Cognitive Computation (2024) 16:191–214
https://doi.org/10.1007/s12559-023-10197-6

TEGAN: Transformer Embedded Generative Adversarial Network for Underwater Image Enhancement

Zhi Gao[1] · Jing Yang[1] · Lu Zhang[2] · Fengling Jiang[3] · Xixiang Jiao[1]

Received: 21 March 2023 / Accepted: 21 August 2023 / Published online: 5 September 2023

Abstract
Underwater robots are widely used in underwater missions. However, due to complex scenes, it is difficult to obtain high-quality underwater images, which usually suffer from severe distortions such as low visibility, blurred edges, and color cast. In this paper, a Transformer embedded generative adversarial network for underwater image enhancement is presented. We propose a window-based dual local enhancement block to compensate for the Transformer's shortcomings in extracting local features and improving image clarity. Convolutional neural network is deployed in sequential and parallel modes for local enhancement. Second, for generator construction, a fusion scheme combining convolutional neural network and Transformer block in units is designed. We exploit a self-attention mechanism to extract long-distance dependencies and fully extract the original features at the initial stage to enhance the image details. Meanwhile, global information is captured through the bottleneck for color correction. Convolutional neural network, which is good at extracting local features, is introduced in Encoder/Decoder units for multiscale feature extraction and reconstruction to effectively reduce edge blurring. Finally, a Transformer embedded generative adversarial network with a two-branch discriminator is established to generate more realistic colors while preserving the image content. Comparative experimental results show that our method can achieve superior results to the state-of-the-art approaches on both paired and unpaired datasets. Excellent learning and generalization ability make it outperform others in subjective perception and overall performance evaluated by image quality metrics. In addition, the enhancement results also show the significant improvement it brings in the downstream visual application tasks.

Keywords Transformer · Generative adversarial network · Underwater image enhancement · Dual local enhancement · Two-branch discriminator

Introduction

Underwater robots with vision guidance have become increasingly common in critical applications in recent years. Examples include underwater exploration [1], monitoring marine species [2], and underwater rescue missions [3]. Approximately 70% of the Earth's area is the sea, which is closely related to human life, but human exploration of the sea is still less than 10%. Unlike in-air images, because of the complex and diverse underwater environments, underwater images suffer from various degradations. According to the principle of underwater imaging, water absorption during light propagation and the forward/back scattering of suspended particles in water are the main factors contributing to degradation. Absorption is mostly responsible for color distortion, while light attenuation is nonlinear and related to the wavelength of light. Due to the large wavelength of red light, it is absorbed faster with depth in water. Thus, most underwater images are greenish and bluish. In addition, forward scattering causes blurred details in underwater images, while backward scattering leads to low contrast and thus a haze effect.

Many traditional methods have been developed for underwater image enhancement (UIE) [4–15]. Although in certain ways, these traditional methods have achieved good results, when dealing with various kinds of underwater environments, there still exist some disadvantages. As shown in

✉ Jing Yang
yangjing@hfuu.edu.cn

1 School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China

2 Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China

3 Hefei Normal University, Hefei 230061, China

Fig. 1, the diversity of image quality degradation usually includes color cast, haze, and blur.

Recently, generative adversarial network (GAN) [16] and Transformer [17] have already been effectively implemented for the translation task of images. The two-player zero-sum game serves as an inspiration for GAN, which is mainly applied to image generation and data enhancement and is further developed for other tasks. It consists of two models: a generator G that captures the data distribution and a discriminator D. The two neural networks are trained simultaneously to finally achieve that the samples generated by the generator are real samples. GAN is also used in unsupervised learning, such as CycleGAN [18], which is used to generate paired datasets to guide deep learning network training. However, standard GAN suffers from mode collapse and vanishing gradients, making training unstable. In addition, because the discriminator only contains one branch and mainly focuses on some of the image's content and details, the color features of the image are challenging to handle. Transformer emerged from the field of NLP. Transformer abandons the traditional convolutional neural network (CNN) and recurrent neural network (RNN). The whole network is composed of self-attention and a feedforward neural network. Due to Transformer's ability to capture long-range dependencies, it has also been successfully applied to the field of computer vision [19]. However, Transformer suffers from high computational consumption and a weak ability to extract local features. In short, Transformer contributes to the network learning capabilities, and GAN contributes to the network learning goals.

To fully utilize the respective advantages of Transformer and GAN, we effectively fuse the two together. First, we propose a window-based dual local enhancement Transformer block (DleWin), which is more suitable for UIE tasks. The DleWin block implements a self-attention mechanism to extract long-range information well. On the other hand, local features are crucial to UIE tasks. We adopt CNN in serial and parallel modes in the DleWin block for local enhancement, and the generator is built based on the DleWin block. Second, we propose a fusion scheme that combines convolutional neural network and Transformer in units. Since Transformer is good at capturing long-range dependencies and extracting raw information, while CNN is good at extracting local features, the two can be effectively fused to correct the color deviation and enhance image clarity. To make it easier for the DleWin block to obtain global information, the generator is designed as a UNet-like network [20]. In the framework of the generator, the DleWin Transformer block is implemented to extract the raw and global information. Finally, we propose a GAN with a two-branch discriminator containing a feature branch and a color branch. The feature branch is used to preserve image features and enhance contrast, while the color branch performs color correction to generate more realistic colors. For the design of the discriminator, we implement it as stacked convolutional layers. The feature branch training is guided by the Wasserstein GAN with gradient penalty (WGAN-GP) [21] loss, and the underwater index loss (Uloss) [22] is used to guide the training of the color branch. Based on the

Fig. 1 Underwater images with various degradations

above three designs, we propose a Transformer embedded generative adversarial network for underwater image enhancement (TEGAN). Comparative experimental results demonstrate that, for both paired and unpaired datasets, our method is superior to the state-of-the-art approaches. It achieves not only the best subjective perception effect but also the overall best performance in terms of image quality evaluation metrics. To show the contribution of each core component, ablation analyses are conducted. In addition, we also test the effect on the downstream tasks. According to the findings, TEGAN can greatly boost the efficiency of visual tasks such as edge detection, underwater object detection, and keypoint matching.

Unlike other comparative methods, TEGAN with a new Transformer block, fusion scheme, and two-branch discriminator is very suitable for solving the degradation problem of underwater images. The contributions are organized as follows:

- We propose a window-based dual local enhancement Transformer block (DleWin) that is more suitable for the UIE task. This novel block can be used to fully extract the original features and global information of the image, alleviate blur, and improve image clarity.
- A fusion scheme that combines convolutional neural network and Transformer in units is designed. According to the fact that CNN is good at extracting local features and Transformer can capture long-distance dependencies well, the two can be effectively fused to correct the color deviation and enhance image clarity.
- A Transformer embedded generative adversarial network with a two-branch discriminator is proposed. The feature branch preserves image features and realizes contrast enhancement, while the color branch rectifies the color cast to generate more realistic colors.
- Extensive experiments demonstrate that TEGAN can achieve superior results compared to the state-of-the-art approaches on public underwater image datasets such as EUVP [24], RUIE [25], and UIEB [26]. In addition, outstanding results reveal that it can significantly facilitate the performance of other downstream visual tasks.

Related Work

For UIE tasks, the existing methods are systematically divided into the following three types. The enhancement methods directly enhance the visual effects. The recovery methods based on the physical model consider the degradation process of underwater images. The deep-learning methods are data-driven.

Enhancement Methods

Enhancement methods directly adjust the pixel values of a given underwater image to achieve contrast enhancement and color correction without considering the degradation process. The enhancement methods reassign the pixel values of a given image without considering the image degradation process for contrast enhancement and color correction. In recent years, fusion-based methods of enhancement have shown promising results. EUF [4] is based on the principle of fusion and does not need professional hardware or learning about underwater scene structure and conditions. Only input and weight metrics are obtained from degraded images. CBFU [5] is proposed to build on the coordination of color compensation and white balance versions of the raw degraded image. It promotes the conversion of edge and color contrast to the enhanced image, and multiscale fusion strategies are employed. ICM [6] is proposed based on shift stretching. The stretches on color contrast, saturation, and intensity are deployed to improve the image quality. Among other enhancement methods, the gamma correction (GC) [7] approach corrects the images with too much gray and too little gray to enhance the contrast. A model that utilizes the features of light scattering is proposed in [8]. First, the RGB channel average ratio is used to categorize the color projection into five different groups. To restore the color projection of underwater images, a multiscene color recovery method is developed using the optical attenuation characteristics to determine the color loss rate of RGB channels in various scenarios. These enhancement methods directly apply image processing by subjectively adjusting the pixel values to eliminate noise, improve edge blur, enhance the features of the target object, and weaken the effect of irrelevant environmental factors on the target. However, since the underwater optical imaging model is not considered, some additional noise is introduced, which can cause severe oversaturation in different image regions.

Recovery Methods

The recovery methods take into account the underwater image degradation process, the imaging principle of the image, and the building of a physical model. DCP [10] is a solution to the image haze reduction issue. Many researchers have created DCP-based underwater recovery methods after observing the resemblance between hazy photos and blurred underwater images in descattering. Because of the particular features of the underwater environment, UDCP [11] is designed to implement DCP within the green and blue channels. IBLA [12] is an underwater scene depth estimation approach according to image blurring and light absorption. It gives more precise background light and depth estimates that may be utilized in the image formation model (IFM).

First, the method selects the background light by the blurred region and obtains a depth map. Then, the transmission map based on the background light is used to recover the scene radiation. The coefficients of ULAP [14] are trained using supervised linear regression based on learning. To recover the real scene radiation, the approach first performs depth estimation. It next estimates the background light and transmission map of RGB relying on the depth map. A new color compensation method is proposed in [15]. The underwater image region with the most severe color distortion is compensated by combining the polarized image with the intensity image. It can improve the exposure of the low-luminance area of the image. The dark channel prior approach is then used to deblur and improve the image. The recovery method recovers degraded images by a priori knowledge, but when the a priori knowledge is inaccurate, it often leads to serious estimation errors. In this area, the absence of reliable priori information about underwater images is now a major obstacle.

Data-Driven Deep Learning Methods

Data-driven deep learning-based approaches are now mainly classified as CNN-based, GAN-based, and Transformer-based. WaterNet [26] takes three images processed by white balance, histogram equalization, and GC as input images and uses the gated fusion network for learning the corresponding confidence map to determine the most important features of the residual inputs in the final result. WaterNet can vastly improve image contrast and correct color cast to some extent. However, for images with a great color cast, the color cast needs to be corrected, and overenhancement is another issue. UWCNN [27] is proposed based on the underwater scene a priori, which may be utilized to create training data. UWCNN reconstructs a clean underwater submerged image without the need for underwater estimate of the imaging model's parameters. MLFcGAN [31] extracts multiscale features and then uses global features to improve local features at each scale. MLFcGAN performs better in off-color correction, but it is challenging to handle hazy images or even introduce false enhancement effects. FUnIEGAN [24] is designed based on a conditional generative adversarial network, formulating an objective function with content-aware loss that assesses perceptual image quality using information about global content, color, local texture, and style. The model uses only a simple 4-layer UNet network to achieve a real-time effect. Since many paired underwater image datasets are generated using CycleGAN [18], CycleGAN can also be used in underwater image enhancement tasks. Uformer [23] is proposed for image restoration with self-attention local enhancement based on a nonoverlapping window (LeWin) Transformer block and a multiscale recovery modulator, which is used to adjust the features in the Uformer decoder layers. Uformer is mainly applied in image enhancement tasks, and its application in underwater image tasks can improve the cast effect and the contrast to some extent. However, it cannot improve the underwater image haze effect. STSC [28] develops an efficient and compact enhancement network in collaboration with a high-level semantic-aware pretrained model, aiming to exploit the hierarchical feature representation as an auxiliary for low-level underwater image enhancement. SCNet [29] focuses on spatial and channel dimensions, with the key idea of learning water type desensitized features. The purpose of this method is to improve the image quality and deal with the degradation diversity of water. TACL [30] achieves both visual-friendly and task-oriented enhancement. The sharpness of the image may be noticeably enhanced, but it is prone to residual water color, and some areas of the image are too bright.

Since the underwater environment is complicated, many methods cannot fully learn the distribution of the target image, so there exists a large deviation between the enhanced image and the target image. Moreover, there are still large differences between synthesized images and real underwater images. The distribution learned on synthesized images by the data-driven deep learning method is difficult to apply to real underwater images, and the processed images still have some defects, such as color cast, missing detail, and overenhancement. How to better solve the aforementioned issues is the focus of this paper.

Proposed Method

Underwater image enhancement learns a mapping from underwater images degraded for various reasons to target clear images. Due to GAN's outstanding performance in the field of image generation, it has drawn increasing attention. As a framework for this paper, we adopt the conditional generative adversarial network (cGAN) [32] and design a proper generator (G) for learning the mapping mentioned above. Recently, Transformer has been increasingly used for visual domain tasks since it can extract long-range dependencies well. This new technology is also incorporated into the construction of TEGAN in this paper.

Here, we introduce a new architecture that contains a well-designed novel generator and a two-branch discriminator. Then, by referring to the LeWin block in Uformer [23] and RPE [33], we propose a new window-based dual local enhancement (DleWin) block that is more suitable for the UIE task. Finally, the WGAN-GP loss, Uloss, and L1 loss are adopted to guide the network training.

Network Architecture

Figure 2 depicts the TEGAN architecture in detail. Elaborately constructed Inception, Bottleneck, and Fusion unit are introduced to the original Encoder-Decoder generator, which is designed as a UNet-like network. The effectiveness of each component will be demonstrated in the "Experiments and Analysis" section. The discriminator includes two branches, namely, a feature branch and a color branch.

