Elsevier required licence: © 2023.

This manuscript version is made available
Under the CC-BY-NC-ND 4.0 license:

http://creativecommons.org/licenses/by-nc-nd/4.0/

The definitive publisher version is available online at:

https://www.sciencedirect.com/science/article/pii/S0923596522001692?via%3Dihub

Cross-Domain Learning for Underwater Image Enhancement

Fei Li[a,*,1], Jiangbin Zheng[a,b], Yuan-fang Zhang[b,c,1], Wenjing Jia[c], Qianru Wei[b] and Xiangjian He[c]

[a]School of Microelectronics and Software, Northwestern Polytechnical University, P. R. China
[b]School of Computer, Northwestern Polytechnical University, P.R.China
[c]Faculty of Engineering and IT, University of Technology Sydney, Australia

ARTICLE INFO

Keywords:
Unsupervised learning, underwater image enhancement, GAN, loss function

ABSTRACT

The poor quality of underwater images has become a widely-known cause affecting the performance of the underwater development projects, including mineral exploitation, driving photography, and navigation for autonomous underwater vehicles. In recent years, deep learning-based techniques have achieved remarkable successes in image restoration and enhancement tasks. However, the limited availability of paired training data (underwater images and their corresponding clear images) and the requirement for vivid colour correction remain challenging for underwater image enhancement, as almost all learning-based methods require paired data for training. In this study, instead of creating the time-consuming paired data, we explore the unsupervised training strategy. Specifically, we introduce a universal cross-domain GAN-based framework to generate high-quality images to address the dependence on paired training data. To ensure the vivid colourfulness, the colour loss is designed to constrain the training process. Also, a feature fusion module (FFM) is proposed to increase the capacity of the whole model as well as the dual discriminator channel adopted in the architecture. Extensive quantitative and perceptual experiments show that our approach overcomes the limitation of paired data and obtains superior performance over the state-of-the-art on several underwater benchmarks in terms of both accuracy and model deployment.

1. Introduction

Recently, underwater image enhancement, which enables intelligent frameworks to acquire and understand underwater information thoroughly, has gradually played a more and more significant role in various real-world applications (*e.g.*, underwater mining, fish detection and underwater archaeology). High-quality images provide essential information that enables the underwater systems to work correctly. Unfortunately, suffering from the severe effects of low visibility, light refraction, absorption, and scattering [19, 30, 54], underwater images mostly witness the obvious quality degradation issues such as color distortion, low contrast, low light and blurring, which brings huge challenges for underwater applications. Naturally, the wavelength related attenuation results in color distortion, as water absorbs red wavelengths quickly, which is followed by the green and blue lights (the wavelengths of the red, green and blue lights are 600nm, 525nm and 475nm, respectively). Thus, underwater images tend to have a green or blue hue. The deeper the diving, the more red light is absorbed and the more devastating the degradation. What is even worse, this distortion is extremely non-linear.

Underwater vision has several specialties, which refers to the sensation that the retina is stimulated by images of underwater objects. The main problems are: (1) Low light vision problem, which is caused by insufficient light; (2) High myopia, due to that the refractive power of water is 3/4 of that of air, the refractive ability of human eyes loses by 2/3; (3) Visual field limitation and glass refractive effect. (4) Water turbidity and chromaticity, which is caused by light transmission of different wavelengths. (5) The illusion of empty vision, which is formed by the loss of structure of underwater vision and can be divided into empty vision myopia, empty vision action and empty vision discoloration, *etc*.

To address the aforementioned issues of underwater image enhancement, a number of approaches for improving the underwater image quality have been proposed. During the past decade, many approaches [41, 8, 12, 31, 55, 2] have attempted to resolve the enhancement issue by specially designed hand-crafted filters to improve the contrast and lightness. These physical models are based on the statistical theory of image dehazing, except that the medium attenuation coefficient is a wavelength dependent, whereas in dehazing it does not depend on the light wavelength.

In contrast, learning based methods such as Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN) based methods [20, 29, 52] have made great progress in underwater image enhancement. Besides, several studies have achieved unsupervised learning by supplying extra data, such as depths [16] and water type [45]. By contrast, these learning-based studies explore the numerous training sets and the robust feature representation capacity of network structures. However, these methods cannot outperform the conventional methods completely, especially in several unique underwater scenes.

One of the challenges in underwater image enhancement is the lack of the paired training data, which consist of low-quality underwater images and their corresponding ground truth (GT) images of better illumination. Moreover, the underwater environment is influenced not only by light absorbing and scattering, but also by water type, water depth and the time of the day, which means it can be extremely time-

*Corresponding author
lflovelife@mail.nwpu.edu.cn (F. Li); zhengjb@nwpu.edu.cn (J. Zheng); zhengjb@nwpu.edu.cn (Y. Zhang); Wenjing.Jia@uts.edu.au (W. Jia); weiqianru@nwpu.edu.cn (Q. Wei); Xiangjian.He@uts.edu.au (X. He)
ORCID(s): 0000-0001-7511-2910 (F. Li)
[1]These authors have contributed equally to this work.

(a) Source Domain X (b) Source Domain Y

Figure 1: Visual illustration of the underwater domain X (a) and its corresponding domain Y (b).

consuming to collect a suitable large-scale underwater dataset. To address this issue, synthetic methods have been proposed. For example, Anwar *et al.* [5] proposed a synthetic method based on the underwater image formation model [11] and created a synthetic dataset of degraded images and their GT. However, the models trained by synthetic datasets lack the real feature representation ability of real underwater scenes and hence experience difficulties when dealing with real, degraded underwater images. Therefore, exploiting solutions addressing the dependence on paired training data is of great significance.

Another more challenging issue is how to retain the vivid color accurately. The current studies [8, 31, 7] all have indicated the significance of color correction in underwater image enhancement. However, the existing learning-based methods have not focused on this issue but simply regarded the whole enhancement process as an end-to-end pixel mapping problem. These methods did not consider whether the recovered color satisfied real situations.

In this work, we develop an unsupervised cross-domain conditional generative adversarial network (CDGAN) to address the dependence of existing learning based methods on paired training data. Our method, inspired by the CycleGANs [57] and EnlightenGANs[22], attempts to build an enhancing model via exploring two unpaired domains (X and Y as shown in Fig. 1) instead of using synthetic data and images with lower diversity of objects. Also, small datasets cannot meet the requirement of advanced learning methods on large training data. According to the less constraint functions and the instability in training unsupervised GANs, we employ several innovative techniques. Firstly, we design a dual-channel discriminator to establish the connection between the local patches and the global content. Besides, we extract the edge information of the underwater images as a self-regularized attention map in each level of deep features to guide the unsupervised learning. Meanwhile, the color cast is an essential feature for enhancement, so we propose an underwater colorfulness loss function to regularize the training process.

