Multimedia Tools and Applications (2023) 82:18691–18708
https://doi.org/10.1007/s11042-022-14228-6

An efficient swin transformer-based method for underwater image enhancement

Rong Wang[1] · **Yonghui Zhang**[1] · **Jian Zhang**[1]

Received: 3 May 2022 / Revised: 13 July 2022 / Accepted: 4 November 2022 /
Published online: 24 November 2022

Abstract
Due to the complex imaging environment of the ocean, the underwater images obtained by optical vision systems are usually severely degraded. Recently, methods for enhancing underwater images are mostly based on deep learning. However, the intrinsic locality of convolution operation makes it difficult to model long-range dependency efficiently, which may lead to the limited performance of these methods. This paper proposes an efficient method for underwater image enhancement by utilizing Swin Transformer for local feature learning and long-range dependency modeling. The network structure of this method is mainly composed of encoder, decoder and skip connections, in which the encoder and decoder take the Swin Transformer block as the basic unit. Specifically, the encoder is used to learn multi-scale feature representations, and the decoder is utilized to upsample the extracted contextual features progressively. Skip connections are used to fuse multi-scale features from the encoder and decoder. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods on different datasets by up to 1.09∼1.64dB (PSNR) and 1.9%∼2.3% (SSIM) in objective metrics, and achieves the best visual effect in subjective comparisons, especially in terms of color cast removal and sharpness enhancement.

Keywords Underwater imaging · Image enhancement · Swin transformer · Deep learning

1 Introduction

Since underwater images carry plenty of ocean information, the quality of underwater images is of great significance to the exploration and utilization of the deep sea. However,

✉ Yonghui Zhang
yhzhang@hainanu.edu.cn

Rong Wang
wrong@hainanu.edu.cn

Jian Zhang
whealther@hainanu.edu.cn

[1] School of Information and Communication Engineering, Hainan University, No. 58 Renmin Avenue, Haikou, 570228, China

due to the absorption and scattering of light by water bodies and the complex biological environment, underwater images captured by the camera usually suffer from severe degradation such as color cast, low sharpness and contrast. Underwater image enhancement (UIE) technology [37, 51] aims to obtain higher quality, more realistic color, and sharper underwater images, which is beneficial for several visual tasks such as object detection and edge detection [4]. This technology has been widely used in underwater robots [20] and underwater archaeology, etc.

The traditional UIE methods mainly include non-physical model-based methods [15, 41, 54, 57] and physical model-based methods [14, 42, 43, 52, 58]. Among them, Jyoti et al. [41] proposed a non-physical model-based method, called Histogram Equalization (HE), which uses pixel value transformation to transform the original image into roughly the same number of pixels in most gray levels. This method is simple and fast, but it is difficult to improve the local contrast of underwater images. On the basis of Dark Channel Prior (DCP) [17], Drews et al. [14] proposed the Underwater Dark Channel Prior (UDCP) to solve the problem that the underwater image is distorted and blue-green due to the serious attenuation of red light in the water. Song et al. [42] proposed the Underwater Light Attenuation Prior (ULAP), which trained a linear model of scene depth estimation based on the underwater light attenuation prior and labeled scene depth data, and the underwater image restoration is realized by estimating scene depth map, atmospheric light value, and transmission image. However, the processed images still have severe color casts. The limitations of traditional methods are mainly in either ignoring the underwater imaging mechanism leading to over/under enhancement, such as HE [41], or being time-consuming or sensitive to the diversity of underwater scenes, such as UDCP [14], ULAP [42], etc. In contrast, our method can produce visually satisfactory enhancement results for underwater images of multiple scenes without being time-consuming.

Compared with the traditional manual setting of spatial features, deep learning can automatically extract spatial features hierarchically [5], which makes it develop rapidly. In recent years, many researchers have begun to apply convolution neural network (CNN) and generative adversarial network (GAN) to underwater image enhancement [11, 16, 19, 20, 25–29, 31, 33, 47, 49]. Among them, Li et al. [28] proposed Water-Net based on CNN, which uses convolution operation to extract underwater image features and learn the mapping relationship between the original underwater image and enhanced underwater image, so as to achieve underwater image enhancement. Islam et al. [20] proposed Fast Underwater Image Enhancement Generative Adversarial Networks (FUnIE-GAN) for enhancing underwater images at high speed, which can be applied to underwater vehicles. Li et al. [29] proposed a multi-color space embedded underwater image enhancement network (Ucolor) based on medium transmission guidance, which combines underwater imaging physical model and deep learning, and uses multi-color space embedding to improve the visual quality of underwater images. Yan et al. [49] proposed a very simple network (MTUR) in which one sub-network is used to predict the media transmission map and the predicted media transmission map is used as guidance to assist another sub-network to enhance the underwater images. This method improves the performance and enables real-time image processing. Compared with traditional methods, deep learning-based methods significantly improve performance. However, the key limitation of CNN or GAN-based methods is that they struggle to effectively model long-range dependency for learning explicit global information interaction due to the use of convolution as the key component for extracting features. In contrast, our method uses Swin Transformer block as the basic unit to efficiently learn local features and model long-range dependency, which can enhance underwater images more accurately.

With the great success of Transformer [46] in the field of natural language processing (NLP), researchers have tried to bring Transformer into the field of computer vision, and have made good progress in several vision problems [7, 13, 45]. However, since vision Transformers usually divide the input image into small patches with a fixed size and process each patch independently, the image generated by vision Transformer may appear with boundary artifacts around each small patch [6, 10]. Recently, Liu et al. [34] proposed Swin Transformer, which integrates the advantages of CNN and Transformer, and has shown great promise. Due to the local attention mechanism, Swin Transformer has the advantage of CNN to learn local features of large-size images. Meanwhile, Swin Transformer can effectively model long-range dependency with the shifted window scheme.

Motivated by the success of Swin Transformer, this paper proposes a novel method based on Swin Transformer for UIE. This method aims to effectively enhance degraded underwater images by exploiting the advantages of Swin Transformer in learning local features on large-size images and modeling long-range dependency. By downsampling in the encoder and upsampling in the decoder, the obtained multi-scale feature maps are used for local details attention and global information interaction via Swin Transformer blocks and feature fusion via skip connections, expecting to be more fully utilized. Figure 1 shows the underwater image enhanced by our method and some comparison methods. As shown, the color of the underwater image enhanced by our method is closest to the reference image and achieves the best visual quality.

The main contributions of this paper can be summarized as follows:

- We introduce Swin Transformer into the underwater image enhancement task and propose an encoder-decoder network with the Swin Transformer block as the basic unit. The Swin Transformer block enables the network to easily achieve local detail attention and global information interaction for underwater images.
- We introduce the HSV color space loss function and use it together with the loss functions in RGB color space for network training, which further improves the ability of the proposed method to remove color casts.

