SINGLE UNDERWATER IMAGE RESTORATION BY CONTRASTIVE LEARNING

Junlin Han, Mehrdad Shoeiby, Tim Malthus, Elizabeth Botha,
Janet Anstee, Saeed Anwar, Lars Petersson, Mohammad Ali Armin

Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

ABSTRACT

Underwater image restoration attracts significant attention due to its importance in unveiling the underwater world. This paper elaborates on a novel method that achieves state-of-the-art results for underwater image restoration based on the unsupervised image-to-image translation framework. We design our method by leveraging from contrastive learning and generative adversarial networks to maximize mutual information between raw and restored images. We also release a large-scale real underwater image dataset to support both paired and unpaired training modules. Extensive experiments with comparisons to recent approaches further demonstrate the superiority of our proposed method.

***Index Terms*—** Underwater image restoration, contrastive learning, underwater image dataset, image-to-image translation.

1. INTRODUCTION

For marine science and ocean engineering, significant applications such as the surveillance of coral reefs, aquatic robot inspection, and inspection of submarine cables require clear underwater images. However, raw underwater images with poor visual quality can not meet the expectations. The quality of underwater images plays an essential role in scientific missions; thus, fast, accurate, and effective image restoration techniques need to be developed to improve the visibility, contrast, and color properties of underwater images for satisfactory visual quality.

In the underwater scene, the visual quality is greatly affected by light refraction, absorption, and scattering. For instance, underwater images usually have a green-bluish tone since red light with longer wavelengths attenuates faster. Underwater image restoration is an ill-posed problem, which requires several parameters (global background light and medium transmission map) that are mostly unavailable in practice. These parameters can be roughly estimated by employing priors and supplementary information. However, due to the diversity of water types and lighting conditions, conventional under image restoration methods fail to rectify the color of underwater images.

Recent advances in deep learning demonstrates dramatic success in different fields. Learning-based models require a large-scale dataset for training, which is often difficult to obtain. Thus, most learning-based models use either small-scale real underwater images [1, 2], synthesized images [3, 4], or natural in-air images [5] instead of restored underwater images as either source domain or target domain of the training set. The aforementioned datasets are limited to capture natural variability in a wide-range of water types.

To overcome the earlier discussed challenges, we construct a large-scale real underwater image dataset with accurate restored underwater images. We formulate the restoration problem as an image-to-image translation problem and propose a novel **C**ontrastive Under**W**ater **R**estoration approach (CWR). Given an underwater image as the input, CWR directly outputs a restored image showing the real color of underwater objects as if the image was taken in-air without any structure and content loss.

The main contribution of our work is summarized as follows:

- We propose CWR, which leverages contrastive learning to maximize the mutual information between corresponding patches of the raw image and the restored image to capture the content and color features correspondences between two image domains.
- We construct a large-scale, high-resolution underwater image dataset with real underwater images and restored images. This dataset supports both paired or unpaired training.

2. A NOVEL DATASET

Heron Island Coral Reef Dataset (HICRD) contains raw underwater images from nine sites with detailed metadata for each site, including water types, maximum dive depth, wavelength-dependent attenuation within the water column, and the camera model. According to raw images' depth information and the distance between objects and the camera, images with roughly the same depth and same distance are labeled as good-quality. Images with sharp depth changes or distance changes are labeled as low-quality. We apply our imaging model described in section 3 to good-quality images,

producing corresponding restored images, and manually remove some restored images with non-satisfactory quality.

HICRD contains 6003 low-quality images, 3472 good-quality images, and 2000 restored images. We use low-quality images and restored images as the unpaired training set. In contrast, the paired training set contains good-quality images and corresponding restored images. The test set contains 300 good-quality images as well as 300 paired restored images as ground truth. All images are in 1842 x 980 resolution. Dataset and code will be released upon acceptance.

3. UNDERWATER IMAGING MODEL

Absorption plays a critical role in an underwater scenario. The absorption coefficient of each channel is wavelength-dependent, being the highest for red and the lowest for blue. The underwater imaging model can be formulated as:

$$I(x) = K * J(x)t(x) + A(1 - t(x)), \tag{1}$$

where $I(x)$ is the observed intensity, $J(x)$ is the scene radiance, $t(x) = e^{-\beta_a d(x)}$ is the medium transmission describing $A(x)$ the portion of light that is not scattered. A is the global atmospheric light, β_a is the light's absorption coefficient. $d(x)$ is the distance between camera and object, while K is a constant. With the assumption that β_a is constant [6], we can simplify equation 1 as:

$$I^c(x) = J^c(x)t^c(x) + A^c(1 - t^c(x)), \quad c \in \{r, g, b\}, \tag{2}$$

where $t^c(x) = e^{-\beta^c d(x)}$ and channels are in RGB space now.

