Semantic-Driven Context Aggregation Network for Underwater Image Enhancement

Dongxiang Shi[1], Long Ma[1], Risheng Liu[2], Xin Fan[2], and Zhongxuan Luo[1(✉)]

[1] School of Software, Dalian University of Technology, Dalian 116024, China
zxluo@dlut.edu.cn
[2] DUT-RU International School of Information Science and Engineering,
Dalian University of Technology, Dalian 116024, China

Abstract. Recently, underwater image enhancement has attracted broad attention due to its potential in ocean exploitation. Unfortunately, limited to the hand-crafted subjective ground truth for matching low-quality underwater images, existing techniques are less robust for some unseen scenarios and may be unfriendly to semantic-related vision tasks. To handle these issues, we aim at introducing the high-level semantic features extracted from a pre-trained classification network into the image enhancement task for improving robustness and semantic-sensitive potency. To be specific, we design an encoder-aggregation-decoder architecture for enhancement, in which a context aggregation residual block is tailored to improve the representational capacity of the original encoder-decoder. Then we introduce a multi-scale feature transformation module that transforms the extracted multi-scale semantic-level features, to improve the robustness and endow the semantic-sensitive property for the encoder-aggregation-decoder network. In addition, during the training phase, the pre-trained classification network is fixed to avoid introducing training costs. Extensive experiments demonstrate the superiority of our method against other state-of-the-art methods. We also apply our method into the salient object detection task to reveal our excellent semantic-sensitive ability.

Keywords: Underwater image enhancement · Semantic feature · Context aggregation network · Feature transformation module
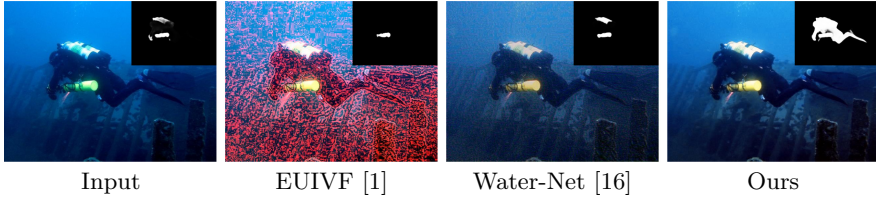
1 Introduction

Underwater images often suffer from severe color casting and contrast decreasing caused by light absorption and scattering. This degradation not only disturbs

| Input | EUIVF [1] | Water-Net [16] | Ours |

Fig. 1. Underwater image enhancement results and the corresponding saliency maps predicted by a salient object detection method F³Net [26] on USOD dataset [13]. Obviously, our method performs better than other state-of-the-art methods, especially without noise or artifact. The superior salient object detection result further reveals semantic-sensitive property of the proposed method in high-level vision tasks.

visual quality of images but also has negative impact on visual tasks, *e.g.*, salient object detection and instance segmentation. Moreover, it is hard for submarine robotic explorers equipped with visual system to autonomously explore underwater environment. Therefore, underwater image enhancement has drawn much attention in recent years.

Underwater image enhancement aims to generate clear and natural images against several degradations (e.g., color cast, low contrast and even detail loss). The existing image underwater enhancement methods can be roughly divided into three categories: non-physical model-based methods, physical model-based methods and data-driven methods. In early enhancement methods, some are directly used to enhance underwater images regardless of underwater imaging model. [12,28] enhanced image contrast by expanding the dynamic range of the image histogram. [6,24] corrected color cast based on color assumption of natural images. To enhance contrast and correct color cast simultaneously, Ancuti *et al.* [1] proposed a fusion-based method to fuse several kinds of enhancement images. From the perspective of underwater physical imaging model, underwater image enhancement is regarded as an inverse problem. As He *et al.* [10] proposed Dark Channel Prior (DCP) which achieved an outstanding performance in single image dehazing, several DCP variants [2,9] were proposed by exploring different underwater prior. However, these methods might be restricted with some assumptions and simple physical imaging model. When the assumption and prior are less adaptive to unseen scene, these methods may generate severe artifacts, as shown in Fig. 1(b).

