JOINT RESIDUAL LEARNING FOR UNDERWATER IMAGE ENHANCEMENT

Minjun Hou[⋆†] *Risheng Liu*[⋆†] *Xin Fan*[*⋆†] *Zhongxuan Luo*[⋆†]

[⋆]DUT-RU International School of Information Science & Engineering, Dalian University of Technology
[†]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province

ABSTRACT

Improving the quality of underwater image has a significant impact on many signal processing and computer vision applications, while haze-effect and color shift are main handicaps need to be surmounted. Due to the complexity of the underwater environmental factors, most existing image enhancement techniques cannot be directly applied to address this task. In this work, we develop a novel framework to jointly performing residual learning on transmission and image domains for underwater scene entrenchment. Indeed, our deep model consists of a data-driven residual architecture for transmission estimation and a knowledge-driven scene residual formulation for underwater illumination balance. Therefore, we can aggregate the prior knowledge and data information to investigate the underlying underwater image distribution. Moreover, by introducing adaptive exposure map, image colors will also be corrected accordingly. Experimentally, both quantitative and qualitative analysis can indicate outstanding effectiveness of the proposed algorithm, against state-of-the-art approaches.

Index Terms— Underwater image enhancement, joint residual learning, transmission map estimation, illumination balance

1. INTRODUCTION

In recent years people pay more and more attention to the development of the ocean, it is important to get clear underwater images in the detection of marine life, assessment of underwater geological environment and other scientific tasks. In reality, there are two main reasons leading to the degradation of underwater image: 1) varying attenuation of light in different wavelengths causes color cast; 2) suspended particles bring about haze-effect as light is absorbed and scattered.

Early image-based enhancement methods always directly use techniques of computer graphics, such as histogram equalization algorithms under different constraints [1], frequency domain filtering method [2], white balance algorithm, multi-scale retinex method [3] and fusion-based methods [4], however, due to abandon the physical imaging model, those image-based enhancement methods can not remove haze effect well.

Another is model-based enhancement methods which aim to restore the original reflection characteristics of scene. According to Jaffe-McGlamery imaging model [5], if the camera is not so far away from the scene, an underwater image can be regarded as a linear superposition of two components: 1) light which has not been scattered or absorbed in the intervening water, called the direct component; 2) light which enters the camera without reflection from the object, called backscatter. It can be formulated as follows:

$$I(x) = J(x)t(x) + B_\lambda(1 - t(x)), \lambda \in \{r, g, b\}, \quad (1)$$

where $I(x)$ is the degraded image, $J(x)$ is the latent scene radiance, B_λ is the global light and λ represent different color channel, $t(x)$ is the medium transmission, showing the portion of the light that does not reach the camera, and it is defined as: $t(x) = e^{-\beta_\lambda d(x)}$, where the β_λ is the medium extinction coefficient, related to the degree of color attenuation in transmission medium, $d(x)$ represents the distance between camera and objects in picture. On account of the complexity of underwater scene, it is difficult to know specific parameters in model (1), therefore the key to solve the problem is to estimate an approximate transmission map by appropriate prior knowledge, such as underwater dark channel [6, 7] and its variant [8], strong difference in attenuation between three color channels in underwater image [9]. Though in many cases those methods can deal with the scattering problem well, relying on certain prior effected by artificial parameter setting, they may be invalid in some specific scenes.

To make up for the above shortcomings of traditional underwater enhancement methods, we design a data-based convolutional neural network (CNN) to learn a more realistic transmission map, and construct scene residual to balance holistic illumination jointly. The contributions of our work are summarized as follows:

In this paper, we provide a new perspective to formulate the enhancement task of underwater image as simultaneously learning transmission and scene residuals within a unified image propagation framework (see Fig. 1). Specifically, we first employ the residual CNN learning strategy to estimate the desired transmission map based on the aggregation of task cues and training data. Furthermore, we regard color cast as a

