Journal of *Marine Science and Engineering*

MDPI

Article

A Two-Stage Network Based on Transformer and Physical Model for Single Underwater Image Enhancement

Yuhao Zhang, Dujing Chen, Yanyan Zhang *, Meiling Shen and Weiyu Zhao

School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 211544, China

* Correspondence: 002243@nuist.edu.cn; Tel.: +86-18795877673

Abstract: The absorption and scattering properties of water can cause various distortions in underwater images, which limit the ability to investigate underwater resources. In this paper, we propose a two-stage network called WaterFormer to address this issue using deep learning and an underwater physical imaging model. The first stage of WaterFormer uses the Soft Reconstruction Network (SRN) to reconstruct underwater images based on the Jaffe–McGramery model, while the second stage uses the Hard Enhancement Network (HEN) to estimate the global residual between the original image and the reconstructed result to further enhance the images. To capture long dependencies between pixels, we designed the encoder and decoder of WaterFormer using the Transformer structure. Additionally, we propose the Locally Intended Multiple Layer Perceptron (LIMP) to help the network process local information more effectively, considering the significance of adjacent pixels in enhancing distorted underwater images. We also proposed the Channel-Wise Self-Attention module (CSA) to help the network learn more details of the distorted underwater images by considering the correlated and different distortions in RGB channels. To overcome the drawbacks of physical underwater image enhancement (UIE) methods, where extra errors are introduced when estimating multiple physical parameters separately, we proposed the Joint Parameter Estimation method (JPE). In this method, we integrated multiple parameters in the Jaffe–McGramery model into one joint parameter (JP) through a special mathematical transform, which allowed for physical reconstruction based on the joint parameter (JP). Our experimental results show that WaterFormer can effectively restore the color and texture details of underwater images in various underwater scenes with stable performance.

Keywords: image enhancement; transformer; underwater imaging model

Citation: Zhang, Y.; Chen, D.; Zhang, Y.; Shen, M.; Zhao, W. A Two-Stage Network Based on Transformer and Physical Model for Single Underwater Image Enhancement. *J. Mar. Sci. Eng.* **2023**, *11*, 787. https://doi.org/10.3390/jmse11040787

Academic Editor: Rafael Morales

Received: 7 March 2023
Revised: 30 March 2023
Accepted: 3 April 2023
Published: 5 April 2023

1. Introduction

Underwater images often exhibit distorted colors, blurred details, and low contrast due to the complex underwater environments. These visual impairments are mainly caused by absorption and scattering in water. Specifically, when the sunlight enters the water, red light is absorbed first, resulting in underwater images appearing bluish and greenish. Additionally, scattering, including forward scattering and background scattering, also affects the visual quality of underwater images. Forward scattering makes underwater images blurry, while background scattering make them hazy. To mitigate these adverse effects, it is necessary to develop new methods for enhancing the visual quality of underwater images.

Existing underwater image enhancement methods can be mainly summarized into three categories: non-physical methods, physical methods, and deep learning methods. Non-physical methods aim to modify pixel value to improve the visual quality. They can improve the contrast and color of underwater images, but the ignorance of the physical imaging process limits the quality of enhancement, making the enhanced images abnormal in some areas. Physical methods aim to establish a hypothetical physical imaging model and then estimate the key parameters to perform inversion of the mathematical formula.

However, the performance of the physical method is restricted to the complicated underwater environments in that hypothesis and prior knowledge do not always make sense in complicated underwater environments and it is challenging to estimate the multiple parameters accurately. A famous UIE physical model is the Jaffe–McGramery model [1], which divides the underwater optical imaging process into three components: direct transmission component, forward scattering component, and background scattering component. In Jaffe–McGramery model, the total energy E_T captured by the camera is defined as follows:

$$E_T = E_d + E_b + E_f \tag{1}$$

where E_d, E_f, and E_b represent the direct transmission component, the forward scattering component, and the background scattering component, respectively. Since the average distance between the underwater scene and the camera is usually large, E_f can usually be ignored. So, only E_d and E_b are considered. After a series of mathematical transformation [1], the Jaffe–McGramery model can be expressed as follows:

$$\begin{aligned} I(x,y) &= J(x,y)e^{-cd(x,y)} + B(x,y)(1 - e^{-cd(x,y)}) \\ &= J(x,y)t(x,y) + B(x,y)(1 - t(x,y)) \end{aligned} \tag{2}$$

where $I(x,y)$ denotes the distorted underwater image, $J(x,y)$ denotes the clear image, c denotes attenuation coefficient, $t(x,y)$ denotes the transmission map, $B(x,y)$ denotes the background light, and (x,y) denotes the coordinates of the pixels in the images.

Recently, researchers have applied deep learning methods in UIE tasks. Existing deep learning methods are mostly CNN and GAN networks. Experiments and previous research [2] have demonstrated that CNN-based models are poor in capturing global information of underwater images due to the fixed receptive field of the convolution kernel, while pixels in the whole image may be related when learning the degrading features. GAN-based models tend to introduce artifacts into the enhanced images and to train a GAN-based model is difficult and unstable [3,4]. Existing deep learning networks are mainly one-stage. Additionally, they seldom pay more attention to the channel-wise degradation features. We found only a one-stage network is insufficient to obtain the desired results. Moreover, channel-wise information is highly significant in UIE tasks because the degradation in the RGB channels is different and correlated, which is an important reference for UIE tasks.

To overcome the problems mentioned above, we propose WaterFormer, a two-stage network based on deep learning and a physical imaging model, the Jaffe–McGramery model. We did not design our model based on CNN or GAN but integrated Transformer architecture into our model. Moreover, we introduced the joint parameter estimation method (JPE) to avoid extra errors when estimating the multiple parameters in the Jaffe–McGramery model separately. Furthermore, we combined *SSIM* loss and L_2 loss to guide our model to learn the detailed features of the underwater images.

The contributions of our paper can be concluded as follows:

1. We proposed a two-stage network, composed of a Soft Reconstruction Network (SRN) and a Hard Enhancement Network (HEN). SRN performs reconstruction via the Jaffe–McGramery model, in which the parameters are estimated through our proposed joint parameter estimation method (JPE). HEN further enhances the images by estimating the global residual.
2. We utilized the Transformer structure to leverage its potential for capturing long-range dependencies. Moreover, to better leverage local information and channel-wise information in underwater images, we propose two novel modules: Locally Intended Multiple layer Perception (LIMP) and the Channel-Wise Self-Attention module (CSA).
3. We introduced a task-oriented loss function for our model, which combines the L_2 loss and *SSIM* loss. By jointly optimizing the L_2 and *SSIM* losses, our model can better capture both the structural and texture details.

2. Related Works

2.1. UIE Methods

There are three main categories of UIE methods [3]: non-physical methods, physical methods, and deep learning methods.