With the success of deep learning in various vision tasks [4,18,20,21] some learning-based underwater image enhancement methods are proposed. Wang *et al.* [25] proposed UIE-Net, which is composed of two branches to estimate attenuation coefficient and transmission map respectively. Li *et al.* [15] directly reconstructed the clear latent natural images from inputs instead of estimating the parameters of underwater imaging model. Recently, some methods [14,17] adopted the generative and discriminative mechanism to improve the capability of network.

However, existing methods for underwater image enhancement still suffer from color distortion and unclear background details in some unknown scenarios, and these algorithms may be adverse to the high-level vision tasks. To settle these issues, we propose a semantic-driven context aggregation network that utilizes multi-scale semantic features to guide detail restoration and color correction. As shown in Fig. 1, our method realizes a stronger adaptation on the unknown scenario and performs a better semantic-sensitive property than other state-of-the-art methods. To be specific, we first build an encoder-aggregation-decoder enhancement network to establish the map between underwater low-quality observations and high-quality images. Then we introduce an encoder-type classification network that has been pre-trained on ImageNet [3], to provide semantic cues for better enhancement. We further construct a multi-scale feature transformation module to convert semantic features to the desired features of the enhanced network. Concretely, our main contributions can be concluded as the following three-folds:

– We successfully incorporate semantic information into a context aggregation enhancement network for underwater image enhancement to achieve robustness towards unknown scenarios and be friendly to semantic-level vision tasks.
– We construct a multi-scale feature transformation module to extract and convert effective semantic cues from the pre-trained classification network to assist in enhancing the low-quality underwater images.
– Extensive experiments demonstrate that our method is superior to other advanced algorithms. Moreover, the application on salienct object detection further reveals our semantic-sensitive property.

2   Method

The overall architecture of our proposed method is shown in Fig. 2. In this section, we begin with describing the overall architecture in Sect. 2.1, then introduce the semantic feature extractor in Sect. 2.2, the proposed multi-scale feature transformation module in Sect. 2.3, and finally the context aggregation enhancement network and the loss function in Sect. 2.4.

2.1   The Overall Architecture

Semantic information extracted from high-level network has potential to facilitate underwater image enhancement with more accurate and robust predictions. Thus, we propose a semantic-driven context aggregation network for underwater image enhancement, as illustrated in Fig. 2. Our whole architecture includes a semantic feature extractor, a multi-scale feature transformation module, and a context aggregation enhancement network. Specifically, we adopt a general VGG16 classification network to extract the semantic features, then the extracted multi-scale semantic features with abundant information are fed into the enhancement network through the multi-scale feature transformation
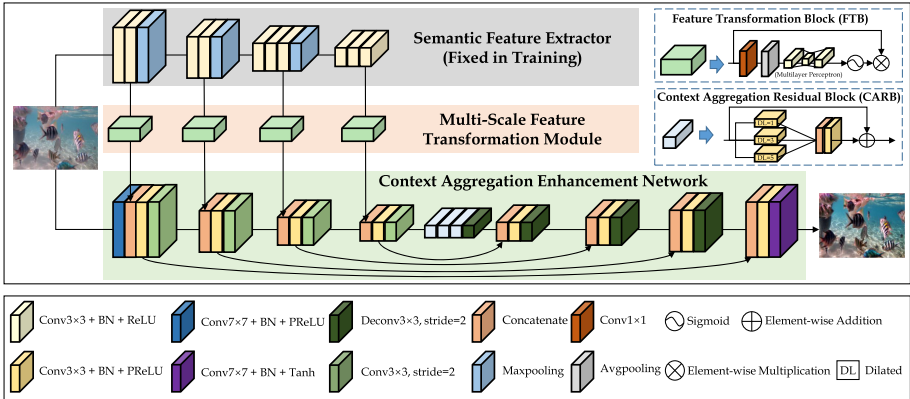
Fig. 2. Overview of the proposed semantic-driven context aggregation network for underwater image enhancement. Our network is composed of three basic modules. a) Semantic feature extractor, which consists of a pre-trained VGG16 classification network to extract semantic features. b) Multi-scale feature transformation module, which is used to concentrate on beneficial semantic features to guide underwater image enhancement. c) Context aggregation enhancement network, which integrates semantic features and enhancement features to generate clear and natural underwater images.