In this paper, we propose an unsupervised Cross-Domain Conditional Generative Adversarial Network (CDGAN) to effectively address the limited availability of paired training data. Our method, inspired by CycleGANs [57] and EnlightenGANs [22], attempts to build an enhancing model via two unpaired domains X and Y, as shown in Fig. 1 , instead of synthesizing data and images of objects with low diversity. Also, a small dataset cannot meet the requirement of learning based methods on large training datasets. Considering the weakly constrained function and the instability in training unsupervised GANs, we develop several innovative techniques. Firstly, we design a dual-channel discriminator to establish the connection between the local patches and the global content. Secondly, we extract the edge information in the input underwater images as a self-regularized attention map in each level of the deep features to guide the unsupervised learning. Last but not the least, the color cast is an essential feature for enhancement. Therefore, we propose a loss function of underwater colorfulness to regularize the training process.

The contributions of this paper are as follows:

- We propose an unsupervised CDGAN for underwater image enhancement, which, to the best of our knowledge, is the first method utilizing unpaired datasets to address the underwater image restoration. This training strategy removes the dependency on paired training data and includes more comprehensive images from different domains.
- Our method introduces the color distance loss, which is a non-reference loss similar to the perceptual loss, to measure the distance between the underwater images and the enhanced images in color space.
- Our proposed method can improve the contrast and recover the image color and achieve the state-of-art

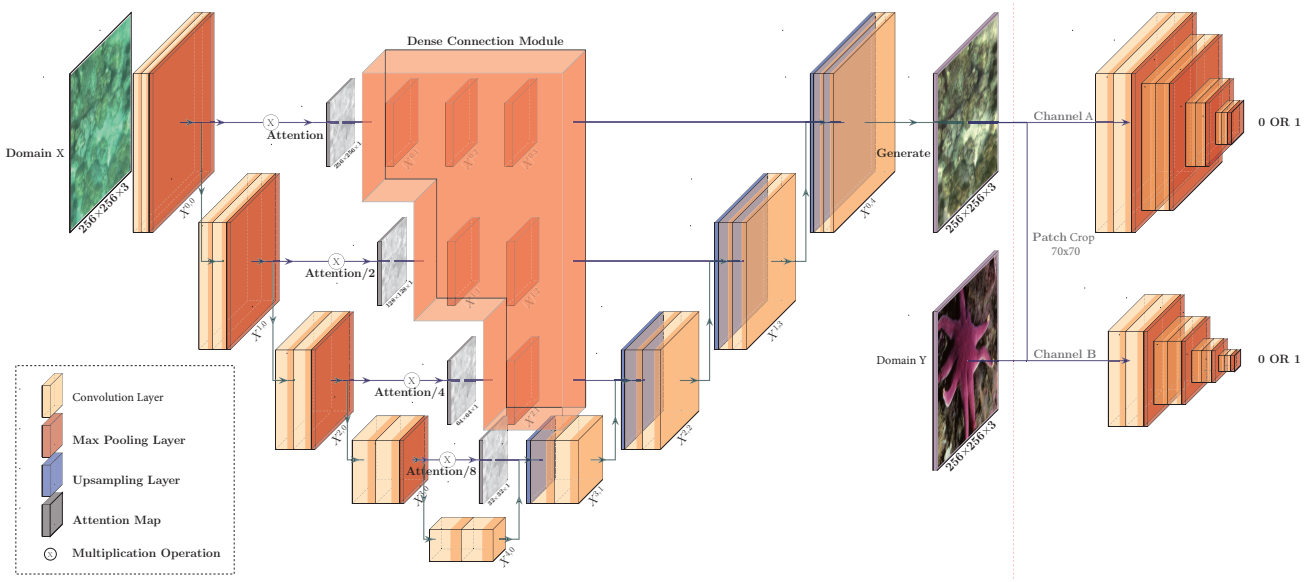


Figure 2: The architecture of our proposed CDGAN network.

performance on several underwater datasets. Furthermore, unpaired training methods prove incredibly easy and stable to be adopted to enhance underwater images from different domains.

2. Related Work

2.1. Low Light Image Enhancement

In general, image enhancement is a challenging research topic in computer vision and there has been increasing attention. To some extent, the low-light image enhancement task is closely related to the enhancement processing in underwater tasks. From the perspective of human vision mechanism under low-light scenario, researchers have investigated a series of research methods to improve the visibility. As an example, Wang *et al.* [46] used the guided filtering and adaptive histogram equalization approach which has improved the visibility remarkably. Moreover, massive mapping functions of the low-light image enhancement have been proposed recently. Spratling *et al.* [44] put forward a method which could effectively conduct object sensing in low light scenario. In [50] and [51], the proposed fusion networks illustrated the principles of the light reflection models.

The conventional approaches commonly use hand-crafted filters to recover the local color constancy and improve the contrast or lightness [41]. Meanwhile, the prior knowledge and statistical assumptions about the environment captured in the scene, the dark channel prior [7], a most effective method, are often adopted to deal with image deblurring and dehazing [17].

2.2. Underwater Image Enhancement

With the rapid development of the deep CNN-based models, the performance for tasks such as image colorization [53], color/contrast adjustment [10], dehazing [9], debluring [43] and rain removing [32], has been improved significantly. These models learn a sequence of non-linear filters from paired training data and produce much better performance than using hand-crafted filters. However, with more complex data distributions and more difficult statistical issues, the CNN-based models seem to have reached their bottleneck. Traditional physics-based methods for removing image haze rely on the atmospheric model, which is used to estimate the transmission and ambient light in a scene [8, 12]. Beside the physics-based models, Lu *et al.* [31] and Zhang *et al.* [55] proposed a series of bilateral and trilateral filters to remove noise and enhance the global contrast.

Apart from the contributions in model structures, there are several tricks and supplement of loss functions proposed. Wasserstein GANs [6] calculated the earth-mover distance to measure the distance between the data distribution and the model distribution to guide the training process. Energy-based GANs [56] improved the training stability by modeling the discriminator as an energy function. The vanishing gradients problem is a critical issue during the training process, and the Least-Squared GANs [33] solved the problem by employing a least-square loss for the discriminator.

The methods mentioned above regard the underwater image enhancement issue as image dehazing and have employed the common atmospheric model to resolve the underwater image enhancement. Then, Akkaynak *et al.* [2] proposed a revised imaging model, which accounted for the unique distortions on underwater light propagation. This method produces more accurate color recovery and better approximation to the ill-posed underwater image enhancement problems. However, the related unique model requires more data as prior, *e.g.*, depth [48, 3], and water type information [45]. Recently, the adversarial learning methods [14, 52, 29] have achieved the state-of-the-art performance.

