Boths: Super Lightweight Network-Enabled Underwater Image Enhancement

Xu Liu, *Student Member, IEEE*, Sen Lin, *Member, IEEE*, Kaichen Chi, Zhiyong Tao, and Yang Zhao, *Member, IEEE*

***Abstract*— Since light is scattered and absorbed by water, underwater images have inherent degradation (e.g., hazing, color shift), consequently impeding the development of remotely operated vehicles (ROVs). Toward this end, we propose a novel method, referred to as *B*est *o*f Bo*th* World*s* (Boths). With parameters of only 0.0064 M, Boths can be considered a super lightweight neural network for underwater image enhancement. On the whole, it has three levels: structure and detail features; pixel and channel dimensions; high- and low-frequency information. Each of these three levels represents "Best of Both Worlds." Initially, by interacting with structure and detail features, Boths can focus on these two aspects at the same time. Further, our network can simultaneously consider channel and pixel dimensions through 3-D attention learning, which is more similar to human visual perception. Lastly, the proposed model can focus on high- and low-frequency information, through a novel loss function based on the wavelet transforms. Upon subsequent analysis and evaluation, Boths has shown superior performance compared with state-of-the-art (SOTA) methods. Our models and datasets are publicly available at: https://github.com/perseveranceLX/Boths.**

***Index Terms*— 3-D attention learning, high- and low-frequency loss functions, structure and detail interaction, underwater image enhancement.**

I. Introduction

A VARIETY of remote sensing techniques are successfully employed in underwater vision scenes [1], notably in visually guided remotely operated vehicles (ROVs) [2]. Compared with autonomous underwater vehicles (AUVs) [3], ROV is more suitable for working in complex, narrow, and unknown environments. Also, it plays a crucial role in practical applications such as underwater archeology [4], marine ecological exploration [5], and deep-sea target detection [6]. During visual-guided research, acquiring images of high quality is an essential step. However, underwater images are often extremely degraded. In the ocean, red, green, and blue light have different attenuation rates, whereas red light is the fastest, so the underwater image generally looks blue or green. Moreover, suspended particles in the water absorb light energy and change its path, resulting in low contrast and blurring. This complex degradation makes it difficult to obtain clear underwater images [7]. Hence, it is urgent and meaningful to design a fast and effective underwater image enhancement algorithm for ROV.

For the above imaging characteristics, two types of underwater image enhancement are commonly employed: prior-driven and deep learning-driven methods. Some of the prior-driven methods apply decomposition and synthesis tools to enhance underwater images through multiple stages. Cho et al. [8] extract the image detail layer through multiband decomposition and refine it with a Laplacian module [model-assisted multiband fusion (MAMF)]. While this algorithm enhances detail, the image has an overall color cast. Yuan et al. [9] adopted contour bougie morphology to separate the scenes, and added various operations to obtain a better outline [Contour Bougie Morphology and Adaptive Contrast Stretch (ACS)]. They also proposed a fusion-based texture enhancement method [texture enhancement model based on blurriness and color fusion (TEBCF)] [10] for real-world underwater images. These two approaches can optimize the structural layer and the texture layer of the image respectively. Despite the favorable results of the prior-driven methods, the estimation of prior conditions limits their performance.

With the advent of deep learning in the 21st century, convolutional neural networks (CNNs) and generative adversarial networks (GANs) are extensively implemented in low-level vision tasks. For the purpose of preserving the image content while also removing the image noise, Chen et al. [11] developed a GAN-based network [GAN-based restoration scheme (GAN-RS)]. Li et al. [12] proposed a CNN framework (Water-Net) with adaptive fusion of multiple preprocessed images. Islam et al. [13] presented a conditional GAN [fully-convolutional conditional GAN-based model (FunIE-GAN)] with a novel loss, which controls color, texture, and style of the generated images. Li et al. [14] designed a CNN-based network (Ucolor) to solve color casts and low contrast, enhancing underwater images by using multicolor space learning. To make the algorithm lightweight, Jiang et al. [15] presented a cascaded CNN [lightweight cascaded network (LCNet)] based on Laplacian pyramids. Recently, considering that underwater images taken in similar scenes tend to degrade generally, Qi et al. [16] proposed a coenhancement network [underwater image co-enhancement network (UICoE-Net)] that relies on CNN-based siamese learning.