Non-physical methods: Non-physical methods modify the pixel value of the underwater images to achieve the enhancement. Hitam et al. [5] used contrast adjustment and adaptive histogram equalization methods in RGB and HSV space to enhance the contrast of underwater images. Ancuti et al. [6] applied white balance and global contrast adjustments to enhance underwater images. Fu et al. [7] proposed a Retinex-based method. Chen et al. [8] noted that better results can be achieved when utilizing multi-frame reconstruction methods. E. Quevedo Gutiérrez et al. [9] proposed a fusion scheme based on a multi-camera environment. Non-physical methods have the advantage of improving the contrast and saturation of the distorted underwater images at a relatively low computational cost, but the underwater imaging process is influenced by physical factors such as light conditions, temperature, and even the turbidity of the water. Non-physical methods take no account of these physical factors. So, it performs poorly on real underwater images with complex underwater conditions.

Physical methods: They perform enhancement with the following steps: (1) establishing a hypothetical physical model and assuming some conditions according to prior experiences; (2) estimating parameters in the model via mathematical methods; (3) reversing the degradation process according to the hypothetical physical model mathematically to enhance performance. Chiang et al. [10] recovered underwater images by combining DCP with a wavelength compensation algorithm. Drews Jr. et al. [11] proposed the underwater dark channel prior algorithm UDCP, based on the introduction of red channel attenuation in underwater images. Carlevaris Bianca et al. [12] considered the attenuation difference prior, between RGB channels, to predict the transmission characteristics of the underwater scene. Physical methods have limitations. Underwater conditions are complex and various, so prior experiences and hypothesis cannot make sense everywhere. Furthermore, existing physical methods estimate the multiple parameters of the physical model separately, thus introducing extra errors.

Deep learning methods: Deep learning has been widely applied in UIE tasks in recent years. By designing an end-to-end network and using data-driven methods, the model can learn the characteristics of underwater images effectively. Li et al. [4] designed a benchmark model Water-Net and proposed a dataset UIEB. Li et al. [13] proposed UWCNN, which is based on CNN. Li et al. [14] proposed a GAN-based underwater image enhancement model WaterGAN. Guo et al. [15] proposed DenseGAN which can utilize multi-scale information in underwater images. Sun et al. [16] proposed UMGAN, in which the feedback mechanism and a noise reduction network are designed to optimize the generator and address the issue of noise and artifacts in GAN-produced images. Existing deep learning methods rarely take physical factors into consideration and are mainly one-stage networks, so they perform poorly in face of various underwater images due to lacking generalization ability.

2.2. Summary

Our proposed network aims at enhancing underwater images by combining physical model and deep learning methods and properly accounting for characteristics of underwater images both space-wise and channel-wise.

3. Proposed Method

The proposed WaterFormer network comprises two stages, as illustrated in Figure 1. The first stage, termed soft reconstruction-network (SRN), reconstructs the images through the Jaffe–McGramery model, where the two key parameters, $t(x, y)$ and $B(x, y)$, are estimated through the joint parameter estimation method (JPE), which can avoid extra errors that can arise from the separate estimation. The outputs of SRN are subsequently fed into the second stage, termed Hard Enhancement Network (HEN), where the images are

further enhanced through the estimation of the global residual. During the whole process, the Channel-Wise Self-Attention module (CSA) pays attention on channel-wise information, and the WaterFormer Block embedded with the Locally Intended Multiple Layer Perceptron (LIMP) learns the features of the distorted underwater images. Additionally, we incorporate the structural similarity index (*SSIM*) loss into our loss function to ensure the consistency of image structure and texture with the desired outputs.

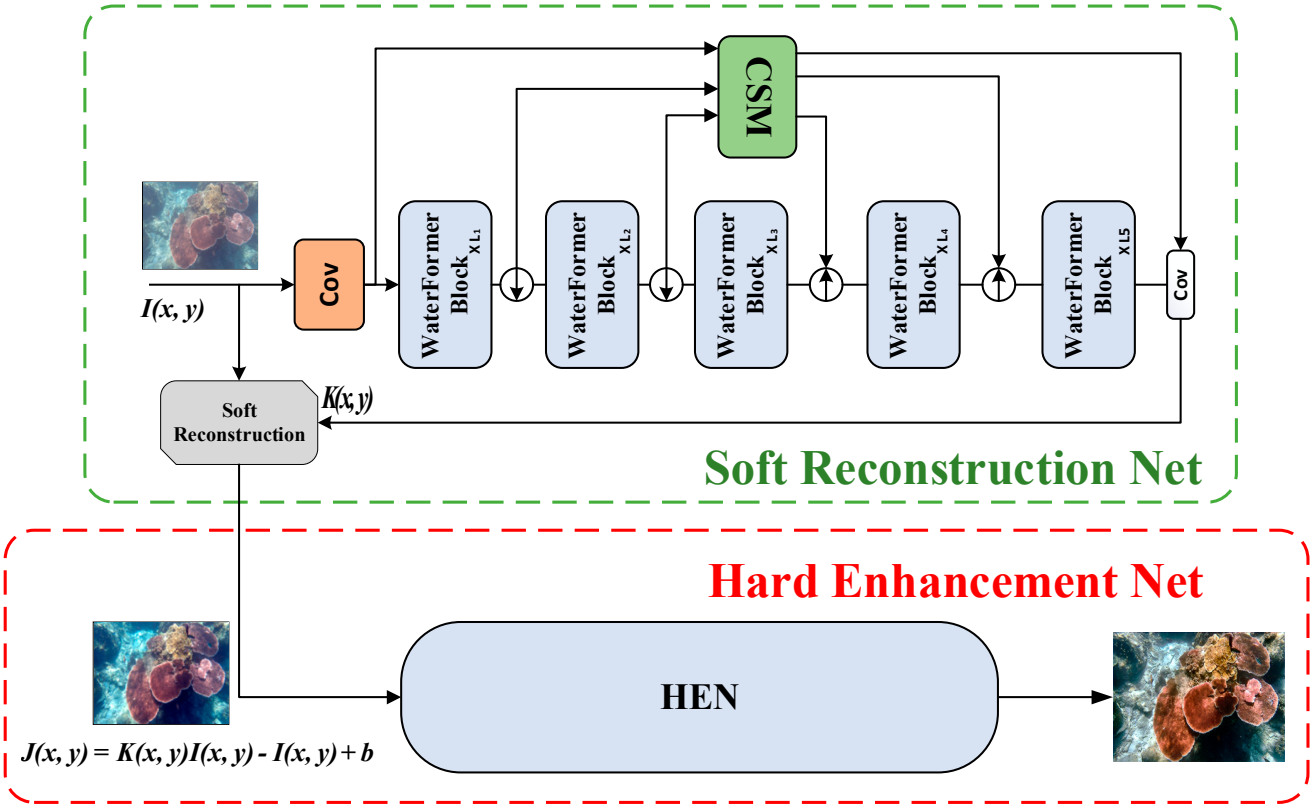


Figure 1. The overall structure of WaterFormer.

In the subsequent sections, we elaborate on SRN, HEN, CSA, WaterFormer Block with LIMP, and the task-oriented loss function.