Generator

Inspired by Uformer, we propose a Transformer embedded generator framework in our underwater image enhancement tasks, but we do not embed Transformer block for each scale as Uformer does. Specifically, a partial fusion scheme is designed to effectively combine Transformer and convolutional neural network. We believe that compared with Transformer, convolutional neural network performs better in multiscale feature extraction, so we use convolutional neural network in the Encoder and Decoder units for multiscale feature extraction and reconstruction, which can effectively reduce the edge blurring and retain more details. Transformer block, due to its expertise in extracting raw and global information from images, is used in Inception and Bottleneck units. The advantage of incorporating global information into each scale is demonstrated in MLFcGAN [31]. We adopt this operation for reference. The global information fully extracted by the Transformer block is integrated into each feature scale, which is particularly effective in solving the problem of color cast in underwater image degradation.

As we can see in Fig. 2a, one DleWin block is embedded in Inception unit to extract the long-range dependencies of the features directly from the original image. Then, it can be used for subsequent feature extraction. We also explore the effect of the number of DleWin blocks in the Inception unit on the model performance, as shown in Fig. 3a. It can be concluded that when there is more than one DleWin block, the time consumption is dramatically increased, and the performance is degraded.

The Encoder unit consists of five encoding layers. Details are shown in Fig. 4 Encoder. It performs multiscale feature extraction on the features preliminarily obtained through the Inception unit and finally inputs the shape of $512 \times 8 \times 8$ feature maps to the Bottleneck unit. In addition, the extracted features of each layer are transferred to the corresponding layer of the Decoder unit through skip connections, as shown in Fig. 2a. Encoder1 contains a convolutional layer, while encoder2-encoder5 contain a Conv + BatchNorm + ReLU (CBR) module. The parameters of all convolutional layers are size = 4×4, stride = 2, and padding = 1, which plays the role of downsampling while extracting features.

The Bottleneck unit embeds two DleWin blocks. When the features extracted by the Inception are downsampled by Encoder to a size of 8×8 (the same size as the window of the DleWin block), the Transformer block can extract global information, such as the overall lighting and image layout. Since Transfomer's self-attention mechanism is good at extracting long-range information, the DleWin block can be used in this unit to achieve a significant performance

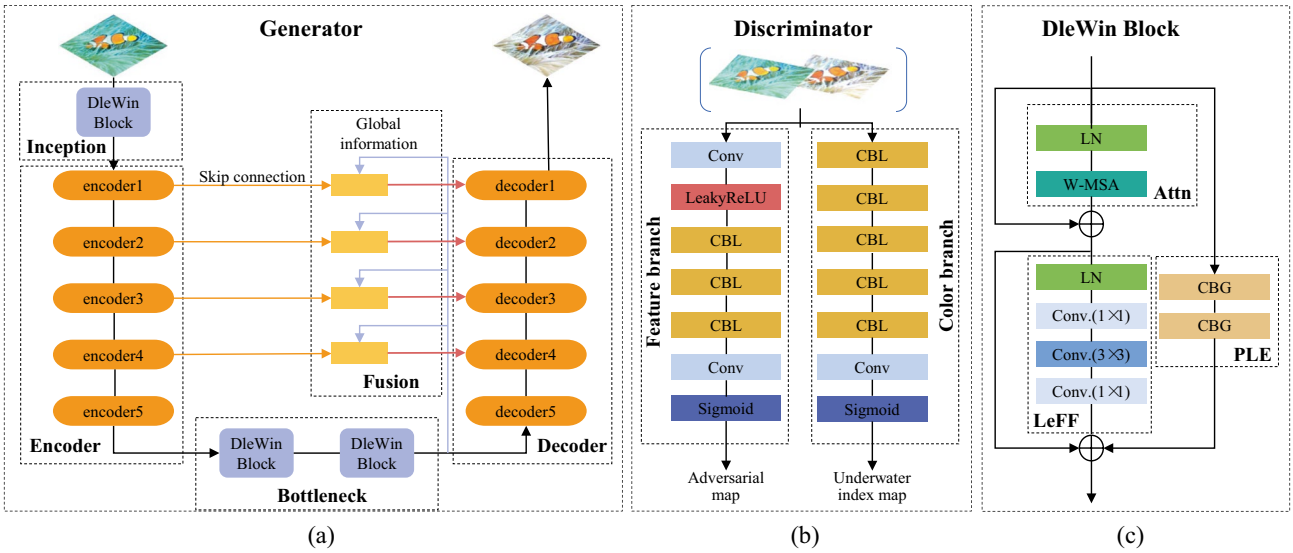


Fig. 2 The architecture of the TEGAN proposed in this paper. From left to right is the framework of the generator (**a**), discriminator (**b**), and DleWin blocks (**c**). The generator is composed of Inception, Encoder, Bottleneck, Decoder, and Fusion units. The discriminator is composed of a feature branch and a color branch. The DleWin block consists of Attn, LeFF, and PLE modules

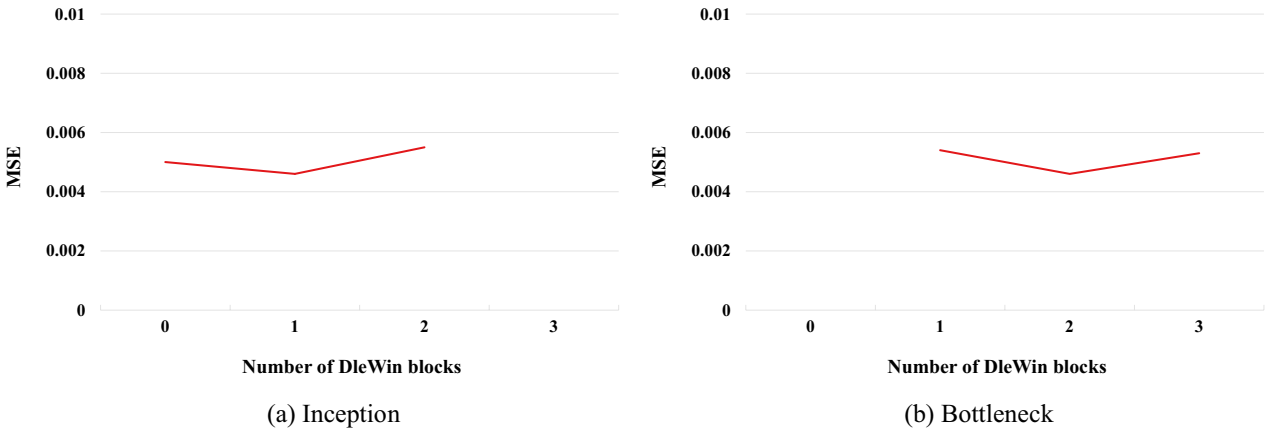


Fig. 3 Effect of the number of DleWin blocks on the overall performance in terms of the image quality evaluation metric MSE. **a** Effect of the number of DleWin blocks in the Inception unit and **b** effect of the number of DleWin blocks in the Bottleneck unit

improvement. We also investigate how the number of DleWin blocks in this unit affects the effectiveness of the model. The number of DleWin blocks is set to 2 for the following reasons. As we can see in Fig. 3b, in the Bottleneck, the optimal performance is achieved when the number of DleWin blocks is 2. As this number increases, network performance deteriorates. An excessive number of DleWin blocks will make the model extract too much global information, resulting in overfitting of training, which will adversely affect the generalization ability and performance of the model. Meanwhile, the time consumption will be dramatically increased due to the deepening of the network. Moreover, the global information extracted by the previous DleWin block will become blurred after the subsequent DleWin processing, thus weakening the positive role of global information in image enhancement. On the other hand, if the number of DleWin blocks is too small (less than two), the extracted global information is not sufficient, and the utilization efficiency is low. In this case, the corresponding network performance is also poor.

The Decoder unit has five decoding layers, as shown in Fig. 4. Decoder receives the global information extracted from the Bottleneck and outputs the enhanced image with a shape of $3\times256\times256$ after five decoding layers. Similar to the Encoder, decoder1 contains a transposed convolution and a tanh activation function. Encoder2-decoder5 each contain a CBR module. All transposed convolution parameters are size $=4\times4$, stride $=2$, and padding $=1$, which plays the role of upsampling while reconstructing features.

The Fusion unit integrates the global information extracted by the Bottleneck unit, such as the overall lighting and layout, into each scale. As shown in Fig. 5, the global information will first go through a F_adjust operation, which is a convolution with size $=1\times1$ and stride $=1$. Through the F_adjust operation, the channels of the global information can be adjusted to correspond to Decoder. Then, the global

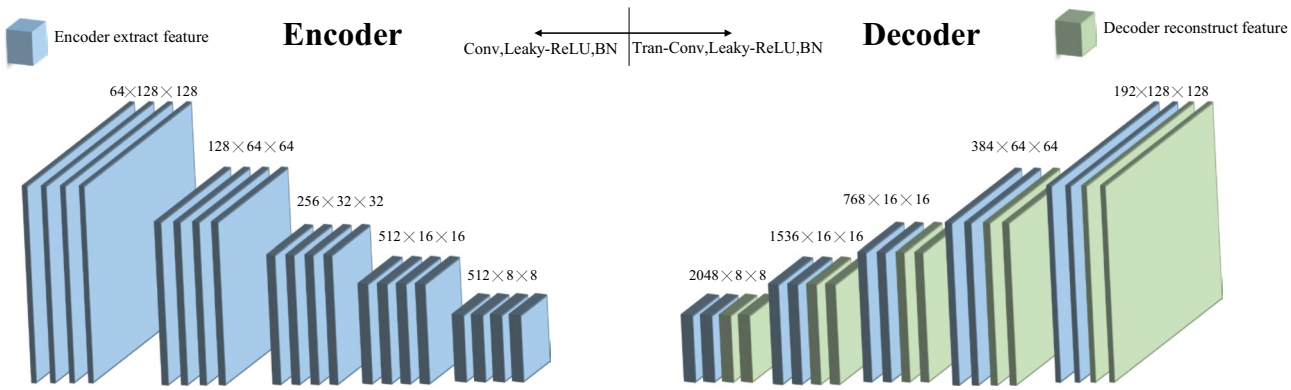


Fig. 4 Network structure of the Encoder and Decoder units in the generator. The blue part in Encoder is the extracted feature map. The corresponding blue part in Decoder represents the feature map from Encoder by skip connection, while the green part represents the feature map reconstructed by Decoder. The numbers on each layer annotate the shape of the features

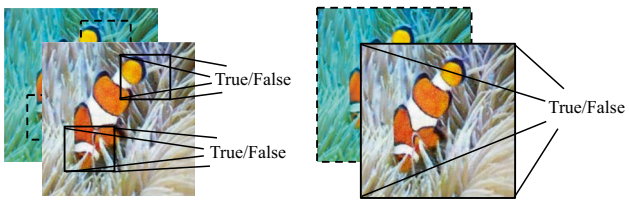
Fig. 5 Schematic diagram of Fusion unit

information will be copied and reshaped by the F_copy and F_reshape operations to finally achieve the effect that the shape of the output fusion information is the same as the feature map of the corresponding layer of Decoder. The fusion of global information to each scale helps to provide more realistic colors and finer details.

Discriminator

We use a two-branch discriminator containing a feature branch and a color branch in Fig. 2b, where the feature branch is used to preserve image features and enhance contrast, while the color branch performs color correction to generate more realistic colors. They both adopt PatchGAN [34], as shown in Fig. 6a. The discriminator of the original GAN evaluates only one value (true or false) for the whole image generated by the generator, as we can see in Fig. 6b. This operation evaluates the overall image quality. However, it lacks image localization evaluation, causing the local details of the image generated by the generator to be blurred. PatchGAN adopts the form of full convolution, and the discriminator evaluates the generated image as a matrix of $N \times N$. Each element in the matrix corresponds to the discriminator's evaluation of a small patch region. The average value of the matrix forms the final evaluation of the discriminator for the whole image. PatchGAN focuses on local information, which can make the generated image have more details and reduce local blur. Moreover, compared to the full-image discriminator, PatchGAN has fewer convolutional layers. In this paper, for a 256×256 generated image, the discriminator forms a 30×30 evaluation matrix, and the perceptual field (patch size) of each evaluation value in the matrix is 70×70.

In detail, the feature branch preserves the image content by one convolution layer. Then, it stacks three layers of Conv + BatchNorm + Leaky-ReLU (CBL) modules and one layer of convolution to identify the authenticity of the image. Finally, it generates an adversarial map for evaluation



(a) PatchGAN (b) Whole image discriminator

Fig. 6 Schematic diagram for different types of discriminators

and facilitates the generator to generate a realistic image. The color branch directly stacks five layers of CBL modules and one convolution layer to discriminate whether the image belongs to the underwater scene. It generates an underwater index map for evaluating the strength of underwater attributes and facilitates the generator to generate colors consistent with the in-air image. The original image and enhanced image or the original image and real image by concatenate operation fed into the discriminator to finally obtain an adversarial map and an underwater index map.

DleWin Block

In contrast to convolutional neural network, Transformer can compute the correlation between each pixel of an image directly without passing through hidden layers. CNN models the relationship between neighborhood pixels, while Transformer pays more attention to the relationship between all pixels. Therefore, we can design strategies to make the two complement each other well.

For underwater image enhancement using Transformer, there are two problems to solve. First, the standard Transformer [17] computes global self-attention among all tokens, which results in a secondary computational cost for tokens and an enormous computational consumption for images. Second, local information is particularly important for vision tasks, especially underwater image enhancement tasks. However, Transform is not good at extracting local information.

