Efficient Wavelet Boost Learning-Based Multi-stage Progressive Refinement Network for Underwater Image Enhancement

Fushuo Huo
School of Electrical Engineering
Chongqing University
20191102013t,zhuxuegui@cqu.edu.cn

Bingheng Li
School of Electronic Engineering
Xidian University
bhlee@stu.xidian.edu.cn

Xuegui Zhu ✉
School of Electrical Engineering
Chongqing University
zhuxuegui@cqu.edu.cn

Abstract

Raw underwater images suffer from low contrast and color cast due to wavelength-selective light scattering and attenuation. The distortions in color and luminance mainly appear at the low frequency while that in edge and texture are mainly at the high frequency. However, the hybrid distortions are difficult to simultaneously recover for existing methods, which mainly focus on the spatial domain. To tackle these issues, we propose a novel deep learning network to progressively refine underwater images by wavelet boost learning strategy (PRWNet), both in spatial and frequency domains. Specifically, the Multi-stage refinement strategy is adopted to efficiently enhance the spatial-varying degradations in a coarse-to-fine way. For each refinement procedure, Wavelet Boost Learning (WBL) unit decomposes the hierarchical features into high and low frequency and enhances them respectively by normalization and attention mechanisms. The modified boosting strategy is also adopted in WBL to further enhance the feature representations. Extensive experiments show that our method achieves state-of-the-art results. Our network is efficient and has the potential for real-world applications. The code is available at: https://github.com/huofushuo/PRWNet.

1. Introduction

Underwater image restoration is a fundamental task for improving advanced marine applications and services like underwater surveillance, image/video compression and transmission, and object detection. However, the poor visibility, blurriness, and color shifts severely degrade the quality of underwater vision. The main reason is that the light propagating through water suffers from wavelength-dependent light scattering and attenuation [2, 4]. Red light is absorbed first because of its longest wavelength, followed by green and blue light. In addition, small particles like micro phytoplankton and non-algal particulate cause light scattering. Besides, attenuation parameters are affected by different optical waters types [9]. Thus, it is difficult but vital to find an effective method to enhance underwater images.

The difficulties of underwater image enhancement may come from two folds: ***First***, spatial-varying hybrid degradations mostly in high frequency (i.e., edge and texture) exist in underwater images. ***Second***, diverse water types [23] show the different distortion representations mostly in low frequency (i.e., color and luminance) [33, 9, 3, 28, 22]. Some researchers propose underwater image enhancement methods [7, 6, 15, 36, 31, 14] while ignore the influence of diverse water types. [33] employs a simple classifier to distinguish water types to facilitate the enhancement procedure. [3] enhances the images with a modified physical model. [9] restores the color of underwater images by considering multiple spectral profiles of different water types. [28] trains the network respectively based on different water types defined by [23]. [22] treats the underwater image enhancement as exemplar-based image style transformation. These methods can eliminate the influence of water types to some extent but could yield suboptimal results in hybrid degradations.

To handle the two problems at the same time, we propose a multi-stage progressive refinement network based on wavelet boost learning (PRWNet), enhancing underwater images both in spatial and frequency domains. Specifically, we propose the multi-stage refinement network. Each

stage is based on the U-shaped architecture to learn multi-scale contextual information. The enhanced features from previous stages are further refined across stages. Based on the observations that the hybrid degradations most arise in high-frequency distortions (i.e., edge and texture), while the diversity of water types is mainly reflected in low frequency (i.e., color and luminance). We decompose high and low frequency of features via wavelet transform and respectively enhance them via novel wavelet boost learning units. For the high frequency, we enhance it with residual learning and attention mechanisms. Meanwhile, we regard the restoration of low frequency as an implicit style transfer problem. Normalization and attention mechanisms are applied to adaptively discriminate and cope with different water types. The modified boost strategy is proposed to further improve the performance. Also, PRWNet is efficient with $6.3M$ parameters and 40+ Frames Per Second (fps). It has the potential for real-world applications which have limited computation resources.

In summary, our contributions can be summarized as follows:

1) We propose PRWNet that progressively refines underwater images both in spatial and frequency domains. The multi-stage progressive refinement network is proposed to acquire rich context and precise spatial features. Wavelet boost learning unit is effective to enhance the images with high fidelity in the frequency domain.

2) We disentangle the features in the frequency domain and respectively enhance them through efficient enhancement methods. The novel disentangling strategy facilitate the network to simultaneously enhance underwater images from the hybrid degradations and influences of water types.

3) PRWNet achieves state-of-the-art results among the popular underwater image datasets. Ablation studies show the effectiveness of each module of our network.

