Incorporating Triple Attention and Multi-scale Pyramid Network for Underwater Image Enhancement

Kaichuan Sun
Ocean College, Jiangsu University of Science and Technology, China
kcsun991@gmail.com

Yubo Tian
School of Information and Communication Engineering, Guangzhou Maritime University, China
tianyubo@just.edu.cn

Abstract: *Clear images are a prerequisite of high-level underwater vision tasks, but images captured underwater are often degraded due to absorption and scattering of light. To solve this issue, traditional methods have shown some success, but often generate unwanted artifacts for knowledge priori dependency. In contrast, learning-based approaches can produce more refined results. Most popular methods are based on an encoder-decoder configuration for simply learning the nonlinear transformation of input and output images, so their ability to capture details is limited. In addition, the significant pixel-level features and multi-scale features are often overlooked. Accordingly, we propose a novel and efficient network that incorporates triple attention and a multi-scale pyramid with an encoder-decoder architecture. Specifically, a triple attention module that captures the channel-pixel-spatial features is used as the transformation of the encoder-decoder module to focus on the fog region; then, a multi-scale pyramid module designed for refining the preliminary defog results are used to improve the restoration visibility. Extensive experiments on the EUVP and UFO-120 datasets corroborate that the proposed method outperforms the state-of-the-art methods in quantitative metrics Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), Patch-based Contrast Quality Index (PCQI) and visual quality.*

Keywords: *Underwater image enhancement, attention mechanism, multi-scale pyramid network, encoder-decoder.*

Received November 22, 2021; accepted December 15, 2022
https://doi.org/10.34028/iajit/20/3/11

1. Introduction

In recent years, underwater robot vision systems have been widely used in the engineering applications such as submarine resource detection, marine species protection, submarine detection for cables and debris [6, 18]. However, due to the effects of physical factors such as light propagation attenuation and underwater media, the raw underwater images captured by the vision system are blurred, contrast decreased and color deviated. This causes gravely negative impacts on the vision tasks of underwater robots including object detection and recognition, object tracking, and scene understanding [4, 23]. Therefore, it is of great value to obtain clear and effective underwater images.

Existing image enhancement methods are mainly divided into traditional and deep learning-based. Traditional methods commonly depend on models proposed by McGlamery [14] and Jaffe [10]. Their studies have demonstrated that the factors affecting imaging are mainly divided into three components, namely direct component, forward scattering, and back scattering. As shown in Figure 1, the left is the absorption of light in an underwater environment, different colors of light are absorbed differently underwater; the right is the imaging process. Although the traditional model-based method can often produce high-visibility images, it is easy to generate the phenomenon of artifacts and overly enhance contrast.

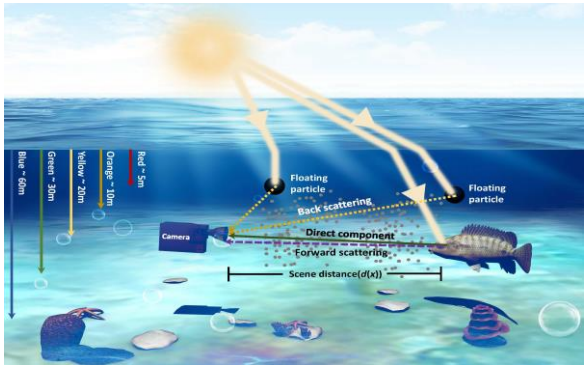


Figure 1. Schematic diagram of underwater optical imaging.

Unlike traditional approaches, deep learning-based methods learn the nonlinear mapping between degraded and clear images. Due to the great performance of the Convolutional Neural Network (CNN) on a mass of training data [34], the focus of underwater image enhancement research is increasingly shifting from artificial features to features based on deep learning. Encoder-decoder network is a very common and effective model framework in computer vision tasks [15, 24] by imitating a process of human cognition. Specifically, this framework has achieved amazing performance in image enhancement and restoration, from which numerous subsequent works are derived [9,

11]. Under the framework of the encoder-decoder, the attention mechanism can obtain more detailed information of the target to be concerned by focusing on the target region, so as to suppress other useless information. With the in-depth study of attention mechanisms by researchers, many works have proposed various forms of attention mechanisms to improve the problems faced by visual tasks, where the channel and spatial attention are used most. However, this attention only considers spatial and channel information, and lacks an in-depth consideration of local details which is important for image enhancement.

Inspired by the work of [21, 28, 29, 30], in this paper we propose a novel Triple Attention Module (TAM), in which channel, spatial, and pixel attention are integrated in a parallel feature fusion way, so as to leverage the features of local details to enhance the representation ability of the network. In addition, multi-scale features are often used to enhance the multi-level feature representation of visual tasks, and in [31, 33] it was verified that these features enhance the non-linear representation capability of the network and improve the performance. In this paper, we further incorporate a Multi-scale Pyramid Module (MPM) into the encoder-decoder and attention module to enhance the processing effect in regions with uneven haze.

To summarize, our contributions are described as follows:

- We propose a novel TAM, which is capable of mining deep hybrid channel, pixel, and spatial attention features. At the same time, we compared three different fusion strategies for the TAM, and the experimental results of quantitative and qualitative show that parallel triple attention is more effective in two datasets compared to other forms.
- We propose an enhancement network incorporating triple attention and multi-scale pyramid modules, called ITAMPN, which embeds the TAM into a simple encoder-decoder architecture to extract effective detail features. This is then combined with the multi-scale pyramid module to enhance the representation ability at different information scales.
- We achieve state-of-the-art results on the EUVP and UFO-120 datasets, and conduct ablation studies to further demonstrate the effect of components.

2. Related Work

Traditional model-based methods mainly use underwater image statistical priors to provide additional constraints for compensating the information loss of images, which are still often used as the auxiliary means to further improve the performance [5, 22]. However, the current works mainly focus on deep learning-based methods. The end-to-end enhancement models learn nonlinear mappings from training samples and have achieved more performance than traditional artificial feature extraction methods. Due to the complexity of the input underwater images, researchers have proposed various optimized neural network models to improve the information representation abilities of these neural networks. Uplavikar *et al*. [24] proposed a novel model capable of learning the context features of the image and handling the diversity of water during the enhancement process. Li *et al*. [12] proposed an underwater image enhancement network (WaterNet), which fused white balance, histogram equalization, and gamma correction algorithms. Meanwhile, they also released an underwater image dataset, which spurred further development in this field. Islam *et al*. [9] proposed a conditional Generative Adversarial Network (GAN)-based model which could be trained on paired and unaired images, and a new underwater image dataset (EUVP) was also presented. When observing an image, humans can quickly focus on the area of interest and ignore other useless information.