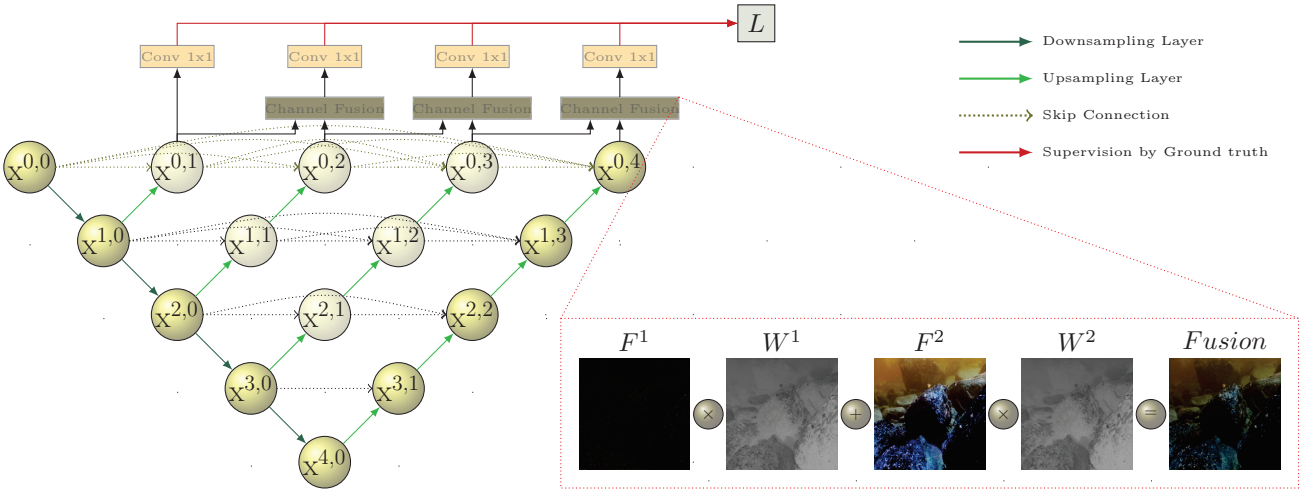


Figure 3: Dense Connection Module

2.3. Unsupervised Learning Methods

Dataset preparation is always the first issue we need to consider before training. Toward this, synthetically distorted images and paired datasets are the most common strategies. However, synthesized data and images of objects are typically with low diversity. Furthermore, one major limitation is that all of these learning models in the above discussed methods require paired training data, which may not be available in most practical applications. And yet the unpaired training on naturally distorted underwater images has not been explored in the literature. The unsupervised-learning based approaches [37, 40] do not require strictly paired data and have provided a promising alternative to address this issue .

Meanwhile, recently the GAN-based models [15] have achieved great success for image-to-image translation problems [21]. Specifically, conditional GANs [35] were proposed to realize the image style transfer, and this model allowed restricting the generator to produce samples that follow a pattern or belong to a particular class. Therefore, [21] was proposed to learn a pixel-to-pixel (Pix2Pix) mapping between the input domain and the desired output domain. Dual-direction GANs, such as CycleGAN [57] and DualGAN [49], introduce the cycle-consistency loss to learn the mutual mappings between different domains by unpaired data.

Inspired by this success, in this paper, we propose an unsupervised cross-domain conditional GAN to effectively address the limited availability of paired training data for underwater image enhancement. Next, the details of our proposed method are presented.

3. Proposed Method

3.1. Overview of the Network Architecture

We propose the Unsupervised Cross-Domain Generative Adversarial Networks (CDGAN), which is an ensemble architecture trained by unpaired underwater images, to address the requirement on paired images with distorted and clean underwater images for training. It is incredibly challenging to simultaneously acquire a degraded and a high-quality photo of the same visual scene in practice. Expressly, we set a source domain X (distorted underwater images) and a desired domain Y (high-quality images) as shown in Fig. 1 , instead of a paired dataset of underwater distorted images and their corresponding high-quality images in the same scene.

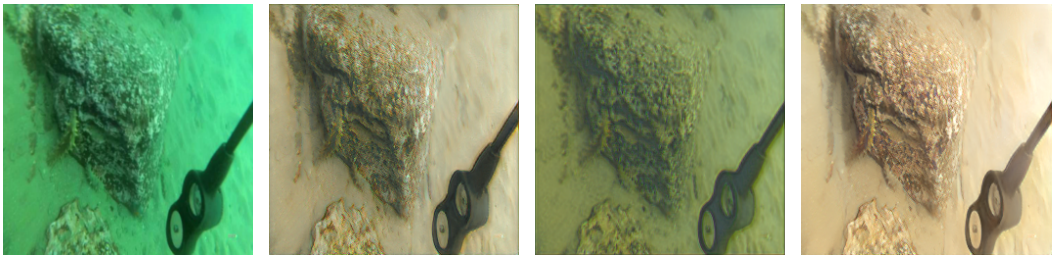
The goal is to estimate the non-linear mapping function $G : \{X, Z\} \rightarrow Y$ by unsupervised learning, to automatically achieve underwater image enhancement. Here, Z denotes the learned parameters. As illustrated in Fig. 2 , our CDGAN model consists of a generator (see Fig. 2 **left**) and a dual-channel discriminator (see Fig. 2 **right**).

Generator: The U-Net [42] model has achieved huge success on semantic segmentation and image reconstruction. Compared with other traditional feature extractors, the U-Net architecture can extract multi-scale features by more skip connections in different depths, including textures and structure. Underwater image enhancement encounters the limitation of available information features among the extreme natural environment. Thus, to extract sufficient details in these extreme scenes, we employ the common encoder-decoder structure to as the backbone of generator and the densely connection module to enhance the feature diversity.

Specifically, the generator consists of two parts, *i.e.*, the encoder layers: $X^{i,0}, i \in \{0, 1, 2, 3, 4\}$, and the decoder layers: $X^{4-j,j}, j \in \{1, 2, 3, 4\}$. Let $x^{i,j}$ indicate the output of node $X^{i,j}$, where i indexes the down-sampling layer along the encoder and j indexes the convolution layer of the dense block. The stack of feature maps represented by $x^{i,j}$ is computed as:

$$x^{i,j} = \begin{cases} \mathcal{H}\left(\mathcal{D}\left(x^{i-1,j}\right)\right), & j = 0 \\ \mathcal{H}\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, \mathcal{U}\left(x^{i+1,j-1}\right)\right]\right), & j > 0 \end{cases}, \quad (1)$$

where the function $\mathcal{H}(\cdot)$ indicates the convolution block, which consists of convolution operations and an activation function, $\mathcal{D}(\cdot)$ and $\mathcal{U}(\cdot)$ represent a down-sampling layer



(a) Input (b) w/o L_{patch} (c) w/o L_{color} (d) Enhanced

Figure 4: Visual comparison from the ablation study on the loss function. Images (b) ∼ (c) demonstrate the effectiveness of each of the components, *i.e.*, L_{patch} and L_{color}, in the loss function.

and an up-sampling layer respectively, and $[\cdot]$ denotes the concatenation operation.

