "Generation-based joint luminance-chrominance learning for underwater image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1123–1139, Mar. 2023.

[38] C. Liu et al., "A dataset and benchmark of underwater object detection for robot picking," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2021, pp. 1–6.

[39] M. Pedersen, J. B. Haurum, R. Gade, and T. B. Moeslund, "Detection of marine animals in a new underwater dataset with varying visibility," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 18–26.

[40] H. Qin, X. Li, J. Liang, Y. Peng, and C. Zhang, "DeepFish: Accurate underwater live fish recognition with a deep architecture," *Neurocomputing*, vol. 187, pp. 49–58, 2016.

[41] "Kaggle starfish," 2021. [Online]. Available: https://www.kaggle.com/competitions/tensorflow-great-barrier-reef/overview/about-co-sponsor-csiro

[42] BT Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, vol. 500, no. 13, 2012.

[43] S. Kumawat, T. Okawara, M. Yoshida, H. Nagahara, and Y. Yagi, "Action recognition from a single coded image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4109–4121, Apr. 2023.

[44] A. Thatipelli et al., "Spatio-temporal relation modeling for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19926–19935.

[45] M. Masana et al., "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5513–5533, May 2023.

[46] J. R. Clough et al., "A topological loss function for deep-learning based image segmentation using persistent homology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8766–8778, Dec. 2022.

[47] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[48] Y. Liu et al., "Quality assessment for real out-of-focus blurred images," *J. Vis. Commun. Image Representation*, vol. 46, pp. 70–80, 2017.

[49] A. Vaswani et al., "Attention is all you need," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 30, 2017.

[50] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.

[51] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[52] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[53] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3667–3676.

[54] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1220–1230.

[55] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

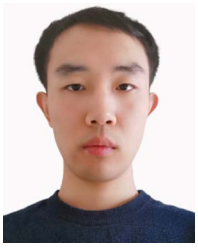
[56] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

[57] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[58] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[59] A. M. Rohaly et al., "Final report from the video quality experts group on the validation of objective models of video quality assessment," ITU-T Standards Contribution COM, vol. 1, pp. 9–80, 2000.

Yutao Liu received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2011, 2013, and 2018, respectively. From 2018 to 2021, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. He is currently an Associate Professor with the Ocean University of China, Qingdao, China. His research interests include image quality assessment, image enhancement, and computer vision.



Baochao Zhang received the B.S. degree from the School of Information Science and Engineering, Yanshan University, Qinhuangdao, China, in 2021. He is currently working toward the master's degree with the School of Computer Science and Technology, Ocean University of China, Qingdao, China. His research interests include image quality assessment, image enhancement and deep learning.

Runze Hu received the B.S. degree in computer science from North China Electric Power University, Baoding, China, in 2014, and the M.Sc. degree in computer science and the Ph.D. degree in electrical and electronics engineering from the University of Manchester, Manchester, U.K., in 2016 and 2020, respectively. His current research interests include image quality assessment, uncertainty quantification techniques, and underwater acoustic ranging.

Ke Gu (Senior Member, IEEE) is currently a Professor with the Beijing University of Technology, Beijing, China. His research interests include environmental perception, image processing, quality assessment, and machine learning. Dr. Gu was the recipient of the Best Paper Award from the IEEE Transactions on Multimedia (T-MM) and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo (ICME) in 2016. He was the Leading Special Session Organizer in the IEEE International Conference on Visual Communications and Image Processing (VCIP) 2016 and the IEEE International Conference on Image Processing (ICIP) 2017. He is also an Associate Editor for *Computer Animation and Virtual Worlds* (CAVW) and *IET Image Processing* (IET-IPR), an Area Editor of *Signal Processing: Image Communication* (SPIC), and an Editor for Applied Sciences, Displays, and Entropy. He is also a reviewer for 20 top SCI journals.

Guangtao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. His research interests include multimedia signal processing and perceptual signal processing. He was the recipient of the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.

Junyu Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees in applied mathematics from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, Edinburgh, U.K., in 2003. He is currently a Professor and the Head of the Department of Computer Science and Technology, Ocean University of China. His research interests include machine learning, Big Data, computer vision, and underwater image processing.