The attention mechanism in deep learning draws on the attention thinking mode of human beings, and has been widely applied in various types of visual tasks such as object detection, image classification, and image enhancement, and achieved remarkable results. Mnih *et al*. [16] proposed the addition of an attention mechanism into the traditional recurrent neural network to focus on important locations, and the resulting network significantly outperformed a convolutional architecture on an object classification task. Woo *et al*. [28] proposed a convolutional block attention module, which inferred the attention map along the channel and spatial directions and then multiplied it by the input feature map for adaptive feature refinement. More recently, Yin *et al*. [29] proposed a novel parallel spatial/channel-wise attention block and integrated it into a complex encoder-decoder module to achieve good dehazing results. Different from the parallel forms of the channel and spatial attention, Qin *et al*. [21] proposed a feature fusion attention network where the fusion channel and pixel attention were implemented in a serial manner, so more weight was placed on important features and the representational capabilities were expanded. Although these attention mechanisms have achieved excellent results in the computer vision domain, they do not leverage the fusion of multiple attention mechanisms fully, which is particularly important for image enhancement tasks. Inspired by the above works, in this study the TAM module is proposed to capture the responses of hybrid channel-pixel-spatial features, and connected with a form of parallel feature fusion. Meanwhile, we fuse a multi-scale pyramid module into the back of the decoding module for better feature representation capabilities.

3. Proposed Method

In this section, the proposed ITAMPN will be presented in detail. We first introduce the overall framework of

our model, then the sub-modules are introduced respectively, and then the loss function is described.

3.1. Overall Architecture

Generally, attention modules can adopt multiple networks as its backbone network, such as ResNet [25], GAN [24], and so on. Due to its excellent performance in feature representation, we employ an encoder-decoder network architecture as our backbone network. As shown in Figure 2, we first encode the given data into a feature map through the encoder. Then, the features are transformed through the TAMs to learn more details from the enhanced features. Subsequently, the enhanced features are decoded by the decoder, and an MPM is integrated with the last layer to achieve feature fusion on multiple levels.

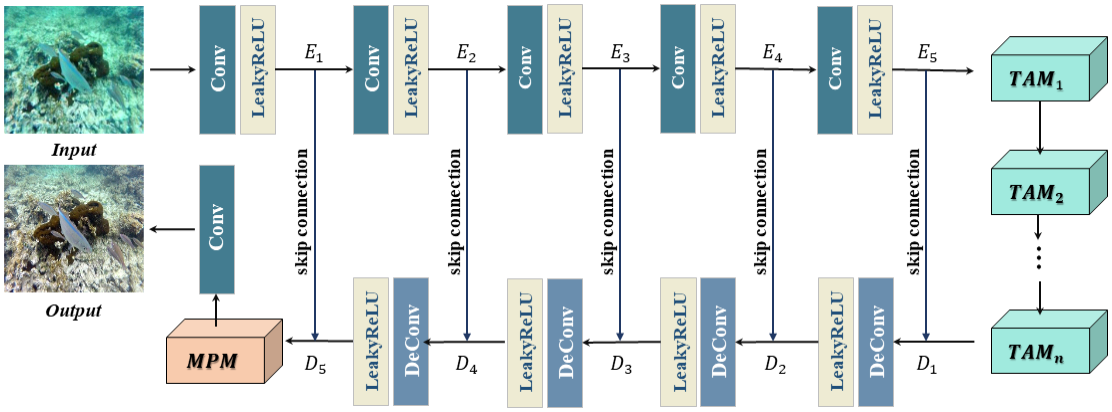


Figure 2. Overview of the proposed ITAMPN network.

3.2. Encoder-Decoder Architecture

Inspired by recent works of neural networks successfully applied in image dehazing and denoising [15, 29], we adopt an encoder-decoder structure as the backbone architecture. Although encoder-decoder networks are used in many enhancement tasks, their forms can vary significantly. The encoder first generates a series of feature maps E_i={E_1, E_2, E_3, E_4, E_5} through a set of convolution and Leaky ReLU operations. Then, the TAM blocks focus on the regions with haze in the feature map and reconstruct the feature map. In the subsequent decoding stage, after deconvolution through the decode layer the feature map is added to the corresponding map of the encoding layer. The output of the decode layer is D_j={D_2, D_3, D_4, D_5}, and provides local and global information and compensates for the loss of high-frequency components during the output process. The transformation process of the feature map is explained using Table 1, where each encoder layer reduces the feature map to a fixed size that is processed at twice the speed of the previous one, and the decoder returns it to the original feature map size via the reverse operation. As illustrated in Figure 2, the process of encoder-decoder module can be expressed as follows:

$$E_i = \Psi_{k,n}(E_{i-1}),\ where\ i=\{1,2,3,4,5\} \quad (1)$$

$$D_{j+1} = \Phi_{k,n}(D_j) + E_{5-j},\ where\ j=\{1,2,3,4\} \quad (2)$$

Where E_0 is the input image, D_1 denotes the feature map obtained from the TAM blocks. $\Psi_{k,n}(\cdot)$ and $\phi_{k,n}(\cdot)$ represent the *Leaky*ReLU(*Conv*(·)) and *Leaky*ReLU(*DeConv*(·)) operation sequence with kernel size k and stride n, respectively. The detailed parameters of the process are shown in Table 1.

Table 1. Architecture layout of our model. The blocks are expressed using the convolution filter size and stride.

Layers	Input size	Blocks	Output Size
$L(E_1)$	$H\times W\times 3$	9×9, 1	$H\times W\times 16$
$L(E_2)$	$H\times W\times 16$	3×3, 2	$H/2\times W/2\times 32$
$L(E_3)$	$H/2\times W/2\times 32$	3×3, 2	$H/4\times W/4\times 64$
$L(E_4)$	$H/4\times W/4\times 64$	3×3, 2	$H/8\times W/8\times 128$
$L(E_5)$	$H/8\times W/8\times 128$	3×3, 2	$H/16\times W/16\times 256$
$L(TAM)$	$H/16\times W/16\times 256$	–	$H/16\times W/16\times 256$
$L(D_2)$	$H/16\times W/16\times 256$	3×3, 2	$H/8\times W/8\times 128$
$L(D_3)$	$H/8\times W/8\times 128$	3×3, 2	$H/4\times W/4\times 64$
$L(D_4)$	$H/4\times W/4\times 64$	3×3, 2	$H/2\times W/2\times 32$
$L(D_5)$	$H/2\times W/2\times 32$	3×3, 2	$H\times W\times 16$
$L(MPM)$	$H\times W\times 16$	–	$H\times W\times 20$
$L(Conv)$	$H\times W\times 20$	3×3, 1	$H\times W\times 3$

3.3. Triple Attention Module

In underwater image enhancement tasks, the haze of the distorted image is very important, because it is uneven and covers the true colors of the raw image. Therefore, the effective capture of the feature information in the hazy regions is important to ensure the success of the deep learning network. To achieve this, a deep feature transformation module is usually added between the encoder and the decoder to perform abstract learning of the encoded features at different levels. Although deep networks have the ability to represent complex information, traditional implementations have the problem of gradient disappearance or expansion. The combination of feature attention and local residual learning proposed in [21] can solve this problem well.