*Corresponding author. E-mail: xin.fan@dlut.edu.cn

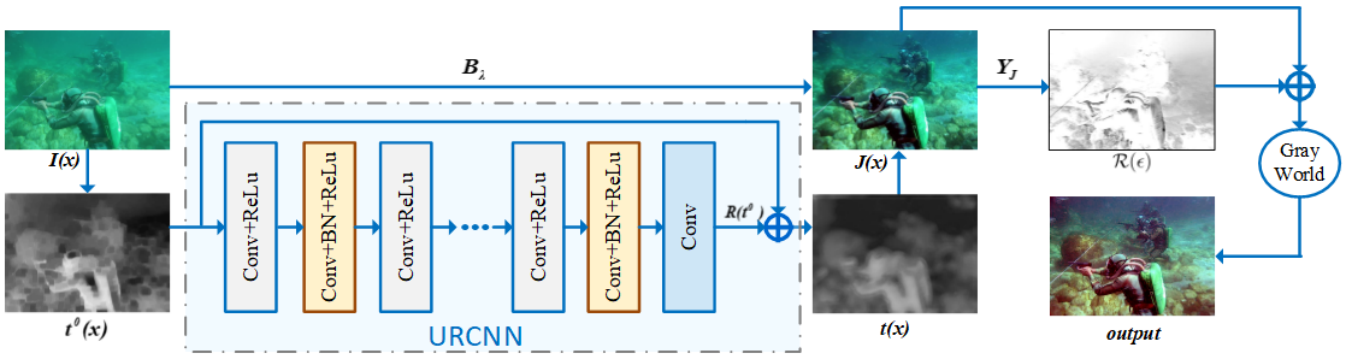


Fig. 1. The flowchart of our proposed underwater image enhancement network.

scene residual and propose an effective solution for illumination balance based on gray-world assumption and the feature of multi-scale local max saturation. Experiences demonstrate that although trained on synthetic dataset, the proposed deep network works well on real-world underwater images. Moreover, no matter on visual effect or scores of image quality evaluation standard, the proposed algorithm can perform excellent capability.

2. THE PROPOSED FRAMEWORK

2.1. Underwater Residual CNN

In this section, we present an underwater residual CNN model (URCNN), then we introduce its network architecture design and learning process. Aggregating both prior knowledge and data information, the proposed URCNN aims to learn a precise transmission map. For network architecture design, we modify the VGG network [10] and set a suitable depth to make it be effective for learning transmission. In learning process, we adopt the residual learning strategy formulation and batch normalization to improve learning performance.

2.1.1. Network Architecture

The input of URCNN is a rough feature map generated by the physical prior, which can be regarded as $t^0 = t + R(t^0)$ (more details about the prior will be illustrated in Sec.2.2). According to [11], when the original mapping is more similar to identity mapping, the residual mapping will be much easier to be optimized. As it is obvious that t^0 is much more like the latent t than the residual image $R(t^0)$, we adopt a residual learning formulation to train a residual mapping $R(t^0) \approx t - t^0$, while transmission learning method [12] use end-to end CNN to learn a mapping function $F(t^0) = t$ to predict transmissions. And the desired transmission can be described as

$$t = t^0 - R(t^0). \tag{2}$$

As shown in Fig. 1, URCNN consists of n blocks. The first block consists of 64 filters and rectified linear units (ReLU, $\max(0, x)$). To guarantee powerful expression of features and reduce number of parameters, we adopt small convolutional filter of size 3×3 to generate feature maps. Therefore, the receptive field of URCNN with depth of d should be $(2d+1) \times (2d+1)$. In the next $2 \sim (n-1)$ blocks, every convolution has 64 filters of size $3 \times 3 \times 64$, and we set BN after convolution in block 2 and block $(n-1)$ to speed up training process. In the last block, we use a filters of size $3 \times 3 \times 64$ to reconstruct the final residual mapping. To make the feature maps of each middle layer have a same size as the input, we use zero padding strategy before convolution. Finally substituting the residual mapping $R(t^0)$ and the feature map t^0 into Eq. (2), we can obtain the refined transmission map.

To sum up, there are two main features of URCNN: applying residual learning strategy to correct handcrafted physical prior, and using batch normalization to improve learning performance as well as training speed. Combining convolution with ReLU, the proposed network can gradually enhance the structure of scene depth from rough input through the hidden layers. Moreover, to make sure that each feature map of the middle layers has the same size as the input, we directly pad zeros before convolution, and this strategy can not result in obvious boundary artifacts.