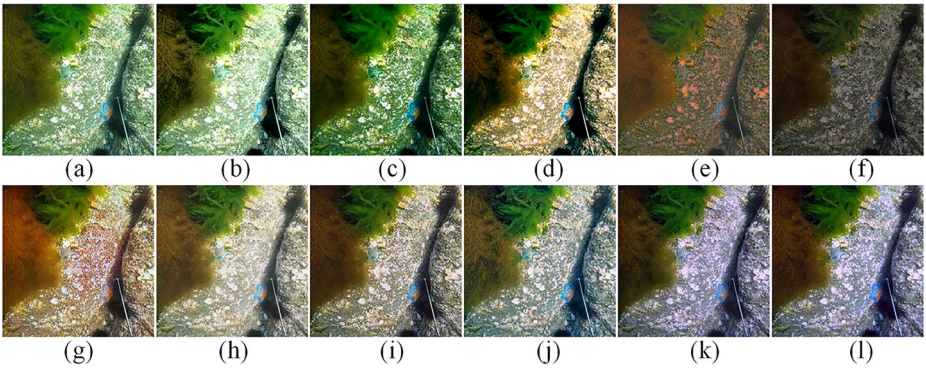

Fig. 1 The underwater image enhanced by our method and several state-of-the-art UIE methods. Our method achieves the best visual quality. (a) Raw image, (b) HE [41], (c) UDCP [14], (d) ULAP [42], (e) UWCNN [27], (f) Water-Net [28], (g) FUnIE-GAN [20], (h) Ucolor [29], (i) Peng et al. [39], (j) MTUR [49], (k) Ours, (l) Reference image

- Extensive experiments on multiple datasets demonstrate that our method achieves superior performance in both objective metrics and visual quality compared to several state-of-the-art UIE methods, especially for color cast removal and sharpness enhancement.

The rest of this paper is organized as follows: Section 2 describes the underwater imaging model and introduces the Vision Transformer, Section 3 illustrates the network architecture and associated inner structure of our proposed method, Section 4 presents the experiment details, experimental results, discussion and ablation studies, and Section 5 gives the conclusion of this work.

2 Related work

2.1 Underwater imaging model

According to the underwater optical imaging model Jaffe-McGlamery proposed by Jaffe et al. [21], the light received by the underwater camera is mainly composed of the direct transmission component $E_d(x, y)$, the forward scattering component $E_f(x, y)$ and the background scattering component $E_b(x, y)$. The underwater imaging model is shown in Fig. 2. The formula for calculating the total light intensity received by the underwater camera is:

$$E_t(x, y) = E_d(x, y) + E_f(x, y) + E_b(x, y). \tag{1}$$

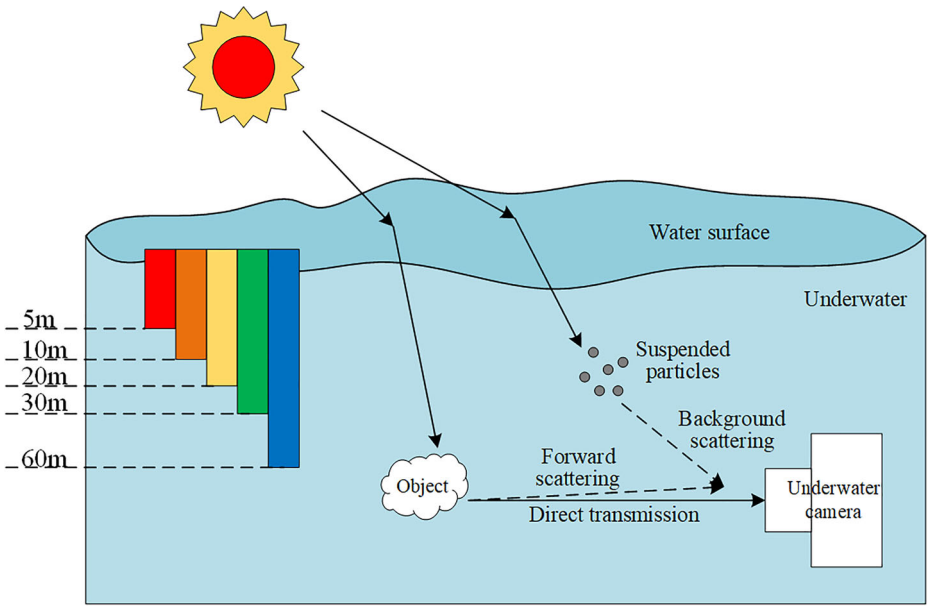

Fig. 2 The underwater imaging model Jaffe-McGlamery

The direct transmission component $E_d(x, y)$ is obtained by the attenuation of the reflected light on the target surface due to the scattering and absorption of water. The calculation formula can be expressed as follows:

$$E_d(x, y) = J(x, y) \cdot t(x, y). \tag{2}$$

where $J(x, y)$ represents the light intensity initially reflected by the imaging object, and $t(x, y)$ represents a transmittance of a medium.

The forward scattering component $E_f(x, y)$ can be expressed as:

$$E_f(x, y) = J(x, y) * h(x, y). \tag{3}$$

where $h(x, y)$ denotes the point spread function.

The background scattering component $E_b(x, y)$ is generally expressed as:

$$E_b(x, y) = E_\infty(1 - t(x, y)). \tag{4}$$

where E_∞ represents the global background light.

Therefore, the underwater image captured by the underwater camera can be expressed as:

$$F(x, y) = J(x, y) \cdot t(x, y) + J(x, y) * h(x, y) + E_\infty(1 - t(x, y)). \tag{5}$$

As shown in Fig. 2, when light travels underwater, different colors of light have different degrees of attenuation in the water. In general, red light is most easily absorbed because of its short wavelength, so it attenuates most rapidly in water. The attenuation of blue light and green light is relatively slow due to their longer wavelengths. As a result, the captured underwater images generally appear green or blue-green. Therefore, the original image taken by the underwater camera needs to be enhanced.

2.2 Vision transformer

Transformer [46] shows an excellent performance in the field of natural language processing (NLP). Many researchers in the field of computer vision have attempted to introduce Transformer which learns to attend to important image regions by exploring the global information interaction between different regions and solved several visual problems, such as image classification [8, 24, 44], object detection [7, 9, 12, 35, 36, 45, 53], segmentation [48, 56] and crowd counting [40], etc. As for underwater image enhancement tasks, Peng et al. [39] introduced several Transformer layers to model global information of the feature maps obtained by four levels of down-sampling, which reinforced the network's attention to seriously degraded parts and contributed to performance improvement. Although many explorations in the field of vision have shown remarkable performance, compared with CNN-based methods, the drawback of Transformer is that it requires pre-training on its own large dataset, which increases the difficulty of training. Recently, Swin Transformer proposed by Liu et al. [34] integrates the advantages of both CNN and Transformer. On the one hand, due to the local attention mechanism, it can easily process images with large sizes, on the other hand, it can effectively realize long-range dependency modeling with the shifted window scheme. In this work, we attempt to use Swin Transformer blocks and skip connections to build an encoder-decoder architecture for underwater image enhancement, thus effectively enhancing degraded real-world underwater images and providing a benchmark comparison for the exploration of Swin Transformer in the field of low-level vision.