However, when the simple serially connected channel and pixel attention are used to capture context features, the spatial feature are ignored to some extent, and this connection form does not maximize the effectiveness of this fusion strategy. In this study, we propose a TAM that connects by channel, pixel, and spatial attention, and fully exploits each attention mechanism to improve the performance of the network. Inspired by a variety of integrated forms of attention mechanisms proposed by researchers [21, 28, 29], three fusion strategies of triple attention were studied, as illustrated in Figure 4. As shown in Section 4, the experimental results show that among the three fusion strategies, the parallel feature fusion strategy achieves a better enhancement effect.

The parallel form of detailed triple attention is illustrated in Figure 3. Before joining the channel-pixel-spatial attention sub-module, a simple local residual block that consists of convolution, $\mathrm{Re}LU$ and a skip connection is added, which allows multiple local residual connections to bypass less important information and focus on essential elements. The feature map F_i is obtained after the residual block is fed into the channel, pixel, and spatial attention in parallel. Channel attention gives different weights to different channels to highlight important feature information. Assuming that the feature map F_i, the feature description of 1×1×c is obtained through global average pooling. The weight coefficients of each channel are obtained using the sigmoid function, and then the channel-oriented feature map $M(C_i)$ of TAM i is obtained by multiplying the weight coefficients with the original feature. Pixel attention is mainly used to deal with the uneven distribution of noise on different pixels, so that the network pays more attention to the noisy pixel areas. The pixel-oriented feature map $M(P_i)$ is obtained through the convolution and activation functions. Spatial attention pays more attention to the spatial location information of heterogeneous haze. An effective feature descriptor is formed by concatenating the Global Average Pooling (GAP) and the GMP, and then the spatial-oriented feature map $M(S_i)$ is obtained through convolution and activation operations. The operations described above can be expressed as follows:

$$F_i = Conv(\Gamma_{k,n}(X_i) \oplus X_i) \tag{3}$$

$$M(C_i) = \sigma(Conv(\Gamma_{k,n}(GAP(F_i)))) \otimes F_i \tag{4}$$

$$M(P_i) = \sigma(Conv(\Gamma_{k,n}(F_i))) \otimes F_i \tag{5}$$

$$M(S_i) = \sigma(Conv([GAP(F_i); GMP(F_i)])) \otimes F_i \tag{6}$$

Where F_i represents the feature map after the local residual block, $M(C_i)$, $M(P_i)$, $M(S_i)$ represent the feature maps obtained after passing through the channel, pixel, and spatial attention sub-modules respectively. X_i is the input feature of the i-th TAM, $\Gamma_{k,n}$ represents the $\mathrm{Re}LU(Conv(\cdot))$ operation sequence with kernel size k and stride n, σ represents the sigmoid activation function.

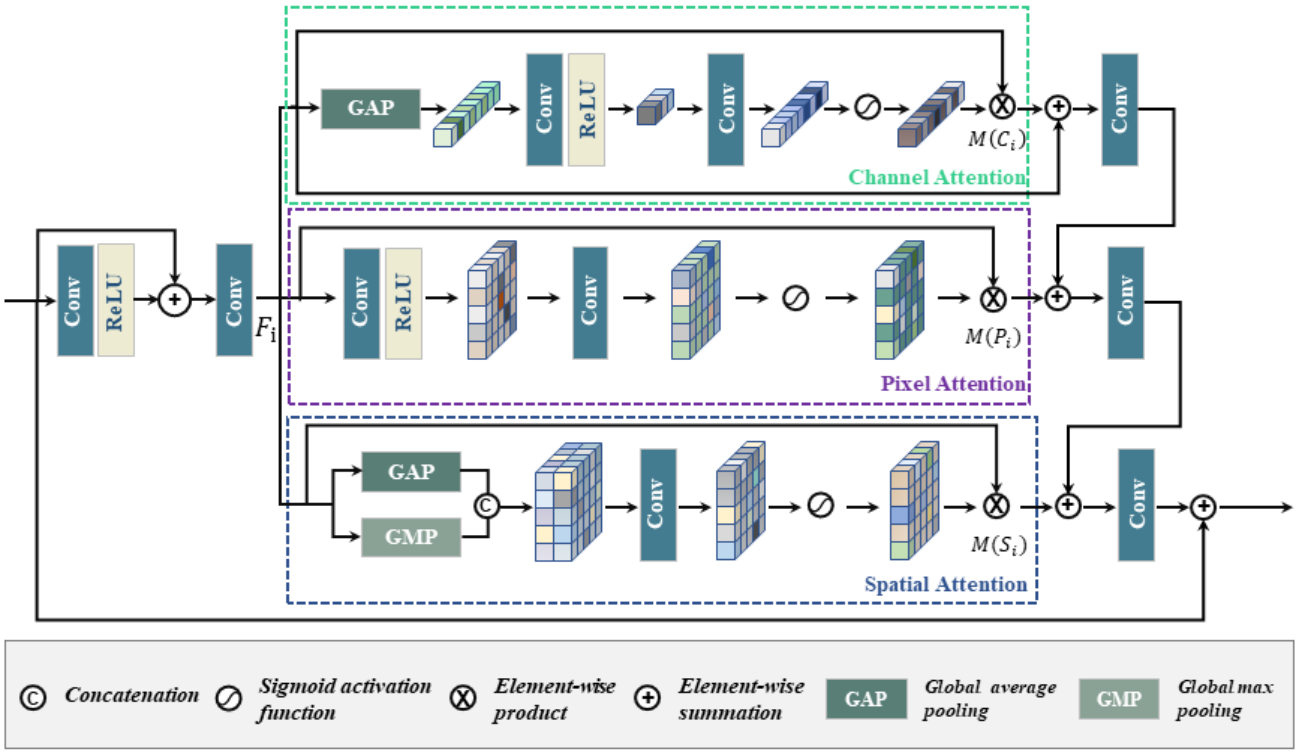
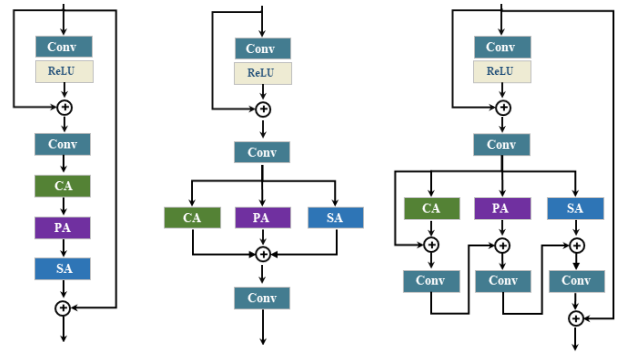


Figure 3. Illustration of TAM.

A parallel feature fusion strategy is adopted to fuse the acquired feature map. The formulation is shown in Equation (7). The effectiveness of this form verified in section 4.

$$Y_i = Conv(Conv(Conv(F_i \oplus M(C_i)) \oplus M(P_i)) \oplus M(S_i)) \oplus X_i \quad (7)$$



a) Serial feature fusion. b) Direct parallel feature fusion. c) Parallel feature fusion.

Figure 4. Three ways of integrating triple attention.