In the dense connection module, we connect every two adjacent nodes in the ensemble to enable gradient back-propagation from the deeper decoders to their shallower counterparts. With the dense connectivity, each node is presented with not only the final aggregated feature maps but also the intermediate aggregated feature maps and the original same-scale feature maps from encoders. As such, the aggregation layer in the decoder node may learn to use only the same-scale encoder feature maps and use all collected feature maps available at the gate.

Specifically, we obtain the intermediate features at each node in the encoding process as:

$$F_{out} = \mathcal{A}\left(x^{i,0}\right) * x^{i,0}, i \in \{0,1,2,3\}, \tag{2}$$

where $\mathcal{A}(\cdot)$ denotes the attention strategy and $x^{i,0}$ indicates the features from the corresponding encoder layer. Moreover, to address the unsuitability of the unsupervised learning, we introduce extra supervision during the training and include a 1×1 convolution layer followed by a TanH activation layer to regulate each output of $\mathrm{X}^{i,0}, j \in \{1,2,3,4\}$. Both are supervised by minimizing the hyper loss function L, which is detailed in Section 3.2 .

Besides, we employ a Feature Fusion Module (FFM) to increase the density of skip-connections, which retains the relationship between the layers of different depths and improves the visual performance of the enhanced outputs, as shown in Fig. 3 . The FFM generates high-quality underwater images with dense skip connections and channel fusion operations. The main goal of this module is to mix features of different depths and channels to regulate the whole training process.

Mathematically, we obtain the intermediate feature maps F^1 and F^2 and the corresponding weight maps W^1 and W^2 as:

$$W^1 = \frac{\exp^{\mathcal{F}(F^1,S)}}{\left(\exp^{\mathcal{F}(F^1,S)} + \exp^{\mathcal{F}(F^2,S)} + \mu\right)} \tag{3}$$

and

$$W^2 = \frac{\exp^{\mathcal{F}(F^2,S)}}{\left(\exp^{\mathcal{F}(F^1,S)} + \exp^{\mathcal{F}(F^2,S)} + \mu\right)}, \tag{4}$$

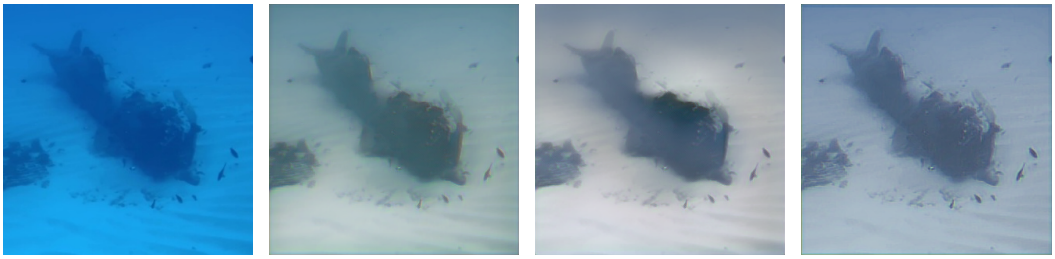
where $\mathcal{F}(\cdot)$ denotes the spatial attention function, S represents the pre-processing method, and μ is a constant value used to avoid dividing by zero. The purpose of the channel fusion strategy is to fuse different channels in the extracted features and enable the generator to extract more stable and semantic features. The fusing feature can be computed as:

$$\text{Fusion} = F^1 * W^1 + F^2 * W^2. \tag{5}$$

In summary, our generator connects the decoders, resulting in densely connected skip connections. Furthermore, it enables dense feature propagation, skip connections and more flexible feature attention fusion at the decoder layers. Horizontally, all layers in the generator combine multi-scale features from all of their preceding nodes at the same resolution; Vertically, they integrate multi-scale features across different resolutions from their preceding nodes.

Specifically, our generator is implemented with eight convolutional blocks. Moreover, each encoder consists of one convolution layer and one max-pooling layer. Whereas, the decoding layer consists of one up-sampling layer and one convolution block layer. And the convolution block consists of one 3×3 convolution layer, one $ReLU$ activation layer, and one batch normalization layer [18].

Discriminator: For the discriminator, we employ a dual-channel discriminator, including the standard channel A and the patch discriminator channel B, to distinguish real from fake images. The standard channel A processes images of the original size generated from the generator, while B receives the cropped patches from the enhanced output to consider the importance of the low-level details (image edges), as shown in Fig. 2 . The purpose of the dual-channel discriminator is to bridge the local detail and global semantic information and enable the discriminator to guide the generator to generate a realistic image. The limitation of the typical single discriminator is that the typical structure often ignores the relationship between the global content and the local details. Thus, in our work we design the dual-channel



(a) Input (b) w/o DC (c) w/o AM (d) Enhanced

Figure 5: Visual comparison from the component ablation study. Images (b) ∼ (c) demonstrate the effectiveness of each of the components DC and AM in the network structure.

structure to process the standard size and cropped patch inputs and guide the network to learn from the global perspective as well as the local perspective.

Specifically, Channel A receives the input of 256×256, and transforms it to the size of 32×32, which represents the average validity response of the discriminator. Moreover, we randomly crop N patches of 70×70 and send them to channel B to compute the probability of the image being real or fake. The impressive patch size was concluded in the work [21], which demonstrated that the patch of 70×70 was similar to the original size in terms of receptive fields. Also, each block in both channels consists of three components, *i.e.*, 3×3 convolutional filters with a stride of 2, the non-linear $ReLU$ and the batch normalization layer (BN).

Attention Strategy: Inspired by the low-light enhancement work in [22], we propose the **attention strategy** to guide the model training. The goal is to recover the color and structure details from the low-quality underwater images with limited contrast and intensity. Unfortunately, the low-light issue occurs in underwater environment frequently. Thus, our model adopts the illumination attention strategy proposed in [22] to guide the model to take more consideration in low-light regions and focus more on dark rather than bright regions. The purpose of the strategy is to avoid both over-enhancement and under-enhancement.

Specifically, the processing flow can be described as follows. First, we convert the input image from the RGB color space to HSI color space, and obtain and normalize the I channel, denoted as y_{in}, to the range of $[0, 1]$ to expand its dynamic range as:

$$y_{\text{out}} = \frac{y_{\text{in}} - y_{\min}}{y_{\max} - y_{\min}}, \tag{6}$$

where y_{max} and y_{min} are the maximum and minimum values of the input image's I channel. Then, we set y_{out} as the image attention map to guide the image enhancement and multiply with the extracted features in $\left[X^{0,0}, X^{1,0}, X^{2,0}, X^{3,0}\right]$.