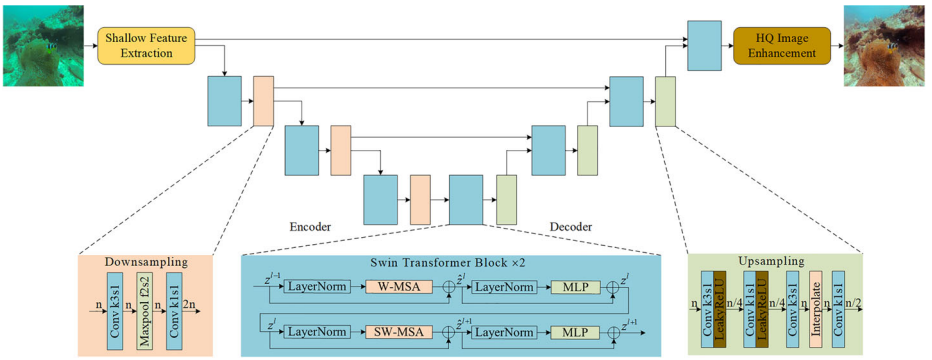

Fig. 3 The network architecture of our proposed method for underwater image enhancement. In the figure, n represents the number of channels of the feature map, the kernel size (k) and stride (s) of each convolutional layer in downsampling and upsampling are provided, and the filter size (f) and stride (s) of maxpool layer in downsampling is also provided

3 Proposed method

We design a Swin Transformer-based encoder-decoder network with skip connections, and the overall architecture of this network is shown in Fig. 3. The network takes the underwater image as input x and learns shallow features of the image by Shallow Feature Extraction, then learns the deep feature representation by the encoder-decoder, and finally outputs the enhanced result y through High Quality (HQ) Image Enhancement. Table 1 shows the description of the proposed algorithm.

3.1 Network architecture

Firstly, a convolutional layer (kernel size 3×3, stride 1, number of channels 32) and a Maxpool layer (filter size 2×2, stride 2) are used to perform the feature extraction process [2]

Table 1 The description of the proposed algorithm

Pseudocode description of the proposed algorithm
Input: degraded image x, Output: enhanced image y
Forward:
$h = Shallow\_Feature\_Extraction(x)$
$h\_down = []$
for i in range(3):
$\quad h\_down.append(h)$
$\quad h = Swin\_Transformer\_Blocks(h)$
$\quad h = Downsampling(h)$
for i in range(3):
$\quad h = Swin\_Transformer\_Blocks(h)$
$\quad h = Upsampling(h)$
$\quad h = Skip\_Connection(h, h\_down[2-i], -1)$
$h = Swin\_Transformer\_Blocks(h)$
$y = HQ\_Image\_Enhancement(h)$

on the input image (size is $W \times H$). Feature extraction of a given image is a critical step in many image analysis and computer vision tasks [3]. For the encoder, the patch tokens transformed from the feature maps output by shallow feature extraction generate hierarchical feature representations via several pairs of two Swin Transformer blocks and downsampling. Where the Swin Transformer blocks are used to capture the local context and model long-range dependency, and downsampling is used to reduce the feature resolution by half and increase the feature dimension by 2 times. For the decoder, upsampling is responsible for increasing the feature resolution by 2 times and reducing the feature dimension by half, and the obtained feature maps are fused with the feature maps of the same size from the encoder through skip connections to complement the loss of spatial information caused by downsampling. The Swin Transformer blocks are used to learn the feature representation of the fused feature maps. Finally, before the convolutional layer (kernel size is 3×3 and the stride is 1) which is used to output the enhanced underwater image, a transposed convolutional layer (kernel size is 4×4, the stride is 2 and channel number is 32) is used for restoring the size of feature maps to the input.

3.2 Swin transformer block

Different from the standard multi-headed self-attention (MSA) module in Transformer, the Swin Transformer block is built based on shifted windows. The internal structure of two consecutive Swin Transformer blocks is shown in Fig. 3. The window-based multi-head self-attention (W-MSA) module and the shifted window-based MSA (SW-MSA) module are applied in two consecutive Swin Transformer blocks, respectively. Each Swin Transformer module also contains a 2-layer MLP with GELU non-linearity and two LayerNorm (LN) layers, one of which is applied before the (S)W-MSA module, the other before the MLP, and the residual connection is applied after each (S)W-MSA module and MLP.

Based on the shifted window partitioning mechanism, the formula for calculating contiguous Swin Transformer blocks can be expressed as follows:

$$\begin{aligned} \hat{z}^{l} &= W - MSA(LN(z^{l-1})) + z^{l-1}, \\ z^{l} &= MLP(LN(\hat{z}^{l})) + \hat{z}^{l}, \\ \hat{z}^{l+1} &= SW - MSA(LN(z^{l})) + z^{l}, \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}. \end{aligned} \tag{6}$$

where $\hat{z}^{l}$ and $\hat{z}^{l+1}$ represent the outputs of W-MSA and SW-MSA modules respectively, z^{l} represents the outputs of the MLP module of the l^{th} block. Similar to works in [30], the attention matrix calculated by the self-attention mechanism is:

$$Attention(Q, K, V) = SoftMax\left(QK^{T} \Big/ \sqrt{d} + B\right) V. \tag{7}$$

where the values in B are taken from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$. Generally, $Q, K, V \in \mathbb{R}^{M^2 \times d}$ and respectively denote the $query$, key and $value$ matrices. M^2 represents the number of patches in a window and d denotes the dimension of the $query$ or key.

3.3 Encoder-decoder

In the encoder, the patch tokens transformed from the feature maps with the size of $W/2 \times H/2$ are fed into two Swin Transformer blocks for representation learning. Meanwhile, the

feature resolution will be down-sampled by half due to the Maxpool layer in downsampling, and the feature dimension will be increased to 2 times the original dimension due to a convolutional layer with a kernel size of 1×1 in downsampling. This will be repeated three times to obtain feature maps with a size of $W/16 \times H/16$.