3.4. Multi-Scale Pyramid Module

The use of only the encoder-decoder structure and the TAM will result in a lack of global structural information for different scale objects. In our model, the global structural information is learned by the network through the addition of an MPM to the last layer of the decoding module. Inspired by the successful application of multi-scale pyramid network in neural networks [29, 31, 33], this module uses an average pooling operation to down-sample features in sizes of 4, 8, 16, and 32, and then the pooled feature information goes through the convolution layer, a $\mathrm{Re}LU$ function, and is finally up-sampled to the original feature size. Finally, it is connected with the D_5 output feature map to form fusion features, and a 3×3 convolution is used for size alignment, as shown in Figure 5. Through multi-scale feature fusion, the network learns the information of different perception domains, and integrates global and local information that offer it better representation abilities. The MPM is expressed as follows.

$$\begin{aligned} F_{in}^{i} &= Avgpool(F_{in}, 2^{i}) \ \ where\ i=\{1,2,3,4\} \\ F_{out} &= [F_{in}, \Omega(\Gamma_{k,n}(F_{in}^{1})), \Omega(\Gamma_{k,n}(F_{in}^{2})), \\ &\quad \Omega(\Gamma_{k,n}(F_{in}^{3})), \Omega(\Gamma_{k,n}(F_{in}^{4}))] \end{aligned} \quad (9)$$

where F_{in}^{i} represents the i-th feature map obtained after the average pooling operation $Avgpool(.,.)$,Ω represents the up-sampling operation, and F_{out} represents the output feature map from the MPM.

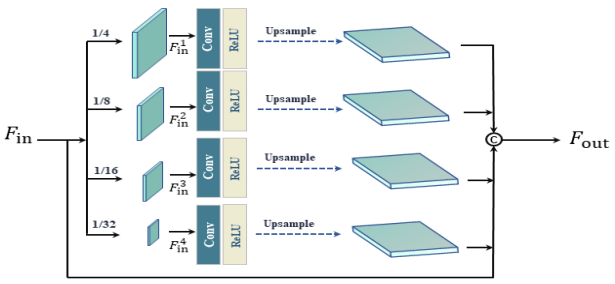


Figure 5. Illustration of MPM. This module is used to capture feature responses at different scales.

3.5. Hybrid Loss Function

In order for the training model to generate a good visual output and achieve high quantitative scores, inspired by [32], we adopt a mixed L1 and MS-SSIM loss function in our network and verify its effectiveness in section 4. Its definition is as follows.

- **L1 loss** is applied to suppress artifacts and color distortion which may be generated by the ITAMPN network. Generally, this term is less sensitive to outliers that deviate from the normal range, and has a stable gradient, which guides the output of the model closer to the reference value sanely. We define $\mathcal{L}_{L1}$ as,

$$\mathcal{L}_{L1} = \frac{1}{N}\sum_{i=1}^{N}\left\| I_{GT}^{i} - ITAMPN(I_{input}^{i}) \right\| \quad (10)$$

- **MS-SSIM loss** is an improved version of Structural Similarity (SSIM), which preserves the edges and details of the restored image. $\mathcal{L}_{MS\text{-}SSIM}$ defined as follows.

$$\begin{aligned} SSIM(x,y) &= \left(\frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}\right)\left(\frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}\right) \\ &= \ell(p)\cdot cs(p) \end{aligned} \quad (11)$$

$$MS\text{-}SSIM(p) = \ell_M^{\alpha}(p)\cdot\prod_{j=1}^{M} cs_j^{\beta_j}(p) \quad (12)$$

$$\mathcal{L}_{MS\text{-}SSIM}(P) = 1 - MS\text{-}SSIM(\tilde{p}) \quad (13)$$

Where μ_x and σ_x represent the mean and variance of the enhanced image, μ_y and σ_y represents the mean and variance of the reference image, σ_{xy} represents the covariance of the enhanced image x and the true value y, c_1 and c_2 are fixed constants. ℓ_M and cs_j are the first and second terms of (11), M and j denote the scales. $\tilde{p}$ is the center pixel of patch P. Further details on MS-SSIM can be found in [32].

In order to retain the high frequency information while keeping the brightness and color unchanged, the complete objective function is formulated as follows:

$$\mathcal{L}_{Loss} = \vartheta\cdot\mathcal{L}_{MS\text{-}SSIM} + (1-\vartheta)\cdot G_{\sigma_G^M}\cdot\mathcal{L}_{L1} \quad (14)$$

Where $G_{\sigma_G^j} = [0.5, 1.0, 2.0, 4.0, 8.0]$, while the value of ϑ is empirically set to 0.025 in this work.

4. Experiments

To demonstrate the effectiveness of the proposed method, in this section we present comprehensive evaluations on real underwater image datasets.

4.1. Datasets

In our assessment experiments, the EUVP dataset [9] is adopted as the training set. The dataset contains three subsets, named Underwater Dark, Underwater ImageNet and Underwater Scenes, and a total of 11, 435 clear / hazy underwater scene image pairs. In order to improve the quality of model training, we augmented

the EUVP dataset by flipping each pair of images horizontally, and then rotated the raw and flipped images by 90°, 180° and 270° degrees, obtaining 8 different pairs of images. In this manner, the 11 thousand pairs of images generated 88 thousand training pairs, and the remaining 3480 pairs were used as the test sets. In order to evaluate the effectiveness of our model on different datasets, we used the open underwater image dataset UFO-120 [8] as the test dataset. This dataset comprises images collected from marine exploration in multiple locations in different water types with a total of 1620 clear/hazy underwater image pairs.

4.2. Implementation Details and Evaluation Metrics

Our experimental configuration used an Intel i9-10900X CPU, 32.0 GB of RAM and a single NVIDIA RTX3090 graphics card with 24GB of video memory. We implemented the proposed network using the Pytorch framework. The size of the images in the training dataset was normalized to 256×256, the batch size was 16, and the number of epochs was empirically set to 60. The Adam optimizer was adopted, which its learning rate is cosine annealing strategy [13] with an initial value of 0.8e-4 and momentum 0.9.

To evaluate the proposed network, we used several reference metrics including the Peak Signal-to-Noise Ratio (PSNR) [7], Structural Similarity (SSIM) [27], the Patch-based Contrast Quality Index (PCQI) [26] and the non-reference Underwater Image Quality Measure (UIQM) [19] metric on the test datasets.

4.3. Comparison with the SOTA Methods

In order to demonstrate the robustness of our method, we compare the proposed model with seven advanced enhancement methods including traditional methods (UDCP [5], IBLA [20], ULAP [22], Ancuti *et al.*'s [1]) and learning-based methods (UGAN [3], Shallow-UWnet [17], and Chen *et al.*'s [2]) through qualitative and quantitative analysis.