3.1. Network Architecture

3.1.1. Soft Reconstruction Network

SRN reconstructs underwater images based on the Jaffe–McGramery model and deep learning. Given a distorted underwater image, SRN estimates the parameters in the Jaffe–McGramery model via JPE, and then performs reconstruction. Given a picture of (512, 512, and 3), the shape of data in each stage can be shown in Table 1.

Table 1. Data flow of SRN.

Id	Layer Names	Input Size	Output Size
1	Conv	(512, 512, 3)	(512, 512, 24)
2	WaterFormer Block $\times$ L_1	(512, 512, 24)	(256, 256, 48)
3	WaterFormer Block $\times$ L_2	(256, 256, 48)	(128, 128, 96)
4	WaterFormer Block $\times$ L_3	(128, 128, 96)	(256, 256, 48)
5	WaterFormer Block $\times$ L_4	(256, 256, 48)	(512, 512, 24)
6	Conv	(512, 512, 24)	(512, 512, 3)

At the end of SRN, we perform reconstruction via the proposed joint parameter estimation method (JPE). In the Jaffe–McGramery model, an underwater image is determined by two key parameters $t(x,y)$ and $B(x,y)$. Generally, existing estimation methods tend to estimate $t(x,y)$ and $B(x,y)$ separately. However, the estimation of $t(x,y)$ relies on a depth map $d(x,y)$ according to Equation (2) and the estimation of $B(x,y)$ depends on certain experiences and prior knowledge. Since experience and prior knowledge cannot make sense everywhere, errors are introduced when estimating $t(x,y)$ and $B(x,y)$ separately. To address this problem, we reformulated Equation (2) to integrate $t(x,y)$ and $B(x,y)$ into a joint parameter, $K(x,y)$. The reformulation can be shown as follows:

Specifically, Equation (2) can be expressed as:

$$\begin{aligned} J(x,y) &= \frac{I(x,y)-B(x,y)}{t(x,y)} + B(x,y) \\ &= \frac{1}{t(x,y)}(I(x,y) - B(x,y)) + (B(x,y) - b) + b \\ &= \frac{\frac{1}{t(x,y)}(I(x,y)-B(x,y))+(B(x,y)-b)}{I(x,y)-1}(I(x,y) - 1) + b \\ &= K(x,y)(I(x,y) - 1) + b \\ &= K(x,y)I(x,y) - K(x,y) + b \end{aligned} \tag{3}$$

where $K(x,y)$ can be expressed as:

$$K(x,y) = \frac{\frac{1}{t(x,y)}(I(x,y) - B(x,y)) + (B(x,y) - b)}{I(x,y) - 1} \tag{4}$$

The process above can realize joint parameter estimation (JPE). SRN estimates joint parameter $K(x,y)$ via JPE and then perform reconstruction according to Equation (4). The whole process of SRN can be depicted in Figure 2.

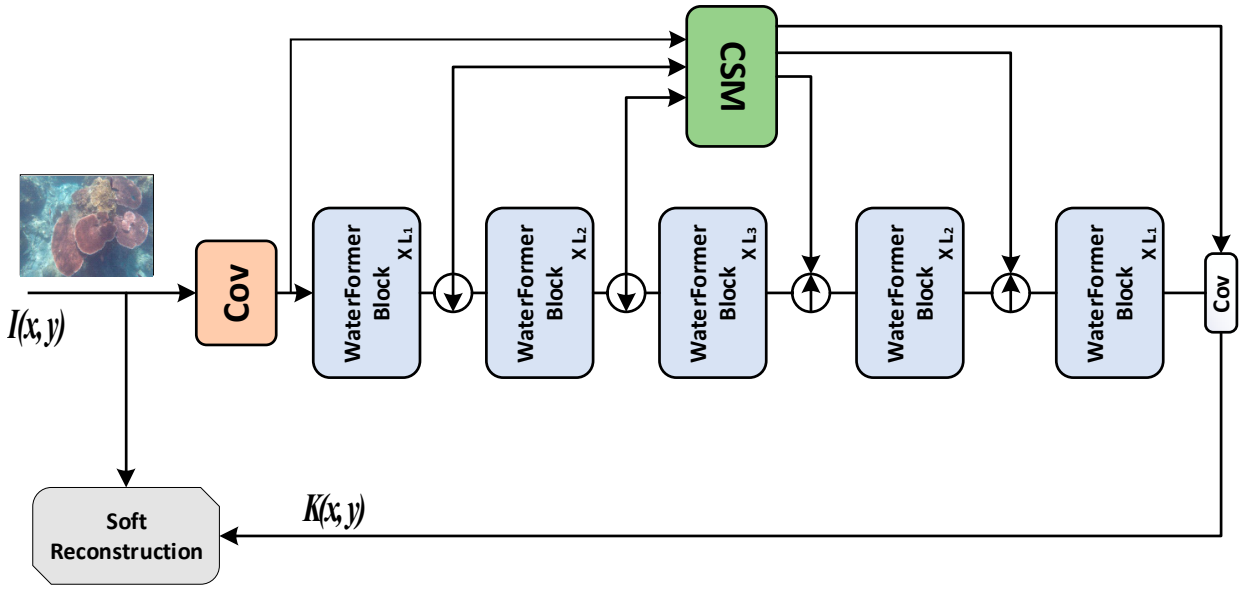


Figure 2. The overall structure of SRN.

According to Equation (5), the joint parameter is a combination of $t(x,y)$ and $B(x,y)$. We randomly selected four images and calculated their corresponding joint parameter map K. We visualize them in Figure 3.

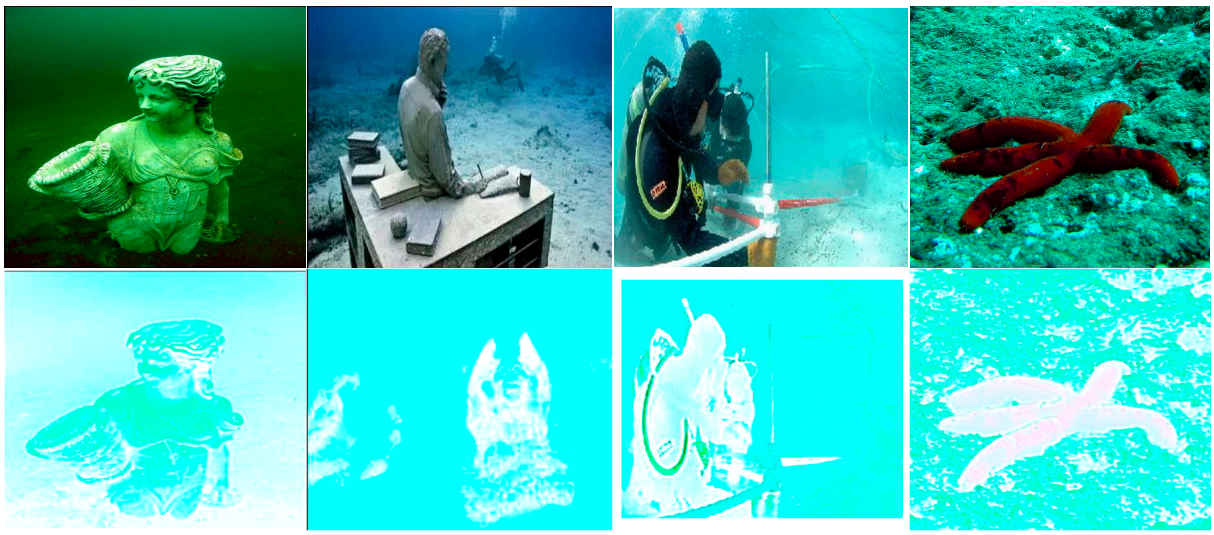


Figure 3. Visualization of the joint parameter map.