2. Related Work

In this section, we discuss the related work about underwater image enhancement. Previously methods are divided into physical model-free and physical model-based types. Physical model-free methods aim to modify image pixel values to improve the visibility, such as multi-scale fusion [8, 7, 6], variational optimization [15], and pixel distribution adjustment [1]. However, Physical model-free methods omit the underwater imaging mechanism thus may produce unpleasant artifacts due to the complex underwater environment. Physical model-based methods [13, 30, 16, 37, 36] regard single underwater image enhancement as an ill-posed inverse problem and estimate the parameters of the underwater image formation model by handcrafts priors. The priors include underwater dark channel prior [13], red channel prior [16], minimum information prior [30], *etc*. These methods take light scattering and attenuation into consideration and achieve promising results. Nevertheless, statistical priors may fail in challenging underwater conditions and the underwater image formation model could yields the error due to diverse scene properties [2, 5]. Akkaynak and Treibitz [2] proposed a revised underwater image formation model which is physically accurate. Based on the revised model, the new underwater image color correction method was proposed based on RGB-D image pairs [3].

Recently, deep learning-based methods have shown remarkable improvements in underwater image enhancement[31, 14, 33, 22, 29, 28, 20, 27]. Due to lacking underwater images and the corresponding clean image pairs, Generative Adversarial Network (GAN) is employed in previous work to synthesize underwater image datasets or conducting unpair learning. Li *et al*. [31] firstly employed Generative Adversarial Network (GAN) to synthesize degraded images and proposed a two-stage refinement network. [20] formulated a multi-modal objective function to improve the perceptual quality of underwater images. [22] regarded underwater image enhancement as the exemplar-based style transfer and utilized wavelet transforms to better reconstruct the signal. Uplavikar *et al*. [33] introduced a simple classifier to make the GAN model more discriminative for diverse water types. To deal with the problem of lacking unpaired datasets for supervised learning, [28] simulated the realistic underwater images based on an underwater imaging physical model and 10 different water types [23]. Then they proposed light-weight CNN models trained on ten water types, respectively. Underwater Image Enhancement Benchmark (UIEB) [29] based on real-world underwater image pairs was proposed to train and evaluate the underwater image enhancement methods. [29] further proposed a gate fusion network to enhance underwater images by fusing three enhanced inputs. Recently, [27] proposed a multi-color space embedding network uniting the advantages of the physical model to deal with color cast and low contrast.

Apart from spatial-varying hybrid degradations, the influences of different water types have been considered by recent underwater image enhancement methods [2, 3, 9, 28, 33]. However, [2, 3] utilizes the RGB-D pairs which need extra devices. [9] consumes much computing resources. [28] proposed the network respectively trained on ten water types but depend on the prior knowledge of the water type for the given images. [33] can learn water-type agnostic features but sometimes produces unstable results due to the disadvantage of the GAN. It is not easy for a single method to enhance underwater images for such multiple image degradations. In this paper, based on the observations that diverse water types are mainly related to color cast and illumination which is low frequency, we disentangle the degradations and deal with them separately. Also,

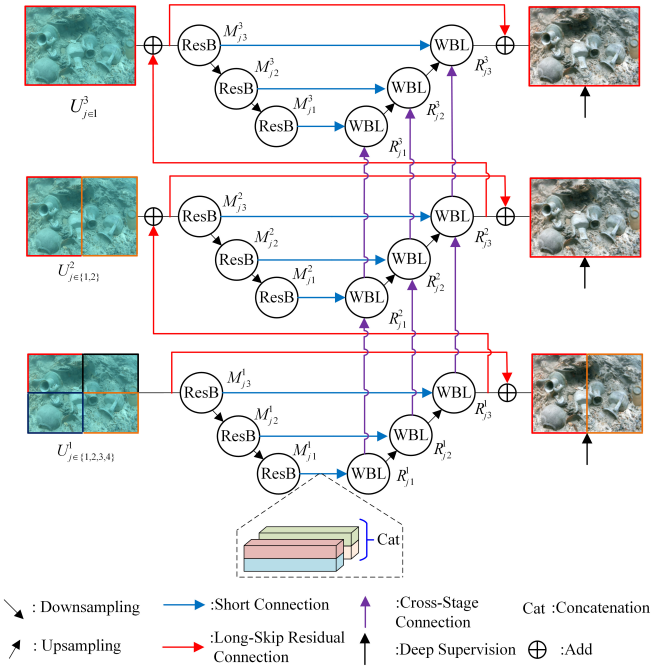


Figure 1. The framework of our network. Earlier stages utilize encoder-decoder structure to extract multi-scale contextualized features while the last stage operates at the original image resolution to generate spatially accurate outputs. At earlier stages, the spatial dimensions of features from encoders and decoders are not the same. They are aligned by the Concatenation operator similar to the dashed area.

multi-stage progressive refinement network is proposed to enhance spatial-varying hybrid degradations. In this way, our network has more generalization ability compared to these methods.