- **Qualitative evaluation**: Figures 6 and 7 shows a qualitative comparison of different algorithms on the UFO-120 and the EUVP (which contains the three subsets of Underwater Dark, Underwater ImageNet, and Underwater Scenes) datasets respectively. As shown in Figures 6 and 7, the raw images are commonly heavily biased with blue-green tones, our method shows an excellent enhancement effect on both datasets and is closer to the reference image. The qualitative results show that images affected by strong blue tones will increase the difficulty of image enhancement. UDCP is incapable of removing the blue-green haze due to their inaccurate color correction method, and even produces a darker background and bluer tones, worsening the color distortion. Among the learning-based method, UGAN can produce more natural scenes, but show little improvement in terms of artifacts and color distortion. Chen et al.'s method integrates deep learning and an image formation model, eliminating the influence of underwater environmental factors and enriching the color and details, but the detail tone improvement is insufficient. Compared with these methods, our model produces a visual appearance most similar to the reference image and shows great improvement in clarity and detail tone restoration.
- **Quantitative evaluation**: In addition to the qualitative evaluation, we also report the quantitative results of the different enhancement algorithms. Tables 2 and 3 lists the quantitative test results obtained using the different underwater enhancement technologies on the EUVP and UFO-120 datasets, and the best results are marked in bold. The results show that our method obtained the highest scores in the PSNR, SSIM, and PCQI metrics, and ranked fifth in the UIQM metric. However, UIQM is a non-reference evaluation metric and does not consider the image structure information. Some methods based on priors and models, such as UDCP, IBLA, and ULAP, have low scores in PSNR, SSIM, PCQI, and UIQM on the two datasets. The reason may be that these methods rely too much on underwater imaging models and prior hypotheses. Ancuti *et al.*'s [1] method achieved the highest scores of UIQM on all datasets, but performed poorly on other metrics, which demonstrates that the corresponding results are good in terms of blue-green tone removal but poor with regard to contrast, brightness, and color correction, which eventually produces a poor visual effect. However, learning-based methods, such as UGAN, Shallow-UWnet, Chen *et al.*'s [2] and our method, are data-driven and do not rely on imaging models, and thus achieve relatively better results. Based on the four quantitative metrics considered, our method shows significant enhancement effectiveness. From the results of our method in different datasets, our method had the highest SSIM and PCQI values and the lowest PSNR value on the Underwater Dark dataset compared to the other datasets. This shows that our method can improve the brightness, contrast and structural information effectively when processing images affected by strong blue tones, but the color correction of details is relatively not good. Our method had the highest PSNR on the UFO-120 dataset, which shows that the results are closer to the reference image.

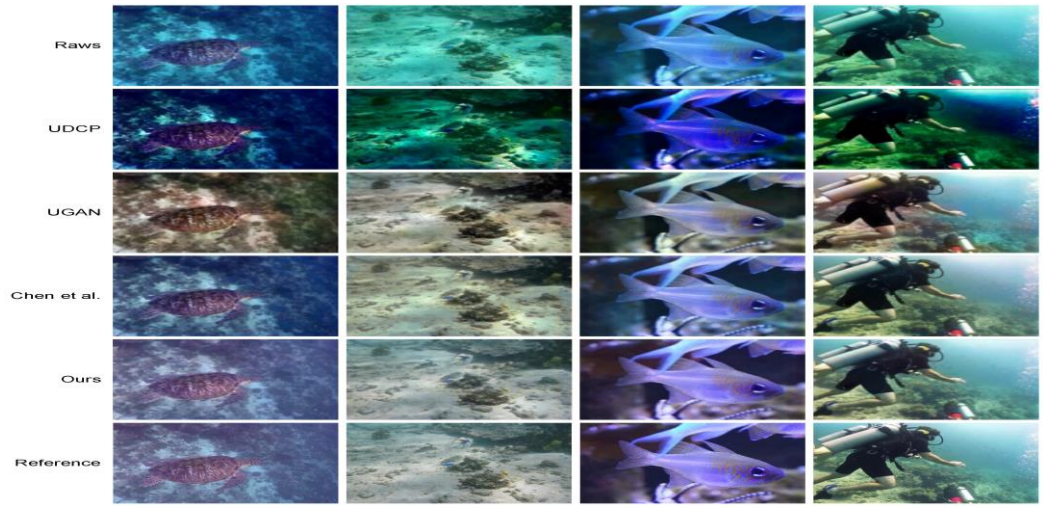


Figure 6. Qualitative comparisons for sample results on the UFO120 dataset.

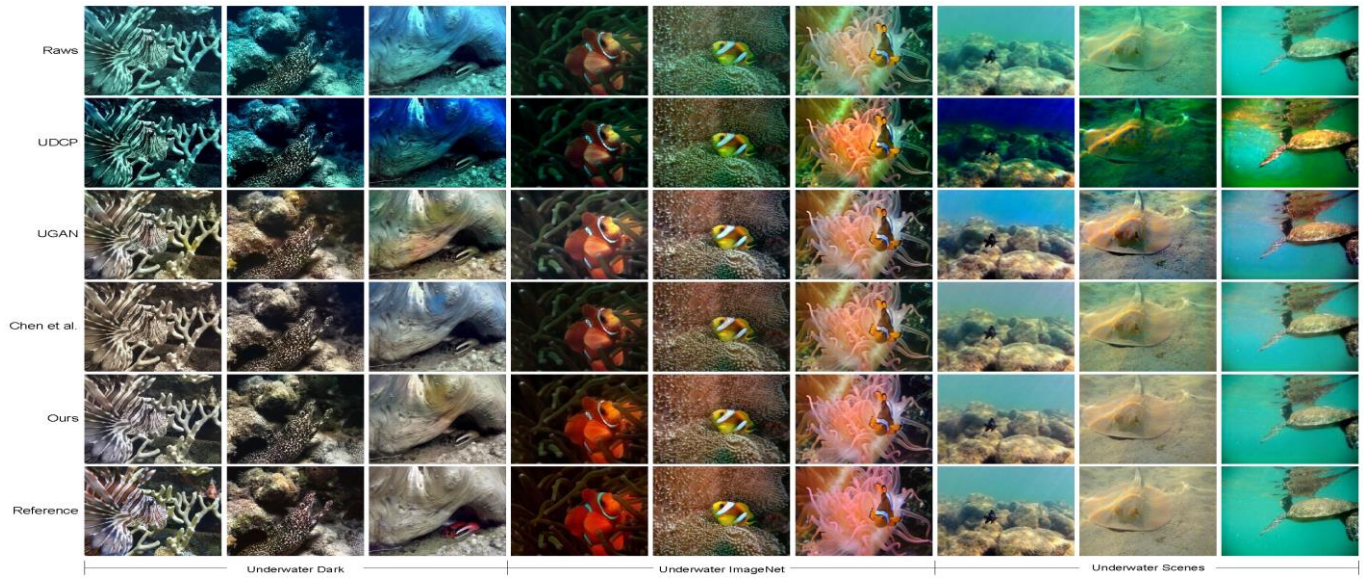


Figure 7. Qualitative comparisons for sample results on the EUVP dataset.

Table 2. Comparison of different models on the EUVP dataset. Higher values mean better results. Bold values show the best performer.

