Underwater Mixed Spatial Attention Network

Qiang Pan¹, Yueting Yang², Chuansheng Yang¹,³, Chao Wang¹,³* and Anhui Tan¹,³

¹ School of Information Engineering, Zhejiang Ocean University, Zhoushan, Zhejiang, 316022, China
² School of Mathematics and Statistics, Beihua University, Jilin, Jilin, 132013, P.R. China
³ Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province, Zhoushan, Zhejiang, 316022, China

* Corresponding author’s e-mail: wangchao@zjou.edu.cn

Abstract. High-quality underwater images play an important role in obtaining and understanding underwater information. However, raw underwater images have many problems, such as low contrast, chromatic aberration, blur and low light, which seriously restrict the progress of other underwater tasks. In this paper, we propose an Underwater Mixed Spatial Attention Network (UMSAN) for underwater image enhancement. The evaluation on the EUVP dataset and some other real underwater images demonstrates that our network is sufficient against several of the advanced models.

1. Introduction
In the exploration of the underwater world, underwater image is an important part of underwater information. However, it’s hard to accurately and reliably extract useful information from underwater images, because images from the complex and changeable underwater environment have some problems, e.g., color offset, blur, low brightness, uneven illumination and so on.

According to the underwater imaging model, the underwater images consist of three parts: floating particles, light reflected from objects, and underwater ambient light:

\[ I(x) = J(x)t(x) + A(1 - t(x)), \quad (1) \]

where \( I(x) \) is the observed underwater image, \( J(x) \) is the recovered image, \( A \) is the background light, and \( t(x) \) is the medium transmission map. Many underwater image enhancement works are based on the above model.

In recent years, with the development of artificial intelligence, many underwater enhancement methods have been presented. However, some methods are not sensitive to the color offset and texture details, which results in image distortion.

Several models [2,3,4,5,6,7,8,9] based on deep Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) show state-of-the-art performance in image enhancement.

Considering the effect of artificial lighting in water, [10] combines the image dehazing algorithm and the characteristics of underwater images for color correction and scattering removal. Li et al. [11] realizes the characteristics of the selective attenuation of different channels and proposes an underwater image restoration method.
In this article, we propose an Underwater Mixed Spatial Attention Network (UMSAN) to solve the underwater image enhancement task. A texture channel is used as additional input at first, aiming to ensure the texture structure of the result. On the whole, we use three groups of densely connected blocks to achieve the enhancement function, and the Dense Block mainly contains the Spatial Channel Attention (SCA) module, which helps learn the global features and color restoration.

In summary, the contributions of our article are as follows:

- An end-to-end underwater mixed spatial attention network is proposed for underwater image enhancement. As far as we know, it achieves better qualitative and quantitative performance than some current state-of-the-art methods.
- The network uses images with an additional texture channel as input to ensure that the output images have a clear texture structure and excellent enhancement effect.
- A hybrid method of SCA is adopted to help the underwater image recovery.

2. Methodology

In this section, the method to generate the texture of images is introduced at first. Then we explain the overall structure of our network and main modules in detail.

2.1. Texture extraction

To maintain a better structure for the haze-free images, the texture generation function is used to get the texture map of an underwater image as the fourth channel of the input. There are various methods to extract texture structure that have been proposed, such as Sobel, Roberts, and Canny. However, these methods still have some disadvantages and limitations, e.g., Sobel and Roberts position the edge inaccurately. The following formula is the method used in this paper:

$$u_i = \frac{\nabla I_{HR}}{\sqrt{1 + |\nabla I_{HR}|^2}}, \quad i \in \{x, y\},$$

where $x$ and $y$ represent the horizontal and vertical directions. $\nabla$ and $\text{div}$ denote the gradient operation and divergence operation, respectively. The proposed formula is sufficient and suitable enough for our work.