To address the first issue, we propose a window-based DleWin block in which CNNs are introduced for local enhancement using both serial and parallel approaches. It implements an efficient mechanism for calculating self-attention in terms of windows. The DleWin block includes three modules, namely, a self-attention module (Attn) for capturing features, a serial local enhancement feedforward network (LeFF), and a parallel local enhancement module (PLE). In Fig. 2c, the input feature maps are subjected to Attn for feature extraction, and then, LeFF performs local enhancement on the features. Meanwhile, PLE performs local enhancement on the features before passing through Attn. A skip connection is added between Attn and LeFF to avoid degradation of the input features.

As we can see in Fig. 2c, Attn contains a layer normalization layer (LN) and a window-based multihead self-attention (W-MSA). LeFF contains a LN and three convolutional layers, where the input tokens are first transformed into tokens by a linear projection as conv.(1×1). The tokens are reshaped into a 2D feature map, which is transformed by a convolutional layer of size $= 3 \times 3$ and then stretched into new tokens. Finally, it

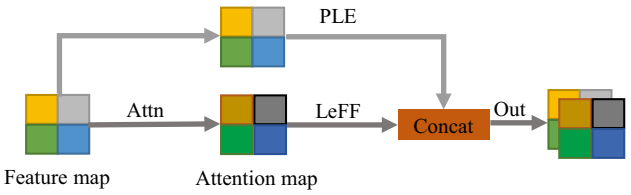


Fig. 7 Schematic diagram for different enhancement modes between PLE and LeFF

is transformed into the same dimension as the input features through linear projection. PLE contains two Conv + BatchNorm + GELU (CBG) modules. Unlike LeFF, the input of PLE is the features not extracted by Attn. LeFF is the long-range dependency captured by Attn for local enhancement, while PLE is a direct local enhancement of the input features. Therefore, they have different local enhancement effects. LeFF is used to compensate for the shortcomings of the Transformer in extracting local features, while PLE is used to further enhance the whole block for local feature extraction. A combination of the two can meet the need for underwater image enhancement tasks for extracting local features and further alleviate the adverse effect of the long-distance dependencies captured by Attn. The differences between PLE and LeFF are shown in Fig. 7. The effectiveness of the embedded DleWin block with PLE and LeFF will be demonstrated in the "Experiments and Analysis" section.

In addition, instead of implementing a global self-attention mechanism, we deploy W-MSA with window-based multi-head self-attention. The input feature matrix $X \in R^{C\times H\times W}$ is partitioned into N feature windows of $M\times M$, where C, H, and W are the number of channels, width, and height, respectively. Then, the transposed and stretched features $X^i \in R^{M^2\times C}$ of each window are obtained. In short, W-MSA encodes every pixel within the window as a token. It performs self-attention within nonoverlapping local windows, which significantly reduces the computational cost. The motivation for using the multi-head self-attention mechanism is that dividing the model into multiple heads and forming multiple subspaces by channels allows the model to focus on different aspects of information. Finally, we combine the information from all aspects. Suppose there are k heads; then, each head has dimension $d_k = C/k$, and the kth head processes a feature map $\hat{X}_k \in R^{M^2*d_k}$. The self-attention of the kth head can be calculated as follows:

$$X = \left\{X^1, X^2, \ldots X^N\right\}, N = HW/M^2 \tag{1}$$

$$Y_k^i = \text{Attention}\left(X^i W_k^Q, X^i W_k^K, X^i W_k^V\right), i = 1, \ldots, N \tag{2}$$

$$\hat{X}_k = \left\{Y_k^1, Y_k^2, \ldots Y_k^M\right\} \tag{3}$$

where $W_k^Q, W_k^K, W_k^V \in R^{C\times d_k}$ represent the projection matrices of query (Q), key (K), and value (V) of the kth head, respectively. The outputs of all heads are then concatenated and linearly mapped to obtain the final results. W-MSA also applies relative position encoding. The attention can be expressed as

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{4}$$

B is the relative position bias, with the value derived from the learnable parameter $\hat{B} \in R^{(2M-1)\times(2M-1)}$. Compared to global attention, W-MSA can decrease the time complexity of the input feature map $X \in R^{C\times H\times W}$ from $O(H^2W^2C)$ to $O(M^2HWC)$.

The so-called self-attention mechanism is depicted in Fig. 8. For the input feature matrix, the query (Q), key (K), and value (V) are generated by the learnable parameter matrices W_q,W_k,W_v, respectively. Then, Q and K are multiplied together with relative position encoding (B) and undergo zero-mean normalization (Z-norm) to obtain the attention matrix Attn. Finally, Attn is activated by Softmax and multiplied by V for output.

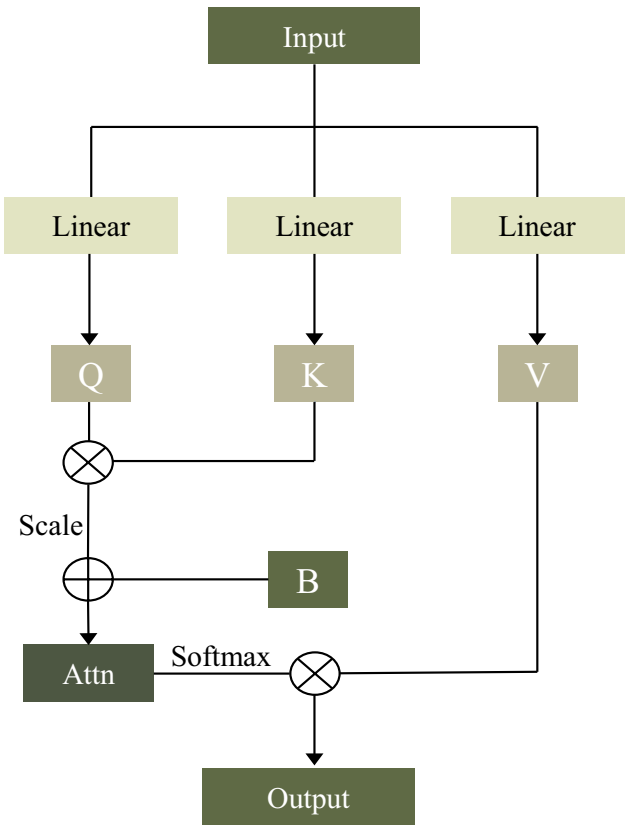


Fig. 8 Illustration of W-MSA's self-attention mechanism

Objective Function

The standard GAN suffers from mode collapse and vanishing gradient. Mode collapse does not occur in underwater image enhancement. To solve the vanishing gradient problem, Martin Arjovsky proposed the Wasserstein GAN (WGAN) [35]. WGAN is needed to compute the Wasserstein distance, requiring that the discriminator satisfies the Lipschitz restriction. First, weights are clipped to a fixed interval $[-c, c]$, but this simple and brutal operation cannot yield better results. Therefore, the WGAN with gradient penalty (WGAN-GP) [21] is introduced, and the equation is transformed as

$$L_{\mathrm{WGAN-GP_D}} = E_{x,y}\left[D_W(x,y)\right] - E_x\left[D_W(x,G(x))\right] + \lambda E_{\hat{x}}[(\|\nabla_{\hat{x}} D_W(\hat{x})\|_2 - 1)^2] \tag{5}$$

$$L_{\mathrm{WGAN-GP_G}} = E_x\left[D_W(x,G(x))\right] \tag{6}$$

where x and y are the degraded image and ground truth (in the air, clear, color balanced target image), respectively. $\hat{x}$ is a linear sample between $G(x)$ and y. In this paper, we use WGAN-GP for the adversarial branch.

To guide the training of the color branch, we introduce Uloss from GAN-RA [22].

As shown in Fig. 9, in the Lab color space, we take the in-air image as the target. The distance between the image and the in-air image is evaluated using the underwater index (U). Its formula is as follows:

$$U = \frac{\sqrt{d_0}}{10 a_l d_a d_b} \tag{7}$$

where d_0 is the distance from the image mean to the center of a and b color channels. a_l denotes the mean of the L channel, while $d_a d_b$ denotes the area of the image pixel value distribution. The higher the value of d_0 is, the more severe the color distortion.

The underwater index loss is designed using the L2 loss:

$$L_{U_D} = E_{x,y}[(D_U(x,y) - U(x,y))^2] + E_x[(D_U(x,G(x)) - U(x,G(x)))^2] \tag{8}$$

$$L_{U_G} = E_x\left[\left(D_U(x,G(x))\right)^2\right] \tag{9}$$

X is the original image, y is the ground truth, and $U(\cdot)$ is the underwater index for computing an image. In the initial stage of training, the color branch is not sufficiently trained to thoroughly distinguish the difference between underwater and in-air images. Thus, it is not sufficient to guide the generator to learn the distribution of in-air images. Therefore, we adopt a two-phase training strategy, i.e., the generator does not add the underwater index loss at the beginning of training but adds it after the color branch is sufficiently trained.

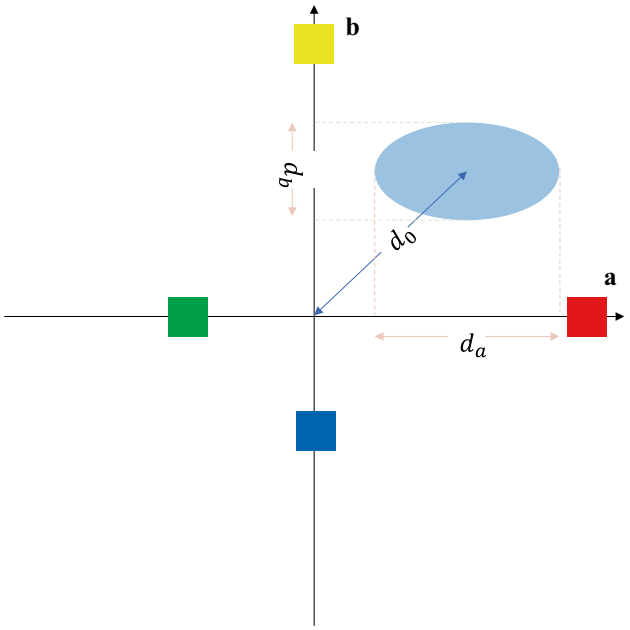


Fig. 9 Underwater index illustration. The elliptical region represents the a-b distribution of the image in the Lab color space. The smaller d_0, larger d_a and d_b indicate that the image is closer to the in-air image [22]

Existing methods show that adding the L1(L2) loss to the objective function allows the generator to learn the distance from the original image to the ground truth directly [36]. It can focus on the low-frequency information of the image, thus reducing blur. Compared to L2 loss, L1 loss is deployed in this paper for blur reduction due to its greater robustness. Its formula is as follows:

$$L_{l1_G} = E_{x,y}\left[\|y - G(x)\|_1\right] \tag{10}$$

The global objective function can be expressed as follows:

$$L_D = L_{\mathrm{WGAN-GP_D}} + L_{U_D} \tag{11}$$

$$L_G = \lambda_W L_{\mathrm{WGAN-GP_G}} + \lambda_U L_{U_G} + \lambda_{l1} L_{l1_G} \tag{12}$$

where λ_W, λ_U, and λ_{l1} are the weight factors. The optimal model is $D^* = \arg_D \min L_D$, $G^* = \arg_G \min L_G$.

Experiments and Analysis

Datasets

Data acquisition for UIE tasks is extremely difficult, especially for paired images with ground truth. We deploy a paired underwater image dataset from UGAN [37] as the

Table 1 Description of the test datasets

Paired test images	Total	Val	Underwater-dark	Underwater-imagenet	Underwater-scenes
	1028	128	300	300	300
Unpaired test images	**Total**	**EUVP**	**RUIE**	**UIEB-Raw**	**UIEB-Challenge**
	7753	2574	4229	890	60

training set and validation set. Simultaneously, several publicly available paired and unpaired underwater image datasets are taken as the test set. All image sizes are adjusted to 256 × 256 using bicubic linear interpolation.

Training Set

To better learn the features from ground truth and preserve the content of the images, our model is trained on paired datasets (including ground truth and degraded underwater images). A total of 6000 pairs of images are randomly selected from 6128 pairs generated by CycleGAN [18] in the literature [37] as our training set. The remaining images serve as the verification set.

Test Set

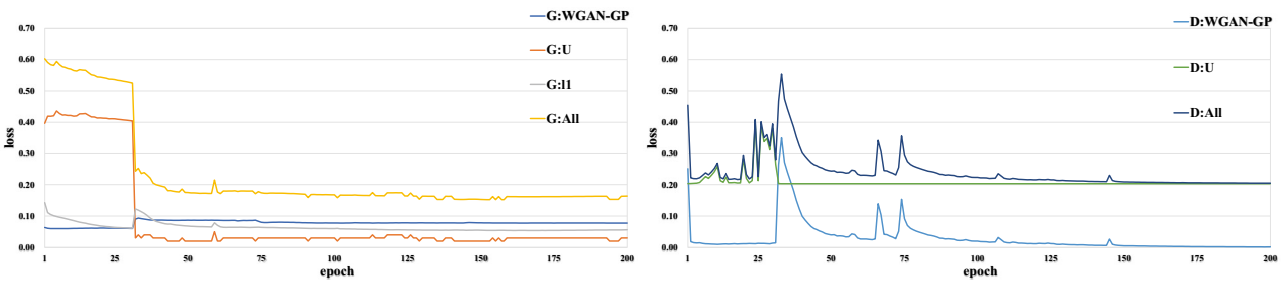
We cite EUVP [24], RUIE [25], and UIEB [26] as our test datasets.