However, deep learning-driven methods are often large and ill-suited to underwater robots (e.g., ROV). Few methods have been devoted to creating powerful models for enhancing. Furthermore, due to underwater imagery representing a more

Manuscript received 19 May 2022; revised 18 August 2022 and 14 October 2022; accepted 11 December 2022. Date of publication 16 December 2022; date of current version 3 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1403303, in part by the National Natural Science Foundation of China under Grant 61972129, and in part by the Key Research and Development Program of Liaoning Province under Grant 2019JH2/10100014. *(Corresponding author: Sen Lin.)*

Xu Liu and Yang Zhao are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: dalong.xu.liu@ieee.org; yzhao@hfut.edu.cn).

Sen Lin is with the School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China (e-mail: lin_sen6@126.com).

Kaichen Chi is with the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: chikaichen@mail.nwpu.edu.cn).

Zhiyong Tao is with the School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China (e-mail: xyzmail@126.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/LGRS.2022.3230049, provided by the authors.

Digital Object Identifier 10.1109/LGRS.2022.3230049

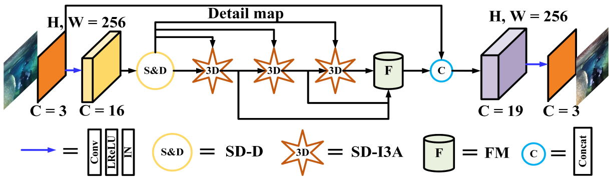

Fig. 1. Overview of the Boths. Conv, Concat, LReLU, and IN means convolution, concatenation, leaky ReLU, and instance normalization. H, W, and C represent height, width, and number of channels, respectively.

challenging issue (than atmospheric imagery), underwater image enhancement by learning still remains a significant area for development.

As shown in Fig. 1, we seek to address these difficulties with a unique super lightweight neural network called ***B****est* ***o****f* *bo****th*** *world****s*** (Boths) that has the potential to enhance underwater images. Following is a summary of the principal implications.

1) ***Structure and Detail Features:*** We propose the structure and detail decomposition (SD-D) which divides the input into two parts. Then, these features are combined through structure and detail interactive 3-D attention (SD-I3A). In addition, we also present a module [fusion module (FM)] to fuse the outputs of multilevel SD-I3A.
2) ***Pixel and Channel Dimensions:*** Typically, attention to pixels and channels are separate concerns. In actuality, they are closely related in terms of human visual perception. Three-dimensional attention learning in Boths is our response. It accounts for the pixel and channel dimensions by a single weight.
3) ***High and Low Frequency Information:*** Based on the wavelet transforms, we design the wavelet mse (WMSE) loss. In conjunction with other losses, the network can focus on high- and low-frequency information when it comes to the training process.

According to our qualitative analysis and quantitative evaluation, Boths is a highly effective model. In light of the modest parameters and floating point of operations (FLOPs) (0.0064 M and 0.4256 G, respectively), our method can be easily deployed in ROV.

II. Proposed Method

The proposed Boths is shown in Fig. 1. SD-D divides the input into two paths, then performs multidimensional learning of the two paths by SD-I3A, and finally fuses the multilevel SD-I3A results by FM.

A. Structure and Detail Decomposition

A retina contains two types of cells [17], midget and parasol cells, whose receptive fields differ. We employ two different dilated convolutions d_1 (kernel size = 3, stride = 1, dilition = 1, padding = 1) and d_2 (kernel size = 3, stride = 1, dilition = 5, padding = 5) to simulate them in Fig. 2. By subtracting them, we obtain the guided map D_g

$$D_g = \sigma(\delta(d_1(A_{RGB})) - \delta(d_2(A_{RGB}))) \tag{1}$$

where σ and δ are the sigmoid and leaky rectified linear unit (LReLU) functions. Structural features have a low contrast, whereas detailed features display a high contrast. So we multiply the guided map with A_{RGB} to obtain the detail map A_{det}