module. The feature transformation blocks process the affluent features in an attentive way, which benefit the enhancement network in restoring details and correcting color casts for underwater images.

2.2   Semantic Feature Extractor

A common perception is that the shallower features from backbone network of high-level task consider texture and local information, while the deeper features focus more on semantic and global information. This motivates us to explore the ability of multi-scale semantic features. Specifically, we extract features of the first four scales (denoted as F_n, $n \in [1, 4]$) from a VGG16 network pre-trained on ImageNet [3]. To avoid information loss caused by pooling operation, we select features before pooling layers of each stage. The abundant use of guidance information from multi-scale features allows us to better handle the challenges in low-quality underwater images. Besides, in order to avoid introducing the training costs, semantic feature extractor is fixed during training phase.

2.3   Multi-scale Feature Transformation Module

With the extracted multi-scale semantic features, we aim at incorporating the abundant semantic information into the enhancement network. A straightfor-ward method is directly combining semantic features with the features in the enhancement network, *e.g.*, addition or concatenation. However, this may ignore the distinguishability of semantic features and introduce redundant information

into the enhancement network. Inspired by the attention mechanism in [11], we propose a feature transformation block (FTB) to attentively select and incorporate the key prior for the enhancement network. We first adopt a 1×1 convolution block to match the channel dimensions of features, then exploit the inter-channel dependences and reweight the importance of each channel to highlight the vital information and suppress the unnecessary ones. The process can be formulated as the following equation:

$$F_o^n = S(MLP(Avgpool(Conv_{1\times1}(F^n)))) \odot F^n, n = 1, 2, 3, 4 \qquad (1)$$

where $Conv_{1\times1}(\cdot)$ denotes a convolution block consisting of 1×1 Conv, BN and PReLU. $Avgpool(\cdot)$ and $MLP(\cdot)$ denote global average pooling and multilayer perceptron respectively, and $S(\cdot)$ denotes a sigmoid function. $\odot$ is pixel-wise multiplication, and F_o^n is the output of feature transformation block.

Through the multi-scale feature transformation module, we can suppress and balance the comprehensive information to guide the enhancement network to achieve finer predictions. The extensive experiments in Sect. 3.3 also demonstrate the effectiveness of our proposed FTB.

2.4   Context Aggregation Enhancement Network and Loss Function

For the enhancement network, we employ an encoder-aggregation-decoder network. On the one hand, in the encoder part, the multi-scale semantic features are attentively incorporated into the corresponding level through FTB. On the other hand, with the goal of extracting global contextual information from the combined features, we adopt Context Aggregation Residual Block (CARB) to further enlarge the respective field following [4]. Finally, the comprehensive and attentive features are fed into the decoder to generate clearer and more natural predictions for underwater images.

In order to apply an appropriate loss function, there are two key factors needed to be considered. First, the widely-used $\mathcal{L}_2$ loss usually leads to over-smoothed results. Thus, we choose to adopt $\mathcal{L}_1$ loss function as a pixel-wise objective function. Second, considering that pixel-wise loss function is not sensitive to image structure characteristics (*e.g.*, luminance, contrast), we simultaneously adopt MS-SSIM [27] to guide the network to focus on image structure information. As a result, the overall loss function can be formulated as

$$\mathcal{L}_{total} = \lambda \mathcal{L}_1 + (1 - \lambda) \mathcal{L}_{MS-SSIM}, \qquad (2)$$

where λ is a balance parameter. In our experiment, λ was empirically set as 0.2.