EUVP subset	Metric	UDCP	IBLA	ULAP	Ancuti *et al.* [1]	UGAN	Shallow-UWnet	Chen *et al.* [2]	Ours
Underwater Dark	**PSNR**	13.23	16.81	17.57	18.98	20.78	20.34	21.80	**22.24**
	SSIM	0.6523	0.7775	0.7753	0.8474	0.8782	0.8756	0.8960	**0.9050**
	PCQI	0.7616	0.7709	0.7540	0.7918	0.8867	0.9014	0.9369	**0.9394**
	UIQM	1.7915	2.0914	2.0296	**2.9704**	2.9501	2.8620	2.9411	2.8568
Underwater ImageNet	**PSNR**	15.57	16.74	18.51	19.28	21.91	22.72	23.20	**24.26**
	SSIM	0.6199	0.6297	0.6604	0.7371	0.7735	0.7914	0.7966	**0.8040**
	PCQI	0.7050	0.6184	0.6313	0.6464	0.7304	0.7562	0.7603	**0.7676**
	UIQM	1.9824	2.0354	1.8778	**3.0942**	3.0787	2.8777	2.8811	2.7867
Underwater Scenes	**PSNR**	13.79	18.68	19.76	17.83	20.61	22.31	23.65	**26.51**
	SSIM	0.5502	0.6991	0.7382	0.7471	0.7698	0.7948	0.8037	**0.8262**
	PCQI	0.6852	0.6924	0.6805	0.6809	0.7398	0.8068	0.8168	**0.8380**
	UIQM	1.8860	2.0709	2.1660	**3.0043**	2.9672	2.8610	2.8676	2.6852

Table 3. Comparison of different models on the UFO-120 dataset. Higher values mean better results. Bold values show the best performer.

Metric	UDCP	IBLA	ULAP	Ancuti *et al.* [1]	UGAN	Shallow-UWnet	Chen *et al.* [2]	Ours
PSNR	14.48	19.14	19.65	18.21	20.40	22.14	24.05	**26.87**
SSIM	0.5350	0.6760	0.7008	0.7261	0.7185	0.7417	0.7883	**0.8085**
PCQI	0.7197	0.7020	0.6890	0.6741	0.7342	0.7997	0.8291	**0.8501**
UIQM	1.9296	2.1889	2.2443	**3.1540**	2.9489	2.8836	2.9885	2.8552

4.4. Component Analysis of Our Model

To further verify the validity of our network, we analyzed the effects of each ITAMPN component and its influence on the results in different combinations, and conducted various ablation studies on the EUVP and UFO-120 datasets.

There are three main components in the proposed ITAMPN, namely the encoder-decoder architecture (EDA), the TAM, and the multi-scale pyramid module (MPM). To investigate the effect of each component on the enhancement results, we divided the models into four categories: EDA, EDA+TAM, EDA+MPM and EDA+TAM+MPM, and used the same training parameters to train the four models and conduct qualitative and quantitative analysis experiments on the two datasets. The average results in terms of PSNR, SSIM and PCQI are shown in Table 4, which shows the relative contribution of the different components on the enhancement performance. It can be seen that EDA+TAM+MPM achieve the highest PSNR, SSIM and PCQI scores except for the suboptimal PCQI score on the EUVP (Underwater Dark) dataset. We also observe that compared with EDA, EDA+MPM achieves better results on PSNR and SSIM, while the score of PCQI was low than or equal to that of EDA. This demonstrates that merely adding the MPM does not enhance the ability of feature representation substantially. When we add the TAM module to the EDA, the overall enhancement effect is significantly improved; especially on the UFO-120 data set the three metrics are significantly improved. In addition, we also illustrate the different performances of the three components in our model qualitatively. As shown in first and second rows of Figure 8, the color of the sea turtle and the jellyfish obtained via the EDS+TAM+MPM combination is closer to the reference image. The enhancement effects of EDA and EDA + MPM are still not very clear and the results have artifacts or blue blocks, while EDA+TAM and EDA+TAM+MPM show a better enhancement effect. In third row of Figure 8, in the results obtained by EDA and EDA+MPM, there are faint artifacts on the divers, while this is not the case for EDA+TAM and EDA+TAM+MPM. From the qualitative and quantitative analyses, we see that EDA+ TAM+MPM show the best results in color restoration, and the performance of ITAMPN is weakened without the TAM or MPM, which verifies their effectiveness in underwater image enhancement tasks.

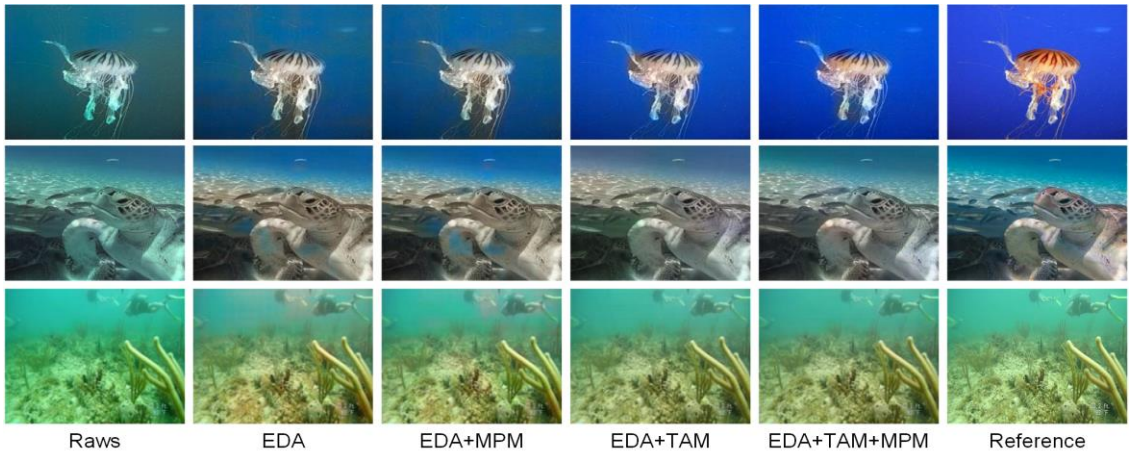


Figure 8. Qualitative comparisons of the contribution of different components of the proposed method on enhancement performance.

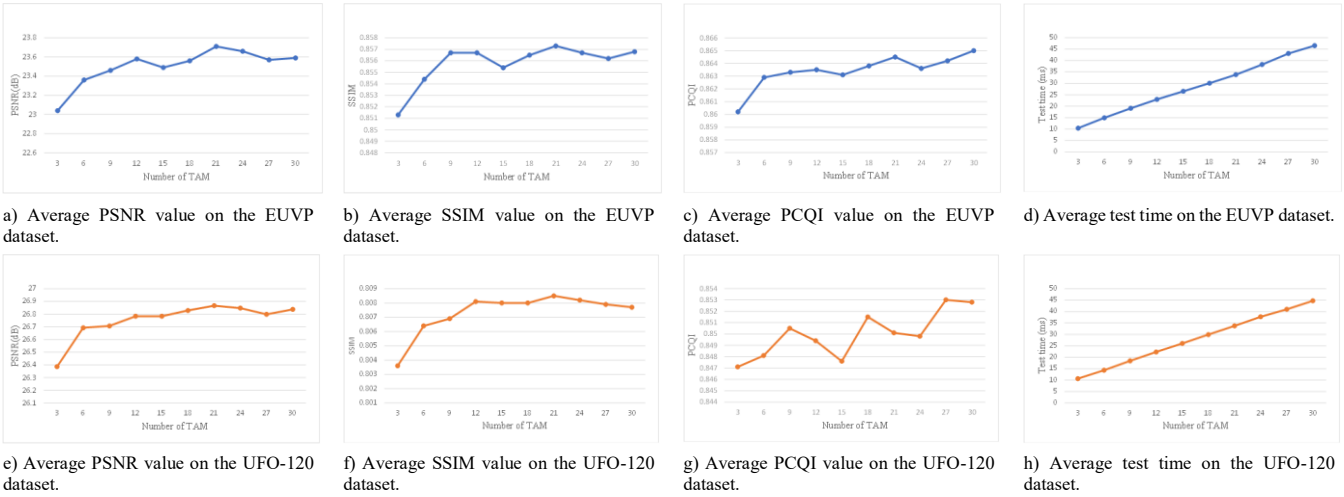
Table 4. Comparison of different component of ITAMPN on the EUVP and UFO-120 datasets.