The EUVP (Enhancing Underwater Visual Perception) dataset contains separate sets of paired and unpaired images with varying degrees of perceptual quality. It mainly contains three subsets: paired, unpaired, and test samples. The paired subset contains dark images, images collected from ImageNet, and bluish and greenish images from real underwater scenes.

The RUIE dataset is a real underwater image dataset without ground truth. It includes three subsets: UCCS, UIQS, and UHTS. Among them, the UCCS subset contains blue, green, and blue-green subsets, corresponding to the common color cast problem in underwater image degradation.

The UIEB dataset achieves the goal of underwater image data collection [26], i.e., diversity of underwater scenes, different characteristics of quality degradation, and a wide range of image contents. It consists of raw and challenge subsets. The raw subset contains 890 underwater images and corresponding reference images. The reference images are the subjective optimal enhancement results selected by using various underwater image enhancement methods. The challenge subset contains 60 underwater images, which have a high degree of degradation and have not achieved satisfactory results by many previous enhancement methods. It should be noted that although UIEB provides reference images, it is only the images generated by other enhancement methods and cannot be considered ground truth.

We construct four separate groups of paired tests containing ground truth and four groups of unpaired tests without ground truth, as shown in Table 1. The paired image test set is divided into four groups, i.e., the Validation set (Val), Underwater-dark, Underwater-imagenet, and Underwater-scenes subsets from EUVP. The total number of paired test images is 1028. For the unpaired image test set, we also divide it into four groups, i.e., all 2574 unpaired images in EUVP, all 4229 real-world underwater images in the three subsets of RUIE, all 890 images in the raw subset of UIEB, and all 60 images in the challenge subset of UIEB. From Table 1, we can see that the unpaired real-world underwater test images are more sufficient, which can reflect the generalization ability of our method.



(a) Generator loss (b) Discriminator loss

Fig. 10 Illustration of the training losses. **a** G: WGAN-GP, G: U, G: l1, and G: T are labeled as the feature branch loss $L_{\text{WGAN-GP}_G}$, color branch loss L_{U_G}, L1 loss L_{l1_G}, and global loss L_G of the generator, respectively. **b** D: WGAN-GP, D: U, and D: T are labeled as the feature branch loss $L_{\text{WGAN-GP}_D}$, the color branch loss L_{U_D}, and the global loss L_D. When L_{U_G} starts to take effect, it can be seen that each loss has a significant change

Training Details

We set a batch size of 32, $\lambda_W = 0.1$, $\lambda_U = 5$, $\lambda_{l1} = 10$ and use Adam optimizer with a learning rate = 0.0002, $\beta 1 = 0.5$, $\beta 2 = 0.999$. The training set images are all first adjusted to 286×286 using bicubic linear interpolation and then randomly cropped to 256×256 to achieve data enhancement. We use PyTorch as a deep learning framework to train 200 epochs on an Inter(R) Xeon Silver 4214R, 4 GB RAM, and GeForce RTX 3090 GPU platform.

Loss curves are shown in Fig. 10. The feature branches $L_{\mathrm{WGAN-GP_G}}$ are gradually in a dynamic equilibrium at the beginning. As mentioned in the "DleWin Block" section, here, we adopted a two-phase training strategy. The generator does not add the underwater index loss at the beginning of training but adds it until the color branch is well trained. In detail, after the 30th epoch, the color branch starts to work due to the addition of L_{U_G} to the generator training, and each loss begins to change dramatically. Then, L_{U_G} and L_{U_D} tend to be in dynamic equilibrium. Moreover, $L_{l1}(G)$ steadily decreases except for the rapid increase when the color branch starts to take effect.

Models for Comparison

Traditional and deep learning-based (data-driven) methods are conducted for comparison to demonstrate the superiority of TEGAN, as shown in Table 2.

The enhancement methods include EUF, CBFU, ICM, and GC (where EUF and CBFU are based on the fusion method). Recovery methods include MIP, DCP, UDCP, and ULAP. Learning-based methods are CycleGAN, FUnIEGAN, MLFcGAN, UWCNN, WaterNet, Uformer-B (the best performance parameter setting in Uformer), STSC, SCNet, and TACL. To compare the performance in an objective way, all learning-based methods except TACL are trained on the same training set. The network parameters for comparison are the recommended settings captured from the original paper to obtain the best enhancement results. It is worth mentioning that the source training code of TACL is not publicly available. Here, we use the trained network parameters provided by the author for comparison.

Results and Analysis

Paired Test Images

The benefits of our method are illustrated by the visual comparisons in Fig. 11. Compared with other methods, the images enhanced by our method are color balanced with higher contrast and better visual effects. Using GT (ground truth) as a reference, it can be seen that some methods have limited quality improvement, while others have obvious quality improvement but still cause overenhancement or wrong color correction. Most traditional methods have difficulty improving the color deviation, which is far from GT.

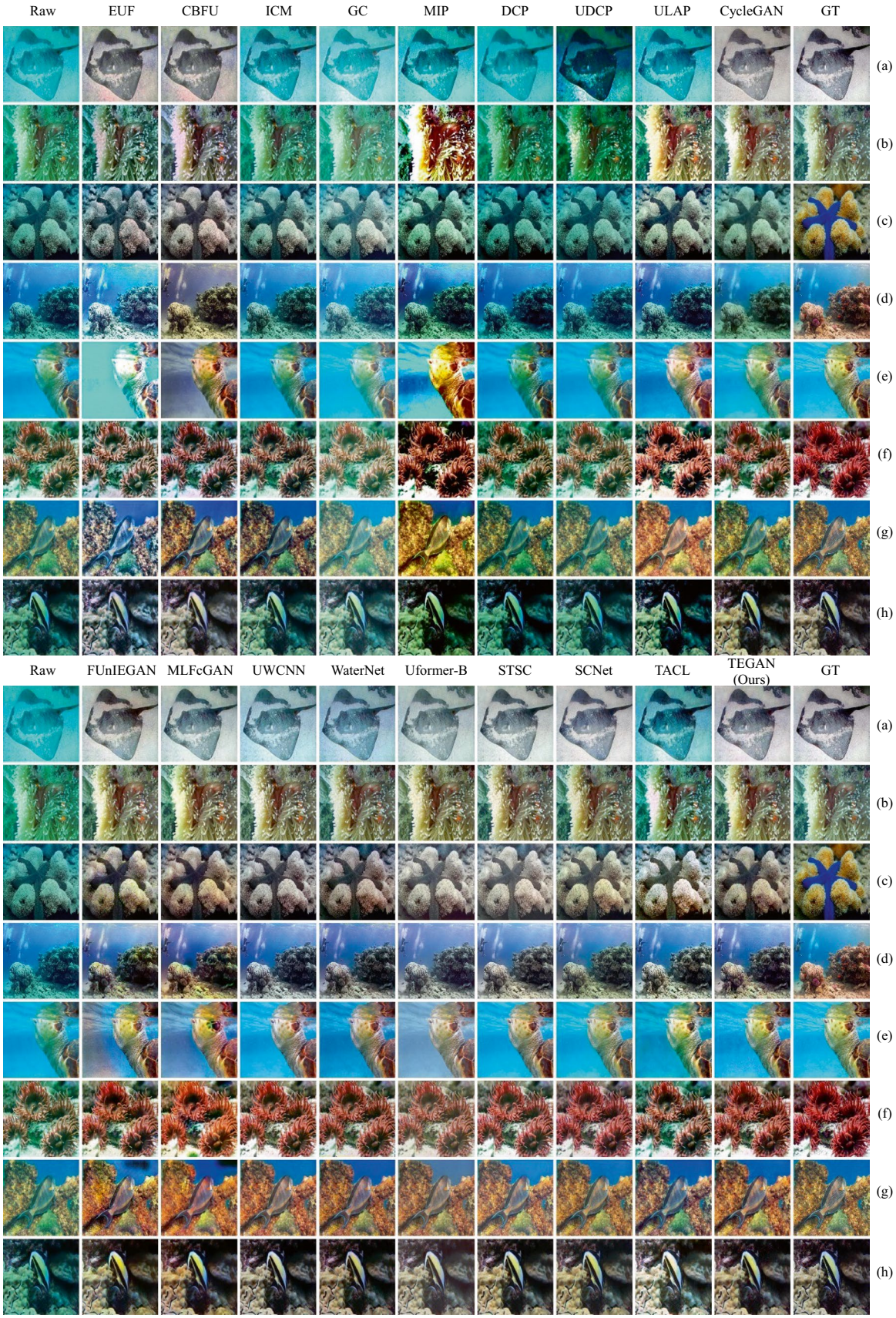
For the enhanced methods, although ICM and GC can reduce blur, in Fig. 11a, c, h, ICM does not reduce the image's color divergence, while GC reduces the image's brightness. The fusion-based methods are more effective in brightness enhancement, but there are still obvious problems, such as overenhancement compared to GT. As we can see, EUF introduces a large amount of red-blue noise in Fig. 11a and is severely exposed in Fig. 11d, e. In Fig. 11b, d, CBFU has serious color distortion compared to GT.

For the recovery methods, the brightness of MIP is improved to some extent, but the yellow compensation is excessive, resulting in the yellow color of the restored image, as shown in Fig. 11b, e, g. DCP can augment the image's contrast, but it cannot solve color divergence, as in Fig. 11a, d, f, g, h. UDCP increases the haze effect while reducing the color cast, in Fig. 11b, d, e, g. Compared with GT, the recovered image still exists residual water color. The quality of ULAP is improved in some images, but still, some images, such as Fig. 11a, d, h, are not good at removing the effects of water bodies and adjusting the color cast.

For the deep-learning based method, improving the color deviation is generally better than the traditional method, but it may not have good results in other aspects of image quality improvement. CycleGAN and Uformer-B retain details well,

Table 2 Models for comparison

Traditional methods									
Model	EUF [4]	CBFU [5]	ICM [6]	GC [7]	MIP [9]		DCP [10]	UDCP [11]	ULAP [14]
Year	2012	2018	2007	2015	2010		2010	2013	2018
Deep learning-based methods									
Model	CycleGAN [18]	FUnIEGAN [24]	MLFcGAN [31]	UWCNN [27]	WaterNet [26]	Uformer-B [23]	STSC [28]	SCNet [29]	TACL [30]
Year	2017	2020	2020	2020	2020	2021	2022	2022	2022


Raw
EUF
CBFU
ICM
GC
MIP
DCP
UDCP
ULAP
CycleGAN
GT
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Raw
FUnIEGAN
MLFcGAN
UWCNN
WaterNet
Uformer-B
STSC
SCNet
TACL
TEGAN (Ours)
GT
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

◂**Fig. 11** Visual comparison of various methods in terms of color, sharpness, and contrast on paired test image sets. Each row is the processing result of the corresponding method. **a**, **b** Selected from Val. **c**, **d** From Underwater-dark. **e**, **f** From Underwater-imagenet. **g**, **h** From Underwater-scenes. Raw denotes the original raw image, and GT denotes the corresponding ground truth

but there is an obvious haze effect, as shown in Fig. 11g. Uformer-B also has a haze effect on Fig. 11f. The effect of noise removal of the water body in CycleGAN is not satisfactory, as shown in Fig. 11c, f. The color saturation of FUNIEGAN enhanced image is improved, but there is still color distortion, such as the background color distortion in Fig. 11e, g. The performance of MLFcGAN is closer to GT but still has a slight color cast. As in Fig. 11a, there is a bluish effect, and the color of the flower is somewhat orange in Fig. 11f. UWCNN and WaterNet contribute to removing blurring, but for images with serious color cast, color deviation still exists. Moreover, STSC and SCNet can largely eliminate the influence of color cast, but STSC still has color deviation, in Fig. 11a, and whitening effects, in Fig. 11c, d. SCNet makes the objects in Fig. 11c, d green, and it cannot remove the haze completely, as shown in Fig. 11g. TACL can improve clarity and brightness to a certain extent, but its performance in color correction is not good, as shown in Fig. 11a. In addition, Fig. 11b, f, g still have serious residual water color, and they are much different from GT. In contrast, TEGAN exhibits competitive performance. As shown in Fig. 11a, e, h, our results are almost the same as GT, while Fig. 11f, c, d are closer to GT. It is worth mentioning that compared with GT, the result in Fig. 11b improves the color cast, while Fig. 11g reduces the haze effect.

Consistent quantitative conclusions can be inferred from Table 3 for the paired image test sets. Since GT images exist, we select the full-reference evaluation metrics MSE, PSNR, and SSIM [38]. Among them, MSE is the expected value of the square of the gap between the enhanced image and GT. PSNR measures the difference between the enhanced image and GT pixels. SSIM is an image quality evaluation criterion that conforms to human intuition. It indicates how close the enhanced image is to GT in structure and texture properties.

Table 3 displays the full-reference evaluation metrics. We can see that traditional methods perform unsatisfactorily in general, and none of them have entered the top two. Some methods perform well on specific test sets, e.g., MLFcGAN achieves the best PSNR and SSIM and the second-best MSE on the Val test set. Notably, UWCNN achieves the best MSE, PSNR, and SSIM on the Underwater-scenes test set. SCNet also seems to perform well in Underwater-dark and has average results. Our TEGAN achieves outstanding results on all data sets. The good generalization capability and competitive performance of TEGAN benefit from the strong learning ability of the DleWin block.