$$A_{det} = D_g \times A_{RGB} \tag{2}$$

then, the remaining structure map A_{str} can be represented as

$$A_{str} = A_{RGB} - A_{det}. \tag{3}$$

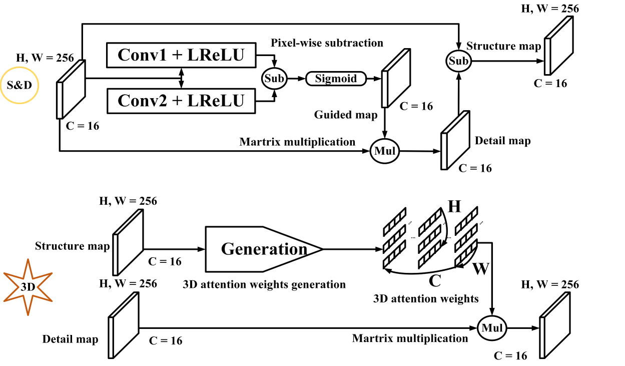

Fig. 2. Architecture of the SD-D, SD-I3A. Conv1 and Conv2 are two different dilated convolutions d_1 and d_2, respectively. H, W, and C represent height, width, and number of channels, respectively.

Due to the lightweight nature of Boths, the SD-D does not employ large-scale convolutions or very deep networks to expand its receptive field.

B. Structure and Detail Interactive 3-D Attention

At present, most of the existing attention mechanisms are concerned with estimating domains separately [18], but these domains often participate in human visual perception simultaneously [19]. Guided by Yang et al. [20] using neural experience to develop the attention module [simple, parameter-free attention module (SimAM)], we adopt their method for generating weights and propose the SD-I3A in Fig. 2. Observation indicates that structural features are often degraded more severely. We obtain 3-D attention weights W_{str} for the structure map A_{str}

$$W_{str} = G(A_{str}) \tag{4}$$

where G means the weights generation. Then, we multiply the 3-D weights W_{str} of the structure map A_{str} by the detail map A_{det}

$$R_{I3A} = W_{str} \times A_{det}. \tag{5}$$

R_{I3A} is the result of SD-I3A. Thus, when processing structural and detailed features, the former can be emphasized. In order to gradually upgrade structural features processing, we use the output of the previous level as a structure map in Fig. 1. Aside from improving results, more SD-I3A will reduce the efficiency of algorithm operation. To achieve higher productivity, we use only a three-level SD-I3A.

C. Fusion Module

We propose a module to fuse the multilevel SD-I3A results as shown in Fig. 1. Specifically, R^1_{I3A}, R^2_{I3A}, R^3_{I3A} are the output from three level SD-I3A. We put the concatenation of them $O_{concat} \in \mathbb{R}^{48\times h\times w}$ into a convolution e_1 (kernel size = 3, stride = 1, padding = 1) to obtain the weight coefficient matrix $W \in \mathbb{R}^{3\times h\times w}$

$$O_{concat} = \mathrm{C}\left[R^1_{I3A}, R^2_{I3A}, R^3_{I3A}\right] \tag{6}$$

$$W = e_1(O_{concat}) \tag{7}$$

where C means concatenation. We obtain the fusion result by calculating the hadamard product of the features R_{I3A} and weight matrix W

$$F_{I3A} = W_1 \circ R^1_{I3A} + W_2 \circ R^2_{I3A} + W_3 \circ R^3_{I3A} \tag{8}$$

where $\circ$ means the hadamard product, R^1_{I3A}, R^2_{I3A}, R^3_{I3A} are the three level outputs, W_1, W_2, $W_3 \in \mathbb{R}^{1\times h\times w}$ represent the weight matrix W of each channel.