2.2. Underwater mixed spatial attention network

The structure of the Underwater Mixed Spatial Attention Network (UMSAN) is shown in figure 1. The changes in the number of channels are also marked in the figure 1, and the size of feature maps will not
change during transmission. As we can see, the network can be divided into the following three parts: head module, body module, and tail module.

The Head Module is composed of a texture generation function and a superficial $3 \times 3$ convolution layer. It can combine the original RGB image with an additional texture channel into a 4-channel image, and then convert the 4-channel image into the 64-channel feature map $x_0$:

$$x_0 = F_{\text{head}}^{4 \rightarrow 64}(I_{\text{in}}),$$

$$I_{\text{out}} = F_{\text{tail}}^{64 \rightarrow 3}(\hat{x}).$$

Similar to the Head Module, the Tail Module uses a Residual Block [13] and two convolution layers to restore the feature map $\hat{x}$ to an intuitive color image.

As is shown in figure 1, the Body Module is composed of three groups of blocks, and each group contains 4 Dense Blocks.

The key point of our network is the Dense Block (DB). In this block, the feature information extracted by previous blocks can be appropriately fused, and the number of channels can keep uniform in the transmission process. These Dense Blocks of each group are densely connected, maintaining the feature information or gradient better during the transmission process and reducing the number of parameters significantly. The input features $x_0, x_1, \ldots, x_{i-1}$ come from different feature spaces, so we need to fuse and project them first. A convolution layer and an activation layer are used to compress these features to 64 channels, and then a specially designed Spatial Channel Attention (SCA) block is adopted to project them into the same feature space. The processed features of the $i$-th fusion part are obtained by:

$$x_{\text{fuse}}^i = F_{\text{fuse}}^{i \times 64 \times 64}([x_0, x_1, \ldots, x_{i-1}]),$$

$$x_{\text{proj}}^i = F_{\text{SCA}}^{64 \rightarrow 64}(x_{\text{fuse}}^i),$$

where $[x_0, x_1, \ldots, x_{i-1}]$ presents the concatenation of $x_0, x_1, \ldots, x_{i-1}$ along the channel. After concatenating the $x_{\text{proj}}^i$ and $x_{i-1}$, a convolution lay is used to integrate information. The output of the $i$-th Dense Block $x_i$ can be got by:

$$x_{\text{conv}}^i = F_{\text{conv}}^{128 \rightarrow 64}([x_{\text{proj}}^i, x_{i-1}]),$$

$$x_i = F_{\text{FA}}^{64 \rightarrow 64}(x_{\text{conv}}^i + x_{i-1}),$$

where $F_{\text{FA}}$ represents the operation of Future Attention from FFA-Net [14].

The SCA block is illustrated in figure 2. We first use a convolution layer to carry out preliminary feature extraction. SA focuses on the significant areas of color and impurities in the water. CA, on the other hand, focuses on useful channels that are sensitive to underwater color and impurities. Therefore, we combine SA and CA to focus on information regions and across channels, enabling the network to effectively remove the underwater fog of different chromatic aberrations and impurities in a single image.

Figure 2. The architecture of Spatial Channel Attention Block (SCA Block).
2.3. Loss function
This article mainly uses $\ell_1$ loss and perceptual loss [15] to train the proposed network:

$$L = L_1 + \lambda L_{per},$$

(9)

where $\lambda$ presents the weight of perceptual loss. The $\ell_1$ loss is used to evaluate the difference between the enhanced image and the ground-truth, while perceptual loss is adopted to compare the deep semantic features such as content and structure.

3. Experimental results
In this section, we cover the experimental setup in detail. Based on indicators such as PSNR and SSIM, our method is compared with other SOTA underwater methods on the EUVP data set.

All the compared networks are trained on $256 \times 256$ patches with batch size of 1 for $1 \times 10^5$ iterations. We use Adam optimizer with the initial learning rate of $1 \times 10^{-4}$ using the cosine annealing strategy. All the experiments are worked using PyTorch 1.7.0 on a GPU Nvidia GeForce RTX 2080Ti (11G) desktop PC. The quantitative experimental results are shown in table 1 below.