3   Experiments

In this section, we first introduce the adopted datasets and implementation details. Next we comprehensively compare the performance of our approach with other state-of-the-art methods. We also perform ablation experiments to analyze the effect of main components in our proposed network. Finally, we evaluate our method in the application of underwater salient object detection.

Table 1. Quantitative evaluation on two underwater benchmark datasets. The best two results are shown in red and blue fonts, respectively.

Dataset	Metric	EUIVF	TSA	UDCP	UWCNN	Water-Net	F-GAN	Ours
UIEB	PSNR ↑	18.300	14.119	12.320	14.891	**19.492**	18.300	**21.611**
	SSIM ↑	0.824	0.631	0.539	0.733	**0.850**	0.812	**0.894**
	UIQM ↑	**2.757**	1.996	1.772	2.428	2.307	2.740	**2.893**
	NIQE ↓	**4.049**	4.166	4.304	4.488	4.485	4.393	**3.923**
	TM ↑	**1.120**	0.747	0.640	0.439	0.912	0.789	**1.184**
UCCS	UIQM↑	**3.073**	2.512	2.098	2.536	**3.150**	2.982	2.983
	NIQE↓	4.542	4.842	5.131	**4.411**	5.696	4.543	**3.574**
	TM ↑	**0.598**	0.432	0.086	0.173	0.527	0.326	**0.613**

3.1   Experimental Setup

Datasets. To evaluate the performance and generalization ability of our model, we conduct experiments on two underwater benchmark datasets: Underwater Image Enhancement Benchmark (UIEB) [16] and Underwater Color Cast Set (UCCS) [19]. The UIEB dataset includes 890 raw underwater images with corresponding high-quality reference images. The UCCS dataset contains 300 real underwater no-reference images in blue, green and blue-green tones. In order to achieve a fair comparison, we randomly selected 712 paired images from 890 paired images on UIEB as the training set. The remaining 178 paired images on UIBE and the UCCS dataset are used for testing.

Implementation Details. We implemented our network using Pytorch toolbox on a PC with an NVIDIA GTX 1070 GPU. The training images were all uniformly resized to 640×480 and then randomly cropped into patches with the size of 256×256. During the training phase, we used the ADAM optimizer and set the parameter β_1 and β_2 as 0.9 and 0.999, respectively. The initial learning rate was set as 5e−4 and decreased by 20% every 10k iterations.