Transmittance is highly related to β^c, which is the absorption coefficient. Unlike previous work [7, 8], instead of assigning a fixed wavelength for each channel containing bias (*e.g.*, 600nm, 525nm, and 475nm for red, green, and blue), we employ the camera sensor response to conduct a more accurate estimation. Figure 1 shows the camera sensor response of camera type CMV2000-QE used in collecting our dataset.

$$p^c = \int_a^b \beta^\lambda S^c(\lambda) d\lambda, \tag{3}$$

where p^c is the total attenuation coefficient, β^λ is the attenuation coefficient of each wavelength, and $S^c(\lambda)$ is the camera sensor response of each wavelength. Following the human visible spectrum, we set a = 400nm and b = 750nm to calculate medium transmission for each channel. We modify $t^c(x)$ in equation 2 leading to a more accurate estimation: $t^c(x) = e^{-p^c d(x)}$.

It is challenging to measure the scene's actual distance from an individual image without a depth map. Instead of using a flawed estimation approach, we assume the distance between the scene and the camera to be small (0.5m - 3m) and manually assign a distance for each good image.

The global atmospheric light, A^c is usually assumed to be the pixel's intensity with the highest brightness value in each channel. However, this assumption often fails due to the presence of artificial lighting and self-luminous aquatic. Since we have access to the diving depth, we can define A^c as follows:

$$A^c = e^{-p^c \phi}, \tag{4}$$

where p^c is the total attenuation coefficient, ϕ is the diving depth.

With the medium transmission and global atmospheric light, we can recover the scene radiance. The final scene radiance $J(x)$ is estimated as:

$$J^c(x) = \frac{I_c(x) - A_c}{max(t_c(x), t_0)} + A_c. \tag{5}$$

Typically, we choose t_0 = 0.1 as a lower bound. In practice, due to image formulation's complexity, our imaging model may encounter information loss, *i.e.*, the pixel intensity values of $J^c(x)$ is larger than 255 or less than 0. This problem is avoided by only mapping a selected range (13 to 255) of pixel intensity values from I to J. However, outliers may still occur; we rescale the whole pixel intensity values to enhance contrast and keep information lossless after restoration.

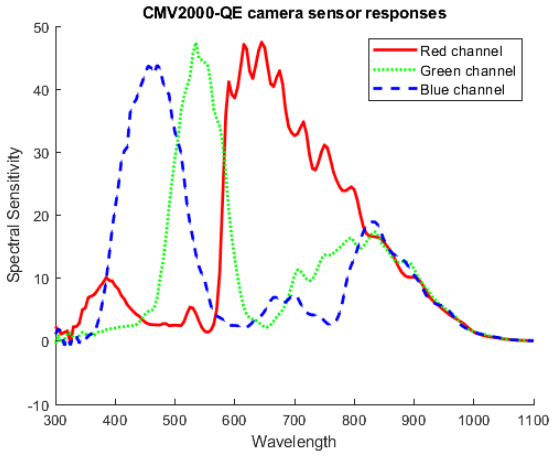

Fig. 1. Camera sensor response for camera type CMV2000-QE which is used in collecting real underwater images.

4. METHOD

Given two domains $\mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ and $\mathcal{Y} \subset \mathbb{R}^{H \times W \times 3}$ and a dataset of unpaired instances X containing raw underwater images x and Y containing restored images y. We denote it $X = \{x \in \mathcal{X}\}$ and $Y = \{y \in \mathcal{Y}\}$. We aim to learn a mapping $G : X \to Y$ to enable underwater image restoration.

Contrastive Under**W**ater **R**estoration (CWR) has a generator G as well as a discriminator D. G enables the restoration process, and D ensures that the images generated by G are undistinguished to domain Y in principle. The first half of

Loss type	Equation No.	Equation
Adversarial	7	$\mathcal{L}_{\text{GAN}}(G, D, X, Y) = \mathbb{E}_{y\sim Y}\left[(D(y))^2\right] + \mathbb{E}_{x\sim X}\left[(1 - D(G(x)))^2\right]$
Cross-entropy	8	$\ell(\boldsymbol{v}, \boldsymbol{v}^+, \boldsymbol{v}^-) = -\log(\frac{\exp(\boldsymbol{cos}(v,\boldsymbol{v}^+)/\tau)}{\exp(\boldsymbol{cos}(v,\boldsymbol{v}^+)/\tau)+\sum_{n=1}^{N}\exp(\boldsymbol{cos}(v,\boldsymbol{v}_n^-)/\tau)})$
PatchNCE	9	$\mathcal{L}_{\text{PatchNCE}}(G, H, X) = \mathbb{E}_{\boldsymbol{x}\sim X}\sum_{l=1}^{L}\sum_{s=1}^{S_l}\ell\left(\hat{z}_l^s, \boldsymbol{z}_l^s, \boldsymbol{z}_l^{S\setminus s}\right)$
Identity	10	$\mathcal{L}_{\text{Identity}}(G) = \mathbb{E}_{y\sim Y}\left[\|G(y) - y\|_1\right]$