In the decoder, the feature resolution of the adjacent dimensions will be up-sampled by 2 times via a bilinear interpolate layer in the upsampling, and the feature dimension will be reduced to half of the original dimension by the convolutional layer with a kernel size of 1×1 in upsampling. Then the up-sampled feature maps will be fused with the multi-scale feature maps from the encoder by skip connections, and two Swin Transformer blocks will be used to learn the feature representation of the fused feature maps. This procedure will also be repeated three times until the resolution of feature maps is $W/2 \times H/2$.

3.4 Loss functions

To take advantage of the HSV color space's more intuitive representation of hue, color saturation, and intensity, we introduce the HSV color space loss function together with loss functions in the RGB color space for our network training. The images from RGB space are firstly converted to HSV color space, as follows:

$$\begin{aligned} H^{N(x)}, S^{N(x)}, V^{N(x)} &= RGB2HSV(N(x)), \\ H^{y}, S^{y}, V^{y} &= RGB2HSV(y). \end{aligned} \tag{8}$$

where x, y and $N(x)$ represent the original input underwater images, the reference underwater images and the enhanced underwater images output by the network, respectively.

The loss function in HSV color space can be expressed as follows:

$$\begin{aligned} Loss_{hsv} = E_{x,y} \Bigg[& -\sum_{i=1}^{n} Q(H_i^y) \log \left(Q\left(H_i^{N(x)} \right) \right) \\ & + \left(S^y - S^{N(x)} \right)^2 + \left(V^y - V^{N(x)} \right)^2 \Bigg]. \end{aligned} \tag{9}$$

where Q stands for the quantization operator.

Mean square error (MSE) loss ($Loss_{mse}$) is the RGB color space loss function to calculate the distance between the predicted images $N(x)$ and the groundtruth images y.

$$Loss_{mse} = E\left[\| N(x) - y \|_2 \right]. \tag{10}$$

We also use structure similarity (SSIM) loss ($Loss_{ssim}$) [55] in the RGB color space to impose the structure and texture similarity on the predicted image. We compute the SSIM score for gray images converted from images in the RGB space. For each pixel x, the SSIM value is calculated within an 11×11 image patch around the pixel. The specific calculation formula is as follows:

$$SSIM(x) = \frac{2\mu_I(x)\mu_{\hat{I}}(x) + c_1}{\mu_I^2(x) + \mu_{\hat{I}}^2(x) + c_1} \cdot \frac{2\sigma_{I\hat{I}}(x) + c_2}{\sigma_I^2(x) + \sigma_{\hat{I}}^2(x) + c_2}. \tag{11}$$

where $\mu_I(x)$ and $\mu_{\hat{I}}(x)$ represent the mean of the image patch from groundtruth image and predicted image, respectively. $\sigma_I(x)$ and $\sigma_{\hat{I}}(x)$ represent the standard deviation of the corresponding image patch, respectively. $\sigma_{I\hat{I}}(x)$ denotes cross-covariance. Here we set $c_1 = 0.02$ and $c_2 = 0.03$.

The calculation formula of SSIM loss can be expressed as:

$$Loss_{ssim} = 1 - \frac{1}{N}\sum_{i=1}^{N} SSIM(x_i). \tag{12}$$

In addition, we use the perceptual loss ($Loss_{perc}$) [22] to compute the distance between the feature representation of the predicted image and the ground-truth image.

Our total loss function can be expressed as follows:

$$L = \alpha Loss_{hsv} + \beta Loss_{mse} + \gamma Loss_{ssim} + \mu Loss_{perc}. \tag{13}$$

where α, β, γ and μ are hyperparameters and represent the weight of each loss term. We set $\alpha = 50.0$, $\beta = 0.001$, $\gamma = 100.0$, and $\mu = 100.0$.

4 Experiments

In order to verify our performance superiority, we compare our method with other 9 different UIE methods qualitatively and quantitatively. These methods include traditional methods based on non-physical model (HE [41]) and physical models (UDCP [14] and ULAP [42]), as well as the recent state-of-the-art methods based on deep learning (UWCNN [27], Water-Net [28], FUnIE-GAN [20], Ucolor [29], Peng et al. [39] and MTUR [49]). We will first supplement the implementation details of the experiment, then introduce the experimental datasets and experimental evaluation metrics, and finally analyze the experimental results.

4.1 Datasets and implementation details

In our experiments, we use five real-world underwater image datasets: LSUI [39] (which contains 5004 pairs of underwater images), UIEBD [28] (which contains 890 pairs of underwater images and 60 challenging unpaired degraded underwater images), EUVP [20], SQUID [1] and RUIE [32]. For training, we randomly extract 4600 pairs of underwater images from the LSUI as the training set to train our network. All images are resized to a fixed size before being input into the network. For testing, the remaining 404 pairs of underwater images in the LSUI are used as the first testing dataset (Test-L404). A random set of 90 pairs of real-world images extracted from the UIEBD is used as the second testing dataset (Test-U90). A random set of 70 pairs of real-world images extracted from the EUVP is used as the third testing dataset (Test-E70). A set of 60 challenging unpaired degraded underwater images in the UIEBD is used as the fourth testing dataset (Test-U60). We use the 16 representative examples presented on the project page of SQUID[1] as the fifth testing dataset (SQUID). A random set of 45 real-world images extracted from the RUIE is used as the sixth testing dataset (Test-R45).

We implement the proposed method by using Pytorch with an NVIDIA RTX 2080TI GPU on Ubuntu 18. During the training period, the initial learning rate is set as 0.0005, and the learning rate decreased 20% every 40 epochs. We set the batch size to 12 and utilize the Adam optimization algorithm for a total of 300 epochs of training. The training time for the model is about three days.

[1] http://csms.haifa.ac.il/profiles/tTreibitz/datasets/ambient_forwardlooking/index.html

4.2 Evaluation metrics

Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR) [23] and Structural SIM-ilarity index (SSIM) [18] are used as full-reference evaluation metrics to assess the proximity to the reference, where a higher PSNR (lower MSE) value and a higher SSIM score represent closer image content and a more similar structure and texture, respectively.

The underwater color image quality evaluation (UCIQE) [50] and the underwater image quality measure (UIQM) [38] are used as non-reference evaluation metrics to comprehensively evaluate underwater image quality by color density, saturation, sharpness and contrast, where the higher UCIQE or UIQM score, the better human visual perception. UIQM is the weighted sum of UICM (colorfulness measure), UISM (sharpness measure) and UIConM (contrast measure), as follows:

$$UIQM = c_1 UICM + c_2 UISM + c_3 UIConM. \tag{14}$$

where c_1, c_2 and c_3 are weight parameters. We set $c_1 = 0.0282$, $c_2 = 0.2953$ and $c_3 = 3.5753$ according to [38].

For full-reference image quality evaluation, we use MSE, PSNR and SSIM metrics to compare the methods on Test-L404, Test-U90 and Test-E70 datasets. For non-reference image quality assessment, we use UCIQE and UIQM metrics to compare the methods on Test-U60, SQUID and Test-R45 datasets.