Dataset	EDA	TAM	MPM	PSNR	SSIM	PCQI
EUVP (Underwater Dark)	✓			22.08	0.9013	0.9318
	✓		✓	22.12	0.9017	0.9318
	✓	✓		22.05	0.9037	**0.9401**
	✓	✓	✓	**22.24**	**0.9050**	0.9394
EUVP (Underwater ImageNet)	✓			23.39	0.7974	0.7616
	✓		✓	23.43	0.7987	0.7611
	✓	✓		24.12	0.8029	0.7637
	✓	✓	✓	**24.26**	**0.8040**	**0.7676**
EUVP (Underwater Scenes)	✓			25.02	0.8169	0.8214
	✓		✓	25.29	0.8152	0.8199
	✓	✓		26.29	0.8244	0.8362
	✓	✓	✓	**26.51**	**0.8262**	**0.8380**
UFO-120	✓			24.71	0.7951	0.8281
	✓		✓	24.95	0.7952	0.8267
	✓	✓		26.82	0.8075	0.8482
	✓	✓	✓	**26.87**	**0.8085**	**0.8501**

In order to further investigate the effectiveness of the feature fusion strategy of TAM, we implemented the three feature fusion methods mentioned in section 3.3, as well as the different dual attention feature fusion combinations. Then, we performed qualitative tests on the EUVP and UFO-120 datasets. The results are shown in Table 5. We see that TAM with parallel feature fusion achieves better PSNR and SSIM scores on both EUVP and UFO-120 datasets, except the SSIM on the EUVP dataset is slightly less than the highest value. This demonstrates that the parallel feature fusion of TAM bestows the model with a stronger ability to extract global context information compared to the other forms.

Table 5. Comparison of different components of attention module on the EUVP and UFO-120 datasets.

Fusion strategy	CA	PA	SA	UFO-120		EUVP	
				PSNR	SSIM	PSNR	SSIM
parallel	✓	✓		26.74	0.8070	23.53	0.8565
parallel	✓		✓	26.76	0.8073	23.04	0.8517
parallel		✓	✓	26.64	0.8067	23.56	**0.8575**
Serial	✓	✓	✓	26.74	0.8071	23.01	0.8516
Direct parallel	✓	✓	✓	24.78	0.7960	23.04	0.8517
Parallel	✓	✓	✓	**26.87**	**0.8085**	**23.71**	0.8573



a) Average PSNR value on the EUVP dataset. b) Average SSIM value on the EUVP dataset. c) Average PCQI value on the EUVP dataset. d) Average test time on the EUVP dataset.

e) Average PSNR value on the UFO-120 dataset. f) Average SSIM value on the UFO-120 dataset. g) Average PCQI value on the UFO-120 dataset. h) Average test time on the UFO-120 dataset.

Figure 9. Quantitative enhancement results with the proposed network under different numbers of TAMs. Average results are shown for the two datasets.

In order to evaluate the influence of the number of TAMs on the model effectively and obtain the best parameters, we conducted model training with 3 TAMs as the base configuration and then increased. The quantitative test results on the EUVP and UFO-120 datasets are shown in Figure 9. As shown in Figure 9-a) 9-b) and 9-e) 9-f), PSNR and SSIM basically show an upward trend with the number of TAMs increasing, and then the trend eases off. The PSNR and SSIM reach their highest values when the number of TAMs is 21.

Figure 9-c) and 9-g) shows the change of PCQI with the number of TAMs. Although there is no obvious trend of continuous increase, the overall change trend is small. Figure 9-d) and 9-h) show how the test time of the model changes with the increase of TAMs. It can be seen that the test time has an obvious upward trend with the increase of attention modules. Based on the values of Figure 9, we believe that 21 TAMs achieve a good balance between enhancement performance and training time.

In order to investigate the effectiveness of different loss functions on our model, we used different loss functions to train our model and tested it on the two datasets. The corresponding quantitative test results are shown in Table 6. The results show that L1 + MS-SSIM lead to the highest PSNR, SSIM, and PCQI values, while the PCQI value on the UFO-120 dataset is slightly below the optimal value. Thus, it can be seen that L1 + MS-SSIM is an effective loss function for our model.

Table 6. Comparison of different loss function evaluation on the EUVP and UFO-120 datasets. Higher values mean better results.

Dataset	Metric	L1	L2	MS-SSIM	L2+MS-SSIM	L1+MS-SSIM
EUVP	**PSNR**	23.56	23.43	23.09	23.49	**23.71**
	SSIM	0.8560	0.8548	0.8555	0.8558	**0.8573**
	PCQI	0.8643	0.8434	0.8557	0.8600	**0.8645**
UFO-120	**PSNR**	26.81	26.62	25.47	26.69	**26.87**
	SSIM	0.8068	0.8035	0.8044	0.8074	**0.8085**
	PCQI	0.8508	**0.8573**	0.8159	0.8384	0.8501

5. Conclusions

In this paper, we propose an ITAMPN network for enhancing underwater images. Our model is based on an encoder-decoder architecture, with an MPM embedded for learning structural information at different scales, and triple attention blocks incorporated for powerful feature representation. Experimental results on real-world datasets show that our method achieves excellent performance in color correction and detail enhancement compared with other methods. Also, an ablation study on the underwater image further shows the effectiveness of our model's components.

However, there are some limitations to our model. Concretely, first, the training sample of the model is large, which will lead to the training needs a lot of time resources. Secondly, it is difficult for our model to obtain satisfactory visual effects when processing some seriously degraded images. In the future, we will try to reduce the training samples without reducing the enhancement effect, and optimize the structure based on the model, so as to improve the performance of model.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province of China under grant no. 2023A1515011272, the Tertiary Education Scientific Research Project of Guangzhou Municipal Education Bureau of China under no. 202234598, the Special Project in Key Fields of Guangdong Universities of China under no. 2022ZDZX1020.

Reference

[1] Ancuti C., Ancuti C., Vleeschouwer C., and Bekaert P., "Color Balance and Fusion for Underwater Image Enhancement," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 379-393, 2018.