3. Proposed Method

In this section, we provide detailed descriptions of the multi-stage refinement network architecture. Then, we introduce the wavelet boost learning strategy. Finally, hybrid loss functions are presented.

3.1. Multi-stage Refinement Network Architecture

As the degradations appear on the underwater images are hybrid and spatial-varying. As shown in Figure 1, we adopt the multi-stage network inspired by [44] to capture multi-scale contextual information and enhance images in a coarse-to-fine way. Different from [44], we employ long-skip residual connection at each stage so that we can imply deep supervision to make the training strategy stable and converge fast. Also, the enhanced features decoders are densely connected to the next stage via Cross-Stage Connection, further boosting the feature information progressively. We discuss the architecture in detail below.

Each stage of the network is composed of U-shaped architecture. Each encoder block consists of one Residual Block [18] (ResB), which preserves the data fidelity and address the gradient vanishing. Downsampling is conducted by 3×3 convolution with the stride of 2 and the number of channels after downsampling doubles. Each decoder block consists of one Wavelet Boost Learning (WBL) unit. Upsampling is conducted by 3×3 transpose convolution with the stride of 2 and the number of channels halves. ***At each stage***, Short Connections facilitate reducing the information flow loss during downsampling and upsampling. Long-skip residual connection is added at each stage to apply deep supervision. ***Between adjacent stages***, residual information is added to the beginning of the next stage to initially refine the input. The features refined by WBL are also densely fed to the corresponding WBL at the next stage via Cross-Stage Connection, enriching the feature progressively. By exchanging information flows at each stage and between stages, the network integrates multi-scale information and refines the underwater images in a coarse-to-fine manner.

Specifically, As shown in Figure 1, the input is divided into four, two, one patch from bottom to top stage respectively. For an input underwater image U_j^i, j and i represent the j-th patch and the i-th stage. $ResB_k^i$ and WBL_k^i represent encoder and decoder, where k means the k-th block. $ResB_k^i$ and WBL_k^i share the same parameters for different input patches. Moreover, M_{jk}^i and R_{jk}^i mean the output features from encoder and decoder, respectively. Considering the 2-nd stage for an example, the 2 divided inputs, U_1^2 and U_2^2, are fed to the encoder $ResB_k^2$:

$$\begin{aligned} M_{j3}^2 &= ResB_3^2(U_j^2),\ j \in \{1,2\} \\ M_{j2}^2 &= ResB_2^2(D(M_{j3}^2)),\ j \in \{1,2\} \\ M_{j1}^2 &= ResB_1^2(D(M_{j2}^2)),\ j \in \{1,2\} \end{aligned} \tag{1}$$

where D means the Downsampling operation. Then we concatenate adjacent features from each encoder block to align spatial dimensions:

$$M_k^2 = \text{Cat}\left(M_{1k}^2, M_{2k}^2\right),\ k \in \{1,2,3\} \tag{2}$$

where Cat means concatenation operation. For the k-th block of decoder WBL_k^2, the refine process is as:

$$R_k^2 = \begin{cases} WBL_k^2\left(R_k^1, U\left(R_{k-1}^2\right), M_k^2\right), k \in \{2,3\} \\ WBL_k^2\left(R_k^1, M_k^2\right), k = 1 \end{cases} \tag{3}$$

where R_k^1 means refined features from WBL_k^1, R_{k-1}^2 means refined features from WBL_{k-1}^2, U represents the Upsampling operation, and R_k^2 is the enhancement results of these aggregation features.

3.2. Wavelet Boost Learning Unit

In this subsection, as shown in Figure 2, we introduce the Wavelet Boost Learning (WBL) unit. We give the detailed

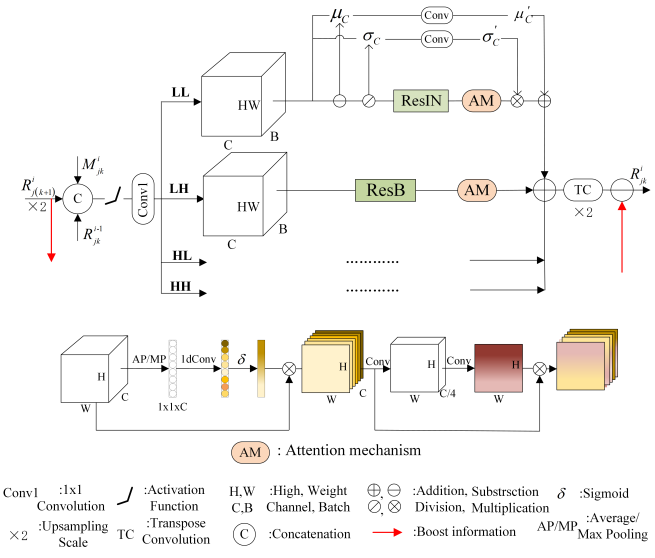


Figure 2. The detailed schematic illustration of Wavelet Boost Learning unit. High-Frequency Subbands (LH, LH, and HH) are enhanced by the same module as LH.

description on the wavelet transform, the enhancement strategy on the high and low subbands, and boost strategy.