4.3 Performance evaluation

4.3.1 Full-reference evaluation

The quantitative evaluation results of different methods on Test-L404, Test-U90 and Test-E70 datasets are shown in Tables 2, 3 and 4, respectively. Visual comparisons of different methods are shown in Fig. 4. For the 6 UIE methods based on deep learning, we use the source codes and pretrained model parameters provided by the corresponding authors.

As shown in Tables 2, 3 and 4, our method achieves the best performance on Test-L404 and Test-E70, and the second-best results on Test-U90. For the PSNR metric, our method outperforms the second-best performer by up to 1.64dB on the Test-L404 dataset and 1.09dB on the Test-E70 dataset. Meanwhile, our SSIM is higher than the compared

Table 2 The full-reference evaluation results of different methods on Test-L404 dataset

Method	MSE($\times 10^3$) ↓	PSNR(dB) ↑	SSIM ↑
HE [41]	2.6933	14.6054	0.6524
UDCP [14]	3.9568	13.3376	0.5513
ULAP [42]	1.7729	17.4179	0.7053
UWCNN [27]	1.5799	16.9803	0.6730
Water-Net [28]	1.4377	17.7641	0.7405
FUnIE-GAN [20]	1.0225	19.6370	0.7363
Ucolor [29]	0.6618	20.8115	0.8008
Peng et al. [39]	0.3758	23.4847	0.8154
MTUR [49]	0.6705	21.0606	0.7902
Ours	0.2417	25.1285	0.8388

Table 3 The full-reference evaluation results of different methods on Test-U90 dataset

Method	MSE($\times 10^3$) ↓	PSNR(dB) ↑	SSIM ↑
HE [41]	2.0908	15.5063	0.7007
UDCP [14]	4.7928	11.8497	0.5112
ULAP [42]	2.5496	15.4050	0.6741
UWCNN [27]	2.6461	14.5801	0.6011
Water-Net [28]	2.1040	15.7125	0.7005
FUnIE-GAN [20]	1.6437	16.9642	0.6778
Ucolor [29]	0.5536	21.5755	0.8094
Peng et al. [39]	0.5609	21.6873	0.7994
MTUR [49]	0.3278	23.7699	0.8285
Ours	0.3999	23.0965	0.8224

Table 4 The full-reference evaluation results of different methods on Test-E70 dataset

Method	MSE($\times 10^3$) ↓	PSNR(dB) ↑	SSIM ↑
HE [41]	2.8499	14.1618	0.6270
UDCP [14]	3.6967	13.2795	0.5506
ULAP [42]	1.0532	18.8865	0.7196
UWCNN [27]	1.5361	17.2613	0.6683
Water-Net [28]	1.4375	17.6695	0.7265
FUnIE-GAN [20]	0.3669	23.1878	0.7756
Ucolor [29]	0.6897	20.2822	0.7744
Peng et al. [39]	0.4216	23.4952	0.7952
MTUR [49]	0.8433	19.4164	0.7615
Ours	0.3086	24.5821	0.8138

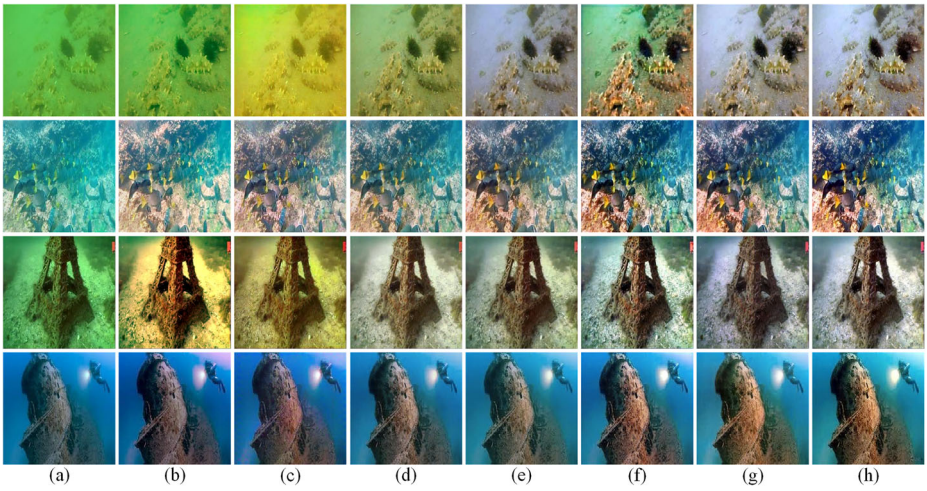

Fig. 4 Visual comparisons of different UIE methods on full-reference images. (a) Raw image, (b) ULAP [42], (c) FUnIE-GAN [20], (d) Ucolor [29], (e) Peng et al. [39], (f) MTUR [49], (g) Ours, (h) Reference image

Table 5 The non-reference evaluation results of different methods

Method	Test-U60		SQUID		Test-R45	
	UCIQE ↑	UIQM ↑	UCIQE ↑	UIQM ↑	UCIQE ↑	UIQM ↑
HE [41]	0.5742	2.1561	0.5560	1.3410	0.5534	2.1582
UDCP [14]	0.5375	1.4508	0.5589	0.9860	0.5509	2.0744
ULAP [42]	0.5424	1.6691	0.4594	0.8914	0.4928	2.4529
UWCNN [27]	0.4668	2.4243	0.4436	2.0590	0.4633	3.0221
Water-Net [28]	0.5305	2.4900	0.5456	2.4047	0.5245	3.0950
FUnIE-GAN [20]	0.5299	2.6497	0.4945	2.1357	0.5035	3.0800
Ucolor [29]	0.5323	2.6159	0.5138	2.1557	0.5226	3.1116
Peng et al. [39]	0.5359	2.5869	0.5278	2.1367	0.5311	3.0797
MTUR [49]	0.5844	2.7538	0.5763	2.3360	0.5558	3.1360
Ours	0.5749	2.7249	0.5774	2.3596	0.5555	3.1403

methods on both Test-L404 and Test-E70 datasets. As shown in Fig. 4, the color of underwater images enhanced by our method is closest to the reference images, and the image visual quality is the best. Some of the underwater images enhanced by ULAP [42] and FUnIE-GAN [20] are yellowish. The underwater images enhanced by Ucolor [29], Peng et al. [39] and MTUR [49] show a relatively good visual effect, but there are still some color casts.

4.3.2 Non-reference evaluation

The quantitative evaluation results of different methods on Test-U60, SQUID and Test-R45 datasets are shown in Table 5. We extract four categories of underwater images (low-illuminated, yellowish, greenish and bluish underwater images) from Test-U60, SQUID and Test-R45 datasets, and subjectively compare the visual quality of the images enhanced by different methods, as shown in Fig. 5.