3.1.2. Hard Enhancement Network

The architecture of HEN is shown in Figure 4. HEN has a similar structure as SRN. HEN estimates the global residual $R(x, y)$ between $J(x, y)$ and $S(x, y)$. After obtaining the global residual $R(x, y)$, HEN performs enhancement via the following formula:

$$S(x, y) = J(x, y) + R(x, y) \tag{5}$$

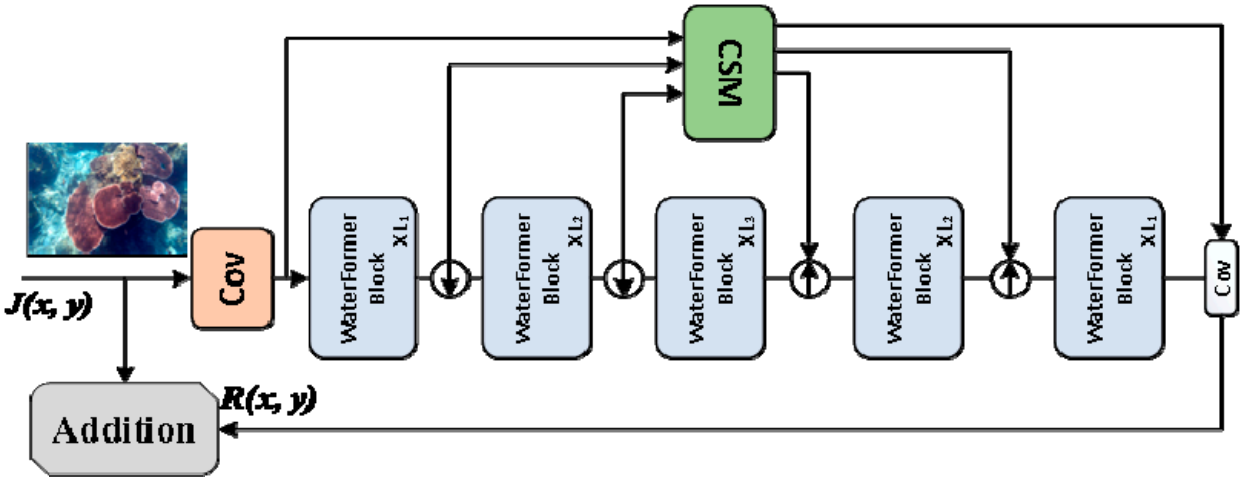
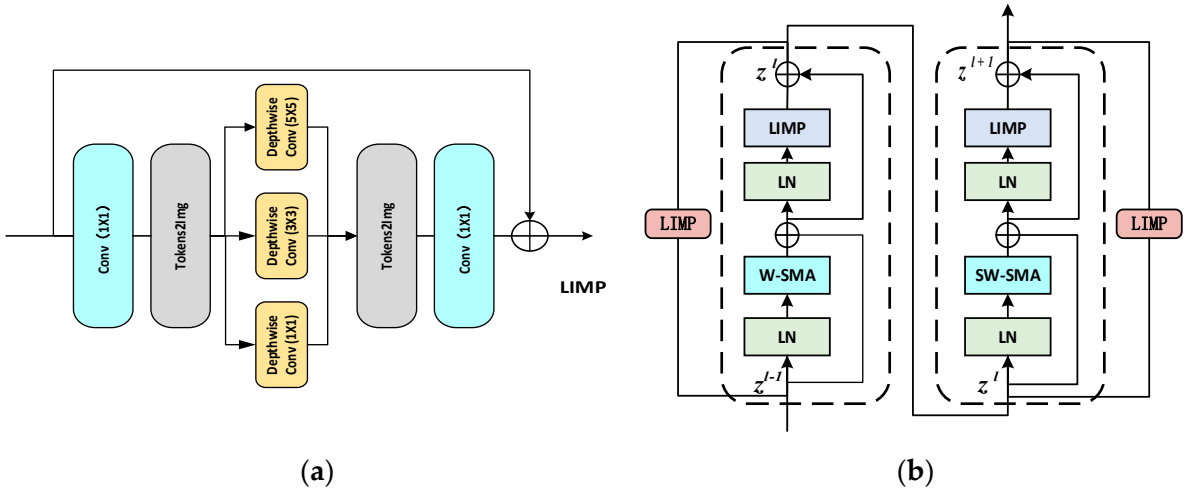


Figure 4. The overall structure of HEN.

3.1.3. WaterFormer Block

The architecture of the WaterFormer Block is shown in Figure 5b.

To cut down the computational cost of self-attention operation, we introduced shifted window self-attention (SWSA), which was applied in SwinTransformer [17]. SWSA divides an image into several patches and computes self-attention operation within each patch. Then, the shifted window scheme is used to fuse the results of each patch. It can be proven that SWSA can compute self-attention at a linear cost [17].



(**a**) (**b**)

Figure 5. Structure in WaterFormer block. (**a**) Structure of LIMP module, (**b**) structure of the WaterFormer Block.

Previous research found that the MLP layer in Transformer has a limited ability to learn the local context [18]. However, UIE images are characteristic of having similar distortion features in adjacent areas, which means the lack of gathering local information can cause the model to perform badly. Therefore, it is necessary to add local attention to our model considering that adjacent pixels are an important reference for reconstructing a distorted underwater image. We substituted the regular MLP in the Transformer structure for our proposed LIMP and added LIMP as a parallel module to WaterFormer Block. As shown in Figure 5, LIMP relies on three different sizes of convolution kernel (1×1, 3×3 and 5×5) to capture local information.

3.1.4. Channel-Wise Self-Attention Module

As color attenuation in the RGB channel varies and is correlated [1], we designed a Channel-Wise Self-Attention module (CSA) to replace the normal skip connection in the regular U-shape architecture. CSA performs Channel-Wise Self-Attention operation, which can pay attention to the more severely attenuated color channels, thereby compensating for the distortion imbalance between RGB channels. The structure of CSA is shown in Figure 6.