Table 1. Components of the full objective. We use least-square adversarial loss for equation 7. For equation 8, we use a noisy contrastive estimation framework to maximize the mutual information between inputs and outputs. The idea behind equation 7 is to correlate two signals, *i.e.*, the "query" and its "positive" example, in contrast to other examples in the dataset (referred to as "negatives"). We map query, positive, and N negatives to K-dimensional vectors and denote them $v, v^+ \in R^K$ and $v^- \in R^{N\times K}$, respectively. Note that $v_n^- \in R^K$ denotes the n-th negative. We set up an $(N+1)$-way classification problem and compute the probability that a "positive" is selected over "negatives". This can be expressed as a cross-entropy loss where $\cos(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ denotes the cosine similarity between $\boldsymbol{u}$ and $\boldsymbol{v}$. τ denotes a temperature parameter to scale the distance between the query and other examples, we use 0.07 as default. For equation 9, We select L layers from $G_{enc}(X)$ and send them to a projection head H_X, embedding one image to a stack of feature $\{\boldsymbol{z}_l\}_L = \left\{H^l\left(G_{\text{enc}}^l(\boldsymbol{x})\right)\right\}_L$, where G_{enc}^l represents the output of l-th selected layers. After having a stack of features, each feature actually represents one patch from the image. We denote the spatial locations in each selected layer as $s \in \{1, ..., S_l\}$, where S_l is the number of spatial locations in each layer. We select a query each time, refer the corresponding feature ("positive") as $\boldsymbol{z}_l^s \in \mathbb{R}^{C_l}$ and all other features ("negatives") as $\boldsymbol{z}_l^{S\setminus s} \in \mathbb{R}^{(S_l-1)\times C_l}$, where C_l is the number of channels in each layer. Two losses are introduced to prevent generator from unnecessary changes. Equation 10 is a ℓ_1 Identity loss preserving the fidelity.

the generator is defined as an encoder while the second half is a decoder and presented as G_{enc} and G_{dec}, respectively.

We extract features from several layers of the encoder and forward them to a two-layer MLP projection head H. Such a projection head learns to project the extracted features from the encoder to a stack of features. CWR combines three losses, including adversarial loss, PatchNCE loss, and Identity loss. The details of our objective are described below.

The restored image should be realistic ($\mathcal{L}_{GAN}$), and patches in the corresponding raw and restored images should share some correspondence ($\mathcal{L}_{\text{PatchNCE}}$). The restored image has an identical structure to the raw image. In contrast, the colors are the true colors of scenes ($\mathcal{L}_{\text{Identity}}$). The full objective is:

$$\begin{aligned}\mathcal{L}(G, D, H) = &\lambda_{GAN}\mathcal{L}_{GAN}(G, D, X) \\ &+ \lambda_{NCE}\mathcal{L}_{\text{PatchNCE}}(G, H, X) \\ &+ \lambda_{IDT}\mathcal{L}_{\text{Identity}}(G).\end{aligned} \tag{6}$$

We set λ_{GAN} = 1, λ_{NCE} = 1, and λ_{IDT} = 10. The details of each component are elaborated in Table 1.

5. EXPERIMENTS

5.1. Baselines and Training Details

We compare CWR to several state-of-the-art baselines from different views, including image-to-image translation approaches (CUT [9] and CycleGAN [10]), underwater image enhancement methods (UWCNN [3], Retinex [11] and Fusion [12]), and underwater image restoration methods (DCP [13], IBLA [7]). We use the pre-trained UWCNN model with water type-3, which is close to our dataset.

We train CWR, CUT, and CycleGAN for 100 epochs with the same learning rate of 0.0002. The learning rate decays linearly after half epochs. We load all images in 800x800 resolution, and randomly crop them into 512x512 patches during training. We load test images in 1680x892 resolution for all methods. The architecture of CWR is is inspired by CUT, a resnet-based generator with nine residual blocks and a PatchGAN discriminator. We use spectral normalization and batch size of one. ADAM optimiser is employed for optimization.

5.2. Evaluation Protocol and Results

To fully measure the performance of different methods, we employ three full-reference metrics: Mean-Square Error (MSE), Peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) as well as a non-reference metric designed for underwater images: Underwater Image Quality Measure (UIQM) [14]. A higher UIQM score refers to the result is more consistent with human visual perception. We additionally use Fréchet Inception Distance (FID) [15] to measure the quality of generated images. A lower FID score means generated images tend to be more realistic.