To obtain each suitable scale attention map, we apply a max-pooling layer to process and multiply the output with all intermediate feature maps. Compared with the existing learning based methods, our procedure is less computationally expensive and easier to implement.

3.2. Objective Function

The proposed generator produces an image to fool the discriminator, which is designed to distinguish between fake and real-world images. For the optimization process, the proposed hyper loss $\mathcal{L}_{CDGAN}$ consists of the GAN loss (L^{Global} and L^{Local}), the perceptual loss L_{per} and the color loss L_{color}, and is defined as:

$$\mathcal{L}_{CDGAN} = \mathcal{L}_G^{Global} + \mathcal{L}_G^{Local} + \lambda_{per}\mathcal{L}_{per} + \lambda_{color}\mathcal{L}_{color}. \tag{7}$$

Here, λ_{per} and λ_{color} are two hyper parameters.

GAN Loss: The unique discriminator allows us to employ two adversarial losses, *i.e.*, L^{Global} and L^{Local}. Recently, the relativistic discriminator structure [24] has been widely adopted in enormous GAN-based research works. This function estimates the probability that the real data is more realistic than the fake data, and also directs the generator to synthesize a fake image that is more realistic than real ones.

The relativistic discriminator structure is defined as:

$$D_{Ra}\left(x_r, x_f\right) = \sigma\left(C\left(x_r\right) - \mathbb{E}_{x_f \sim \mathbb{P}_{\text{fake}}}\left[C\left(x_f\right)\right]\right) \tag{8}$$

and

$$D_{Ra}\left(x_f, x_r\right) = \sigma\left(C\left(x_f\right) - \mathbb{E}_{x_r \sim \mathbb{P}_{\text{real}}}\left[C\left(x_r\right)\right]\right) \tag{9}$$

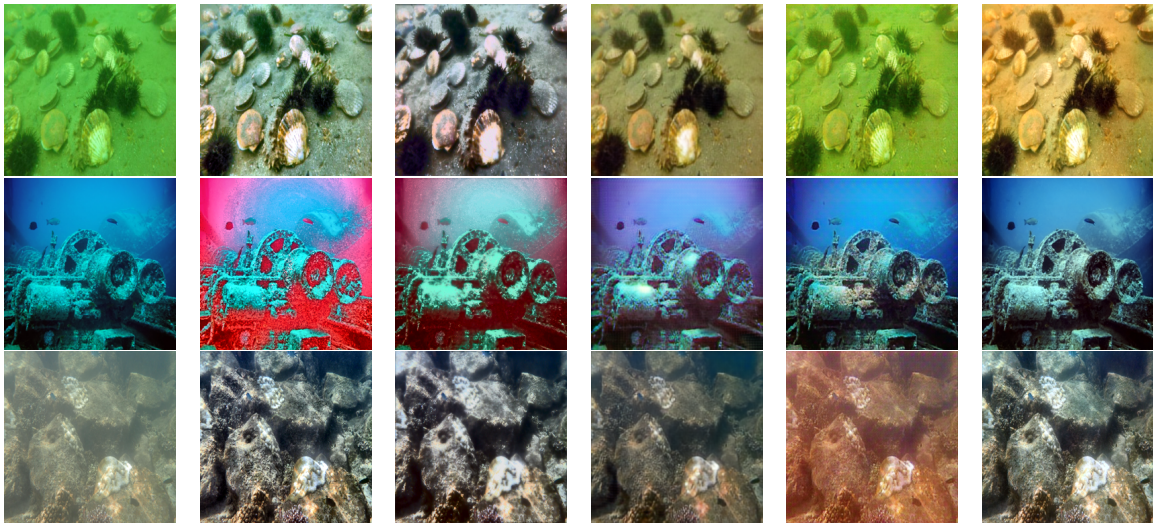
where C indicates the discriminator, x_r and x_f are samples selected from the real $\mathbb{P}_{\text{fake}}$ and fake distribution $\mathbb{P}_{\text{fake}}$, respectively, and σ represents the activation function.

In our work, we employ the relativistic discriminator and take the least square GAN (LSGAN) [33] as the activation function of our global discriminator. Thus, the $\mathcal{L}^{Global}$ for the generator G and global discriminator D are defined as:

$$\begin{aligned}\mathcal{L}_D^{Global} =& \mathbb{E}_{x_r \sim \mathbb{P}_{real}}\left[\left(D_{Ra}\left(x_r, x_f\right) - 1\right)^2\right] \\ &+ \mathbb{E}_{x_f \sim \mathbb{P}_{\text{fake}}}\left[D_{Ra}\left(x_f, x_r\right)^2\right]\end{aligned} \tag{10}$$

and

$$\begin{aligned}\mathcal{L}_G^{Global} =& \mathbb{E}_{x_f \sim \mathbb{P}_{fake}}\left[\left(D_{Ra}\left(x_f, x_r\right) - 1\right)^2\right] \\ &+ \mathbb{E}_{x_r \sim \mathbb{P}_{\text{real}}}\left[D_{Ra}\left(x_r, x_f\right)^2\right]\end{aligned} \tag{11}$$



(a) Input (b) FE [4] (c) RB [34] (d) CycleGAN [27]* (e) FUNIE [20]* (f) Ours*

Figure 6: Visual comparison with state-of-the-art methods on an over-blue image (a) from the U45 dataset. The super-scripted asterisk (*) denotes unpaired training.

Note that, different from the $\mathcal{L}^{Global}$, $\mathcal{L}^{Local}$ computes the similarity between the paired patches cropped from the enhanced output and the desired images. Moreover, to reduce the computational cost, we adopt the original LSGAN as the local adversarial loss, which is defined as:

$$\begin{aligned}\mathcal{L}_D^{Patch} =& \mathbb{E}_{x_r\sim\mathbb{P}_{\text{real-patches}}}\left[\left(D\left(x_r\right)-1\right)^2\right]+ \\ & \mathbb{E}_{x_f\sim\mathbb{P}_{\text{fake-patches}}}\left[\left(D\left(x_f\right)-0\right)^2\right]\end{aligned} \tag{12}$$

and

$$\mathcal{L}_G^{Patch} = \mathbb{E}_{x_r\sim\mathbb{P}_{\text{fake-patches}}}\left[\left(D\left(x_f\right)-1\right)^2\right] \tag{13}$$

Feature Loss: In order to preserve the high-level feature from the original input domain X, we employ the perceptual loss [23]. Perceptual loss computes the distance between the enhanced image and the desired image in feature space. It is regarded as the most effective strategy to regularize unsupervised training, which is widely adopted in many low-level vision tasks. Specifically, perceptual loss computes the distance between the enhanced image and desired image in feature space, and is defined as:

$$\mathcal{L}_{per}\left(I^U\right) = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}\left(\phi_{i,j}\left(I^U\right)-\phi_{i,j}\left(G\left(I^U\right)\right)\right)^2 \tag{14}$$

where I^U denotes the input underwater image, $G(I^U)$ indicates the enhanced output from the generator G, $\phi_{i,j}$ represents the feature map with i representing its i-th pooling layer and j denoting the convolution layer after i-th pooling layer, and $W_{i,j}$ and $H_{i,j}$ denote the width and height of the extracted feature maps, respectively. Specifically, the value of i is set as 5 and j is set as 1. In this work, the feature maps are extracted with a VGG-Net model pre-trained on ImageNet [26].