Algorithm 1 WMSE Loss

```
input : Two images $B(I)$, $GT$, factor $f$, times $n$
output: $\mathcal{L}_{\text{WMSE}}$
$(B(I))_0^{LL}, (GT)_0^{LL} = B(I), GT$;
$\mathcal{L}_{\text{WMSE}} = 0$;
$a = f$;
for $i \leftarrow 1$ to $n$, $i++$, do
    // Wavelet transfoms
    $(B(I))_i^{\text{LL}}, (B(I))_i^{\text{LH}}, (B(I))_i^{\text{HL}}, (B(I))_i^{\text{HH}} = \text{DWT}((B(I))_{i-1}^{\text{LL}})$;
    $(GT)_i^{\text{LL}}, (GT)_i^{\text{LH}}, (GT)_i^{\text{HL}}, (GT)_i^{\text{HH}} = \text{DWT}((GT)_{i-1}^{\text{LL}})$;
    // Calculate MSE loss for LH, HL,
       HH parts
    $\mathcal{L}_{\text{WMSE}} += [\mathcal{L}_{\text{MSE}}((B(I))_i^{\text{LH}}, (GT)_i^{\text{LH}}) + \mathcal{L}_{\text{MSE}}((B(I))_i^{\text{HL}}, (GT)_i^{\text{HL}}) + \mathcal{L}_{\text{MSE}}((B(I))_i^{\text{HH}}, (GT)_i^{\text{HH}})] \cdot a$;
    $a = a \times a$;
// Calculate MSE loss for the LL part
$\mathcal{L}_{\text{WMSE}} += \mathcal{L}_{\text{MSE}}((B(I))_i^{\text{LL}}, (GT)_i^{\text{LL}}) \cdot a$;
return $\mathcal{L}_{\text{WMSE}}$;
```

D. Loss Function

In our network, we aim to learn mappings between input I and ground truth (GT), the loss function consists of two terms: WMSE loss and other losses.

1) WMSE Loss: High- frequency information is often ignored, Boths enjoys a WMSE loss $\mathcal{L}_{\text{WMSE}}$ to consider high- and low-frequency information. Algorithm 1 shows the detailed steps of $\mathcal{L}_{\text{WMSE}}$. Four parts are derived using wavelet transforms

$$(M)^{\text{LL}}, (M)^{\text{LH}}, (M)^{\text{HL}}, (M)^{\text{HH}} = \text{DWT}(M) \tag{9}$$

where $(M)^{\text{LH}}$, $(M)^{\text{HL}}$, and $(M)^{\text{HH}}$ are high frequency parts, $(M)^{\text{LL}}$ is the low frequency part. $B(I)$ are images generated by Boths. We set the factor f to 0.25, times n to 2.

2) Other Losses: Following [13], we use $\mathcal{L}_1$ loss and $\mathcal{L}_{VGG}$ loss to measure the pixel and content similarity respectively,

$$\mathcal{L}_1 = \|B(I) - GT\|_1 \tag{10}$$

$$\mathcal{L}_{\text{VGG}} = \|\phi_{\text{k}}(B(I)) - \phi_{\text{k}}(GT)\|_2. \tag{11}$$

$\phi_{\text{k}}(.)$ represents the features extracted by VGG19 [21].

3) Asynchronous Training Mode: Inspired by adaptive learning attention network (LANet) [18], we employ their mode to make the network converge faster. Our Boths is trained through the following two stages. In the first stage, we utilize L_{WMSE}

$$\mathcal{L}_{\text{I}} = L_{\text{WMSE}}. \tag{12}$$

In the second stage, the loss is the linear superposition of $\mathcal{L}_1$ and $\mathcal{L}_{\text{VGG}}$

$$\mathcal{L}_{\text{II}} = L_1 + \lambda L_{\text{VGG}} \tag{13}$$

where λ is a constant, we set it to 0.1.

TABLE I

NUMBER OF TRAINING AND TESTING SAMPLES IN MIX BENCHMARK

Number	Datasets in MIX benchmark		
	UVE-38K	EUVP	UIEB
Training samples	3647	11435	9790
Testing samples	272	515	890

III. EXPERIMENT

A. Implementation Details

1) Datasets: For training and full-reference assessment, we use the UVE-38K [16], enhancement of underwater visual perception dataset (EUVP) [13] and underwater image enhancement benchmark (UIEB) [12] datasets. Due to insufficient samples in UIEB, we expanded the original dataset. We rotate the image at various angles. The angles A are 0, $\pi/2$, π, and $3\pi/2$. Later, we mirror the four images. The flips F are $NoFlip$, $HorizontalFlip$ and $VerticalFlip$. Hence, each image will get 12 augmented results. We finally get a large-scale benchmark containing 26 549 paired images. We name our benchmark as MIX.[1] Table I shows the division of training and testing samples. To assess the results of Boths in complex environments, we use T40,[2] U45 [22] and C60 [12] for no-reference assessment. Among them, T40[2] is an underwater sensing scene image dataset which contains 40 real-world images collected by us from ROVs. It is tough to enhance.