Wavelet Transform: Wavelet Transforms (WT) have been employed to augment regular deep learning networks in low-level computer vision tasks. Inspired by photo-realistic style transfer [43], [22] leveraged WT to reduce noise amplification for exemplar-based underwater image enhancement. Guo *et al.* [17] proposed a deep wavelet super-resolution network to recover missing details on frequency subbands. [32] further employed multi-level WT to enlarge receptive field without information loss. Considering the frequency characteristics of hybrid degradations, WT is used in this paper to decompose the features in the frequency domain. Note that we reconstruct the frequency subbands by addition operator and transpose convolution while do not use Inverse Wavelet Transform (IWT). Experiments in ablation studies show our reconstruction process achieves comparable results while reduces the training and inference time. Specifically, The 2D fast WT [34] is used to calculate Haar wavelets. As shown in Figure 3, for the pixels in a 2×2 patch (denoted as a, b, c, and d), the calculation process of 2D Haar wavelet coefficients are defined as:

$$\begin{aligned} A &= (a+b+c+d)/4 \\ B &= (a-b+c-d)/4 \\ C &= (a+b-c-d)/4 \\ D &= (a-b-c+d)/4 \end{aligned} \tag{4}$$

A contains low-frequency information. B, C, and D represent the high-frequency in horizontal, vertical, and diagonal orientation. The height and width of the decomposed

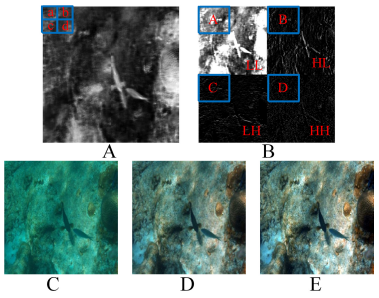


Figure 3. A visual example of Wavelet Transform. A is the average values of feature map from the encoder (M_{j3}^{1}). We normalize the values of A to the range of [0,1] for visualization. LL is the low-frequency subband. HL, LH, and HH are the high-frequency subbands. C, D, and E are the underwater image, the enhanced image, and the reference image, respectively.

bands are half of the original image size. Then we enhance high-frequency subbands (HL, LH, and HH) and the low-frequency subband (LL) respectively, as shown in Figure 2.

High-Frequency Subbands: The High-frequency signal mainly contains texture and edge information, as shown in Figure 3B, which is mainly degraded by haze, blurriness, and noise. [22] extracts high-frequency information and connects to the decoder to preserve edge-like information. [20] enforces the high-frequency information consistency in adversarial fashion via Markovian Patch-GANs [21]. We employ one residual block (ResB) with Attention Module (AM) to enhance high-frequency subbands. Concretely, AM consists of channel attention (CA) [41] model and spatial attention (SA) [38] model. As the high-frequency information is sparse, we replace the average-pooling with max-pooling to emphasize the important channels. Then the SA makes the network pay more attention to spatial informative features.

Low-Frequency Subband: We consider the elimination of the effects of diverse water types (mostly in color and illumination) as an implicit style transfer in low frequency. We propose a novel low-frequency enhancement branch based on normalization and attention mechanisms. Normalization schemes (mean and standard deviation) have widely used in style transfer tasks [40, 19, 26]. Ulyanov *et al.*[40] firstly proposed instance normalization (IN) to style transfer due to its invariance to the contrast of the content in the spatial space. [19] transferred an image to an arbitrary style via an adaptive instance normalization layer, which aligns the mean and variance of the content features with those of the style features. Li *et al.*[26] proposed a position normalization (PN) in the channel space roughly capture style and shape information of an image. Recently, [22] regarded underwater enhancement as a photo-realistic style transfer problem. However, they rely on a high-quality underwater image and transfer the image online, which may hinder real-world applications. Inspired by these previous work-

s, we imply the IN and PN to adaptively adjust the learned mean and standard deviation of inputs, both on spatial and channel space. The Attention Module (AM) [41, 38] is applied to emphasize the important information. Concretely, PN and IN mechanisms are in the form of:

$$PN/IN = \sigma^{'}\left(\frac{LL-\mu}{\sigma}\right)+\mu^{'} \tag{5}$$

where μ and σ are mean and standard deviation of feature statistics. $\mu^{'}$ and $\sigma^{'}$ are affine parameters learned from the data. For each LL subband (LL), we firstly extract the mean ($\mu_C = 1/C\sum_{c=1}^{C} LL_{B,C,H,W}$) and standard deviation ($\sigma_C = \sqrt{1/C\sum_{c=1}^{C}(LL_{B,C,H,W}-\mu_C)^2}$) across channels. Then the normalized LL are fed to the ResIN block, which consists of Conv, IN, active function, and residual connection. IN discards the extracted spatial statistics (i.e., μ_{HW} and σ_{HW}). As Equation (5), the learned $\mu^{'}_{HW}$ and $\sigma^{'}_{HW}$ eliminate the influence of diverse water types across spatial space by affine transformation. Then AM further discriminates the importance of features. Finally, the channel-wise information of the features are adjusted by $\mu^{'}_{C}$ and $\sigma^{'}_{C}$, which are learned by vanilla convolutions.

Boost Strategy: To further enhance the enhanced information, inspired by Strengthen-Operate-Subtract (SOS) boosting strategy [39, 12], we propose a simple boosting strategy. For WBL unit in the k-th block at the i-th stage, the boost strategy is defined as:

$$R^i_{jk} = \mathrm{WL}\left[C_{1\times1}\left(\varphi\left(\mathrm{Cat}\left(R^{i-1}_{jk}, R^i_{j(k+1)}, M^i_{jk}\right)\right)\right)\right] - R^i_{j(k+1)} \tag{6}$$

when k=3, Equation (6) does not have $R^i_{j(k+1)}$. Cat, WL, and φ mean the concatenation operation, wavelet learning module, and activate function, respectively. The channels of concatenated features are reduced to the original number by 1×1 convolution. We concatenate the features and emphasize the informative features via active function, while do not simply use the addition operator like [12]. The ablation studies show that our method improves 0.3 PSNR compared to [12].

3.3. Hybrid Loss Functions

Our hybrid loss functions are defined as the summation of the overall outputs:

$$L = \sum_{k=1}^{3} \alpha_k l^k \tag{7}$$

where l^k is the loss of the k-th output of the stage. α_k denotes the weight of each loss. In this experiment, we define α_k as 1 without fine-tuning. l^k consists of three loss functions as:

$$l^k = l^k_{pix} + \alpha l^k_{edge} + \beta l^l_{per} \tag{8}$$

Table 1. Qualitative Comparisons on T-1000 and T-90

	T-1000		T-90	
Methods	PSNR	SSIM	PSNR	SSIM
Inputs	12.96	0.61	16.11	0.72
Aucuti *et al.* [6]	13.27	0.78	19.19	0.72
Peng *et al.* [37]	13.04	0.65	15.77	0.79
Berman *et al.* [9]	15.09	0.71	17.41	0.77
AIO-GAN [33]	11.81	0.41	10.46	0.27
FUnIE-GAN [20]	14.83	0.73	16.97	0.73
WaterNet [29]	15.47	0.83	19.81	0.85
UWCNN [28]	19.14	0.80	18.14	0.77
Ucolor [27]	**23.05**	**0.86**	**20.63**	**0.87**
PRWNet	**24.37**	**0.90**	**21.72**	**0.89**

Table 2. Qualitative Comparisons on T-90, T-60, and SQUID

	T-90	T-60	SQUID
Methods	UIQM[35]		
Inputs	2.57	2.08	0.77
Aucuti *et al.* [6]	2.99	2.76	1.75
Peng *et al.* [37]	2.41	2.06	1.03
Berman *et al.* [9]	2.47	2.02	1.43
AIO-GAN [33]	3.13	**2.88**	**2.94**
FUnIE-GAN [20]	**3.20**	**3.01**	1.92
WaterNet [29]	3.04	2.64	2.30
UWCNN [28]	2.90	2.45	2.01
Ucolor [27]	3.15	2.75	2.21
PRWNet	**3.18**	2.80	**2.36**

where l^k_{pix}, l^k_{edge}, and l^l_{per} denote pixel loss [10], edge loss, and perceptual loss [24], respectively. Especially, pixel loss helps the network generate the enhanced image close to the ground truth in the pixel level. The pixel loss is defined as:

$$l_{pix} = \sqrt{\|R-G\|^2+\varepsilon^2} \tag{9}$$

where R and G denotes the enhanced images and ground truth image. ε is set to 10^{-3}. The edge loss is defined as:

$$l_{edg} = \sqrt{\|\Delta R-\Delta G\|^2+\varepsilon^2} \tag{10}$$

where Δ denotes the laplacian operator. In addition, l_{per} is the perceptual loss given by:

$$l_{per} = \|\phi(R)-\phi(G)\|_2 \tag{11}$$

ϕ denote relu1_2, relu2_2, and relu3_3 layers of the VGG-16 network pre-trained on the ImageNet. In this paper, we set α and β as 0.05 and 0.01.