Table 2 provides quantitative evaluation, where no method always wins in terms of all metrics. However, CWR performs stronger than all the baseline. Figure 2 presents the randomly selected qualitative results. Conventional methods produce

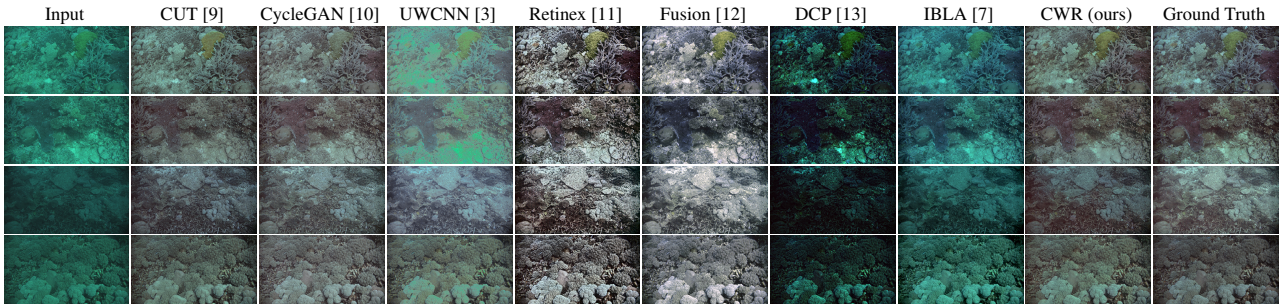

Fig. 2. Qualitatively results on HICRD test set. CWR shows visual satisfactory results without any content and structure loss.

Method	**MSE**↓	**PSNR**↑	**SSIM**↑	**UIQM**↑	**FID**↓
CUT [9]	170.27	26.30	0.796	5.26	22.35
CycleGAN [10]	448.16	21.81	0.591	5.27	**16.74**
UWCNN [3]	775.81	20.20	0.754	4.18	33.43
Retinex [11]	1227.19	17.36	0.722	**5.43**	71.90
Fusion [12]	1238.60	17.53	0.783	5.33	58.57
DCP [13]	2548.20	14.27	0.534	4.49	37.52
IBLA [7]	2366.42	14.49	0.192	3.63	23.06
CWR (ours)	**127.23**	**26.88**	**0.834**	5.25	18.20

Table 2. Comparisons to baselines on HICRD dataset. We show five metrics for all methods. CWR is in the first place for MSE, PNSR, and SSIM while the second place for FID.

blurry and unrealistic results, while learning-based methods tend to rectify the distorted color successfully. CWR performs better than other learning-based methods in keeping the restored images' structure and content identical to raw images with negligible artifacts.

6. CONCLUSION

This paper presents an underwater image dataset HICRD that offers large-scale underwater images and restored images to enable a comprehensive evaluation of existing underwater image enhancement & restoration methods. We believe that HICRD will make a significant advancement for the use of learning-based methods. A novel method, CWR employing contrastive learning is proposed to capitalize on HICRD. Experimental results show that CWR significantly performs better than several conventional methods while showing more desirable results compared to state-of-the-art learning-based methods.

7. REFERENCES

[1] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao, "An underwater image enhancement benchmark dataset and beyond," *Transactions on Image Processing*, 2019.

[2] Md Jahidul Islam, Youya Xia, and Junaed Sattar, "Fast underwater image enhancement for improved visual perception," *Robotics and Automation Letters*, 2020.

[3] Chongyi Li, Saeed Anwar, and Fatih Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognition*, 2020.

[4] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar, "Enhancing underwater imagery using generative adversarial networks," in *Int. Conf. on Robot. and Automat.*, 2018.

[5] Chongyi Li, Jichang Guo, and Chunle Guo, "Emerging from water: Underwater image color correction based on weakly supervised color transfer," *Signal processing letters*, 2018.

[6] Seiichi Serikawa and Huimin Lu, "Underwater image dehazing using joint trilateral filter," *Comp. & Elect. Engg.*, 2014.

[7] Yan-Tsung Peng and Pamela C Cosman, "Underwater image restoration based on image blurriness and light absorption," *Transactions on image processing*, 2017.

[8] John Y Chiang and Ying-Ching Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Transactions on Image Processing*, 2011.

[9] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*, 2020.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Int. Conf. on Comp. Vis.*, 2017.

[11] Xueyang Fu, Peixian Zhuang, Yue Huang, Yinghao Liao, Xiao-Ping Zhang, and Xinghao Ding, "A retinex-based enhancing approach for single underwater image," in *International Conference on Image Processing*, 2014.

[12] Codruta O Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, and Philippe Bekaert, "Color balance and fusion for underwater image enhancement," *Transactions on image processing*, 2017.

[13] Kaiming He, Jian Sun, and Xiaoou Tang, "Single image haze removal using dark channel prior," *Transactions on pattern analysis and machine intelligence*, 2010.

[14] Karen Panetta, Chen Gao, and Sos Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, 2015.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017.