2.1.2. Training

During the training, learning the residual mapping is achieved by minimizing the loss function as follows:

$$l(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|R(t_i; \Theta) - (t_i^0 - t_i^*)\|_F^2, \tag{3}$$

where N is the number of training data, Θ is trainable parameters to be learned. We minimize the loss using the SGD method with the backpropagation learning rule [13].

Furthermore, moderate number of batch normalization (BN) layers can change the distributions of internal non-

linearity inputs to avoid the internal covariate shift [14]. Through experiments we find that just inserting the BN between convolution layer and ReLU just in the first and last blocks of deep CNN is adequate to lead to a stable training.

2.2. Underwater Haze Remove

As blue and green light have less attenuation, the corresponding color channels can reserve more intact haze information. Considering that the pixels of background (water) are usually the pixels with the strongest background scattering light, we select the set of the first 0.1% brightest pixels from blue and green channel to ensure the stability and robustness of the algorithm and represent it as $\Omega(x)$. Then to exclude the interference of the black or blue objects, we locate the pixel that has maximum blue-red or green-red difference as the point where the background light is, which can be formulated as follows:

$$B_\lambda = \max_{x\in\Lambda}[I_g(x) - I_r(x), I_b(x) - I_r(x)], \tag{4}$$

To incorporate the physical prior into the learning process of the proposed network, inspired by dark channel prior [15] we extract the image feature t^0 which implies probable structural information of depth as follows:

$$t^0(x) = 1 - \min_{y\in\Omega(x)}[\min_{\lambda} J_\lambda(y)]/B_\lambda, \lambda \in \{g, b\}, \tag{5}$$

where $\Omega(x)$ is a local patch centered at the pixel x. Afterwards taking t^0 as the input of URCNN, we obtain the desired transmission from Eq (2). Since B_λ and $t(x)$ were estimated. To avoid the case that the $t(x)$ is close to zero, the scene radiance $J(x)$ can be recovered as follows:

$$J(x) = I(x) - B_\lambda/\max(\xi, t(x)), \tag{6}$$

where the small constant ξ is used to prevent division by zero.

2.3. Scene Residual Reduction

Till now, color cast problem is still not be properly addressed in the haze-removed $J(x)$. So in this part, we introduce a new scene residual component into the solution of colour cast and regard observation as $J(x) = \mathcal{H}C(x) + \mathcal{R}(\epsilon)$, where $\mathcal{H}$ is a degradation matrix, $C(x)$ is the latent clean image, $\mathcal{R}(\epsilon)$ is scene residual, representing local deviation of color feature ϵ (such as hue, saturability and illuminance etc).

We experimentally find that generally there are local exaggerated exposure exiting in haze-removed $J(x)$, this may be due to the over-estimation of depth information. Consequently we select the illuminance as scene feature to compensate the exposure deviation. According to [16], the scene residual can be estimated as follows:

$$\mathcal{R}(\epsilon) = J. * \left[1 - f_{guide}(\frac{Y_J Y_I + \lambda Y_I^2}{Y_J^2 + \lambda Y_I^2})\right], \tag{7}$$

where Y_J and Y_I are the illumination intensity of the restored image $J(x)$ and input image $I(x)$ respectively, $f_{guide}(\cdot)$ represents guidefilter [17]. Inspired by the Gray World assumption in color constancy [18], we define $\mathcal{H}$ as a linear constraint to balance the wavelength absorption in different color channels, and by bringing $\{\sum_r J^c = \sum_g J^c = \sum_b J^c\}$ into existence we can enhance the picture $J^c = J(x) - \mathcal{R}(\epsilon)$, of which the exposure is normal.

3. EXPERIMENT RESULTS

In this section, we conduct extensive experiments to evaluate our proposed underwater image enhancement framework.

3.1. Data Generalization

To train the proposed URCNN, we generate a dataset with ground truths of transmission and feature maps. Firstly, we randomly select 1000 samples from the NYU Depth dataset [20] to synthesize the ground truth t^* by $t^*(x) = e^{-\beta_\lambda d(x)}$, where $d(x)$ is the realistic depth. Considering that light with different color has different attenuation degree, we set random medium extinction coefficient $\beta = [\beta_r, \beta_g, \beta_b]$ ($\beta_r \in [0.05, 0.15], \beta_g \in [0.6, 0.9], \beta_b \in [0.7, 1.0]$) for each depth.