Color Loss: As mentioned in the recent underwater image enhancement works [19, 30, 54], the most significant challenge for underwater image enhancement is color casts and correction, especially the greenish and bluish phenomenon.

Hence, it is essential to design suitable objective functions in unsupervised learning. In this work, we introduce the unsupervised color loss to recover rich color. Since the training process is an unsupervised process, we cannot apply strong supervised regulations such as $L1$ or $L2$ loss. So, we propose a color loss to formulate the color vectors of the input underwater image and the desired output, and measure the spatial distance in color space. Moreover, this distance can be regarded as an effective loss function to guide the training process, simple and fast for computation. Besides, the color difference ensures that the enhanced image contains realistic color distribution.

Therefore, we employ the color loss $\mathcal{L}_{color}$ to constrain the training process and improve the colorfulness quantity in visual presentation. Specifically, we formulate the color vector distance in color (R, G, B) space between the generated image $\mathcal{F}(I_i)$ and the input underwater image I_i as:

$$\mathcal{L}_{color} = \sum_P \angle\Big(\left(\mathcal{F}\left(I\right)\right)_p, \left(\tilde{I}\right)_p\Big), \tag{15}$$

where $(\cdot)_p$ denotes the pixel and $\angle(\cdot)$ represents an operator that calculates the angle between the two color vectors, which regards the RGB channel vector as a 3D vector. The color loss defines the sum of the angles between the color vectors for every pixel pair in $\mathcal{F}(I_i)$ and I_i.

Table 1
Quantitative performance comparison of our proposed approach with the state of the arts on the U45 [28] dataset. (The super-scripted asterisk (*) denotes unpaired training)

Methods	UCIQE	UIQM	UICM	UISM	UIConM	NIQE
FE [4]	**31.5788**	3.9347	-31.5702	6.9856	0.7726	7.1416
RB [34]	26.8432	**4.7289**	-2.1708	6.8958	0.7702	5.0520
UDCP [13]	31.5356	3.7108	-38.0269	6.7520	0.7802	4.1871
CycleGAN [27]*	30.6228	4.2121	-15.2623	6.7015	0.7450	4.6558
FUNIE [20]*	27.8998	3.9924	-28.0229	6.8322	0.7734	4.2588
Ours*	29.9686	4.2343	-14.9563	6.8980	0.7326	**3.8835**

4. Implementation Details

In this work, we introduce an unsupervised approach to address the limitation of paired distorted and good images. We select 2K underwater images from the EVUP dataset [20] as domain X, and take 1K high-quality underwater images from the EVUP dataset and 1K terrestrial images from DIVIK [1] as domain Y, without the need of keeping them in pair or alignment. The unpaired domain image selection is employed to ensure that the domain Y contains the realistic feature distribution of higher resolution images. Besides, we take the validation dataset in EVUP as the test set.

The hyper parameters of $\mathcal{L}_{per}$ and $\mathcal{L}_{color}$ in the loss function are set as 0.5 and 0.1, respectively. Furthermore, we train all models for 200 epochs with a batch-size of 16, and the loss is minimized using the Adam [25] optimizer with a learning rate of 10^{-4}. We use the Pytorch [39] libraries to implement the CGAN model. Besides, two NVIDIA GeForce GTX 1080Ti GPUs are used for training.

5. Experimental Results and Analysis

In this section, we firstly present the visual effect and qualitatively analyze the visual performance of the enhanced outputs from the aspects of the colourfulness and sharpness. Then, we discuss the performance of our approach in comparison with the state of the arts qualitatively and quantitatively. Furthermore, we conduct ablation studies with respect to each of the employed components and loss parts.

5.1. Qualitative and Quantitative Comparison

We first show the visual comparison in Fig. 6 on three challenging cases, *i.e.*, over-blue colour, over-green colour and hazy images, in the U45 dataset from top to bottom. We compare the proposed model with several state-of-the-art methods on the public underwater benchmark U45. These competitive methods include FusionEnhance (FE) [4], RB [34], UDCP [13], CycleGAN [27] and FUNIE [20].

As shown in Fig. 6 , FUNIE [20] generally fails to rectify the greenish hue in images, while CycleGAN [27] performs reasonably well and its enhanced outputs are comparable to the images produced by our proposed approach. Moreover, we observe that achieving color consistency and hue rectification is relatively more challenging through unpaired learning. This is mostly because of the limitation of strong constraints among color and texture information in the training process. By introducing $\mathcal{L}_{color}$ and $\mathcal{L}_{per}$, our method can recover more realistic colour distribution as illustrated in Fig. 7 , which reveals rich and natural colour compared with other methods that have adopted unpaired training data. Besides, as previously mentioned, the physical model based methods usually rely on extra inputs, such as scene depth and prior knowledge. Overall, our CDGAN using unpaired images outperforms the existing unsupervised learning methods and addresses the limitation of data preparation effectively.

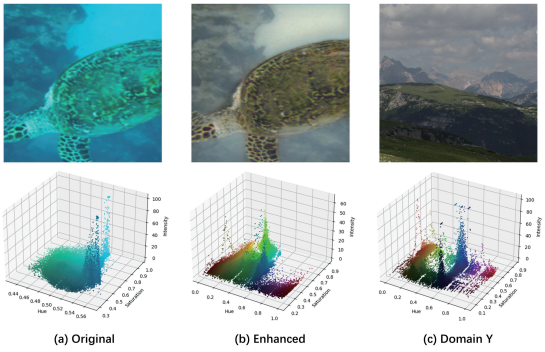


Figure 7: Comparison of 3D color spaces HSI for original underwater image (a), enhanced image (b) and terrestrial example in Domain Y(c), respectively.

Besides, we conduct the qualitative comparison of perceptual image enhancement by our model and the above methods. To evaluate the improved performance comprehensively, we consider three commonly-used underwater image quality measure metrics, namely Underwater Image Quality Measure (UIQM) [38], underwater color image quality evaluation (UCIQE) [47] and Natural Image Quality Evaluator (NIQE) [36].