[2] Chen X., Zhang P., Quan L., Yi C., and Lu C., "Underwater Image Enhancement based on Deep Learning and Image Formation Model," *arXiv preprint arXiv:2101.00991*, 2021.

[3] Fabbri C., Islam M., and Sattar J., "Enhancing Underwater Imagery using Generative Adversarial Networks," *in Proceedings of IEEE International Conference on Robotics and Automation*, Brisbane, pp. 7159-7165, 2018.

[4] Gaude G. and Borkar S., "Fish Detection and Tracking for Turbid Underwater Video," *in Proceedings of the International Conference on Intelligent Computing and Control Systems*, Madurai, pp. 326-331, 2019.

[5] Guo Q., Xue L., Tang R., and Guo L., "Underwater Image Enhancement Based on the Dark Channel Prior and Attenuation Compensation," *Journal of Ocean University of China*, vol. 16, no. 5, pp. 757-765, 2017.

[6] Haulsee D., Breece M., Miller D., Wetherbee B., Fox D., and Oliver M., "Habitat Selection of a Coastal Shark Species Estimated from an Autonomous Underwater Vehicle," *Marine Ecology Progress Series*, vol. 528, pp. 277-288, 2015.

[7] Hore A. and Ziou D., "Image Quality Metrics: PSNR vs. SSIM," *in Proceedings of the International Conference on Learning Representations*, Istanbul, pp.2366-2369, 2010.

[8] Islam M., Luo P., and Sattar J., "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception," *in Proceedings of the Robotics: Science and Systems*, Corvalis, 2020.

[9] Islam M., Xia Y., and Sattar J., "Fast Underwater Image Enhancement for Improved Visual Perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227-3234, 2020.

[10] Jaffe J., "Computer Modeling and the Design of Optimal Underwater Imaging Systems," *IEEE Journal of Oceanic Engineering*, vol. 15, no. 2, pp. 101-111, 1990.

[11] Jamadandi A., and Mudenagudi U., "Exemplar-based Underwater Image Enhancement Augmented by Wavelet Corrected Transforms," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, pp. 11-17, 2019.

[12] Li C., Guo C., Ren W., Cong R., Hou J., Kwong S., and Tao D., "An Underwater Image Enhancement Benchmark Dataset and Beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376-4389, 2019.

[13] Loshchilov I. and Hutter F., "SGDR: Stochastic Gradient Descent with Warm Restarts," *in Proceedings of the International Conference on Learning Representations*, San Juan, 2016.

[14] McGlamery B., "A Computer Model for Underwater Camera Systems," *Ocean Optics VI. International Society for Optics and Photonics*, vol. 208, pp. 221-231, 1980.

[15] Mei K., Jiang A., Li J., and Wang M., "Progressive Feature Fusion Network for Realistic Image Dehazing," *in Proceedings of the Asian Conference on Computer Vision*, Cham, pp. 203-215, 2018.

[16] Mnih V., Heess N., Graves A., and Kavukcuoglu K., "Recurrent Models of Visual Attention," *in Proceedings of the Advances in Neural Information Processing Systems*, Montreal, pp. 2204-2212, 2014.

[17] Naik A., Swarnakar A., and Mittal K., "Shallow-UWnet: Compressed Model for Underwater Image Enhancement," *in Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, vol. 35, no. 1, pp. 15853-15854, 2021.

[18] Ortiz A., Simó M., and Oliver G., "A Vision System for an Underwater Cable Tracker," *Machine Vision and Applications*. vol. 13, no. 3, 129-140, 2002.

[19] Panetta K., Gao C., and Agaian S., "Human-Visual-System-Inspired Underwater Image Quality Measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541-551, 2016.

[20] Peng Y. and Cosman P., "Underwater Image Restoration Based on Image Blurriness and Light Absorption," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1579-1594, 2017.

[21] Qin X., Wang Z., Bai Y., Xie X., and Jia H., "FFA-Net: Feature Fusion Attention Network for Single Image Dehazing," *in Proceedings of AAAI Conference on Artificial Intelligence*, New York, pp. 11908-11915, 2020.

[22] Song W., Wang Y., Huang D., and Tjondronegoro D., "A Rapid Scene Depth Estimation Model Based on Underwater Light Attenuation Prior for Underwater Image Restoration," *in Proceedings of Pacific Rim Conference on Multimedia*, Hefei, pp. 678-688, 2018.

[23] Teixeira B., Silva H., Matos A., and Silva E., "Deep Learning Approaches Assessment for Underwater Scene Understanding and Egomotion Estimation," *in Proceedings of IEEE OCEANS*, Seattle, pp. 1-9, 2019.

[24] Uplavikar P., Wu Z., and Wang Z., "All-in-One Underwater Image Enhancement Using Domain-Adversarial Learning," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Seoul Long Beach, pp. 1-8, 2019.

[25] Wang F., Jiang M., Qian C., Yang S., Li C., Zhang H., Wang X., and Tang X., "Residual Attention Network for Image Classification," *in Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, Honolulu, pp. 3156-3164, 2017.

[26] Wang S., Ma K., Yeganeh H., Wang Z., and Lin W., "A Patch-Structure Representation Method for Quality Assessment of Contrast Changed Images," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2387-2390, 2015.

[27] Wang Z., Bovik A., Sheikh H., and Simoncelli E., "Image Quality Assessment from Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[28] Woo S., Park J., Lee J., and Kweon I., "CBAM: Convolutional Block Attention Module," *in Proceedings of the European Conference on Computer Vision*, Munich, pp. 3-19, 2018.

[29] Yin S., Wang Y., and Yang Y., "A Novel Image-Dehazing Network with a Parallel Attention Block," *Pattern Recognition*, vol. 102, pp. 107255, 2020.

[30] Zamir S., Arora A., Khan S., Hayat M., Khan F., Yang M., and Shao L., "Learning Enriched Features for Real Image Restoration and Enhancement," *in Proceedings of the European Conference on Computer Vision*, Glasgow, pp. 492-511, 2020.

[31] Zhang H. and Patel V., "Densely Connected Pyramid Dehazing Network," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, pp. 3194-3203, 2018.

[32] Zhao H., Gallo O., Frosio I., and Kautz J., "Loss Functions for Image Restoration with Neural Networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47-57, 2016.

[33] Zhao H., Shi J., Qi X., Wang X., and Jia J., "Pyramid Scene Parsing Network," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 2881-2890, 2017.

[34] Zhou W. and Jia J., "Training Convolutional Neural Network for Sketch Recognition on Large-Scale Dataset," *The International Arab Journal of Information Technology*, vol. 17, no. 1, pp. 82-89, 2020.

Kaichuan Sun is currently a computer science Ph.D. student at Jiangsu University of Science and Technology, China. His research interests include computer vision and deep learning, particularly in the domains of image processing.

Yubo Tian received a Ph.D. degree in radio physics from the department of electronic science and engineering, Nanjing University, in 2004. He is currently with the school of information and communication engineering, Guangzhou Maritime University. His current research interest is machine learning methods and their applications in electronics and images.