Unpaired Test Images

Experimental results for the unpaired images are depicted in Fig. 12, where every two images are selected from one test set. For the original images, two images from the EUVP are blue-green casted. For the two images from RUIE, Fig. 12c is heavily greenish, and Fig. 12d has greenish haze and blurred details. The two images from UIEB-Raw, exhibited in Fig. 12e, f, show a slight blue cast. In the UIEB-Challenge dataset, Fig. 12g shows a slight haze effect, and Fig. 12h shows a haze and green cast effect.

The traditional methods fail to solve the color cast problem better, except for EUF and CBFU. However, both EUF and CBFU have color distortion caused by color oversaturation and excessive color compensation, as shown in Fig. 12c–e. In Fig. 12a, b, d, f, the other two enhancement methods, GC and ICM, contribute to blur reduction, but the color distortion and haze effect are not significantly improved.

Among the recovery methods, MIP can improve the color saturation to some extent, but it will cause overenhancement and under-enhancement for slightly degraded and severely degraded images, respectively, such as the overenhancement effect in Fig. 12a and almost no improvement in Fig. 12d. DCP can correct color distortion, but it will cause lower image brightness, as shown in Fig. 12f, and it will not improve the color deviation of the seriously degraded image, as shown in Fig. 12d. UDCP can improve the brightness, but it will cause a serious haze effect. ULAP improves the cast color correction, but the blue color is overcompensated, and the recovered image is bluish, in Fig. 12g, h.

For the deep-learning based method, CycleGAN enhanced image still has color deviation. The color correction effect of FUnIEGAN and MLFcGAN on the image with slight color deviation is relatively good, as shown in Fig. 12a, b. However, there will be wrong halos for the hazy image, as shown in Fig. 12d, h. UWCNN and Uformer-B are also ideal for color correction of images with slight color deviation, but they are unable to improve the haze effect. WaterNet works well for detail processing, as in Fig. 12a, b, but it also shows the wrong enhancement effect, such as the wrong green lines on top of Fig. 12f. STSC and SCNet have a satisfactory improvement on some images with less serious color distortion, such as Fig. 12a–c. However, the color saturation and contrast generated in Fig. 12e are low, and the haze removal effect is not satisfactory in Fig. 12d. The sharpening effect of TACL on the edge of the object is relatively obvious, such as the echinus in Fig. 12c, d. However, on the whole, the color deviation is nonnegligible, as shown in Fig. 12a, b, d–f. In contrast, TEGAN achieves the best effect both from the perspective of color correction and blur elimination, making the enhanced image richer in color, higher in contrast, and more distinct in detail. It is worth

Table 3 Full-reference image quality evaluation for paired image test sets.

Model	Val			Underwater-dark			Underwater-imagenet			Underwater-scenes			Average		
	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑
EUF [4]	0.0420	15.2108	0.6849	0.0278	16.0742	0.7264	0.0527	14.6220	0.6757	0.0211	17.1565	0.7880	0.0359	15.7659	0.7188
CBFU [5]	0.0145	19.0985	0.7737	0.0140	18.9166	0.7964	0.0169	18.8727	0.7834	0.0122	19.7021	0.8364	0.0144	19.1475	0.7975
ICM [6]	0.0197	17.6900	0.7105	0.0204	17.3201	0.7242	0.0202	17.4193	0.7311	0.0106	20.0073	0.8218	0.0177	18.1092	0.7469
GC [7]	0.0324	15.2375	0.6384	0.0320	15.1930	0.6620	0.0336	15.0466	0.6587	0.0229	16.5034	0.7549	0.0302	15.4951	0.6785
MIP [9]	0.1757	7.9269	0.3679	0.0268	16.0345	0.6801	0.0398	15.1060	0.6376	0.0316	15.9692	0.7008	0.0685	13.7592	0.5966
DCP [10]	0.0415	14.5672	0.6307	0.0440	13.8761	0.6273	0.0368	14.8865	0.6649	0.0214	17.2620	0.7565	0.0359	15.1480	0.6698
UDCP [11]	0.0598	13.2493	0.6073	0.0288	15.6992	0.6791	0.0267	16.3734	0.6960	0.0116	19.6371	0.8049	0.0317	16.2397	0.6968
ULAP [14]	0.0232	17.5730	0.7418	0.0196	17.4960	0.7315	0.0202	17.8384	0.7666	0.0132	19.1099	0.8223	0.0191	18.0043	0.7656
CycleGAN [18]	0.0063	22.8608	0.8653	0.0152	18.7521	0.7506	0.0063	22.5958	0.8796	0.0083	21.2465	0.8609	0.0090	21.3638	0.8391
FUnIEGAN [24]	0.0062	22.8698	0.8638	0.0136	19.0435	0.7645	0.0097	20.8387	0.8409	0.0106	20.1491	0.8394	0.0100	20.7253	0.8271
MLFcGAN [31]	*0.0047*	**24.0191**	**0.8908**	0.0118	19.6285	0.7910	0.0082	21.6599	0.8625	0.0098	20.5239	0.8521	0.0087	21.4579	0.8491
UWCNN [27]	0.0121	20.9314	0.8473	0.0101	20.5414	0.7893	0.0056	23.2709	0.8864	**0.0052**	**23.3712**	**0.9025**	0.0083	22.0287	0.8564
WaterNet [26]	0.0060	22.8399	0.8695	0.0114	19.9268	0.7945	0.0062	22.6749	0.8850	0.0058	22.5844	*0.9010*	0.0074	22.0065	0.8625
Uformer-B [23]	0.0060	22.8766	0.8662	0.0114	20.0979	0.7918	0.0116	20.1118	0.8353	0.0105	20.3713	0.8475	0.0099	20.8644	0.8352
STSC [28]	0.0077	22.8446	0.8735	0.0127	19.5803	0.7919	0.0059	*23.5815*	*0.9011*	0.0075	22.0023	0.8887	0.0085	22.0022	0.8638
SCNet [29]	0.0058	23.1616	0.8780	*0.0094*	*20.9345*	*0.8108*	*0.0053*	23.3853	0.8977	0.0058	22.8325	0.8967	*0.0065*	*22.5785*	*0.8708*
TACL [30]	0.0161	18.5745	0.7590	0.0193	17.9269	0.7537	0.0168	18.4264	0.7838	0.0115	19.9484	0.8392	0.0159	18.7190	0.7839
TEGAN(Ours)	**0.0046**	*24.0110*	*0.8877*	**0.0090**	**20.9761**	**0.8186**	**0.0036**	**24.9123**	**0.9199**	*0.0054*	*22.9551*	0.8996	**0.0056**	**23.2136**	**0.8814**

Bold indicates the best, *Italic* indicates the second best, and *Underline* indicates the third best

mentioning that, in contrast to CycleGAN, we can see that the supervised learning used in our paper is more suitable for underwater image enhancement tasks than unsupervised learning. Compared with MLFcGAN, it can be concluded that using Transformer block is more effective than using CNN in extracting long-distance or even global information. More importantly, the comparison with Uformer reveals that the unit fusion scheme we proposed is superior to the method of using Transformer block in each layer of the network.

For the unpaired image test set, since there is no GT as a reference, we choose nonreference evaluation metrics: UIQM [39], UCIQE [40], NIQE [41], BRISQUE [42], FRIQUEE [43], information entropy (Entropy), and underwater index (U) [22].

UIQM consists of three underwater image attribute metrics: UICM, UISM, and UIConM. Each attribute is selected to evaluate one aspect of underwater image degradation. UCIQE uses the Lab color space to linearly combine color density, saturation, and contrast to quantitatively evaluate underwater images for nonuniform color cast, blurring, and low contrast. NIQE extracts features using a multivariate Gaussian model and then combines them with quality distributions using an unsupervised approach. It is concluded that BRISQUE and FRIQUEE have a high consistency of human subjective perception and allow objective evaluation of images [44]. Due to the high time complexity of FRIQUEE, only the first 100 images are evaluated for each test group at most. Information entropy reflects the richness of the image, and the underwater index can be considered as the characteristic image intensity.

The numerical comparison results shown in Table 4 demonstrate the excellent performance of TEGAN. According to the average results of the five test groups, our method is superior to others in UIQM, NIQE, BRISQUE, and FRIQUE. UIQM is at least 0.3175 higher, UCIQE is at least 0.0054 higher, NIQE is at least 0.4465 lower, BRISQUE is at least 8.1168 lower, and FRIQUEE is at least 2.4478 higher. The performance of TEGAN in Entropy and underwater index needs to be improved, but it still ranks at the forefront among various comparison methods.

Ablation Study

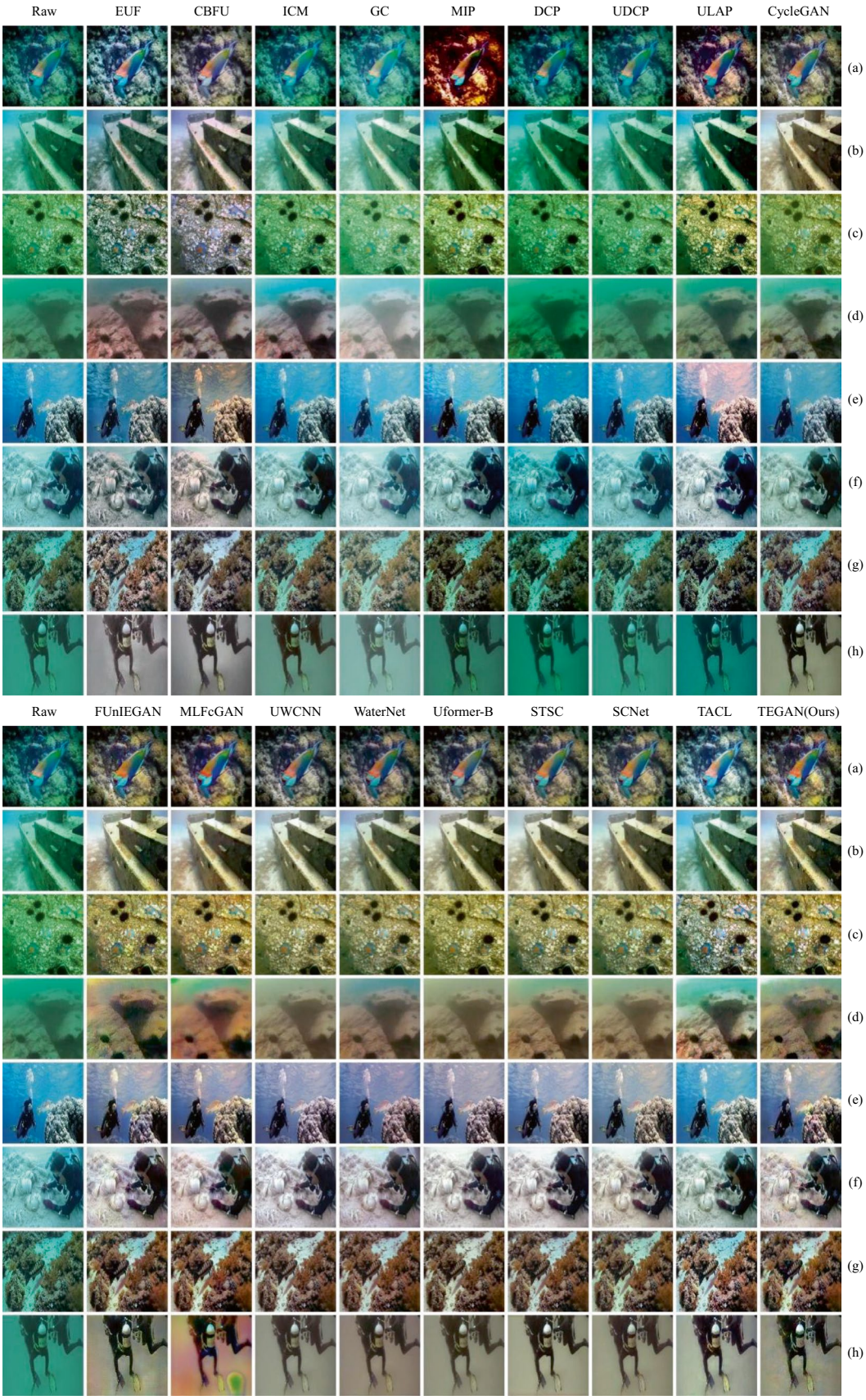
To confirm the efficacy of our strategy, several ablation studies are carried out. The color branch of the discriminator aims to generate more realistic colors than the GT. Therefore, for the validation set with GT, to evaluate and demonstrate the learning ability of the DleWin block and other units of the generator, in this part of the experiment, the discriminator only uses the feature branch to guide the generator's training (see Table 5).

1. The generator deploys only the Encoder + Decoder unit, noted as ED.
2. The generator adds a Bottleneck unit in the middle of the Encoder and Decoder based on ED, denoted as ED-B.
3. The generator adds the fusion unit to ED-B and fuses the features extracted by the Bottleneck unit into each layer of the Decoder, which is named ED-BF.
4. Replace the DleWin block with W-MSA on the generator framework proposed in this paper, denoted as Ours-W.
5. On the generator framework proposed in this paper, replace the DleWin block with the LeWin block, denoted as Ours-L.
6. The generator framework of TEGAN proposed in this paper is denoted as Ours.

Here, each generator structure is trained on the training set, and the model with the largest PSNR on the validation set is taken as a comparison. Detail is crucial for improving the quality of the underwater image. We compare the detailed enhancement effect of different generator frameworks and Transformer blocks in Fig. 13. From a global perspective, TEGAN's generator framework improves the input image significantly in terms of brightness, color, and contrast and is closest to the GT. Locally, our generator enhances the structural details well, as shown by the enlarged areas in the red and blue boxes in Fig. 13.