2) Training Details: The network is implemented on PyTorch and employs RMSprop [22] for model optimization. In addition, the batchsize and epoch are set to 6 and 100, respectively. For the first 30 epochs, we set the learning rate to 0.0001, then 0.0000001 for the remaining epochs. We normalize all pixels to $[-1, 1]$. We train UVE-38K, EUVP and UIEB separately due to the different image mappings.

3) Comparison Methods: To compare the enhanced results between our Boths and state-of-the-art (SOTA) models, we use seven deep learning-driven methods: GAN-RS [11], Water-Net [12], FUnIE-GAN [13], Ucolor [14], LCNet [15], UICoE-Net [16], LANet [18]; and three prior-driven methods: MAMF [8], ACS [9], TEBCF [10] for evaluation.

B. Discussion of Complexity

First, we perform a complexity comparison. It is an appropriate standard to measure the ability of algorithm deployment. As shown in Table II, we calculated the parameters and FLOPs of SOTA and our method. It is noteworthy that the parameters and FLOPs of Boths (0.0064 M and 0.4256 G, respectively) are far lower than those of other approaches. In our view, this relates to our model using 3-D attention. The parameter of our attention weight generation process is 0, while that in Ucolor is $2C^2/r$ (C represents the channels, r is the reduction ratio). As the proposed network is extremely lightweight, it can be implemented on mobile devices (e.g., ROV) that do not have a strong computing capability.

C. Quantitative Evaluation

Throughout this part of the experiment, we use metrics of mse, root mean square error (RMSE), peak signal to noise ratio (PSNR), structure similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS) [24] for full-reference image quality assessment in the UVE-38K, EUVP

[1] Our MIX benchmark: https://github.com/perseveranceLX/MIX

[2] Our T40 dataset: https://github.com/perseveranceLX/T40

TABLE II

ALGORITHM COMPLEXITY COMPARISON, NUMBER HIGHLIGHTED WITH RED, BLUE, AND BROWN TO INDICATE THE BEST THREE RESULTS

Complexity	Methods										
	MAMF	ACS	TEBCF	GAN-RS	Water-Net	FUnIE-GAN	Ucolor	LCNet	UICoE-Net	LANet	***Boths***
Parameters (M) ↓	–	–	–	11.3830	**1.0906**	7.0195	148.7712	**0.0233**	14.5751	5.1488	**0.0064**
FLOPs (G) ↓	–	–	–	–	142.9038	**10.2317**	2805.3399	139.0550	**64.6267**	355.7255	**0.4256**

TABLE III

QUANTITATIVE RESULTS EVALUATED BY FULL-REFERENCE IMAGE QUALITY ASSESSMENT, NUMBER HIGHLIGHTED WITH RED, BLUE, AND BROWN TO INDICATE THE BEST THREE RESULTS