4. Experiments

In this section, we first introduce the experiment settings, then compare our method with other state-of-the-art methods qualitatively and quantitatively. The ablation studies are conducted to validate the effectiveness of each module of our network.

4.1. Experiment Settings

To train our network, we adopt the same training set as [29, 28, 27], which is composed of 800 pairs of underwater images from [29] and 1250 synthetic pairs (consisting of

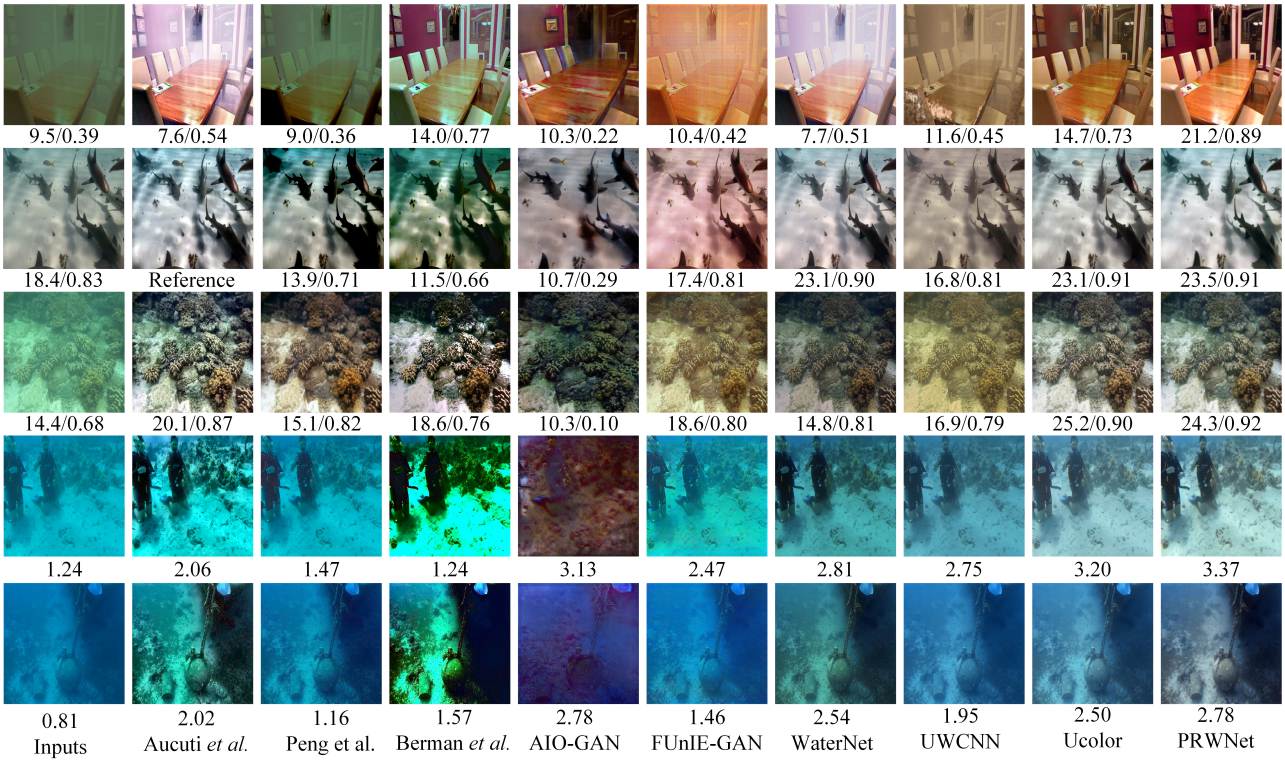


Figure 4. Quantitative Comparisons on typical underwater images. The Row 1 is a synthesized image. Row 2, 3, 4, and 5 are real-world images. The values of PSNR/SSIM or UIQM are marked below for reference.

10 water types) from [28]. We deploy the PyTorch framework with a single NVIDIA 1080Ti GPU. During the training phase, the initial learning rate is set as 10^{-4} and it is decayed by a cosine learning rate scheduler. Adam [25] with default settings are set as our optimizer. The network is trained for 60K iterations with a batch size of 16. The inference time of one 256x256 image is $0.022s$ and The network is with $6.3M$ parameters.