Table 6 exhibits the quantitative evaluations of different generators on the validation set. Due to the robustness and excellent learning ability of our proposed generator framework, it achieves the best results on MSE, PSNR, and SSIM. In addition, from Table 6, we can derive the following conclusions:

1. Transformer can extract global information, which is very important for UIE. The role of Transformer can be clearly seen from the comparison between ED and ED-B.
2. Fusion unit has an excellent ability to fuse features across different scales. ED-B and ED-BF show that the Fusion unit has a positive effect. It integrates global information, such as overall lighting and image layout, into each scale. The fusion of global and local information at different scales facilitates the generation of images with more natural colors and better details.
3. Transformer can effectively extract the original features. The comparison between ED-BF and Ours shows that the extraction of dependency between original features is helpful to improve the enhancement results.
4. The DleWin block is more effective for underwater image enhancement tasks. Numerical comparisons between Ours-W, Ours-L, and Ours show that local enhancement is particularly important. Notably, compar-


Raw
EUF
CBFU
ICM
GC
MIP
DCP
UDCP
ULAP
CycleGAN
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Raw
FUnIEGAN
MLFcGAN
UWCNN
WaterNet
Uformer-B
STSC
SCNet
TACL
TEGAN(Ours)
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

◂**Fig. 12** Visual comparison of various methods in terms of color, sharpness, and contrast on unpaired test image sets. Each row is the processing result of the corresponding method. **a**, **b** Selected from EUVP. **c**, **d** From RUIE. **e**, **f** From UIEB-Raw. **g**, **h** From UIEB-Challenge. Raw indicates the original image

ing Ours-W and Ours-L shows that LeFF can effectively improve the model's metric performance. Comparing Ours-L and Ours, we can also see the positive effect of PLE. Dual local enhancement can effectively improve image clarity and significantly correct the overall color of the image.

Running Time Comparison

The average running times for different methods on the Intel(R) Core i5-9th CPU and GeForce RTX 3090 GPU platforms are illustrated in Table 7. The image resolution is 256×256. Among them, only GC, FUnIEGAN, and UWCNN are faster than ours because these three methods mainly pursue time performance. We can conclude that TEGAN not only achieves superior image quality improvements but also has a relatively high speed to meet real-time processing requirements.

Table 4 Nonreference image quality evaluation for unpaired image test sets

Dataset	Model	UIQM↑	UCIQE↑	NIQE↓	BRISQUE↓	FRIQUEE↑	Entropy↑	U↓
EUVP	EUF [4]	4.3605	0.4569	6.1210	49.8592	60.5756	7.2300	5.2731
	CBFU [5]	*4.8559*	0.4689	4.8432	44.5544	57.8182	*7.4979*	0.8520
	ICM [6]	4.3895	0.4708	4.9388	44.6138	55.6038	7.3679	1.3448
	GC [7]	4.5395	0.4400	4.9167	45.4318	56.7953	7.1317	1.6230
	MIP [9]	3.1248	**0.4870**	5.6630	49.4842	57.3332	6.9780	1.3280
	DCP [10]	3.6444	0.4770	5.1293	44.7957	55.1869	7.0497	1.3332
	UDCP [11]	4.2555	0.4612	5.0583	44.0932	54.8832	7.0843	1.5592
	ULAP [14]	3.5666	0.4804	5.2165	47.5604	60.2092	7.2905	0.8728
	CycleGAN [18]	4.7943	0.4678	4.9743	38.9983	55.8087	7.3096	1.4579
	FUnIEGAN [24]	4.5453	0.4785	4.6151	*36.3961*	62.8247	7.2652	**0.5127**
	MLFcGAN [31]	4.6513	0.4758	5.8194	39.5810	51.8567	7.2509	0.6951
	UWCNN [27]	4.4731	0.4629	6.0721	47.7350	59.2801	7.2498	0.7432
	WaterNet [26]	4.7310	0.4612	5.2515	40.7045	54.4222	7.1574	1.1002
	Uformer-B [23]	4.1923	0.4597	5.2934	51.8378	*63.5998*	7.2652	0.6358
	STSC [28]	4.6549	0.3883	5.7503	43.0354	54.7920	7.2488	1.8633
	SCNet [29]	4.6623	0.3875	5.4085	43.2540	55.7166	7.2024	1.3603
	TACL [30]	4.8071	0.4019	*4.4481*	41.0130	58.5231	**7.5932**	0.7890
	TEGAN(Ours)	**5.0046**	*0.4832*	**4.0557**	**29.9324**	**68.3314**	7.4676	*0.6093*
RUIE	EUF [4]	4.6946	0.4254	5.5569	46.5171	64.0904	6.7802	6.1083
	CBFU [5]	*4.8801*	**0.4574**	4.8306	44.7774	64.6524	7.2173	2.2157
	ICM [6]	4.7870	*0.4525*	4.7716	43.8303	63.9481	*7.3133*	1.3726
	GC [7]	3.8810	0.4085	5.1634	44.8073	65.0544	6.5722	2.9404
	MIP [9]	4.1749	0.4412	5.2502	43.8162	63.9635	6.6336	3.8282
	DCP [10]	4.0101	0.4497	5.4381	42.6637	63.9148	6.5310	3.8083
	UDCP [11]	4.0145	0.4394	5.5014	42.6025	64.0881	6.4275	4.1596
	ULAP [14]	4.1918	0.4476	5.0156	44.6175	*65.3616*	6.8414	2.9148
	CycleGAN [18]	4.6778	0.4483	4.9908	39.0995	58.1584	7.1954	2.0706
	FUnIEGAN [24]	4.1129	0.4397	5.3971	40.0753	61.8528	6.7025	*1.0773*
	MLFcGAN [31]	4.0349	0.4509	5.8126	*37.4295*	53.5573	6.7106	1.0855
	UWCNN [27]	3.9704	0.4337	6.3955	43.4842	53.1582	6.6867	5.6886
	WaterNet [26]	4.5347	0.4274	5.5067	41.9502	54.5429	6.5759	2.4009

Table 4 (continued)

Dataset	Model	UIQM↑	UCIQE↑	NIQE↓	BRISQUE↓	FRIQUEE↑	Entropy↑	U↓
	Uformer-B [23]	4.0131	0.4313	5.4822	39.6656	56.8438	6.6721	3.6738
	STSC [28]	4.3872	0.3478	5.9195	42.2562	55.8793	6.8004	3.6151
	SCNet [29]	4.1591	0.3553	5.4893	39.5819	58.0078	6.6918	3.4528
	TACL [30]	4.8640	0.3890	*4.2395*	38.7441	64.8604	**7.6556**	**0.9030**
	TEGAN(Ours)	**5.1054**	0.4510	**4.0887**	**29.5649**	**66.8172**	7.1121	1.4617
UIEB-Raw	EUF [4]	4.3533	0.4443	5.6150	49.6060	*66.6609*	7.0417	3.8549
	CBFU [5]	*4.7121*	0.4585	4.2984	43.2827	66.5699	*7.3973*	1.1379
	ICM [6]	4.2979	0.4631	4.4754	42.0219	64.5000	7.2737	1.5602
	GC [7]	3.9741	0.4251	4.3815	41.2776	65.2016	6.8713	2.8155
	MIP [9]	3.6387	0.4684	4.7268	44.6651	63.4432	6.9466	2.0195
	DCP [10]	3.8068	0.4655	4.7534	41.7284	62.9893	6.8783	1.8785
	UDCP [11]	4.0091	0.4467	4.6350	41.0713	63.3175	6.8696	2.7133
	ULAP [14]	3.7549	0.4652	4.4575	44.3742	65.6067	7.2090	1.0840
	CycleGAN [18]	4.6478	0.4612	4.6871	38.7810	56.9937	7.2008	1.5918
	FUnIEGAN [24]	4.4433	*0.4736*	4.5994	39.1750	62.3954	7.1242	**0.5619**
	MLFcGAN [31]	4.3514	0.4634	5.5019	*37.6960*	53.8039	7.0965	0.7319
	UWCNN [27]	4.3294	0.4492	5.7197	44.3048	58.5633	7.0170	1.3305
	WaterNet [26]	4.6768	0.4489	4.9557	41.6345	57.8682	6.9959	1.3131
	Uformer-B [23]	4.1303	0.4430	4.6986	44.9932	61.1022	7.0378	1.3397
	STSC [28]	4.4883	0.3718	5.7175	47.6483	58.4120	7.1524	2.0971
	SCNet [29]	4.3964	0.3694	5.1116	44.7554	58.9690	7.0257	1.8784
	TACL [30]	4.6030	0.3933	*4.1232*	41.3565	66.2499	**7.5516**	0.9533
	TEGAN(Ours)	**5.0246**	**0.4771**	**3.6101**	**29.1717**	**68.4747**	7.3943	*0.7216*
UIEB-Challenge	EUF [4]	3.9284	0.4329	6.4262	49.7073	**64.1524**	6.7990	1.9861
	CBFU [5]	4.3636	0.4499	5.3194	45.1126	*63.9500*	*7.1765*	1.7696
	ICM [6]	3.8398	0.4491	5.9938	43.1220	58.5856	6.9660	2.2257
	GC [7]	3.4450	0.4139	6.1840	43.1688	61.6943	6.6510	3.5586
	MIP [9]	2.7937	0.4450	6.5929	46.3823	57.3661	6.2865	6.2615
	DCP [10]	3.1293	0.4482	6.5111	43.0522	56.5807	6.3873	4.8595
	UDCP [11]	3.4065	0.4362	6.6717	43.2497	57.8285	6.4536	5.0141
	ULAP [14]	2.9505	0.4488	6.1821	45.7979	60.9854	6.6874	6.5051
	CycleGAN [18]	4.1463	0.4517	5.7245	38.9674	50.0609	6.9731	2.3589
	FUnIEGAN [24]	3.7477	0.4544	5.0228	44.5457	55.6403	6.6343	*1.1139*
	MLFcGAN [31]	3.6526	*0.4555*	6.5267	*37.7937*	47.8029	6.7071	1.2746
	UWCNN [27]	3.6827	0.4375	7.2701	44.3438	51.9532	6.5983	4.6327
	WaterNet [26]	4.0754	0.4402	6.1286	42.5714	54.2292	6.6321	2.5510
	Uformer-B [23]	3.3948	0.4348	5.8971	46.9003	60.2016	6.7350	1.8592
	STSC [28]	3.8233	0.3692	6.8951	46.5192	56.3617	6.7155	4.1391
	SCNet [29]	3.7517	0.3698	6.4181	42.0599	56.5097	6.6261	3.5771
	TACL [30]	*4.3727*	0.3907	*4.5992*	40.3940	60.8208	**7.4045**	1.1685
	TEGAN(Ours)	**4.9469**	**0.4565**	**3.8694**	**31.3636**	61.6470	7.0560	**1.0573**
Average	EUF [4]	4.3342	0.4399	5.9298	48.9224	*63.8698*	6.9627	4.3056
	CBFU [5]	*4.7029*	0.4587	4.8229	44.4318	63.2476	*7.3223*	1.4938

Table 4 (continued)

Dataset	Model	UIQM↑	UCIQE↑	NIQE↓	BRISQUE↓	FRIQUEE↑	Entropy↑	U↓
	ICM [6]	4.3286	0.4589	5.0449	43.3970	60.6594	7.2302	1.6258
	GC [7]	3.9599	0.4219	5.1614	43.6714	62.1864	6.8066	2.7344
	MIP [9]	3.4330	0.4604	5.5582	46.0870	60.5265	6.7112	3.3593
	DCP [10]	3.6477	0.4601	5.4580	43.0600	59.6679	6.7116	2.9699
	UDCP [11]	3.9214	0.4459	5.4666	42.7542	60.0293	6.7088	3.3616
	ULAP [14]	3.6160	0.4605	5.2180	45.5875	63.0407	7.0071	2.8442
	CycleGAN [18]	4.5666	0.4573	5.0942	38.9615	55.2554	7.1697	1.8698
	FUnIEGAN [24]	4.2123	*0.4616*	4.9086	40.0480	60.6783	6.9316	**0.8165**
	MLFcGAN [31]	4.1726	0.4614	5.9152	*38.1250*	51.7552	6.9413	*0.9468*
	UWCNN [27]	4.1139	0.4458	6.3644	44.9670	55.7387	6.8879	3.0987
	WaterNet [26]	4.5045	0.4444	5.4606	41.7152	55.2656	6.8403	1.8413
	Uformer-B [23]	3.9326	0.4422	5.3428	45.8492	60.4368	6.9275	1.8771
	STSC [28]	4.3384	0.3693	6.0706	44.8648	56.3612	6.9793	2.9287
	SCNet [29]	4.2424	0.3705	5.6069	42.4128	57.3008	6.8865	2.5672
	TACL [30]	4.6617	0.3937	*4.3525*	40.3769	62.6135	**7.5512**	0.9535
	TEGAN(Ours)	**5.0204**	**0.4670**	**3.9060**	**30.0082**	**66.3176**	7.2575	0.9625

Bold indicates the best, *Italic* indicates the second best, and *Underline* indicates the third best

Enhancement Effect for High-Resolution Images

High-resolution images cited from SUIM dataset [45] with 512×512 are tested to verify the enhancement effect of TEGAN. Several representative comparison methods and evaluation metrics are selected. The enhancement effects are shown in Fig. 14. As we can see, CBFU easily causes color bias. The color corrections of ULAP, WaterNet, Uformer-B, and SCNet are incomplete, and the enhanced image still has a thin veil effect. TACL results in incorrect enhancement, such as white patches appearing in the lower areas of the turtle. Additionally, there still exists residual water color in the TACL-enhanced image of the second row. The TEGAN-enhanced image has high clarity and realistic color. The advantages of TEGAN in terms of evaluation metrics are exhibited in Table 8. These results indicate that TEGAN can handle high-resolution images well. In addition, the average running times of different methods on high-resolution images are shown in Table 9. Among them, TEGAN has the fastest processing speed, which can facilitate many practical applications.