Evaluation Metrics. To comprehensively evaluate the performance of various underwater image enhancement methods, we adopt five evaluation metrics, including two widely-used evaluation metrics for data with reference, *i.e.*, Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM); and three reference-free metrics, *i.e.*, Naturalness Image Quality Evaluator (NIQE) [22], Underwater Image Quality Measure (UIQM) [23], and Twice-Mixing (TM) [8].
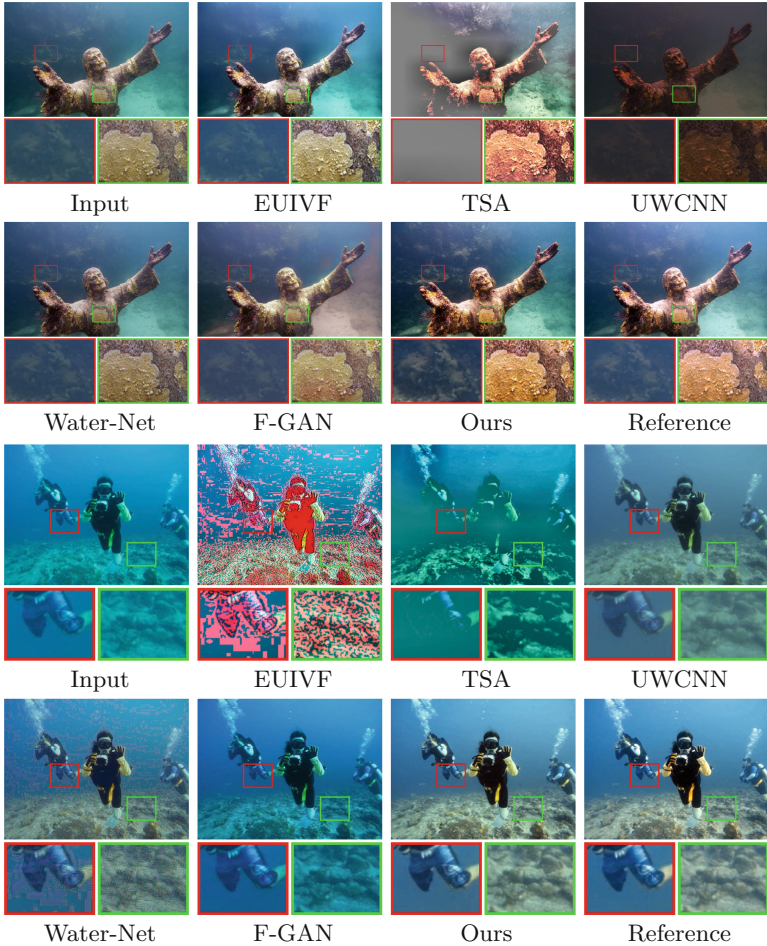
Fig. 3. Visual results of our method and top-ranking methods on UIBE dataset.

3.2 Comparison with the State-of-the-Arts

To fully evaluate the performance of our method, we compare our method with other six state-of-the-art underwater image enhancement methods. There are three conventional methods, *i.e.*, EUIVF [1], TSA [7] and UDCP [5], and three learning-based methods, *i.e.*, UWCNN [15], Water-Net [16] and F-GAN [14].

Quantitative Evaluation. Table 1 shows the validation results of all the competing methods on the UIEB dataset and the UCCS dataset. It is noted that our method outperforms other advanced methods in terms of all the evaluation metrics across the two datasets except the Top-3 UIQM on UCCS dataset. Specially, our method respectively boosts the PSNR and SSIM by 10.87% and 5.17%, compared to the suboptimal method Water-Net on the UIEB dataset.
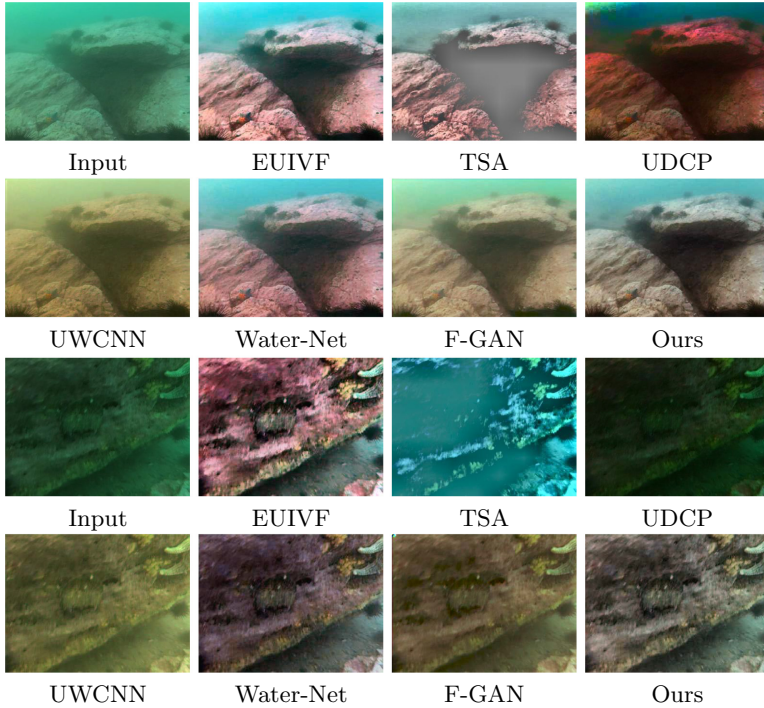
Fig. 4. Visual results of our method and top-ranking methods on UCCS dataset.