4.2. Comparisons to SOTAs

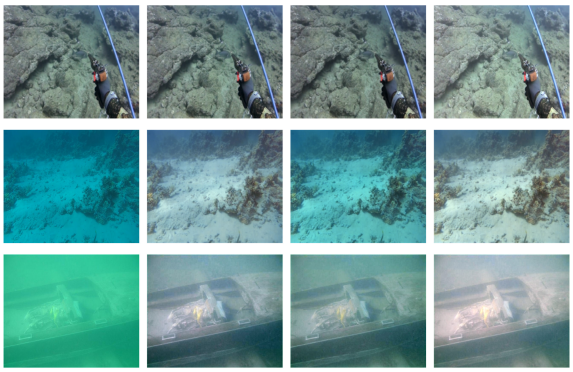
We compare our method with 8 state-of-the-art (SOTA) methods, including one physical model-free method (Ancuti *et al.* [6]), two physical model-based methods (Peng *et al.* [37] and Berman *et al.* [9], two GAN-based methods (AIO-GAN [33] and FUnIE-GAN [20]), three CNN-based methods (WaterNet [29], UWCNN [28], and Ucolor [27])

Qualitative Comparison: We consider two full-reference image quality assessment (IQA) metrics including PSNR and SSIM [42]. To evaluate the real-world images, no-reference underwater IQA metrics, UIQM [35], are utilized to conduct a comprehensive evaluation. For all the three metrics, the higher score means better image quality. 1000 synthesized images in 10 water types (T-1000) [28], 90 natural underwater images with professional- generated reference images (T-90) [29], 60 real-world images without reference images (T-60), and 16 representative images presented on the project page of SQUID [9] are used as our test datasets. As we can see from Table 1, PRWNet achieves the best scores between synthesized and real-world images in different water types. Meanwhile, our network only uses $6.3M$ parameters compared with $38.8M$ of the second-best method, which owes to the efficient progressive refinement strategy. As for real-world datasets, our methods also achieve competitive results. Interestingly, the results of the no-reference IQA metric UIQM are not consistent with the full-reference IQA metrics. The results of AIO-GAN in Figure 4 show examples of the inconsistency. It may because the UIQM metric is somewhat heuristic and has limited applicability [9, 29, 27].

Quantitative Comparison: To visualize the enhancement results of our method, we conduct visual comparisons on five images from the T-1000, T-90, T-60, and SQUID. From Row 1, our method can restore the bright red color of the desk. As for different water types, our method can accurately discriminate and faithfully enhance the images. The edges and textures in Row 3, 4 are appropriately enhanced while the physical model-based (i.e., Peng *et al.* [37] and Berman *et al.* [9]) and GAN-based methods (i.e., AIO-GAN [33] and FUnIE-GAN [20]) tend to over-enhance the details. Our method can also achieve visually pleasant re-

Table 3. Ablation studies

	Network Architecture			Wavelet Learning						Boost Strategy		
	Plain [44]	w/ CSC	w/ DS	w/o WL	w/ OC	w/o AM	w/o IN	w/o PN	w/ IWT	w/o BS	w/ addition	our
PSNR	20.59	21.32	21.03	21.51	24.12	22.93	23.13	23.73	24.39	23.86	24.02	**24.37**
SSIM	0.78	0.81	0.80	0.83	0.88	0.86	0.86	0.88	0.89	0.87	0.88	**0.90**



A. Input B. w/o PN C. w/o IN D. Result

Figure 5. Visual examples of normalization mechanisms. A is the original input. B is the result of the network without PN. C is the result of the network without IN. D is the result of our network.

sults even for the seriously degraded image in Row 3, 4, and 5.

4.3. Ablation Studies and Analysis

In Table 3, we analyze the contributions of each module in our network based T-1000 dataset. We divide the ablation studies into three parts: Network Architecture, Wavelet Learning (WL), and Boost Strategy (BS). We adopt the same training settings as mentioned.

Network Architecture: Apart from the plain network architecture from [44], we add Cross-Stage Connections (w/ CSC) between adjacent stages. We also deploy the long-skip residual connection to conduct Deep Supervision (w/ DS) at each stage. CSC improves the performances a lot because the enhanced information is progressively refined across the stages. DS also improves the performances of plain [44]. These results validate the effectiveness of our new architecture.

Wavelet Learning (WL): To validate the effectiveness of the Wavelet Learning (WL) strategy, we firstly replace the WL with ResBlock (w/o WL). We also deploy Octave Convolution (OC) operation [11] (w/ OC) to disentangle features into high and low-frequency subbands, validating the superiority of the Haar wavelet decomposition. Then we further study enhancement methods in frequency subbands. We ablate the Attention Mechanism (AM) (w/o AM), Instance Normalization (IN) (w/o IN), and Position Normalization (PN) (w/o PN), respectively. Besides, Inverse Wavelet Transform (w/ IWT) is employed to reconstruct the different subbands. Table 3 shows that each component of WL improves the performances of PRWNet. We can observe that WL improves the network by almost 3 PSNR in total. Normalization and attention mechanisms play an important role in WL. As Haar wavelet extract contextual high-frequency features from three orientations (i.e., HL, LH, and HH), the performances of (w/ OC) are slightly worse than ours. As shown in Figure 5, Normalization schemes help to adaptively eliminate the influence of diverse water types via learned mean and standard deviation parameters. Besides, reconstruction subbands by IWT does not significantly boost the performance while increases the computational burden (i.e., the training time increases almost 50% and the inference time increases to $0.038s$).