Table 1. Quantitative comparison for average PSNR and SSIM values on 1K paired test images of EUVP dataset.

| Methods     | Uw-HL  | Mband-EN | Res-WGAN | Res-GAN | LS-GAN | Pix2Pix | UMSAN(ours) |
|-------------|--------|----------|----------|---------|--------|---------|-------------|
| PSNR        | 18.85  | 12.11    | 16.46    | 14.75   | 17.83  | 20.27   | 21.68       |
| SSIM        | 0.7722 | 0.4565   | 0.5762   | 0.4685  | 0.6725 | 0.7081  | 0.8073      |

As shown in figure 3, underwater images processed by our network maintain better color and contrast. Our network also shows a good ability to restore the image distortion.

![Figure 3. Enhanced image presentation, performance in color, shape and structure.](image-url)
Qualitative analysis is shown in figure 4 below. The results obtained by these methods are qualitatively compared with some blur image samples. Underwater images tend to be bluer because blue light has a shorter wavelength, we believe that the less blue in the image, the better the model effect will be. Compared with the other eight methods, our results are closer to the ground truth.

Figure 4. We compared eight groups of experiments. As can be seen from the figure, we handle the color offset problem well.

4. Conclusion
In this paper, we propose an Underwater Mixed Spatial Attention Network that uses spatial attention to learn global features and feature recovery, and channel attention to learn cross-scale features. We use dense connections to improve information utilization and we use texture information to help restore the image to better detail. The underwater image is well recovered. A large number of experiments show that our method can produce better visibility and achieve better color restoration.

Although our method achieves better performance than the most advanced underwater image enhancement methods, it still has some limitations. First, our method mainly focuses on removing the color projection and improving the contrast; However, real-world underwater images are subject to all kinds of degradation, such as noise and uneven light. Second, due to limited test data and reliable evaluation indicators, the true gap between the enhanced results and the real underwater image is not clear.

Acknowledgments
This work was supported by the Key Project of Natural Science Foundation Joint Fund of Jilin Province (2020122367JC); Science and Technology Development Project Program of Jilin Province (20190303132SF); National Natural Science Foundation of China (12001005).

References
[1] McCartney, E.J. (1976) Optics of the atmosphere: scattering by molecules and particles. New York.
[2] Zhu, J.Y., Park, T., Isola, P., Efros, A. A. (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision. pp. 2223-2232.
[3] Fabbri, C., Islam, M.J., Sattar, J. (2018, May) Enhancing underwater imagery using generative adversarial networks. In 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 7159-7165.
[4] Isola, P., Zhu, J.Y., Zhou, T., Efros, A. A. (2017) Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125-1134.
[5] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S. (2017) Least squares generative adversarial networks. In Proceedings of the IEEE international conference on
computer vision. pp. 2794-2802.

[6] Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S. (2017) Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1222-1230.

[7] Arjovsky, M., Chintala, S., Bottou, L. (2017, July) Wasserstein generative adversarial networks. In International conference on machine learning. pp. 214-223.

[8] 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(4): 2822-2829.

[9] 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.

[10] Chiang, J. Y., & Chen, Y. C. (2011) Underwater image enhancement by wavelength compensation and dehazing. IEEE transactions on image processing, 21(4):1756-1769.

[11] Li, C., Quo, J., Pang, Y., Chen, S., Wang, J. (2016, March) Single underwater image restoration by blue-green channels dehazing and red channel correction. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1731-1735.

[12] He, K., Zhang, X., Ren, S., Sun, J. (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778.

[13] Johnson, J., Alahi, A., Fei-Fei, L. (2016, October) Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision. pp. 694-711.

[14] Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H. (2020, April) FFA-Net: Feature fusion attention network for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence Vol. 34, No. 07, pp. 11908-11915.

[15] Simonyan, K., Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.