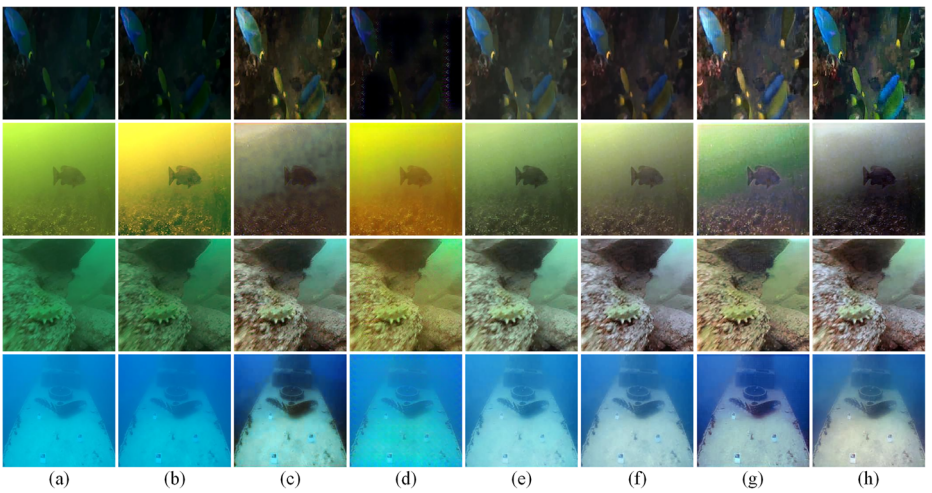

Fig. 5 Subjective comparisons of different UIE methods on greenish, bluish, yellowish and low-illuminated underwater images. (a) Raw image, (b) ULAP [42], (c) Water-Net [28], (d) FUnIE-GAN [20], (e) Ucolor [29], (f) Peng et al. [39], (g) MTUR [49], (h) Ours

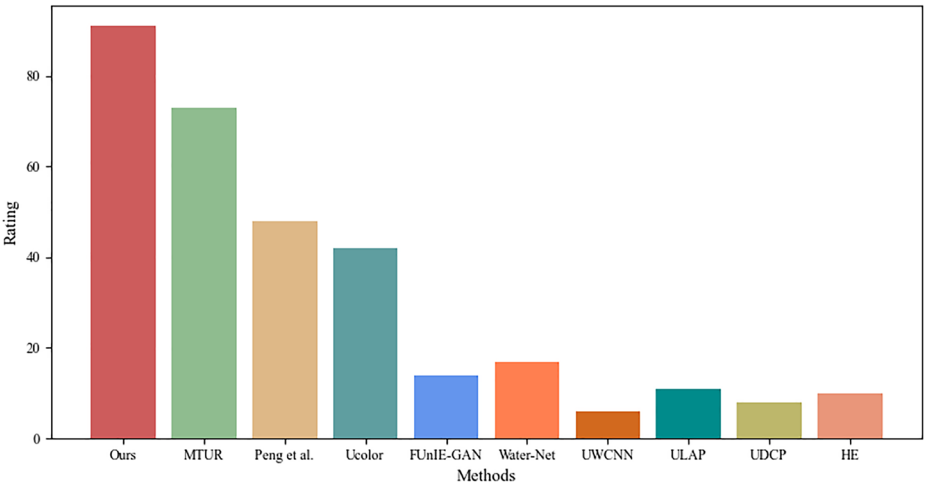

Fig. 6 The generated image equality evaluation results of different methods on SQUID

As in Table 5, our method achieves the highest UCIQE and UIQM score on SQUID, the highest UIQM score on Test-R45, and the second-best results on Test-U60. As shown in Fig. 5, the visual perception of underwater images is greatly affected by color deviation. ULAP [42] enhanced the sharpness of the greenish underwater image, but over-enhanced the yellowish image and under-enhanced the low-illumination and bluish underwater image. Water-Net [28] corrected the color deviation of the underwater image well, but the enhanced image were darker overall. FUnIE-GAN [20] improved the sharpness of the underwater images, but there was an obvious color cast in the enhanced greenish and yellowish underwater images. Ucolor [29], Peng et al. [39] and MTUR [49] improved the visual quality of underwater images, but there were still some color casts. By contrast, our method effectively removed the color cast of greenish, bluish, and yellowish underwater images and improved the brightness of low-illumination underwater images. This demonstrates that our method has good performance in color cast removal and sharpness enhancement.

To further verify the effect of our method and avoid the influence of our subjective judgment on the visualization results, we prepared 160 pictures generated by 10 methods (HE, UDCP, ULAP, UWCNN, WaterNet, FunieGAN, Ucolor, Peng et al., MTUR, and Ours) on SQUID, and then invited 20 volunteers to compare the image in terms of chromatic aberration, sharpness, contrast, etc., and select the best image without knowing the corresponding method. The statistical results are shown in Fig. 6. As shown in this graph, we can observe that our method received the highest number of best ratings.

4.4 Discussion

HE [41] crudely modified the pixel values of underwater images using pixel value transformations, which improved the contrast but caused color casts due to the introduction of excessive red components, resulting in high UCIQE and UIQM scores, while the quality of the enhanced images was quite poor. UDCP [14] and ULAP [42] can improve the sharpness of greenish and bluish underwater images, while over-enhancing yellowish images as well as under-enhancing low-illumination images, probably because they rely on a fixed underwater imaging model and cannot be applied well in various scenarios. The results produced

by Water-Net [28] are visually darker overall, probably due to the introduction of a white balance channel in the enhancement process, which is not always dependable for underwater images. FUnIE-GAN [20] is a lightweight model that can process images quickly. However, the fewer parameters of the model make it easy to reach the bottleneck during learning features from complex underwater images, which may lead to color deviation in the enhanced results. The enhanced results of Ucolor [29] present a relatively high quality visually, but there are still some color casts, which may be due to the introduction of multi-color space in the network without full utilization. Peng et al. [39] achieved high results for both full-reference and non-reference evaluations, but the enhanced results are still visually slightly color biased, probably because the method uses multi-scale features but does not simultaneously implement local context capture and global information interaction. MTUR [49] achieved good performance on UIEBD, but the results on EUVP and LSUI were not as good, probably because the lightweight model design makes the method unable to handle multiple types of underwater images efficiently. By comparing the full-reference evaluation results on Test-L404, we can observe that our method improves 7.0% and 2.9% in PSNR and SSIM metrics compared to the second-best method. The non-reference evaluation results on SQUID and Test-R45 show a slight advantage of our method compared to the second-best method. From the subjective comparison shown in Fig. 5, we can observe that the enhanced results of our method have the best visual quality and the least color deviation. These demonstrate the superiority of our method compared to other state-of-the-art methods. Furthermore, in terms of computational complexity, our model costs about 3.3M parameters which are 78% of the FUnIE-GAN. The testing time of 71 FPS for the images (size of 256×256) indicates that our method outperforms several state-of-the-art deep learning-based methods in processing efficiency while maintaining superior performance.