- **UIQM:** The UIQM metric comprises three properties of underwater images, *i.e.*, the underwater image colourfulness measures (UICM), the underwater images sharpness measures (UISM), and the underwater image contrast measure (UIConM), and is defined as:

$$UIQM = c_1 \times UICM + c_2 \times UISM + c_3 \times UIConM,$$

Table 2
Quantitative performance comparison of our proposed approach with the state of the arts on the EVUP [20] dataset.

Methods	UCIQE	UIQM	UICM	UISM	UIConM	NIQE
FUNIE [20] *	28.3378	**3.9708**	-23.3969	6.8912	0.7260	4.3032
Ours*	**29.7354**	3.5364	-34.0610	6.8074	0.6950	**4.1368**

(16)

where the hyper-parameters c_1, c_2 and c_2 are set as 0.0282, 0.2953 and 3.5753, respectively.

- **UCIQE:** We choose UCIQE as another underwater enhancement metric, which was introduced by Yang and Sowmya [47]. The metric conducts a liner combination of chroma, saturation and contrasts to measure the nonuniform colour cast, blurring, and low contrast, respectively.

- **NIQE:** Besides, we consider the standard non-referenced image quality metric NIQE in order to compare with other methods quantitatively. The metric is a well-known, no-reference image quality assessment for evaluating real image restoration without ground truth.

Table 1 compares the average UCIQE, UIQM and NIQE scores obtained by our CDGAN and other approaches. The results indicate the FE [4] performs best on UCIQE while RB [34] leads on UIQM, and our CDGAN performs best in terms of NIQE. From the results, these physical methods outperform the learning based methods. However, our CDGAN outperforms all unsupervised learning methods, including CycleGAN [27] and FUNIE [20]. In contrast, unsupervised methods resolve the limitation of paired training data and can ignore the shortage of paired dataset.

Furthermore, we adopt another dataset with more images and more complex scenes to demonstrate the expansibility and robustness of these methods. Table 2 compares the results. Note that, due to the limited availability of the source codes, we adopt the FUNIE [20] as the main comparative method. Also, our CDGAN is superior to FUNIE [20] on both metrics of UCIQE and NIQE. Hence, we can conclude that our CDGAN not only effectively learns the photographic adjustment for enhancing the colour cast but also generates the images full of texture details.

5.2. Ablation Studies

We conduct ablation studies to reveal the effectiveness of each component we propose in our network, discussed in loss function and network structure, respectively.

5.2.1. Ablation Study on the Loss Function

Fig. 4 shows ablation study results that demonstrate the effectiveness of the each component in our proposed loss function. The result without local feature similarity loss L_{patch} has relatively lower contrast in the overall result, which shows the importance of L_{patch} in preserving the local details between the input and generated images. Removing the colourfulness measure loss L_{color} fails to recover the rich colour and appears severe colour casts. Based on the visual results in Fig. 4 , we can conclude that all these loss components play a critical role in the overall architecture. The results guided by the whole loss function indicate clearer details and better contrast. The significance is that the colour loss added in the whole function can produce rich colour over others. Furthermore, Table 3 also presents the corresponding contribution of each component in terms of the three image metrics.

5.2.2. Ablation Study on the Network Modules

For the structure of the whole network, we consider these modules as variants of the proposed method and use the following abbreviations:

- 'w/o DC': without the dual channel operation;

- 'w/o DSC': without the dense skip connection operation;

- 'w/o AM': with the attention map operation;

As shown in Table 3 , we notice that both the dual-channel (DC) operation and the dense skip connection (DSC) operation can improve the UCIQE and UIQM of the underwater images. Compared with the proposed network dropping the attention map (AM), we extract the illumination map from the original input and add these maps with intermediate feature maps to improve the NIQE. As shown in Fig. 4 , our method with the attention map operation has improved the sharpness by feeding more details.

6. Conclusion, Limitations and Future Work

In this paper, we have proposed an unsupervised conditional GAN-based model for underwater image enhancement. The proposed model formulates a colour loss function by evaluating the image colour vector difference. We have also performed extensive qualitative and quantitative evaluations and conducted the ablation study, which demonstrated the effectiveness of the proposed additive components in the loss function.

Although our approach can enhance the visual effect of the resultant images effectively, it remains a significant challenge. We observe a couple of challenges for our model that is not effective for enhancing over-degraded and texture-less

Table 3
Ablation study for various loss parts and structure components ('w/o' means 'without'.)

Methods	UCIQE	UIQM	UICM	UIConM	UISM	NIQE
Ours w/o L_{patch}	29.2079	2.5995	-63.3253	6.8976	0.6568	4.3105
Ours w/o L_{color}	**29.9247**	3.4796	-55.4484	6.8919	0.8413	4.2269
Ours w/o DC	27.9993	3.9678	-8.4463	6.7018	0.6229	4.8348
Ours w/o DSC	28.4485	**4.4835**	6.6532	7.0175	0.6219	4.5277
Ours w/o AM	29.1673	4.4326	-1.6108	6.8824	0.6840	**4.2269**

images. Shallow light conditions and noise amplification often blur these over-degraded images in such cases and their colour and texture recovery remains poor.

In future work, we intend to provide new insights into underwater image sequences and videos. Due to the limitation of information extracted from the underwater images, the multi-source fusion vision will be an understudied and important potential next frontier.

Acknowledgments

This work was sponsored by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX201959) and Synergy Innovation Foundation of the University and Enterprise for Graduate Students in Northwestern Polytechnical University (XQ201910). This work was also supported in part by the National Natural Science Foundation of China under Grant 61972321.

CRediT authorship contribution statement

Fei Li: Conceptualization of this study, Methodology, Writing - Original draft preparation. **Jiangbin Zheng:** Methodology, Writing - Original draft preparation. **Yuan-fang Zhang:** Data curation, Writing - Original draft preparation. **Wenjing Jia:** Writing - Original draft preparation. **Qianru Wei:** Writing - Original draft preparation. **Xiangjian He:** Writing - Original draft preparation.

References

[1] Agustsson, E., Timofte, R., 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) , 1122–1131.

[2] Akkaynak, D., Treibitz, T., 2018. A revised underwater image formation model. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 6723–6732.

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

[4] Ancuti, C., Ancuti, C.O., Haber, T., Bekaert, P., 2012. Enhancing underwater images and videos by fusion. 2012 IEEE Conference on Computer Vision and Pattern Recognition , 81–88.

[5] Anwar, S., Li, C., Porikli, F., 2018. Deep underwater image enhancement. ArXiv abs/1807.03528.

[6] Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: ICML.