Downstream Application Test

In this section, several typical downstream visual tasks are selected to prove the effectiveness of our model. We test SIFT keypoint matching [46], Canny edge detection

Table 5 Model description with different generator structures

Units/blocks							Model
Encoder	**Decoder**	**Bottleneck**	**Fusion**	**Inception**	**W-MSA**	**LeWin**	
√	√	×	×	×	×	×	**ED**
√	√	√	×	×	×	×	**ED-B**
√	√	√	√	×	×	×	**ED-BF**
√	√	√	√	√	×	×	**Ours**
√	√	√	√	√	√	×	**Ours-W**
√	√	√	√	√	×	√	**Ours-L**

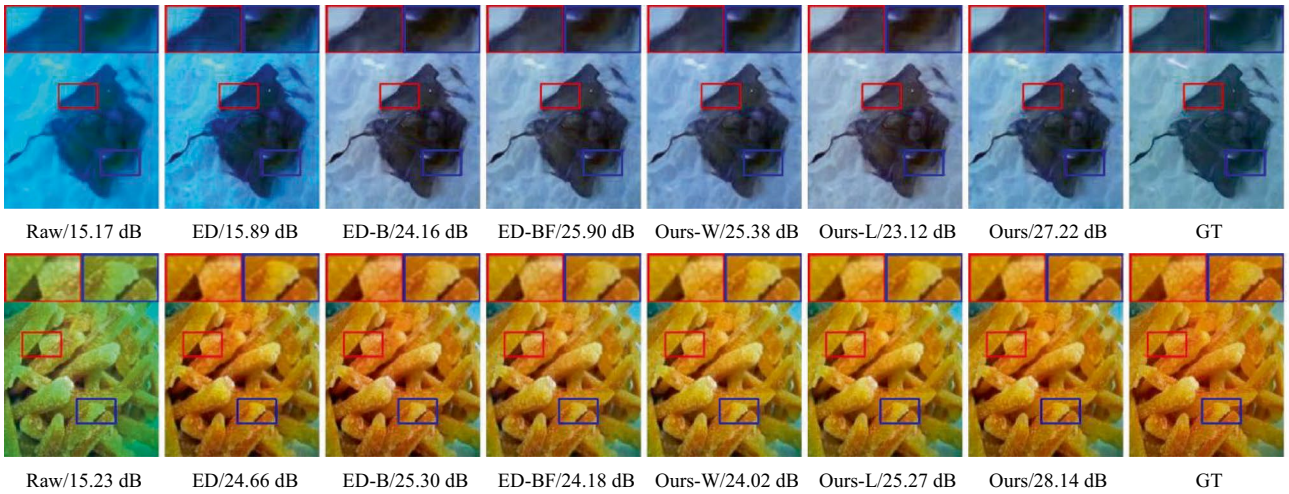


Fig. 13 Ablation study of the contributions of each unit/block in terms of color, sharpness, and contrast on the validation set. Red and blue areas in each image are enlarged and displayed above to indicate the details. The number on the bottom of each image refers to its PSNR (dB)

Table 6 Quantitative evaluations of the ablation study on the validation set

	Raw	ED	ED-B	ED-BF	Ours-W	Ours-L	Ours
MSE↓	0.0277	0.0056	0.0043	0.0042	0.0042	0.0042	**0.0039**
PSNR↑	16.3432	23.3505	24.8078	24.8716	24.5245	24.5896	**25.1048**
SSIM↑	0.6664	0.8750	0.9030	0.9025	0.9014	0.8986	**0.9034**

Table 7 Average running times of different methods (in seconds)

Model	Runtimes↓	Model	Runtimes↓	Model	Runtimes↓	Model	Runtimes↓
DCP [10]	3.1561	ULAP [14]	0.3195	SCNet [29]	0.0606	UWCNN [27]	0.0129
UDCP [11]	2.7793	Uformer-B [23]	0.3141	TACL [30]	0.0528	FUnIEGAN [24]	0.0118
MIP [9]	2.4418	WaterNet [26]	0.2409	CycleGAN [18]	0.0444	GC [7]	0.0108
ICM [6]	0.6752	STSC [28]	0.1692	MLFcGAN [31]	0.0319		
CBFU [5]	0.3550	EUF [4]	0.1344	**TEGAN (Ours)**	**0.0291**		

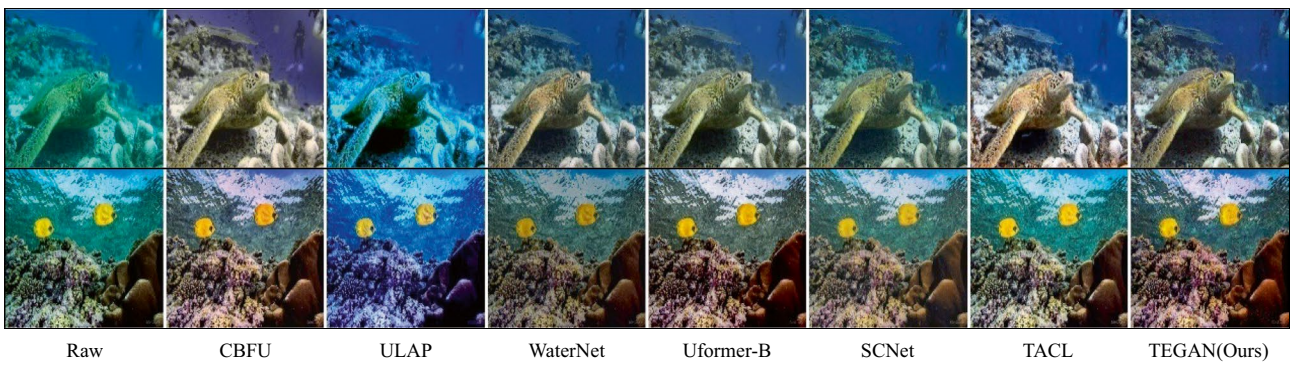


Fig. 14 Enhancement effect for high-resolution images on SUIM

Table 8 Nonreference image quality evaluation for high-resolution images on SUIM

	CBFU [5]	ULAP [14]	WaterNet [26]	Uformer-B [23]	SCNet [29]	TACL [30]	TEGAN (Ours)
UIQM↑	4.0765	*4.0786*	2.7571	3.3433	3.8316	3.9983	**4.5687**
UCIQE↑	**0.3927**	0.3778	0.3757	0.3784	0.3678	*0.3861*	0.3851
NIQE↓	6.33252	5.463042	**5.094818**	6.527694	*5.17847*	5.390188	5.208449
BRISQUE↓	44.9696	41.2959	39.4789	*37.5817*	41.5359	39.1594	**34.0691**

Bold indicates the best, *Italic* indicates the second best, and *Underline* indicates the third best

Table 9 Average running time of different methods on high-resolution images (in seconds)

Model	CBFU [5]	ULAP [14]	WaterNet [26]	Uformer-B [23]	SCNet [29]	TACL [30]	TEGAN (Ours)
Runtimes	0.8785	1.1938	0.4412	0.4804	0.3153	0.0862	**0.0471**

[47], and underwater object detection and then compare them with some representative underwater image processing methods. As shown in Fig. 15, the SIFT algorithm has only a few matched keypoints on the original image, while other enhanced methods have improved the number of matches. In the images enhanced by our model, significant features have been extracted, and a large number of accurate matchings can be attained. The same conclusion can be seen from Fig. 16. Canny detection only has relatively few edges on the original image. Compared

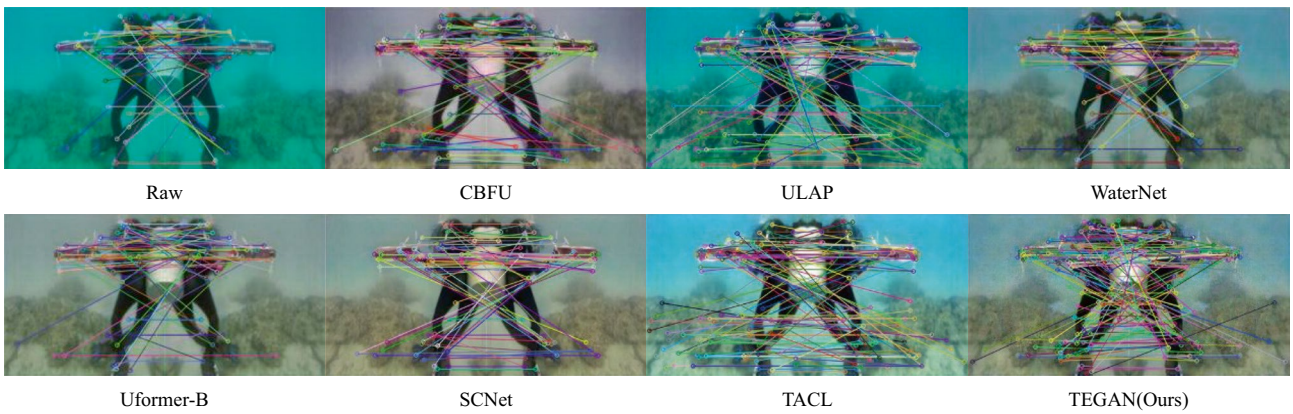


Fig. 15 SIFT keypoint matching results with different methods. The original image and its flipped mirror image are used to exhibit the feature point matching performance

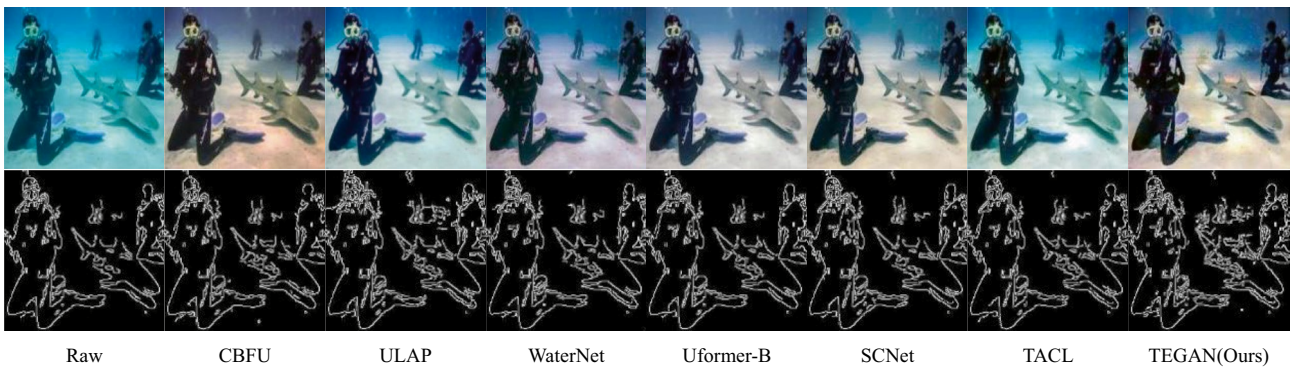


Fig. 16 Canny edge detection results with different methods. The upper row represents the images to be detected, and the lower row represents the edge

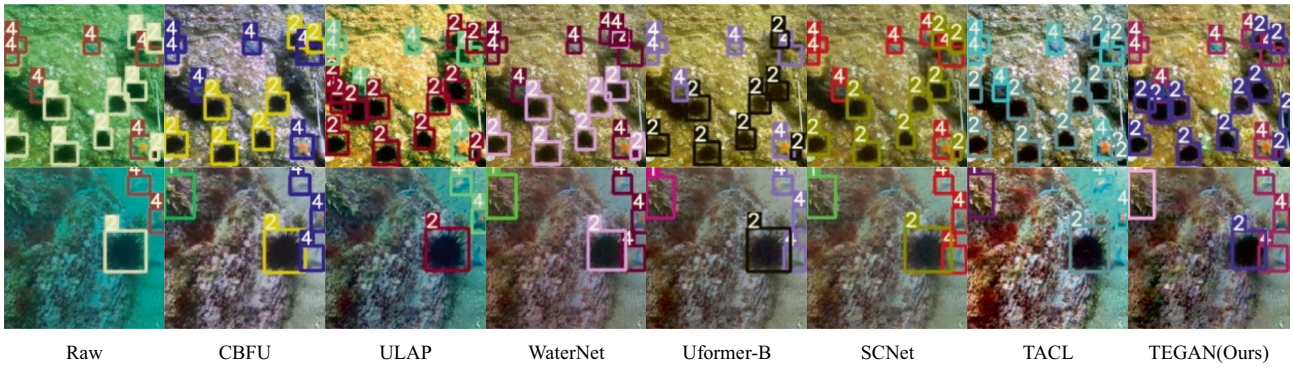


Fig. 17 Underwater object detection results by YOLOv5 with different methods. The labeled bounding boxes 1, 2, 3, and 4 represent the object categories of holothurian, echinus, scallop, and starfish, respectively

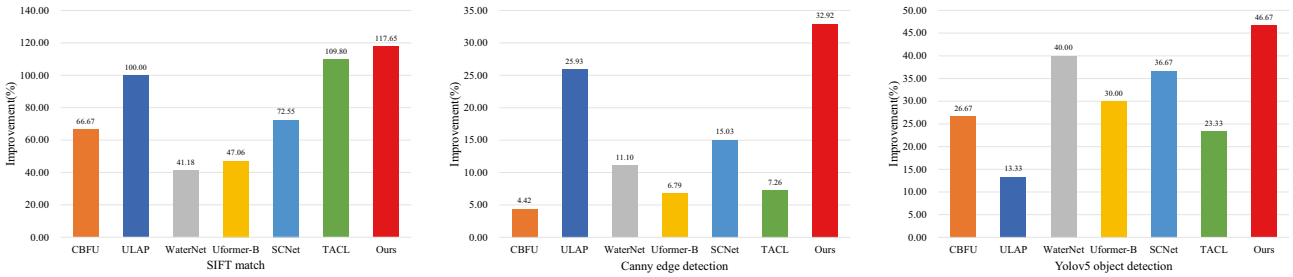


Fig. 18 Percentage of performance improvement tested on SIFT keypoint matching, Canny edge detection, and underwater object detection with different methods

with others, the image enhanced by ours can significantly obtain more edges. In Fig. 17, we use YOLOv5 [48] model for underwater object detection, which is trained on a dataset containing 300 labeled images. By comparing with original images and other enhancement methods, TEGAN can achieve significant and competitive improvement in detection accuracy. The numerical experiments shown in Fig. 18 indicate the percentage of performance improvement tested on the above three downstream applications. Although the results vary depending on different tasks, we observe approximately 41–117, 4–32, and 13–46% improvements, respectively. The outstanding results reveal that our model can facilitate the performance of other visual tasks.

