A ROBUST OBJECT SEGMENTATION NETWORK FOR UNDERWATER SCENES

Ruizhe Chen[1,2], *Zhenqi Fu*[1,2], *Yue Huang*[1,2,3], *En Cheng*[1,2], *Xinghao Ding*[1,2,3⋆]

[1]Key Laboratory of Underwater Acoustic Communication and Marine Information Technology,
Ministry of Education, Xiamen University
[2]School of Informatics, Xiamen University, China
[3]Institute of Artificial Intelligence, Xiamen University, China
⋆dxh@xmu.edu.cn

ABSTRACT

Underwater object segmentation is one of the key technologies in the fields of marine biology research and autonomous underwater vehicles. The challenges of underwater object segmentation originate from two aspects, 1) the complex underwater environment and 2) the camouflage characteristics of marine animals. In this paper, we propose WaterSNet, an underwater object segmentation network to address these challenges. Specified, we propose a random style adaption (RSA) module as well as a siamese structure to reduce the impact of water degradation diversity. We also extract multi-scale features via the receptive field block (RFB) module, and then fuses multi-level features to better utilize global context information via the attention fusion block (AFB) module. Experimental results on marine animal dataset MAS3K demonstrate that the proposed method outperforms other state-of-the-art methods significantly. The code will be available at: https://github.com/ruizhechen/WaterSNet/

Index Terms— underwater image, image segmentation, camouflaged objects, attention, feature fusion

1. INTRODUCTION

Underwater object segmentation aims at detecting important underwater objects, such as marine animals, from the complex underwater environment. It plays a very important role in the visual system of autonomous underwater vehicles, marine biology and archaeology, as well as marine environmental monitoring. Recently, deep learning provides rich solutions for object segmentation [1], including salient object detection (SOD) [2, 3, 4, 5] and camouflaged object detection (COD). Among them, SOD is based on the attention mechanism of human vision to perceive the most attractive area in the scene, so as to segment the salient objects. As contradicting categories to salient objects, camouflaged objects have high intrinsic similarities to background, making it far more challenging than the salient object detection. For example, Fan *et al.* [6] developed a simple but effective framework for COD, and released a novel COD dataset in various natural scenes. Sun *et al.* [7] proposed a novel context-aware cross-level fusion network exploiting multi-scale feature and rich global context information. These methods have been applied to natural images and achieved remarkable results.

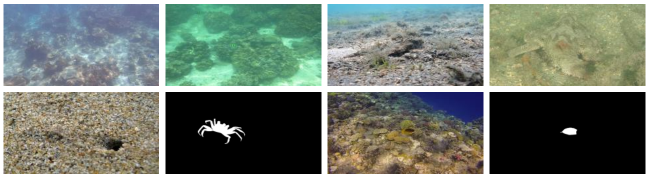

Fig. 1. Examples of underwater images with different degradation styles (first row), and objects with camouflaged characteristics (second row).

Unfortunately, current image segmentation models have great limitations in the task of underwater object segmentation, which is embodied in the following challenges: first, due to the turbid water medium and insufficient illumination of the underwater scenes, underwater images suffer from the quality degradation problem. Low contrast and saturation blur the boundary of the objects to be segmented. As shown in Fig.1, the diversity of light conditions and water degradation styles make some salient objects not salient anymore. Second, many marine animals have camouflaged properties. Some look for habitats similar to their body color, such as crabs. Others can change their body color similar to the surrounding environment, such as butterfly fish. Fig.1 shows an example of the two camouflaged marine animals. The challenges mentioned above make it much more difficult to segment marine objects from underwater images.

To this end, we propose an underwater object segmentation model, called WaterSNet. The model is organized as a

The study is supported partly by the National Natural Science Foundation of China under Grants 61971369, 52105126, 82172033, U19B2031, China Postdoctoral Science Foundation (No. 2021M702726), Science and Technology Key Project of Fujian Province(No. 2019HZ020009) and Fundamental Research Funds for the Central Universities 20720200003.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