Qualitative Evaluation. From a more intuitive view, we visualize some representative results generated from our method and other top-ranking approaches across UIBE and UCCS datasets in Figs. 3 and 4, respectively. It can be easily seen from Fig. 3 that our method can simultaneously restore clearer details of the background and more natural color. Moreover, we can observe from the visual results in Fig. 4 that our method achieves more visually pleasing results on UCCS dataset. For instance, other methods have trouble in restoring natural images (*e.g.*, the reddish color of row 1, column 4, and the yellowish color of row 2, column 1 in Fig. 4), while our results are closer to the real scenarios with lighter color casts.
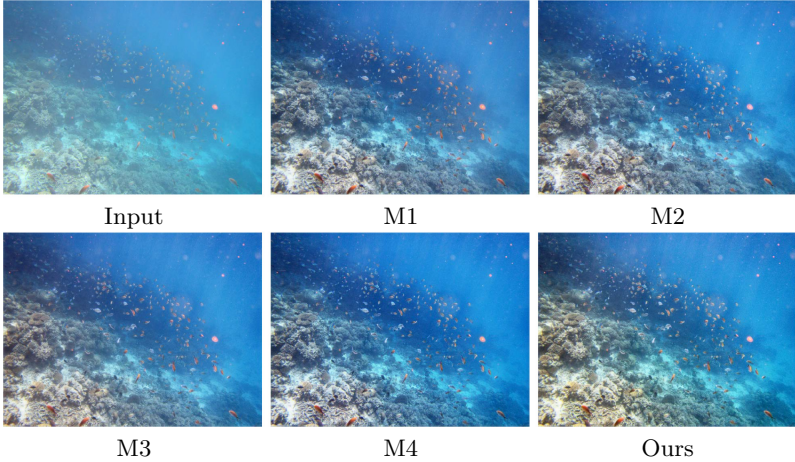
3.3   Ablation Study

In this section, we conduct ablation studies to validate the effectiveness of the key components proposed by our method.

Effectiveness of Semantic Features and FTB. First, we apply the encoder-aggregation-decoder network as our baseline (denoted as "M1" in Table 2). For the sake of investigating the effectiveness of introducing semantic features into the underwater image enhancement task, we directly concatenate the semantic features from the pre-trained VGG16 network and the enhancement features

Table 2. Ablation study of our method on UIEB dataset. "w/" means with the corresponding experimental setting. The best result is shown in **bold** font.

| Method	Baseline	w/F1, F2	w/F3, F4	w/FTB	PSNR/SSIM
M1	✓				20.519/0.883
M2	✓	✓	✓		20.846/0.887
M3	✓	✓		✓	21.014/0.884
M4	✓		✓	✓	21.056/0.887
Ours	✓	✓	✓	✓	**21.611/0.894**

Input   M1   M2

M3   M4   Ours

Fig. 5. Visual results of ablation study on UIEB dataset.

(denoted as "M2"). The comparison results of "M1" and "M2" in Table 2 demonstrate that directly introducing semantic features can bring 0.33dB performance gains towards PSNR on the UIEB dataset. Moreover, we further employ our proposed FTB to attentively extract and transmit the semantic features (denoted as "Ours"). We can see from the comparison results of "M2" and "Ours" that after applying the proposed FTB, our network obtains consistent performance gains (*e.g.*, 0.76dB towards PSNR). In addition, the corresponding visual results shown in Fig. 5 also demonstrate that our FTB is beneficial for color correction and detail restoration.