Full-reference		Methods										
		MAMF	ACS	TEBCF	GAN-RS	Water-Net	FUnIE-GAN	Ucolor	LCNet	UICoE-Net	LANet	***Boths***
UVE-38K	MSE ↓	1119.3661	434.3175	503.1908	744.3462	**269.3165**	392.0454	**225.4618**	344.4753	294.1896	424.4879	**216.3530**
	RMSE ↓	33.4569	20.8403	22.4319	27.2817	**16.4109**	19.8001	**15.0154**	18.5600	17.1520	20.6031	**14.7089**
	PSNR ↑	17.9595	21.9420	21.3853	20.4522	**25.7128**	23.7391	**25.8397**	23.6221	24.1768	23.4940	**27.6369**
	SSIM ↑	0.8148	0.8443	0.8038	0.7916	**0.9187**	0.8751	0.9174	0.9069	**0.9188**	0.9074	**0.9230**
	LPIPS ↓	0.3130	0.2518	0.2805	0.2873	**0.1576**	0.2838	**0.1456**	0.2207	0.1662	0.1580	**0.1275**
EUVP	MSE ↓	907.2823	645.7779	657.7667	1221.1948	264.3434	**155.2618**	362.2267	**249.5698**	297.5698	468.4888	**155.2910**
	RMSE ↓	30.1211	25.4122	25.6470	34.9456	16.2584	**12.4604**	19.0323	**15.7978**	17.2576	21.6446	**12.4616**
	PSNR ↑	19.3417	20.6694	20.5276	17.9556	**25.6456**	**27.5260**	23.7140	24.7166	23.9734	22.6675	**27.6592**
	SSIM ↑	0.7632	0.7902	0.7858	0.7248	**0.8836**	0.8720	**0.8806**	0.8635	0.8708	0.8612	**0.8931**
	LPIPS ↓	0.4790	0.3536	0.3461	0.3814	0.2797	**0.2001**	**0.2752**	0.3251	0.2832	0.2869	**0.2557**
UIEB	MSE ↓	972.7319	572.8534	584.0434	764.1076	416.9423	548.4220	**242.9328**	598.8168	352.4565	**252.7175**	**257.0951**
	RMSE ↓	31.1886	23.9344	24.1670	27.6425	20.4192	23.4184	**15.5863**	24.4707	18.7738	**15.8917**	**16.0342**
	PSNR ↑	19.5702	21.4553	21.3225	19.9606	23.8237	22.5354	**25.8479**	22.1681	24.8269	**25.8020**	**25.6789**
	SSIM ↑	0.8183	0.8392	0.8425	0.7779	0.8766	0.8581	0.8952	0.8672	**0.9206**	**0.9054**	**0.9046**
	LPIPS ↓	0.3880	0.2795	0.2528	0.3103	0.1679	0.3074	**0.1435**	0.2438	0.1649	**0.1396**	**0.1490**

TABLE IV

QUANTITATIVE RESULTS EVALUATED BY NO-REFERENCE IMAGE QUALITY ASSESSMENT, NUMBER HIGHLIGHTED WITH RED, BLUE, AND BROWN TO INDICATE THE BEST THREE RESULTS

No-reference		Metrics	
		UCIQE ↑	UIQM ↑
Methods	GAN-RS	0.5550	**1.5206**
	Water-Net	**0.5922**	**1.3328**
	FUnIE-GAN	0.5273	1.2674
	Ucolor	0.5474	1.2230
	LCNet	0.5411	**1.3024**
	UICoE-Net	0.5477	1.2619
	LANet	**0.5813**	1.2877
	Boths	**0.5873**	1.2876

and UIEB datasets (see Table III), underwater color image quality evaluation metric (UCIQE) [25], human-visual-system-inspired underwater image quality measures (UIQM) [26] for no-reference image quality assessment in the T40, U45, and C60 datasets (see Table IV). The number in Table IV represents the average value of three datasets. Among them, the PSNR and SSIM evaluate the image similarity, and the other three full-reference metrics evaluate the discrepancy. UCIQE and UIQM can comprehensively measure the quality of underwater images. UCIQE relies on CIELab space, which assesses the color cast, blur, and low contrast. UIQM considers three attributes when measuring the image quality on the HSV model-color, sharpness, and contrast. In summary, a comparison of these metrics reveals that our method can adapt to multiple datasets, and the enhanced results are more similar to GT with less noise. The brightness, chroma, and contrast of the enhanced image are better than most models.

D. Qualitative Evaluation

As can be seen from Figs. 3 and 4, we make a full-reference qualitative performance comparison and a no-reference qualitative performance comparison. It is apparent that most enhanced images processed by SOTA and our algorithm can improve the visual effect to a certain extent. Comparing the 15 results, it can be seen that our method can better restore the image detail and balance the image color. Specifically, MAMF improves the contrast of results, but the image is prone to color cast. After ACS, TEBCF, and GAN-RS enhancement, some images have a lot of small noise and overexposure. FUnIE-GAN can not effectively remove haze in the image. Five learning-driven models (Water-Net, Ucolor, LCNet, UICoE-Net, and LANet) based on CNN are similar to our network, but their performances on some images are inferior to ours. Taken together, these results suggest that the Boths achieves the best visual effect.