Boost Strategy (BS): We ablate the BS (w/o BS) and also adopt the strategy as [12] (w/addition). The ablation studies show that the BS is effective and the BS outperforms [12] (w/addition) both in PSNR and SSIM. These two experiments validate the effectiveness of the proposed BS.

5. Conclusions

In this paper, we propose a novel network named PRWNet to enhance underwater images. The main contributions are that we not only explore the spatial-varying information but also disentangle the degradations and enhance them separately in the frequency domain. Concretely, we propose the novel multi-stage network to progressively refine the hybrid degradations. To eliminate the effect of diverse water types and enhance the detail simultaneously, we decompose the features via wavelet transform. Then we imply normalization and attention mechanisms to enhance them separately. Comprehensive experiments show the PRWNet achieves the SOTA results. The network is efficient ($6.3M$ parameters and $40+fps$) and has the potential for real-world tasks.

References

[1] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Underwater image quality enhancement through integrated color model with rayleigh distribution. *Applied Soft Computing*, 27:219–230, 2015. 2

[2] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6723–6732, 2018. 1, 2

[3] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691, 2019. 1, 2

[4] Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. What is the space of attenuation coefficients in underwater computer vision? In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2017. 1

[5] Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. What is the space of attenuation coefficients in underwater computer vision? In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2017. 2

[6] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–88, 2012. 1, 2, 5, 6

[7] Codruta O. Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, and Philippe Bekaert. Color balance and fusion for underwater image enhancement. *IEEE Transactions on Image Processing*, 27(1):379–393, 2018. 1, 2

[8] Codruta Orniana Ancuti, Cosmin Ancuti, Tom Haber, and Philippe Bekaert. Fusion-based restoration of the underwater images. In *2011 18th IEEE International Conference on Image Processing*, pages 1557–1560, 2011. 2

[9] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 2, 5, 6

[10] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172 vol.2, 1994. 5

[11] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Yan Shuicheng, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3434–3443, 2019. 7

[12] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2154–2164, 2020. 5, 7

[13] Paulo L.J. Drews, Erickson R. Nascimento, Silvia S.C. Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE Computer Graphics and Applications*, 36(2):24–35, 2016. 2

[14] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165, 2018. 1, 2

[15] Xueyang Fu, Peixian Zhuang, Yue Huang, Yinghao Liao, Xiao-Ping Zhang, and Xinghao Ding. A retinex-based enhancing approach for single underwater image. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4572–4576, 2014. 1, 2

[16] Adrian Galdran, David Pardo, Artzai Picn, and Aitor Alvarez-Gila. Automatic red-channel underwater image restoration. *Journal of Visual Communication and Image Representation*, 26:132–145, 2015. 2

[17] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1100–1109, 2017. 4

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 4

[20] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020. 2, 4, 5, 6

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 4

[22] Adarsh Jamadandi and Uma Mudenagudi. Exemplar-based underwater image enhancement augmented by wavelet corrected transforms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1, 2, 4

[23] N. G. Jerlov. Marine optics, 231 pp. 1976. 1, 2

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 5

[25] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[26] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pages 1620–1632, 2019. 4

[27] Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, 30:4985–5000, 2021. 2, 5, 6

[28] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020. 1, 2, 5, 6

[29] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020. 2, 5, 6

[30] Chong-Yi Li, Ji-Chang Guo, Run-Min Cong, Yan-Wei Pang, and Bo Wang. Underwater image enhancement by de-

hazing with minimum information loss and histogram distribution prior. *IEEE Transactions on Image Processing*, 25(12):5664–5677, 2016. 2

[31] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation Letters*, 3(1):387–394, 2018. 1, 2

[32] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 886–88609, 2018. 4

[33] Pritish M Uplavikar, Zhenyu Wu, and Zhangyang Wang. All-in-one underwater image enhancement using domain-adversarial learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1, 2, 5, 6

[34] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 4

[35] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2016. 5, 6

[36] Yan-Tsung Peng, Keming Cao, and Pamela C. Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, 27(6):2856–2868, 2018. 1, 2

[37] Yan-Tsung Peng and Pamela C. Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE Transactions on Image Processing*, 26(4):1579–1594, 2017. 2, 5, 6

[38] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:11908–11915, 04 2020. 4, 5

[39] Yaniv Romano and Michael Elad. Boosting of image denoising algorithms. *SIAM Journal on Imaging Sciences*, 8:1187–1219, 06 2015. 5

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. 4

[41] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539, 2020. 4, 5

[42] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[43] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9035–9044, 2019. 4

[44] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5971–5979, 2019. 3, 7