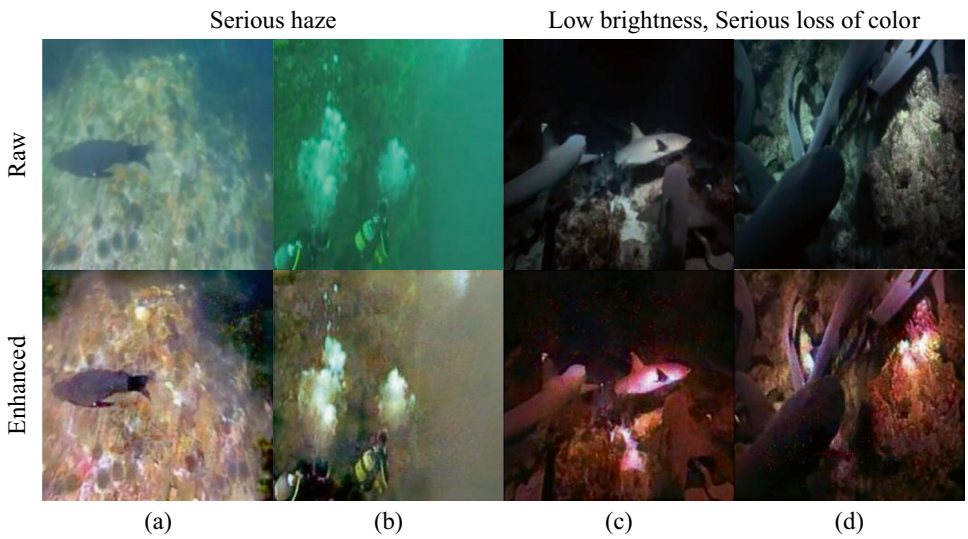


Fig. 19 Failure enhancement results. The upper row represents the raw images with serious haze, low brightness, and serious loss of color. The lower represents the less satisfactory enhancement results by our method

Failure Case Analysis

Our method still has some shortcomings. For images with severe haze, after enhancement, some noise and blur will be introduced, as shown in Fig. 19a, b, respectively. For images with very low brightness and serious color loss, the brightness has not been greatly improved while suffering from excessive color enhancement, as we can see in Fig. 19c, d.

These failure cases are mainly caused by the fact that our training set does not contain these severely degraded and distorted images. For the deep learning method, it is difficult to handle images with large differences from the training set. In addition, since all our images are tested on the size of 256×256, for some very large size images, some edge features will be lost after the scaling down operation, which also leads to blurring.

Conclusions and Future Work

In this paper, we propose a Transformer embedded generative adversarial network for underwater image enhancement. A DleWin Transformer block that can adapt well to the high demands of underwater image enhancement tasks for local feature extraction is designed. We also fuse Transformer with CNN in units, which allows our model to focus on local information and capture long-range or even global dependencies. The proposed TEGAN with a two-branch discriminator can preserve the image content by the feature branch and restore the image color by the color branch. Compared with other methods, TEGAN achieves the best results in terms of comprehensive performance, whether on paired or unpaired datasets. Moreover, it can significantly facilitate the performance of other downstream visual tasks. Future works can be carried out in the following aspects. Other attention mechanisms can be integrated into Transformer to further improve the downstream application tasks. Combining unsupervised and supervised methods for training to solve the problem of insufficient paired datasets will be another focus. In addition, there are still some problems with the mainstream evaluation metrics. In some cases, there is a deviation between image quality metrics and subjective perception. The inconsistency between the two is also an urgent issue to be studied and improved.

Acknowledgements The authors would like to thank the anonymous reviewers for their insightful comments and suggestions.

Author Contribution Zhi Gao: conceptualization, methodology, writing—original draft. Jing Yang: data curation, software, writing—review and editing, project administration, supervision. Lu Zhang: visualization, investigation. Fengling Jiang: supervision. Xixiang Jiao: writing—review and editing.

Funding This work was supported by the Nature Science Foundation of Anhui Province, China (Grant No. 2108085MF195), the Talent Research Foundation of Hefei University (Grant No. 20RC16), the Natural Science Foundation of Education Bureau of Anhui Province (2022AH052130), the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation (MMC202007), and the University Humanities and Social Sciences Research Project of Anhui Province (KJ2021A0992).

Data Availability The datasets generated and analyzed during the current study are available in EUVP http://irvlab.cs.umn.edu/resources/euvp-dataset, RUIE https://github.com/dlut-dimt/Realworld-Underwater-Image-Enhancement-RUIE-Benchmark, UIEB https://li-chongyi.github.io/proj_benchmark.html, and SUIM https://irvlab.cs.umn.edu/resources/suim-dataset.

Declarations

Ethics Approval This article does not contain any studies involving human participants and/or animals by any of the authors.

Competing Interests The authors declare no competing interests.

References

1. Bingham B, Foley B, Singh H, Camilli R, Delaporta K, Eustice R, et al. Robotic tools for deep water archaeology: surveying an ancient shipwreck with an autonomous underwater vehicle. J Field Robot. 2010;27(6):702–17.
2. Shkurti F, Xu A, Meghjani M, Higuera J C G, Girdhar Y, Giguere P, et al. Multi-domain monitoring of marine environments using a heterogeneous robot team. IEEE/RSJ Int Conf Intell Robots Syst. 2012. p. 1747–1753.
3. Wu J, Song C, Ma J, Wu J, Han G. Reinforcement learning and particle swarm optimization supporting real-time rescue assignments for multiple autonomous underwater vehicles. IEEE Trans Intell Transp Syst. 2021;23(7):6807–20.
4. Ancuti C, Ancuti C O, Haber T, Bekaert P. (2012, June). Enhancing underwater images and videos by fusion. In: 2012 IEEE Conf Comput Vis Pattern Recognit (CVPR). 2012. p. 81–88.
5. Ancuti CO, Ancuti C, Vleeschouwer CD, Bekaert P. Color balance and fusion for underwater image enhancement. IEEE Trans Image Process. 2018;27(6):379–93.
6. Iqbal K, Salam R A, Osman A M, Talib A Z. Underwater image enhancement using an integrated colour model. IAENG Int J Comput Sci. 2007;34(2).
7. Babakhani P, Zarei P. Automatic gamma correction based on average of brightness. Adv Comput Sci : Int J. 2015;4(6):156–9.
8. Zhou J, Wei X, Shi J, Chu W, Zhang W. Underwater image enhancement method with light scattering characteristics. Comput Electr Eng. 2022;100: 107898.
9. Carlevaris-Bianco N, Mohan A, Eustice R M. Initial results in underwater single image dehazing. In: Oceans 2010 MTS/IEEE Seattle. 2010. p. 1–8.
10. He K, Sun J, Tang X. Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell. 2010;33(12):2341–53.
11. Drews P, Nascimento E, Moraes F, Botelho S, Campos M. Transmission estimation in underwater single images. In: Proceedings of the IEEE Int Conf Comput Vis Workshops. 2013. p. 825–830.
12. Peng YT, Cosman PC. Underwater image restoration based on image blurriness and light absorption. IEEE Trans Image Process. 2017;26(4):1579–94.
13. Chao L, Wang M. Removal of water scattering. In: 2010 2nd Int Conf Comput Eng Technol. 2010. p. V2–35-V2–39.

14. Song W, Wang Y, Huang D, Tjondronegoro D. A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In: Advances in Multimedia Information Processing-PCM 2018: 19th Pacific-Rim Conference on Multimedia. 2018. p. 678–688.
15. Gong K, Hua D. Research on the method of color compensation and underwater image restoration based on polarization characteristics. In: 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA). 2022. p. 746–751.
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Neural Inf Process Syst. 2014. p. 2672–2680.
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017. 30.
18. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE Int Conf Comput Vis. 2017. p. 2223–2232.
19. Jiang X, Zhu Y, Cai G, Zheng B, Yang D. MXT: a new variant of pyramid vision transformer for multi-label chest X-ray image classification. Cogn Comput. 2022;14(4):1362–77.
20. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th Int Conf. 2015. p. 234–241.
21. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. Adv Neural Inf Process Syst. 2017;2017:30.
22. Chen X, Yu J, Kong S, Wu Z, Fang X, Wen L. Towards real-time advancement of underwater visual quality with GAN. IEEE Trans Industr Electron. 2019;66(12):9350–9.
23. Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H. Uformer: a general U-shaped transformer for image restoration. In: 2022 IEEE/CVF Conf Comput Vis Pattern Recog (CVPR). 2022. p. 17662–17672.
24. Islam MJ, Xia Y, Sattar J. Fast underwater image enhancement for improved visual perception. IEEE Robot Autom Lett. 2020;5(2):3227–34.
25. Liu R, Fan X, Zhu M, Hou M, Luo Z. Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light. IEEE Trans Circuits Syst Video Technol. 2020;30(12):4861–75.
26. Li C, Guo C, Ren W, Cong R, Hou J, Kwong S, et al. An underwater image enhancement benchmark dataset and beyond. IEEE Trans Image Process. 2019;29:4376–89.
27. Li C, Anwar S, Porikli F. Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recogn. 2020;98: 107038.
28. Wang D, Ma L, Liu R, Fan X. Semantic-aware texture-structure feature collaboration for underwater image enhancement. In: 2022 Int Conf Robot Autom (ICRA). 2022. p. 4592–4598.
29. Fu Z, Lin X, Wang W, Huang Y, Ding X. Underwater image enhancement via learning water type desensitized representations. In: ICASSP 2022–2022 IEEE Int Conf Acoust, Speech Signal Process (ICASSP). 2022. p. 2764–2768.
30. Liu R, Jiang Z, Yang S, Fan X. Twin adversarial contrastive learning for underwater image enhancement and beyond. IEEE Trans Image Process. 2022;31:4922–36.
31. Liu X, Gao Z, Chen BM. MLFcGAN: multilevel feature fusion-based conditional GAN for underwater image color correction. IEEE Geosci Remote Sens Lett. 2020;17(9):1488–92.
32. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. 2014.
33. Wu K, Peng H, Chen M, Fu J, Chao H. Rethinking and improving relative position encoding for vision transformer. 2021 IEEE/CVF Int Conf Comput Vis (ICCV). 2021. p. 10033–10041.
34. Isola P, Zhu J Y, Zhou T, Efros A A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conf Comput Vis Pattern Recognit. 2017. p. 1125–1134.
35. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv preprint arXiv:1701.07875. 2017.
36. Han R, Guan Y, Yu Z, Liu P, Zheng H. Underwater image enhancement based on a spiral generative adversarial framework. IEEE Access. 2020;8:218838–52.
37. Fabbri C, Islam M J, Sattar J. Enhancing underwater imagery using generative adversarial networks. 2018 IEEE Int Conf Robot Autom (ICRA). 2018. p. 7159–7165.
38. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.
39. Panetta K, Gao C, Agaian S. Human-visual-system-inspired underwater image quality measures. IEEE J Oceanic Eng. 2015;41(3):541–51.
40. Yang M, Sowmya A. An underwater color image quality evaluation metric. IEEE Trans Image Process. 2015;24(12):6062–71.
41. Mittal A, Soundararajan R, Bovik AC. Making a "completely blind" image quality analyzer. IEEE Signal Process Lett. 2013;20(3):209–12.
42. Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. IEEE Trans Image Process. 2012;21(12):4695–708.
43. Ghadiyaram D, Bovik A C. Live in the wild image quality challenge database. 2015. http://live.ece.utexas.edu/research/ChallengeDB/index.html.
44. Gu YS, Jiang QP, Shao F, Gao W. A real-world quality evaluation dataset for enhanced underwater images. J Image Graph. 2022;27(05):1467–80.
45. Islam M J, Edge C, Xiao Y, Luo P, Mehtaz M, Morse C, et al. Semantic segmentation of underwater imagery: dataset and benchmark. IEEE/RSJ Int Conf Intell Robot Syst. 2020, pp. 1769–1776.
46. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vision. 2004;60:91–110.
47. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;6:679–98.
48. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430. 2021. https://github.com/ultralytics/yolov5.