[7] Berman, D., Levy, D., Avidan, S., Treibitz, T., 2020. Underwater single image color restoration using haze-lines and a new quantitative dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1.

[8] Bryson, M., Johnson-Roberson, M., Pizarro, O., Williams, S.B., 2016. True color correction of autonomous underwater vehicle imagery. J. Field Robotics 33, 853–874.

[9] Cai, B., Xu, X., Jia, K., Qing, C., Tao, D., 2016. Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing 25, 5187–5198.

[10] Cheng, Z., Yang, Q., Sheng, B., 2015. Deep colorization. 2015 IEEE International Conference on Computer Vision (ICCV) , 415–423.

[11] Chiang, J.Y., Chen, Y., 2012. Underwater image enhancement by wavelength compensation and dehazing. IEEE Transactions on Image Processing 21, 1756–1769.

[12] Cho, Y., Jeong, J., Kim, A., 2018. Model-assisted multiband fusion for single image enhancement and applications to robot vision. IEEE Robotics and Automation Letters 3, 2822–2829.

[13] Drews, P., Nascimento, E.R., Botelho, S., Campos, M., 2016. Underwater depth estimation and image restoration based on single images. IEEE Computer Graphics and Applications 36, 24–35.

[14] Fabbri, C., Islam, M.J., Sattar, J., 2018. Enhancing underwater imagery using generative adversarial networks. 2018 IEEE International Conference on Robotics and Automation (ICRA) , 7159–7165.

[15] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets, in: NIPS.

[16] Gupta, H., Mitra, K., 2019. Unsupervised single image underwater depth estimation. 2019 IEEE International Conference on Image Processing (ICIP) , 624–628.

[17] He, K., Sun, J., Tang, X., 2011. Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[18] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

[19] Islam, M.J., 2019. Understanding human motion and gestures for underwater human-robot collaboration. J. Field Robotics 36, 851–873.

[20] Islam, M.J., Xia, Y., Sattar, J., 2020. Fast underwater image enhancement for improved visual perception. IEEE Robotics and Automation Letters 5, 3227–3234.

[21] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 5967–5976.

[22] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z., 2019. Enlightengan: Deep light enhancement without paired supervision. arXiv preprint arXiv:1906.06972 .

[23] Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European conference on computer vision, Springer. pp. 694–711.

[24] Jolicoeur-Martineau, A., 2018. The relativistic discriminator: a key element missing from standard gan. arXiv preprint arXiv:1807.00734 .

[25] Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. CoRR abs/1412.6980.

[26] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks, in: CACM.

[27] Li, C., Guo, J., Guo, C., 2018a. Emerging from water: Underwater image color correction based on weakly supervised color transfer. IEEE Signal Processing Letters 25, 323–327.

[28] Li, H., Li, J., Wang, W., 2019. A fusion adversarial underwater image enhancement network with a public test dataset. arXiv: Image and Video Processing .

[29] Li, J., Skinner, K.A., Eustice, R., Johnson-Roberson, M., 2018b. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. IEEE Robotics and Automation Letters 3, 387–394.

[30] Lu, H., Li, Y., Serikawa, S., 2013a. Underwater image enhancement using guided trigonometric bilateral filter and fast automatic color correction. 2013 IEEE International Conference on Image Processing , 3412–3416.

[31] Lu, H., Li, Y., Serikawa, S., 2013b. Underwater image enhancement using guided trigonometric bilateral filter and fast automatic color correction. 2013 IEEE International Conference on Image Processing , 3412–3416.

[32] Luo, Y., Xu, Y., Ji, H., 2015. Removing rain from a single image via discriminative sparse coding. 2015 IEEE International Conference on Computer Vision (ICCV) , 3397–3405.

[33] Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) , 2813–2821.

[34] Marques, T.P., Albu, A., Hoeberechts, M., 2019. A contrast-guided approach for the enhancement of low-lighting underwater images. J. Imaging 5, 79.

[35] Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .

[36] Mittal, A., Soundararajan, R., Bovik, A., 2013. Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters 20, 209–212.

[37] Ni, Z., Yang, W., Wang, S., Ma, L., Kwong, S., 2020. Towards unsupervised deep image enhancement with generative adversarial network. IEEE Transactions on Image Processing 29, 9140–9151.

[38] Panetta, K., Gao, C., Agaian, S., 2016. Human-visual-system-inspired underwater image quality measures. IEEE Journal of Oceanic Engineering 41, 541–551.

[39] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.

[40] Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434.

[41] Rahman, Z., Jobson, D., Woodell, G.A., 2002. Retinex processing for automatic image enhancement, in: IST/SPIE Electronic Imaging.

[42] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: MICCAI.

[43] Schuler, C., Hirsch, M., Harmeling, S., Schölkopf, B., 2016. Learning to deblur. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1439–1451.

[44] Spratling, M.W., 2016. A hierarchical predictive coding model of object recognition in natural images. Cognitive Computation 9, 151 – 167.

[45] Uplavikar, P.M., Wu, Z., Wang, Z., 2019. All-in-one underwater image enhancement using domain-adversarial learning., in: CVPR Workshops, pp. 1–8.

[46] Wang, Z., Luo, J., Qin, K., Li, H., Li, G., 2017. Model based edge-preserving and guided filter for real-world hazy scenes visibility restoration. Cognitive Computation 9, 468–481.

[47] Yang, M., Sowmya, A., 2015. An underwater color image quality evaluation metric. IEEE Transactions on Image Processing 24, 6062–6071.

[48] Ye, X., Li, Z., li Sun, B., Wang, Z., Xu, R., Li, H., Fan, X., 2020. Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks. IEEE Transactions on Circuits and Systems for Video Technology , 1–1.

[49] Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation. 2017 IEEE International Conference on Computer Vision (ICCV) , 2868–2876.

[50] Ying, Z., Li, G., Gao, W., 2017a. A bio-inspired multi-exposure fusion framework for low-light image enhancement. ArXiv abs/1711.00591.

[51] Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W., 2017b. A new low-light image enhancement algorithm using camera response model. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) , 3015–3022.

[52] Yu, X., Qu, Y., Hong, M., 2018. Underwater-gan: Underwater image restoration via conditional generative adversarial network, in: CVAUI/IWCF/MIPPSNA@ICPR.

[53] Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization, in: ECCV.

[54] Zhang, S., Wang, T., Dong, J., Yu, H., 2017a. Underwater image enhancement via extended multi-scale retinex. Neurocomputing 245, 1–9.

[55] Zhang, S., Wang, T., Dong, J., Yu, H., 2017b. Underwater image enhancement via extended multi-scale retinex. Neurocomputing 245, 1–9.

[56] Zhao, J., Mathieu, M., LeCun, Y., 2017. Energy-based generative adversarial networks, in: ICLR.

[57] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) , 2242–2251.