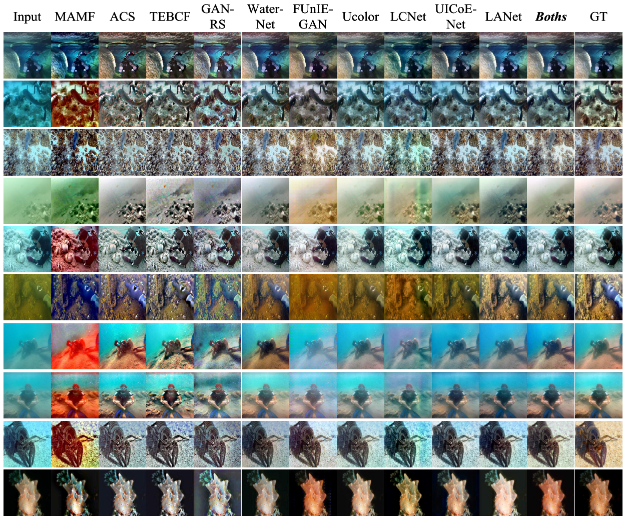

Fig. 3. Full-reference qualitative performance comparison in UVE-38K, EUVP, and UIEB datasets, GT is the ground truth.

TABLE V

QUANTITATIVE RESULTS OF THE ABLATION STUDY, NUMBER HIGHLIGHTED WITH RED, BLUE, AND BROWN TO INDICATE THE BEST THREE RESULTS

Ablation study	Baselines			
	A	B	C	***Boths***
MSE ↓	**267.3836**	**279.1683**	280.5072	**257.0951**
LPIPS ↓	**0.1529**	0.1699	**0.1533**	**0.1490**

E. Ablation Study

Ablation studies are intended to demonstrate the superiority of core components in our method. We train and test three

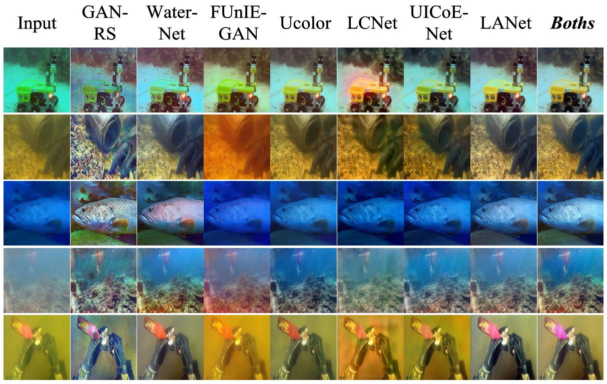

Fig. 4. No-reference qualitative performance comparison of learning-driven methods in T40, U45, and C60 datasets.

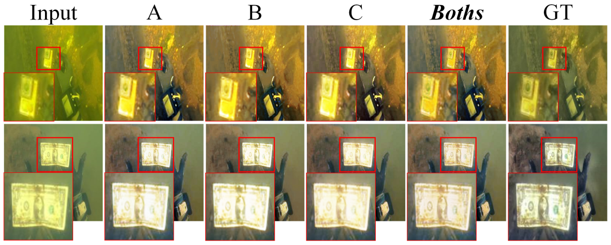

Fig. 5. Qualitative results of the ablation study, GT means ground truth. Compared to the partial view, the full-model image has superior visuals. A lacks some features, B omits high-frequency information, and C shows unclear outlines and details.

baselines: A) Without FM; B) Without WMSE loss; and C) Without SD-D + SD-I3A + FM in the UIEB dataset. Then, we use MSE and LPIPS to test. The quantitative and qualitative results obtained from the ablation study are summarized in Table V and Fig. 5. It can be seen that removing any components in Boths decreases the effect of underwater image enhancement.

IV. Conclusion

Presented in this letter is a novel method for enhancing underwater images using a super lightweight neural network, which is suitable for implementing in ROVs. It generates clear images by interacting structure and detail features, 3-D attention learning, high- and low-frequency loss functions. The parameters and FLOPs of our approach are only 0.0064 M and 0.4256 G, respectively. And a large number of quantitative and qualitative experiments comparing with SOTA methods demonstrate that our network has a surprising performance in several datasets. Therefore, the proposed method is super lightweight but extremely powerful. In future research, our Boths will be applied on a large scale to power several underwater robotic vision platforms.