Combining the ground truth t^* with the corresponding clean image, we can synthesize underwater images using Eq. (1), here we generate random background light $B = [B_r; B_g; B_b]$ ($B_r \in [0.05, 0.2], B_g \in [0.6, 0.8], B_b \in [0.7, 1.0]$), then we can estimate feature map t^0 through the method put forward by section 2.2, next we crop two 180×180 regions from each t^0 and corresponding t^* to build up training set $\{t_i^0; t_i^*\}_{i=1}^{2000}$. Finally we randomly select 1800 pairs of data for training and the other 200 pairs for testing.

3.2. Parameter Setting and Network Training

We initialize the weights of deep network by the method in [21] and use SGD with the weight decay of 0.0001, the momentum of 0.9 and the mini-batch size of 4. 60 epochs are

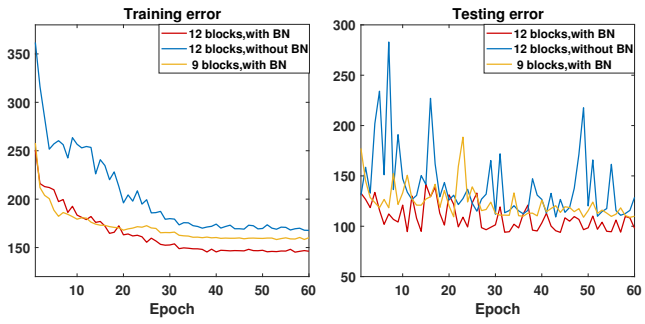
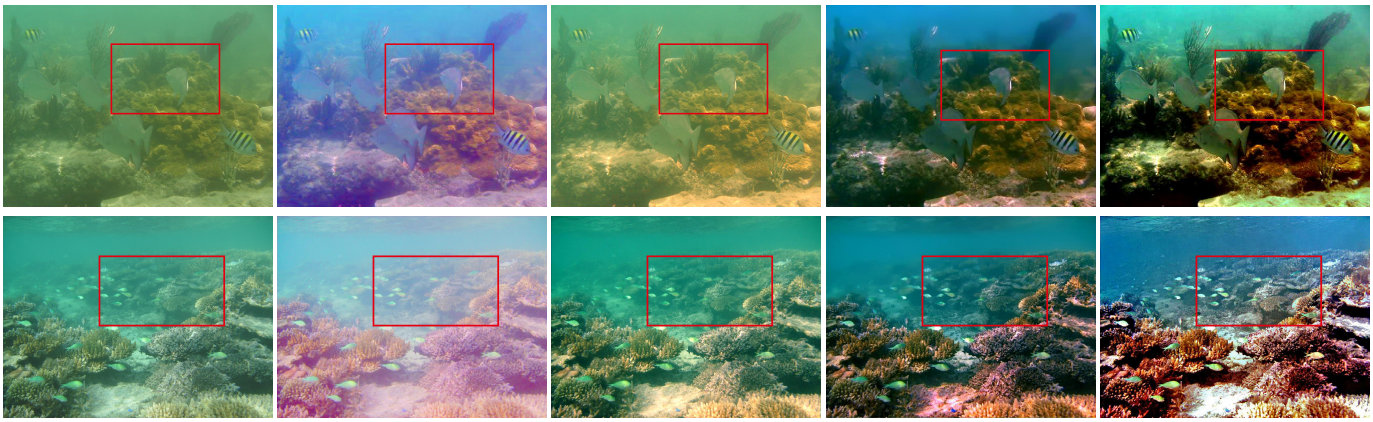


Fig. 3. Training error and testing error of different network structure.



Input Jobson *et al.* [19] Chiang *et al.* [6] Li *et al.* [7] Ours

Fig. 2. Subjective result comparison.

Table 1. Average BM, SSEQ and NIQE scores.

Method	[19]	[6],	[7]	[8]	Ours
BM	0.34	0.33	0.30	0.33	**0.26**
SSEQ	35.57	34.97	33.81	35.42	**32.57**
NIQE	4.87	4.33	3.83	4.64	**3.62**

trained. The learning rate decays exponentially from 0.1 to 0.0001 for the 60 epochs. We also experimentally find that a 12-block (25-layer) depth is appropriate. Fig. 3 illustrates the convergence performance of our network with different architecture, it can be inferred that a suitable depth and BN can bring about better learning performance.