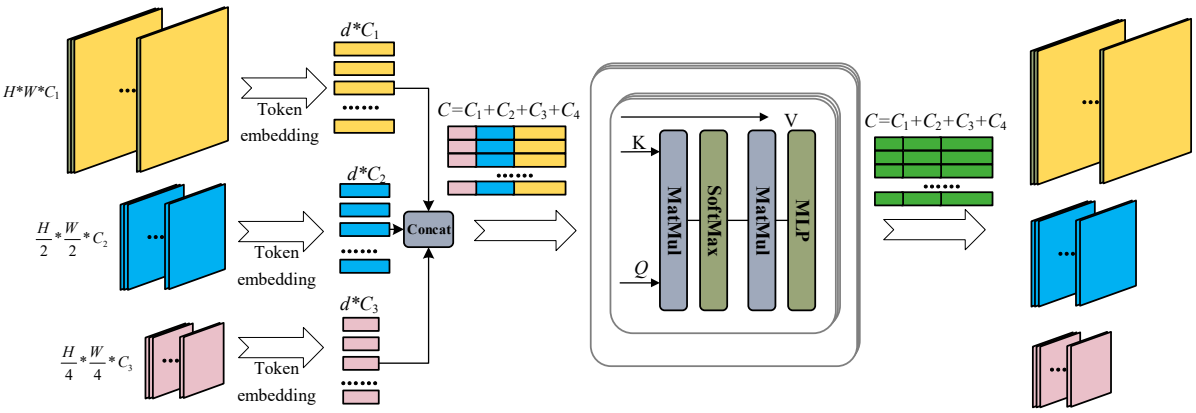


Figure 6. Channel-Wise Self-Attention module.

The inputs of CSA are the feature maps $F_i \in \mathbb{R}^{C_i \times \frac{H_i}{2^i} \times \frac{W_i}{2^i}}$ with different scales in the encoding stage. After we embedded the feature maps to tokens via a linear projection, we obtain three token sequences $S_i \in \mathbb{R}^{d \times C_i} (i = 1, 2, 3)$, where $d = HW$. Similarly, $Q \in \mathbb{R}^{d \times C_i} (i = 1, 2, 3)$, $K \in \mathbb{R}^{d \times C}$, and $V \in \mathbb{R}^{d \times C}$ can be obtained by

$$\begin{aligned} Q &= S \cdot W_Q \\ K &= S \cdot W_K \\ V &= S \cdot W_V \end{aligned} \tag{6}$$

where $W_Q \in \mathbb{R}^{d \times C}$, $W_K \in \mathbb{R}^{d \times C}$, and $W_V \in \mathbb{R}^{d \times C}$ represent learnable parameter matrices. S is generated by concatenating $S_i \in \mathbb{R}^{d \times C_i} (i = 1, 2, 3)$ channel-wisely. So, the output $O \in \mathbb{R}^{C \times d} (i = 1, 2, 3)$ can be obtained by:

$$O = SoftMax(LN(\frac{Q^T K}{\sqrt{C}}))V^T \tag{7}$$

Then, we transform O to feature maps via a linear mapping operation and attach them to outputs in each decoding each stage.

3.2. Loss Function

Loss function is an objective measurement between output $J(x, y)$ and ground truth $R(x, y)$. To obtain results with more details, we designed a loss function based on L_2 loss and *SSIM* loss.

L_2 loss function can be expressed as the following formula:

$$L_2(J(x), J^*(x)) = \frac{1}{N} \sum_{i=1}^{N} ||J(x) - J^*(x)||^2 \tag{8}$$

where N is the total number of images in the training set, and $||\cdot||$ stands for L_2 norm.

SSIM [19,20] is used to measure the similarity between two images. *SSIM* considers the following factors in a single picture: brightness, contrast, and structure contrast. The *SSIM* value between $J(x)$ and $J^*(x)$ can be expressed as:

$$SSIM(J(x), J^*(x)) = \frac{2\mu_{J(x)}\mu_{J^*(x)} + C_1}{{\mu_{J(x)}}^2 {\mu_{J^*(x)}}^2 + C_1} \cdot \frac{2\delta_{J(x)}\delta_{J^*(x)} + C_1}{{\delta_{J(x)}}^2 {\delta_{J^*(x)}}^2 + C_1} \tag{9}$$

where $C_1 = (K_1 + L)^2, C_2 = (K_2 + L)^2, K_1 = 0.01, L = 1$ and μ, δ represent the mean and standard deviation of an image, respectively. $\delta_{J(x)}\delta_{J^*(x)}$ is the covariance of a gray image. Then, the *SSIM* loss can be expressed as follows:

$$L_{SSIM}(J(x), J^*(x)) = 1 - \frac{1}{N} \sum_{i=1}^{N} SSIM(J(x), J^*(x)) \tag{10}$$

Therefore, the joint loss function used in this paper can be expressed as follows:

$$L_{total} = \omega_1 \cdot L_2(J(x), J^*(x)) + \omega_2 \cdot L_{SSIM}(J(x), J^*(x)) \tag{11}$$

where $\omega_1 = 0.7, \omega_2 = 0.3$, respectively.

4. Experiment and Results Analysis

4.1. Experimental Environment Configuration and Datasets Preparation

The experiments are performed on Pycharm. The hardware environments are listed as follows: Intel Xeon E5-2600 V3 processor (CPU), 32 G memory, NVIDIA GeForce GTX 1080 Ti (11 G) graphics processor (GPU), and the operating system is Ubuntu16.04. The

environment is Python version 3.7.4, and CUDA version 10.1. In addition, the experimental hyperparameter settings are shown in Table 2.

Table 2. Experimental hyper-parameter settings.

Hyperparameter	Parameter Setting
The sample size was trained	$256 \times 256 \times 3$
learning rate	0.0001
Batch size	16
Optimizer/momentum	Adam W/0.5

We used the UIEB dataset to verify the practical effect of WaterFormer on the UIE task. The UIEB dataset was first proposed by Li et al. [4] with their benchmark UIE network, WaterNet. The UIEB dataset contains 950 underwater images in multiple underwater scenes and with various underwater features with degradation (hazy, fog, low-contrast, and insufficient exposure). A total of 890 of them have corresponding high-quality reference images, and the remaining 60 images have no reference. We randomly selected 800 images in the UIEB dataset as the training set and the remaining 90 images as testing set.

4.2. Experimental Results and Analysis

We conducted an experiment on the UIEB dataset to compare the practical effect with other algorithms, including five traditional algorithms and four deep learning methods (CLAHE, Fusion [21], UWCNN [13], Water-Net [4], WaterGAN [14], CylcleGAN [22], etc.) to evaluate the practical effectiveness of WaterFormer.

4.2.1. Qualitative Evaluation

We randomly selected a few images in the UIEB dataset and perform enhancement via the different UIE methods mentioned above. Figure 7 shows the experiment results.