Effectiveness of Multi-scale Features. In order to further study the effectiveness of multi-scale semantic features, we carefully divide the multi-scale features into two groups and specified a series of experimental settings, *i.e.*, the shallower group (denoted as "M3") with the 1st scale and the 2nd scale features, and the deeper group (denoted as "M4") with the 3th scale and the 4th scale features. The comparison results of "M3", "M4" and "Ours" in Table 2 indicate that the incorporation of deeper-scale features and shallower-scale features both obtain

Table 3. Application of the top-ranking image enhancement methods and ours to the saliency detection task evaluated on USOD dataset. The best result is shown in **bold** font.

Method	F-measure↑	S-measure↑	MAE↓
Original F^{3}Net	0.837	0.822	0.085
+ EUIVF	0.850	0.833	0.081
+ Water-Net	0.852	0.833	0.080
+ F-GAN	0.851	0.830	0.082
+ Ours	**0.860**	**0.843**	**0.078**

Input	EUIVF	Water-Net

F-GAN	Ours	Groud truth
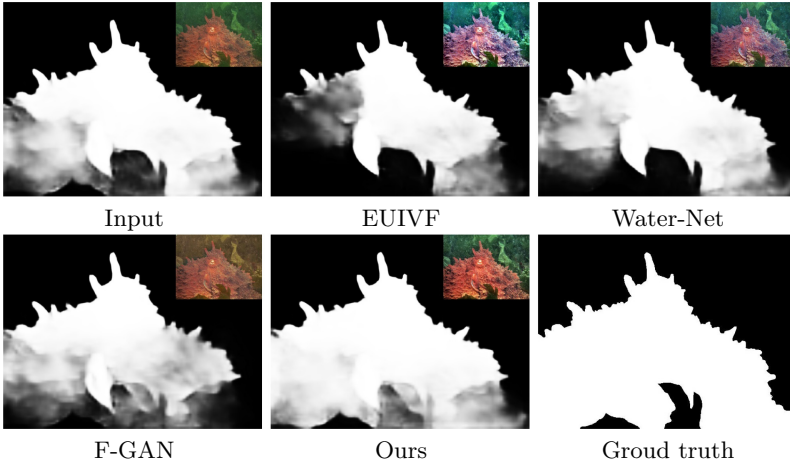
Fig. 6. Visualization of the application of the top-ranking image enhancement methods and ours to the saliency detection task evaluated on the USOD dataset. The corresponding enhanced images are shown in the upper right corner.

much performance gains. And when the multi-scale features are fully incorporated, we achieve the best results. Besides, the visual comparisons shown in Fig. 5 also present the consistent performance.

3.4 Application on Salient Object Detection

To further verify the effectiveness and applicability of our proposed network, we also apply our method to the underwater salient object detection task. Specifically, we first adopt the pre-trained salient object detection network F^{3}Net [26] and evaluate it on an underwater salient object detection dataset (USOD) [13] (denoted as "Original input" in Table 3). We employ several top-ranking image enhancement networks and our proposed network to conduct image enhancement on the inputs and made saliency predictions through the F^{3}Net. The quantitative results of the predicted saliency maps are tabulated in Table 3. It is obvious that

our method shows performance gains against other image enhancement methods. Meanwhile, we can note that after applying our method, the images achieve finer details and more natural colors, and further facilitate F^{3}Net to predict saliency maps with superior consistence and improved robustness in Fig. 6.

4   Conclusion

In this paper, we presented a semantic-driven context aggregation network to cooperatively guide detail restoration and color correcting. Multi-scale semantic features extracted from a pre-trained VGG16 network are fused into the encoder-aggregation-decoder architecture to explore the ability of semantic features. We further proposed a multi-scale feature transformation module which attentively concentrates on the key priors and suppresses the unhelpful ones. Moreover, experimental results conducted on two real datasets demonstrate that our method outperforms the state-of-the-art methods. Additionally, our method can also help salient object detection to achieve better performance.