4.5 Ablation study

To demonstrate the effect of skip connections in our model and the effectiveness of the loss function for joint RGB and HSV color spaces, we conduct ablation studies with Test-L404 and Test-U90 datasets.

4.5.1 Effect of skip connections

By removing or retaining skip connections, we explore the influence of skip connections on the performance of the proposed model. As in Table 6, the performance of the model without skip connections is degraded. It can also be seen from Fig. 7 that the quality of the image enhanced by the model without skip connections is poor.

Table 6 Image quality assessment of different ablation models

Model	Test-L404		Test-U90	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Without skip connections	22.6201	0.5632	21.2869	0.5407
Without $Loss_{hsv}$ for training	24.7371	0.8352	22.8459	0.8186
Our full model	25.1285	0.8388	23.0965	0.8224

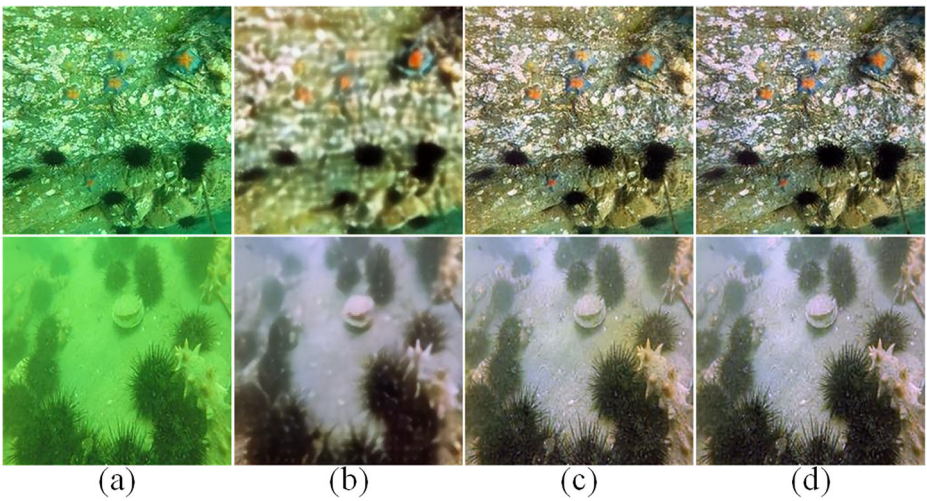

Fig. 7 Visual comparisons of model without skip connections or loss function in HSV color space for training. (a) Raw image, (b) Without skip connections, (c) Without loss function in HSV color space for training, (d) Our full model

4.5.2 Effect of loss function for joint RGB and HSV color spaces

By training model with/without loss function in HSV color space $Loss_{hsv}$, we explore the impact of loss function for joint RGB and HSV color spaces on the performance of the proposed model. As in Table 6, training with loss function for joint RGB and HSV color spaces further improves the quality of the underwater images. As shown in Fig. 7, the image enhanced by model training without loss function in HSV color space shows relatively good visual quality, but still suffers from a slight color cast.

5 Conclusion

In this paper, we propose an efficient Swin Transformer-based method for underwater image enhancement. The network of the proposed method is mainly composed of encoder, decoder and skip connections, where the encoder and decoder take Swin Transformer blocks as the basic unit, and skip connections are used to fuse multi-scale features from the encoder and decoder. The local attention mechanism of the Swin Transformer makes it easier to learn the local details of underwater images. Long-range dependency modeling with the shifted window scheme by Swin Transformer enables efficient explicit global information interaction. Extensive experiments on two real-world underwater image datasets demonstrate the superiority of our method for underwater image enhancement, especially in color cast removal and sharpness enhancement. There are also some limitations of our model. First, since the number of heads and some other parameters in the Swin Transformer block need to be set according to the image size before training, the trained model can only satisfy the input with a specific size and cannot process images with arbitrary size. Second, it is difficult to enhance underwater images in real-time because of the relatively large number of parameters and complicated computation of the model. Therefore, we will aim to transform the

standard Swin Transformer block and appropriately reduce the redundant parameters of the model for further improvement.

Acknowledgements This work was supported by the Key Research and Development Project of Hainan Province (No. ZDYF2019024).

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare no conflict of interest.

References

1. Berman D, Levy D, Avidan S, Treibitz T (2021) Underwater single image color restoration using haze-lines and a new quantitative dataset. IEEE Trans Pattern Anal Mach Intell 43(8):2822–2837
2. Bhatti UA, Huang M, Wu D, Zhang Y, Mehmood A, Han H (2019) Recommendation system using feature extraction pattern recognition in clinical care systems. Enterp Inf Syst 13(3):329–351
3. Bhatti UA et al (2020) Geometric algebra applications in geospatial artificial intelligence and remote sensing image processing. IEEE Access 8:155783–155796
4. Bhatti UA et al (2021) Advanced color edge detection using clifford algebra in satellite images. IEEE Photonics J 13(2):1–20
5. Bhatti UA et al (2022) Local similarity-based spatial–spectral fusion hyperspectral image classification with deep CNN and gabor filtering. IEEE Trans Geosci Remote Sens 60:1–15
6. Cao J et al (2021) Video super-resolution transformer. arXiv:2106.06847
7. Carion N et al (2020) End-to-end object detection with transformers. In: Eur conf comput vis. Springer, Cham, pp 213–229
8. Chen C-FR, Fan Q, Panda R (2021) Crossvit: cross-attention multi-scale vision transformer for image classification. In: IEEE int conf comput vis, pp 357–366
9. Chen D-J, Hsieh H-Y, Liu T-L (2021) Adaptive image transformer for one-shot object detection. In: IEEE conf comput vis pattern recognit, pp 12242–12251
10. Chen H et al (2021) Pre-trained image processing transformer. In: IEEE conf comput vis pattern recognit, pp 12299–12310
11. Chen L et al (2021) Perceptual underwater image enhancement with deep learning and physical priors. IEEE Trans Circuits Syst Video Technol 31(8):3078–3092
12. Dai Z, Cai B, Lin Y, Chen J (2021) UP-DETR: unsupervised pre-training for object detection with transformers. In: IEEE conf comput vis pattern recognit, pp 1601–1610
13. Dosovitskiy A et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. In: Int conf learn represent
14. Drews PJ, Do Nascimento E, Moraes F, Botelho S, Campos M (2013) Transmission estimation in underwater single images. In: IEEE int conf comput vis workshops, pp 825–830
15. Gao S-B, Zhang M, Zhao Q, Zhang X-S, Li Y-J (2019) Underwater image enhancement using adaptive retinal mechanisms. IEEE Trans Image Process 28(11):5580–5595
16. Guo Y, Li H, Zhuang P (2020) Underwater image enhancement using a multiscale dense generative adversarial network. IEEE J Oceanic Eng 45(3):862–870
17. He K, Sun J, Tang X (2011) Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell 33(12):2341–2353
18. Hore A, Ziou D (2010) Image quality metrics: PSNR vs. SSIM. In: Int conf pattern recognit, pp 2366–2369
19. Hu J, Jiang Q, Cong R, Gao W, Shao F (2021) Two-branch deep neural network for underwater image enhancement in HSV color space. IEEE Signal Process Lett 28:2152–2156
20. Islam MJ, Xia Y, Sattar J (2020) Fast underwater image enhancement for improved visual perception. IEEE Rob Autom Lett 5(2):3227–3234
21. Jaffe JS (1990) Computer modeling and the design of optimal underwater imaging systems. IEEE J Oceanic Eng 15(2):101–111
22. Johnson J, Alahi A, Li F (2016) Perceptual losses for real-time style transfer and super-resolution. In: Eur conf comput vis. Springer, Cham, pp 694–711