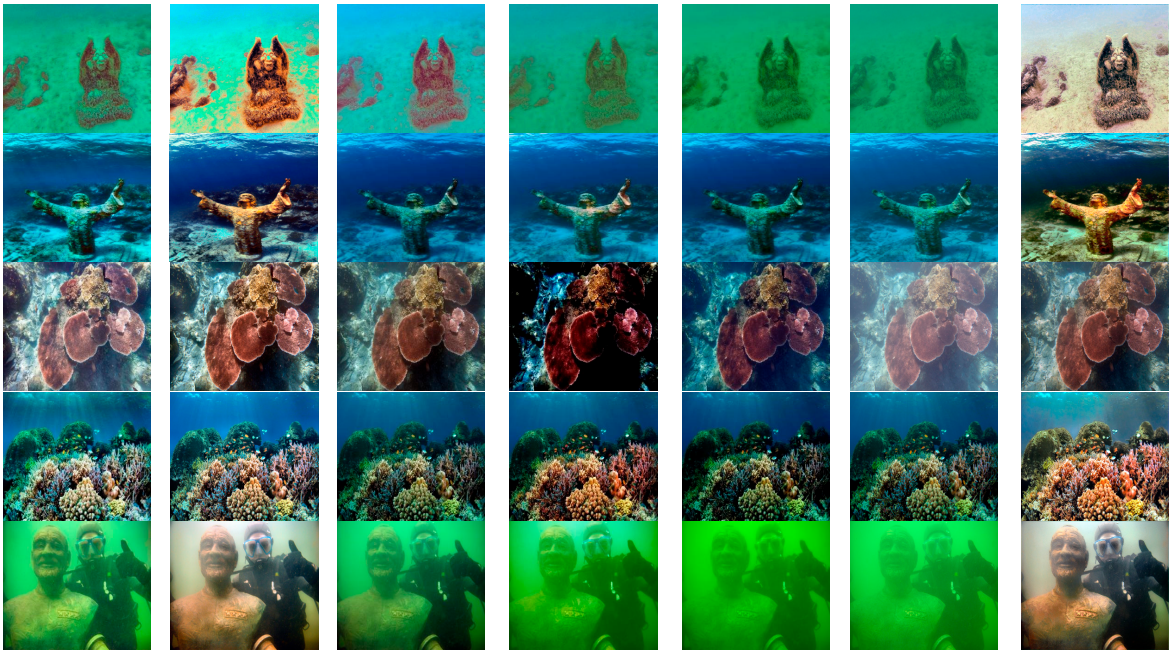


Figure 7. Comparison with traditional methods.

As can be seen from Figure 7, traditional UIE methods can improve the color and contrast to some extent, but they have limitations, especially in the greenish scene.

As shown in Figure 8, UWCNN can compensate for the red channel to some extent, but its ability to correct other channels is poor. Water-Net enhances underwater images through gating and fusion schemes, but it introduces noise and artifacts. WaterGAN and DenseGAN, which are GAN-based models, tend to overcompensate for the red channel and introduce artifacts. Compared with CNN-based model and GAN-based model, WaterFormer can restore underwater images with good visual quality and rich details.

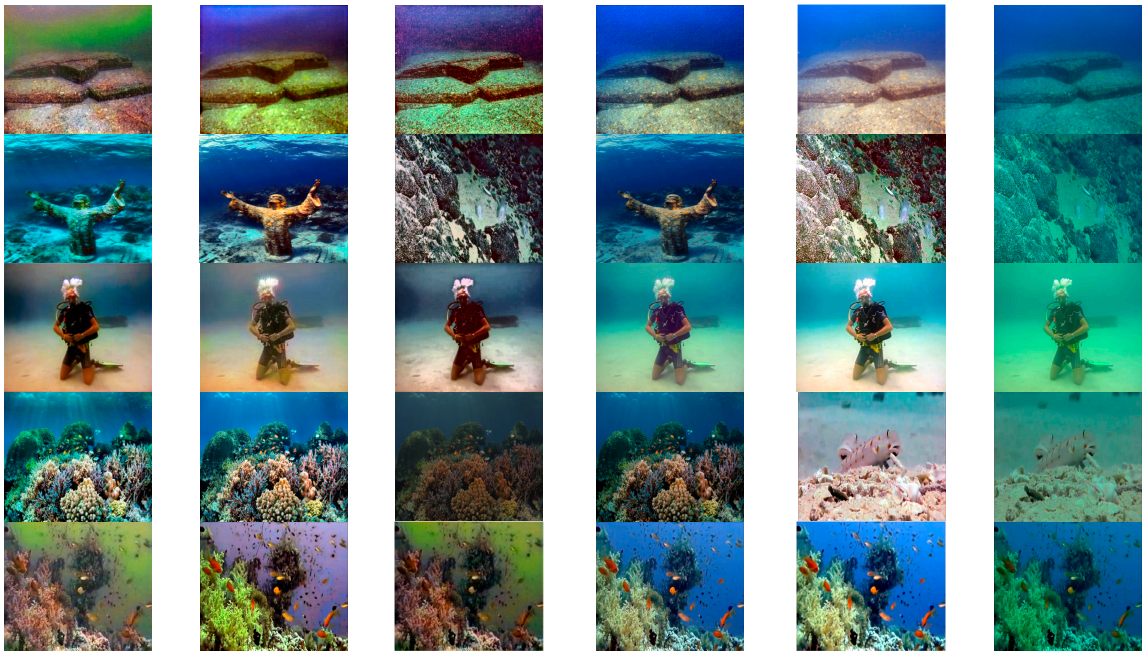


Figure 8. Comparison with deep learning methods.

4.2.2. Quantification Evaluation

In order to quantitatively analyze the performance of the algorithm on the underwater image enhancement, *PSNR* [20] and *UCIQE* [23] are selected as the quantitative measurements.

Given two images, f and g, the size of $M \times N$, the *PSNR* value between f and g is defined as follows:

$$PSNR(f,g) = 10\log_{10}(\frac{255^2}{MSE(f,g)}) \tag{12}$$

where *MSE* can be expressed as:

$$MSE(f,g) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(f_{ij} - g_{ij})^2 \tag{13}$$

UCIQE, a linear combination of color concentration, contrast, and saturation, evaluates the quality of a single degraded underwater image. The larger the value, the better the visual quality of the underwater image. According to [23], *UCIQE* can be expressed as follows:

$$UCIQE = c_1 \cdot \sigma_c + c_2 \cdot con_l + c_3 \cdot \mu_s \tag{14}$$

where σ is the standard deviation of the image, and it can represent the average of saturation.

The results of the quantitative experiment are shown in Table 3. The time cost of each model is also recorded. A total of 300 underwater images of different styles are randomly selected for quantitative analysis. Among all the methods, WaterFormer showed the best *PSNR*, *SSIM*, and *UCIQE* scores by 23.82%, 0.91, and 0.632, respectively with acceptable time cost.

Table 3. Quantitative comparison of the enhanced performance on the UIEB datasets.

Method	PSNR	SSIM	UCIQE	Time Cost (s)
CLAHE	16.67	0.66	0.567	0.0139
IBLA	16.88	0.63	0.611	28.12
Fusion	16.75	0.73	0.654	0.152
UWCNN	16.22	0.80	0.464	1.21
WaterNet	18.14	0.77	0.570	1.03
UWGAN	19.05	0.74	0.502	1.58
WaterGAN	16.85	0.62	0.603	1.67
CycleGAN	15.75	0.51	0.511	1.96
Ours	**23.82**	**0.91**	**0.632**	**1.57**

4.3. Ablation Experiments

To demonstrate the effectiveness of each component in WaterFormer, we conducted ablation experiments.