Acknowledgment

The authors would like to express their gratitude to Nanfeng Jiang, Weiling Chen, Yuting Lin, Tiesong Zhao, and Chia-Wen Lin for providing their test code [15]; Jieyu Yuan, Wei Cao, Zhanchuan Cai, and Binghua Su for sharing their results [9], [10]; and Qi Qi, Yongchang Zhang, Fei Tian, Q. M. Jonathan Wu, Kunqian Li, Xin Luan, and Dalei Song for sharing their results [16]. They would also like to thank Chongyi Li for providing the code of an evaluation metric.

References

[1] D. D'Alimonte, T. Kajiyama, and A. Saptawijaya, "Ocean color remote sensing of atypical marine optical cases," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6574–6586, Nov. 2016.

[2] E. Ashford, T. L. Flanagan, N. Ashford, and E. Ashford, "Championing the future of ghost pot recovery through the implementation of remotely operated vehicles and community science models," in *Proc. OCEANS San Diego Porto*, Sep. 2021, pp. 1–4.

[3] M. Shinohara et al., "Development of a high-resolution underwater gravity measurement system installed on an autonomous underwater vehicle," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1937–1941, Dec. 2018.

[4] H. Renkewitz, S. Matz, S. Thomas, J. Schwendner, and J. Albiez, "Evaluation of a high-end laserscanner for underwater archeology," in *Proc. OCEANS San Diego Porto*, Sep. 2021, pp. 1–5.

[5] S. L. Danielson et al., "Collaborative approaches to multi-disciplinary monitoring of the Chukchi shelf marine ecosystem: Networks of networks for maintaining long-term Arctic observations," in *Proc. OCEANS Anchorage*, Sep. 2017, pp. 1–7.

[6] Z. Yan, J. Ma, J. Tian, H. Liu, J. Yu, and Y. Zhang, "A gravity gradient differential ratio method for underwater object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 4, pp. 833–837, Apr. 2014.

[7] X. Liu, Z. Gao, and B. M. Chen, "MLFcGAN: Multilevel feature fusion-based conditional GAN for underwater image color correction," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1488–1492, Sep. 2020.

[8] Y. Cho, J. Jeong, and A. Kim, "Model-assisted multiband fusion for single image enhancement and applications to robot vision," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 2822–2829, Oct. 2018.

[9] J. Yuan, W. Cao, Z. Cai, and B. Su, "An underwater image vision enhancement algorithm based on contour bougie morphology," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8117–8128, Oct. 2021.

[10] J. Yuan, Z. Cai, and W. Cao, "TEBCF: Real-world underwater image texture enhancement model based on blurriness and color fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[11] X. Chen, J. Yu, S. Kong, Z. Wu, X. Fang, and L. Wen, "Towards real-time advancement of underwater visual quality with GAN," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9350–9359, Dec. 2019.

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

[13] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3227–3234, Apr. 2020.

[14] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Trans. Image Process.*, vol. 30, pp. 4985–5000, 2021.

[15] N. Jiang, W. Chen, Y. Lin, T. Zhao, and C.-W. Lin, "Underwater image enhancement with lightweight cascaded network," *IEEE Trans. Multimedia*, vol. 24, pp. 4301–4313, 2022, doi: 10.1109/TMM.2021.3115442.

[16] Q. Qi et al., "Underwater image co-enhancement with correlation feature matching and joint learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1133–1147, Mar. 2022.

[17] F. Soto et al., "Efficient coding by midget and parasol ganglion cells in the human retina," *Neuron*, vol. 107, no. 4, pp. 656–666, Aug. 2020.

[18] S. Liu, H. Fan, S. Lin, Q. Wang, N. Ding, and Y. Tang, "Adaptive learning attention network for underwater image enhancement," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5326–5333, Apr. 2022.

[19] M. Carrasco, "Visual attention: The past 25 years," *Vis. Res.*, vol. 51, no. 13, pp. 1484–1525, 2011.

[20] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[22] H. Li, J. Li, and W. Wang, "A fusion adversarial underwater image enhancement network with a public test dataset," 2019, *arXiv:1906.06819*.

[23] T. Kurbiel and S. Khaleghian, "Training of deep neural networks based on distance measures using RMSProp," 2017, *arXiv:1708.01911*.

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

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

[26] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541–551, Jul. 2015.