The MatConvNet package [22] is used to train the URCN-N models, which takes about 6 hours approximately. In addition, all experiments are carried out on the Matlab (R2016a) environment, PC with Intel(R) Core(TM) i5-6300HQ CPU 2.30GHz and Nvidia GeForce GTX 960M GPU.

3.3. Quantitative Evaluations

Unlike the common image quality assessment or image restoration areas, it is difficult to obtain underwater ground truth image, thus on an underwater image dataset provided by [8], we employ non-reference image quality evaluation methods to compare the efficiency of the proposed algorithm with other typical methods, including the image-based enhancement algorithms Jobson *et al.* [19] and the prior-based algorithms [6], Li *et al.* [7] and Galdran *et al.* [8]. The criterion we adopt are three popular metrics: the Blur Metric (BM) [23], Spatial-Spectral Entropy-based Quality (SSEQ) [24] and Natural Image Quality Evaluator (NIQE) [25]. B-M evaluates the image quality in term of blur perception, SSEQ utilizes local spatial and spectral entropy features to assess image quality, NIQE makes score by calculating the distance between feature model and the established model. Respectively lower score means higher image quality. Table 1 shows the average scores on dataset and it is obvious that our method receives the highest on the different criterions.

3.4. Qualitative Evaluations

To verify the effectiveness of the algorithm, we collect degraded underwater images in different scene[1], on which we make visual comparisons with state-of-the-art approaches [19, 6, 7]. As shown in Fig. 2, it can be observed that results of Jobson can remove the blue and green tones well, but the haze effect still exists. The choice of parameters in algorithm strongly limits results of Chiang, so the adaptability needs to be improved. And results of Li tends to over-smooth local details especially in scene far away from the camera, so as to lose some image imformation in detail. Comparatively, it is obvious that the proposed algorithm can remove haze effect well, enhance the detailed information as clean as possible and correct color cast at the meantime.

4. CONCLUSION

This paper proposes a residual convolutional neural network to underwater image enhancement, of which the outstanding learning capacity leads to an appropriate transmission. Cooperated with the solution of scene residual, the proposed method can remove haze effect, increasing contrast and restore natural appearance effectively. Consequently experiences demonstrate that the property of our method is better than others compared. Moreover, utilizing the scene residual to reduce the color shift may provide a feasible research thought waited to explore deeply.

[1]All the test images and more underwater images captured in underwater target identification tasks can be found on https://github.com/dlut-dimt/Underwater-Database.

5. REFERENCES

[1] M. S. Hitam, E. A. Awalludin, W. N. J. H. W. Yussof, and Z. Bachok, "Mixture contrast limited adaptive histogram equalization for underwater image enhancement," in *International Conference on Computer Applications Technology*, 2013.

[2] S. Bazeille, I. Quidu, L. Jaulin, and J.-P. Malkasse, "Automatic underwater image pre-processing," in *Caracterisation Du Milieu Marin*, 2006.

[3] Z.-u. Rahman, D. J. Jobson, and G. A. Woodell, "Multiscale retinex for color image enhancement," in *Image Processing*, 1996.

[4] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Conference on Computer Vision and Pattern Recognition*, 2012.

[5] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE Journal of Oceanic Engineering*, vol. 15, no. 2, pp. 101–111, 1990.

[6] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1756–1769, 2012.

[7] Y. Li, F. Guo, R. T. Tan, and M. S. Brown, "A contrast enhancement framework with jpeg artifacts suppression," in *European Conference on Computer Vision*, 2014.

[8] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *Journal of Visual Communication and Image Representation*, vol. 26, pp. 132–145, 2015.

[9] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice, "Initial results in underwater single image dehazing," in *OCEANS*, 2010.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*, 2016.

[12] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European Conference on Computer Vision*, 2016.

[13] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *International Conference on Computer Vision*, 2011.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[15] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[16] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *Computer Vision and Pattern Recognition*, 2014.

[17] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[18] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin institute*, vol. 310, no. 1, pp. 1–26, 1980.

[19] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," *European Conference on Computer Vision*, 2012.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision*, 2015.

[22] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM International Conference on Multimedia*, 2015.

[23] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric.," in *Human Vision and Electronic Imaging*, 2007.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[25] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.