4.3.1. Two-Stage Structure

The results of the experiments are shown in Figure 9 and Table 4. We remove SRN and HEN separately and compare the results with that of the two-stage network. HEN significantly improves the color and contrast of the underwater images, but extra noise and artifacts are introduced due to the ignorance of physical imaging process. When the images are processed by these two networks together, color and contrast are significantly improved with little noise and are more in line with humans' visual sense.

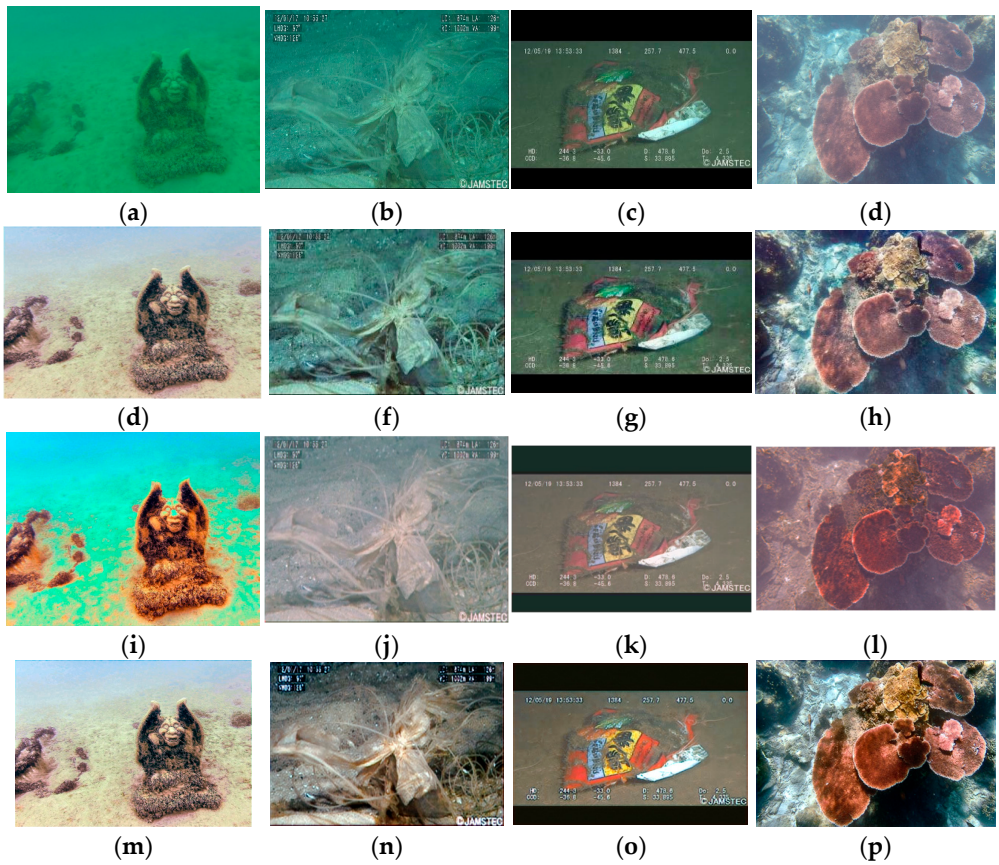


Figure 9. Results of ablation experiments on the two-stage network (**a**–**p**).

Table 4. Results of the ablation experiments on the two-stage network.

SRN	HEN	*PSNR*	*UCIQE*
√	-	18.89	0.596
-	√	21.96	0.611
√	√	**23.82**	**0.623**

4.3.2. Channel-Wise Self-Attention Module

To verify the effect of CSA, we replaced CSA with a simple skip connection. Experiment results show that *PSNR* and *UCIQE* decrease without CSA. Results of the ablation experiments on CSA are shown in Table 5.

Table 5. Results of the ablation experiments on CSA.

CSA	Simple Skip Connection	*PSNR*	*UCIQE*
-	√	22.91	0.601
√	-	**23.82**	**0.623**

4.3.3. SSIM Loss

We added *SSIM* loss to the total function to guide our model to learn the proper texture and structure of the desired images. To demonstrate the effectiveness of the *SSIM* component, we removed *SSIM* loss in the total loss function and compared their results. Experiment results in Table 6 show that *PSNR* and *UCIQE* decrease without *SSIM* loss.

Table 6. Verification of *SSIM* loss component.

SSIM Loss	L_2 Loss	*PSNR*	*UCIQE*
-	√	21.31	0.612
√	√	**23.82**	**0.623**

4.3.4. SWSA

To lower the computational cost of self-attention, WaterFormer is designed based on shifted windows self-attention (*SWSA*), where self-attention operations are performed within each window and the information of each window is fused through a shifting scheme. Through *SWSA*, we can perform self-attention operation in a linear computational cost. To be specific, given an image size of $h \times w$, the computational complexity of a standard multi-head self-attention operation (*MSA*) and a *SWSA* operation can be shown as follows:

$$\begin{aligned} Complexity(MSA) &= 4hwC^2 + 2(hw)^2C \\ Complexity(SWSA) &= 4hwC^2 + 2M^2hwC \end{aligned} \tag{15}$$

The complexity of *MSA* is quadratic to the size of an underwater image while the computational cost is linear to hw when *SWSA* is applied. To further verify the efficiency of the *SWSA* in WaterFormer, we replaced *SWSA* with standard multi-head self-attention module (*MSA*) and compared their computational time cost. The experimental results are shown in Table 7.

Table 7. Computational cost of different UIE methods.

MSA	*SWSA*	Time Cost (s)
√	-	6.25
-	√	1.57

5. Application

WaterFormer has a wide range of applications, one of which is underwater detection. When applied in the underwater robot, underwater detection becomes easier and more convenient. We applied WaterFormer to the Trash_ICRA19 dataset. Trash_ICRA19 was proposed by Fulton et al. [24], and it contains plastic (marine waste and all plastic materials), remote submersible (remote, remote submersible, and sensor), and biological (all natural biological materials, including fish, plants, and biological debris) debris. The enhancement results on the Trash_ICRA19 dataset are shown in Figure 10.