23. Korhonen J, You J (2012) Peak signal-to-noise ratio revisited: is simple beautiful? In: 2012 Fourth international workshop on quality of multimedia experience (QoMEx), pp 37–38
24. Lanchantin J, Wang T, Ordonez V, Qi Y (2021) General multi-label image classification with transformers. In: IEEE conf comput vis pattern recognit, pp 16473–16483
25. Li Y, Chen R (2021) UDA-Net: densely attention network for underwater image enhancement. IET Image Proc 15(3):774–785
26. Li H, Zhuang P (2021) Dewaternet: a fusion adversarial real underwater image enhancement network. Signal Process Image Commun, vol 95(116248)
27. Li C, Anwar S, Porikli F (2020) Underwater scene prior inspired deep underwater image and enhancement. Video Pattern recognit, vol 98(107038)
28. Li C et al (2020) An underwater image enhancement benchmark dataset and beyond. IEEE Trans Image Process 29:4376–4389
29. Li C, Anwar S, Hou J, Cong R, Guo C, Ren W (2021) Underwater image enhancement via medium transmission-guided multi-color space embedding. IEEE Trans Image Process 30:4985–5000
30. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) Swinir: image restoration using swin transformer. In: IEEE int conf comput vis, pp 1833–1844
31. Liu P, Wang G, Qi H, Zhang C, Zheng H, Yu Z (2019) Underwater image enhancement with a deep residual framework. IEEE Access 7:94614–94629
32. Liu R, Fan X, Zhu M, Hou M, Luo Z (2020) Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light. IEEE Trans Circuits Syst Video Technol 30(12):4861–4875
33. Liu X, Gao Z, Chen BM (2020) MLFCGAN: multilevel feature fusion-based conditional GAN for underwater image color correction. IEEE Geosci Remote Sens Lett 17(9):1488–1492
34. Liu Z et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE int conf comput vis, pp 10012–10022
35. Mao J et al (2021) Voxel transformer for 3D object detection. In: IEEE int conf comput vis, pp 3144–3153
36. Misra I, Girdhar R, Joulin A (2021) An end-to-end transformer model for 3D object detection. In: IEEE int conf comput vis, pp 2906–2917
37. Moghimi MK, Mohanna F (2021) Real-time underwater image enhancement: a systematic review. J Real-Time Image Process 18(5):1509–1525
38. Panetta K, Gao C, Agaian S (2016) Human-visual-system-inspired underwater image quality measures. IEEE J Oceanic Eng 41(3):541–551
39. Peng L, Zhu C, Bian L (2021) U-shape transformer for underwater image enhancement. arXiv:2111.11843
40. Sajid U, Chen X, Sajid H, Kim T, Wang G (2021) Audio-visual transformer based crowd counting. In: IEEE int conf comput vis workshops, pp 2249–2259
41. Singhai J, Rawat P (2007) Image enhancement method for underwater, ground and satellite images using brightness preserving histogram equalization with maximum entropy. In: IEEE int conf comput intell multimed appl, pp 507–512
42. Song W et al (2018) A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In: Pacific rim conf multimed. Springer, Cham, pp 678–688
43. Song W, Wang Y, Huang D, Liotta A, Perra C (2020) Enhancement of underwater images with statistical model of background light and optimization of transmission map. IEEE Trans Broadcast 66(1):153–169
44. Srinivas A, Lin T-Y, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck transformers for visual recognition. In: IEEE conf comput vis pattern recognit, pp 16519–16529
45. Touvron H et al (2021) Training data-efficient image transformers & distillation through attention. In: Int conf mach learn, pp 10347–10357
46. Vaswani A et al (2017) Attention is all you need. In: Adv neural inf process syst, pp 5998–6008
47. Wang J et al (2020) CA-GAN: class-condition attention GAN for underwater image enhancement. IEEE Access 8:130719–130728
48. Wang Y et al (2021) End-to-end video instance segmentation with transformers. In: IEEE conf comput vis pattern recognit, pp 8737–8746
49. Yan K et al (2022) Medium transmission map matters for learning to restore real-world underwater images. Appl Sci 12(11):5420
50. Yang M, Sowmya A (2015) An underwater color image quality evaluation metric. IEEE Trans Image Process 24(12):6062–6071
51. Yang M, Hu J, Li C, Rohde G, Du Y, Hu K (2019) An in-depth survey of underwater image enhancement and restoration. IEEE Access 7:123638–123657
52. Yu H, Li X, Lou Q, Lei C, Liu Z (2020) Underwater image enhancement based on DCP and depth transmission map. Multimed Tools Appl 79:20373–20390

53. Zhang Z, Lu X, Cao G, Yang Y, Jiao L, Liu F (2021) ViT-YOLO: transformer-based YOLO for object detection. In: IEEE int conf comput vis, pp 2799–2808
54. Zhang W et al (2021) Enhancing underwater image via color correction and bi-interval contrast enhancement. Signal Process Image Commun, vol 90(116030)
55. Zhao H, Gallo O, Frosio I, Kautz J (2017) Loss functions for image restoration with neural networks. IEEE Trans Comput Imaging 3(1):47–57
56. Zheng S et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: IEEE conf comput vis pattern recognit, pp 6877–6886
57. Zhuang P, Ding X (2020) Correction to: underwater image enhancement using an edge-preserving filtering Retinex algorithm. Multimed Tools Appl 79(25):17257–17277
58. Zhuang P, Li C, Wu J (2021) Bayesian retinex underwater image enhancement. Eng Appl Artif Intell, vol 101(104171)