References

1. Ancuti, C., Ancuti, C.O., Haber, T., Bekaert, P.: Enhancing underwater images and videos by fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 81–88 (2012)
2. Chiang, J.Y., Chen, Y.C.: Underwater image enhancement by wavelength compensation and dehazing. IEEE Trans. Image Process. **21**(4), 1756–1769 (2011)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition, pp. 248–255 (2009)
4. Deng, S., et al.: Detail-recovery image deraining via context aggregation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 14560–14569 (2020)
5. Drews, P.L., Nascimento, E.R., Botelho, S.S., Campos, M.F.M.: Underwater depth estimation and image restoration based on single images. IEEE Comput. Graphics Appl. **36**(2), 24–35 (2016)
6. Ebner, M.: Color constancy, vol. 7. John Wiley & Sons (2007)
7. Fu, X., Fan, Z., Ling, M., Huang, Y., Ding, X.: Two-step approach for single underwater image enhancement. In: 2017 International Symposium on Intelligent Signal Processing and Communication Systems, pp. 789–794 (2017)
8. Fu, Z., Fu, X., Huang, Y., Ding, X.: Twice mixing: a rank learning based quality assessment approach for underwater image enhancement. arXiv preprint arXiv:2102.00670 (2021)
9. Galdran, A., Pardo, D., Picón, A., Alvarez-Gila, A.: Automatic red-channel underwater image restoration. J. Vis. Commun. Image Represent. **26**, 132–145 (2015)
10. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2341–2353 (2010)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

12. Hummel, R.: Image enhancement by histogram transformation. Comput. Graph. Image Process. **6**(2), 184–195 (1977)

13. Islam, M.J., Wang, R., de Langis, K., Sattar, J.: Svam: saliency-guided visual attention modeling by autonomous underwater robots. arXiv preprint arXiv:2011.06252 (2020)

14. Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE Robot. Automation Lett. **5**(2), 3227–3234 (2020)

15. Li, C., Anwar, S., Porikli, F.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recogn. **98**, 107038 (2020)

16. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE Trans. Image Process. **29**, 4376–4389 (2019)

17. Li, C., Guo, J., Guo, C.: Emerging from water: underwater image color correction based on weakly supervised color transfer. IEEE Signal Process. Lett. **25**(3), 323–327 (2018)

18. Liu, R., Fan, X., Hou, M., Jiang, Z., Luo, Z., Zhang, L.: Learning aggregated transmission propagation networks for haze removal and beyond. IEEE Trans. Neural Networks Learn. Syst. **30**(10), 2973–2986 (2018)

19. Liu, R., Fan, X., Zhu, M., Hou, M., Luo, Z.: Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light. IEEE Trans. Circuits Syst. Video Technol. **30**(12), 4861–4875 (2020)

20. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10561–10570 (2021)

21. Ma, L., Liu, R., Zhang, X., Zhong, W., Fan, X.: Video deraining via temporal aggregation-and-guidance. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)

22. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2012)

23. Panetta, K., Gao, C., Agaian, S.: Human-visual-system-inspired underwater image quality measures. IEEE J. Oceanic Eng. **41**(3), 541–551 (2015)

24. Van De Weijer, J., Gevers, T., Gijsenij, A.: Edge-based color constancy. IEEE Trans. Image Process. **16**(9), 2207–2214 (2007)

25. Wang, Y., Zhang, J., Cao, Y., Wang, Z.: A deep cnn method for underwater image enhancement. In: IEEE International Conference on Image Processing, pp. 1382–1386 (2017)

26. Wei, J., Wang, S., Huang, Q.: F^{3}net: Fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12321–12328 (2020)

27. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging **3**(1), 47–57 (2016)

28. Zuiderveld, K.: Contrast limited adaptive histogram equalization. Graphics gems, pp. 474–485 (1994)