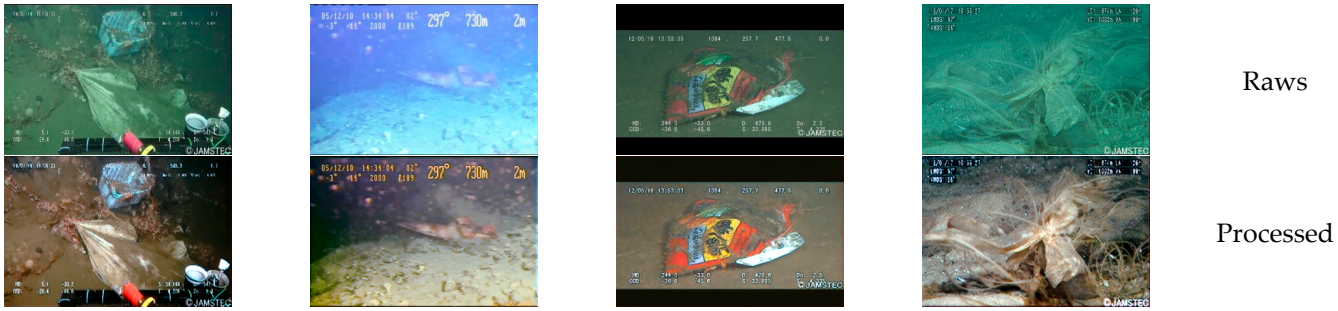


Figure 10. Enhancement on Trash_ICRA19.

6. Conclusions

WaterFormer is a two-stage network that utilizes deep learning and an underwater physical imaging model to enhance underwater images. Soft Reconstruction Network (SRN) reconstructs the underwater images based on the Jaffe–McGramery model. The joint parameter estimation method (JPE) is also proposed to reduce extra error when estimating multiple parameters in the Jaffe–McGramery model. Hard Enhancement Network (HEN) further enhances the images by estimating the global residual between the original image and the reconstructed results. The encoder and decoder of WaterFormer are based on the Transformer structure, which is designed to capture long dependencies between pixels. Additionally, the Locally Intended Multiple Layer Perceptron (LIMP) and Channel-Wise Self-Attention module (CSA) are proposed to effectively process local and channel-wise information, respectively. A task-oriented loss function with *SSIM* loss also adds to the enhancement effects. Experimental results demonstrate that WaterFormer performs well in restoring color and texture details in various underwater scenes.

Author Contributions: Conceptualization, Y.Z. (Yuhao Zhang) and Y.Z. (Yanyan Zhang); methodology, Y.Z. (Yuhao Zhang); software, Y.Z. (Yuhao Zhang); validation, Y.Z. (Yuhao Zhang), W.Z., Y.Z. (Yanyan Zhang) and D.C.; formal analysis, Y.Z. (Yanyan Zhang); investigation, Y.Z. (Yanyan Zhang); resources, Y.Z (Yanyan Zhang); data curation, Y.Z. (Yanyan Zhang); writing—original draft preparation, M.S.; writing—review and editing, Y.Z. (Yanyan Zhang); visualization, Y.Z. (Yanyan Zhang); supervision, Y.Z. (Yanyan Zhang); project administration, Y.Z. (Yanyan Zhang); funding acquisition, Y.Z. (Yanyan Zhang); All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in reference number [4,23].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jaffe, J.S. Underwater Optical Imaging: The Past, the Present, and the Prospects. *IEEE J. Ocean. Eng.* **2015**, *40*, 683–700. [CrossRef]
2. Peng, L.; Zhu, C.; Bian, L. U-shape Transformer for Underwater Image Enhancement. In *Computer Vision—ECCV 2022 Workshops*; Springer: Cham, Switzerland, 2023; pp. 290–307.
3. 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*, 4861–4875. [CrossRef]
4. Li, C.; Chunle, G.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Trans. Image Process.* **2019**, *29*, 4376–4389. [CrossRef] [PubMed]
5. Hitam, M.; Yussof, W.; Awalludin, E.; Bachok, Z. Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In Proceedings of the 2013 International conference on computer applications technology (ICCAT), Sousse, Tunisia, 20–22 January 2013; pp. 1–5.
6. Ancuti, C.; Codruta, A.; Haber, T.; Bekaert, P. Enhancing Underwater Images and Videos by Fusion. In Proceedings of the 2012 IEEE conference on computer vision and pattern recognition, Providence, RI, USA, 16–21 June 2012; pp. 81–88.
7. Fu, X.; Zhuang, P.; Huang, Y.; Liao, Y.; Zhang, X.P.; Ding, X. A retinex-based enhancing approach for single underwater image. In Proceedings of the 2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, 27–30 October 2014; pp. 4572–4576. [CrossRef]
8. Chen, Y.; Li, W.; Xia, M.; Li, Q.; Yang, K. Super-resolution reconstruction for underwater imaging. *Opt. Appl.* **2011**, *41*, 841–853.
9. Quevedo Gutiérrez, E.; Delory, E.; Marrero Callico, G.; Tobajas, F.; Sarmiento, R. Underwater video enhancement using multi-camera super-resolution. *Opt. Commun.* **2017**, *404*, 94–102. [CrossRef]
10. Chiang, J.Y.; Chen, Y.-C. Underwater Image Enhancement by Wavelength Compensation and Dehazing. *IEEE Trans. Image Process.* **2012**, *21*, 1756–1769. [CrossRef] [PubMed]
11. Drews, P., Jr.; Nascimento, E.; Botelho, S.; Campos, M. Underwater Depth Estimation and Image Restoration Based on Single Images. *IEEE Comput. Graph. Appl.* **2016**, *36*, 24–35. [CrossRef] [PubMed]
12. Carlevaris-Bianco, N.; Mohan, A.; Eustice, R. Initial Results in Underwater Single Image Dehazing. In Proceedings of the Oceans 2010 Mts/IEEE Seattle, Seattle, WA, USA, 20–23 September 2010; pp. 1–8.
13. Li, C.; Anwar, S. Underwater Scene Prior Inspired Deep Underwater Image and Video Enhancement. *Pattern Recognit.* **2019**, *98*, 107038. [CrossRef]
14. Li, J.; Skinner, K.; Eustice, R.; Johnson-Roberson, M. WaterGAN: Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394. [CrossRef]
15. Guo, Y.; Li, H.; Zhuang, P. Underwater Image Enhancement Using a Multiscale Dense Generative Adversarial Network. *IEEE J. Ocean. Eng.* **2019**, *45*, 862–870. [CrossRef]
16. Sun, B.; Mei, Y.; Yan, N.; Chen, Y. UMGAN: Underwater Image Enhancement Network for Unpaired Image-to-Image Translation. *J. Mar. Sci. Eng.* **2023**, *11*, 447. [CrossRef]
17. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
18. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
19. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
20. Horé, A.; Ziou, D. Image quality metrics: *PSNR* vs. *SSIM*. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369.
21. Chen, Q.; Zhang, Z.; Li, G. Underwater Image Enhancement Based on Color Balance and Multi-Scale Fusion. *IEEE Photonics J.* **2022**, *14*, 1–10. [CrossRef]
22. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv* **2017**, arXiv:1703.10593.
23. Yang, M.; Sowmya, A. An Underwater Color Image Quality Evaluation Metric. *IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc.* **2015**, *24*, 6062–6071. [CrossRef] [PubMed]
24. Fulton, M.; Hong, J.; Islam, M.; Sattar, J. Robotic Detection of Marine Litter Using Deep Visual Detection Models. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019.