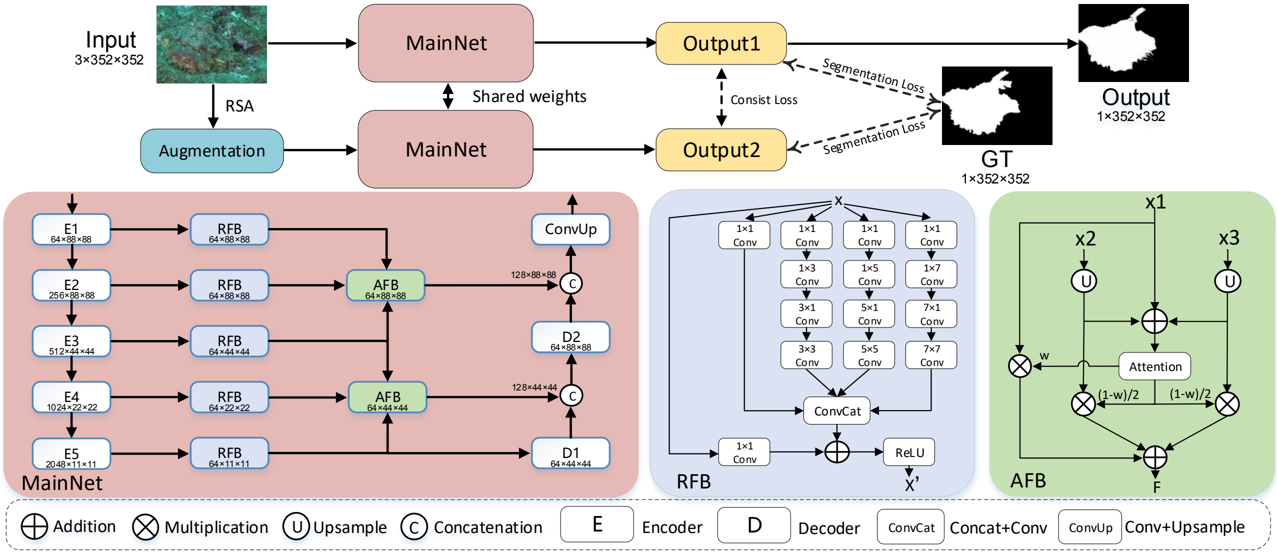

Fig. 2. Overall framework of our method. The whole pipeline is based on the siamese network. The MainNet contains two key blocks, i.e., RFB and AFB. We use different colors to distinguish different modules.

siamese network with a shared main network. We propose a random style adaption (RSA) module to generate augmentation of a image. The main network assembles an asymmetric encoder-decoder structure. We first use five modified receptive field blocks (RFB) to enhance the multi-scale features extracted from the backbone network. The enhanced multi-scale features are then fed into attention fusion blocks (AFB), the AFB module compute both channel and spatial attention to fuse the features, exploiting the rich context information. The fused feature is then used for decoder by concatenation. The main contributions of this work are:

- We propose a robust underwater object segmentation network to segment both salient objects and camouflaged objects, which fuses multi-scale and multi-level features to better utilize global context information.
- In order to reduce the impact of water degradation diversity, we propose a random style adaption (RSA) module and a siamese structure to make the network focus more on the high frequency feature.
- Evaluation of the proposed method that includes comparison with 7 state-of-the-art SOD and COD methods on marine animal segmentation dataset shows that our method outperforms other SOTA models in all metrics.

2. PROPOSED METHOD

The proposed method is illustrated in Fig.2. The overall architecture is a siamese network with a shared main network. The main network is based on an asymmetric encoder-decoder structure. We adopt 5 different layers of Res2Net-50 [8] to extract features from images. As shown in Fig.2, the decoder consists of bilinear upsampling modules and basic convolution modules with batch normalization and ReLU activation, aims to predict the mask from the features of RFB and AFB. In the training phase, a batch of input images are augmented by RSA, the original image and the augmented image are inputted to the network getting two outputs. In the testing phase, the RSA module is removed, and the input image directly passes the MainNet to get the prediction mask.

2.1. Receptive Field Block

As suggested in Liu *et al.* [9], setting the receptive fields of deep learning models at the same size with a regular sampling grid on a feature map may reduce feature discriminability and robustness, therefore they proposed RFB, a multi-branch convolution block with different kernels and trailing dilated pooling or convolution layers. For underwater object segmentation task, it is necessary to increase the feature robustness. Therefore, we adopt a modified RFB block which is also adopted in [7], as shown in Fig.2. In each branch, the first convolution layer has kernel size 1×1 and output channel size 64. The followed 1D asymmetric convolutions can not only enrich the feature space, but also save the number of parameters. The concatenate of the right four branches is then convoluted and fused with the original feature via 1×1 convolution. The RFB module achieves multi-scale feature extraction and improves feature representation ability.

2.2. Attention Fusion Block

Salient objects are salient in both global and local contexts, but for camouflaged objects, their characteristic makes them

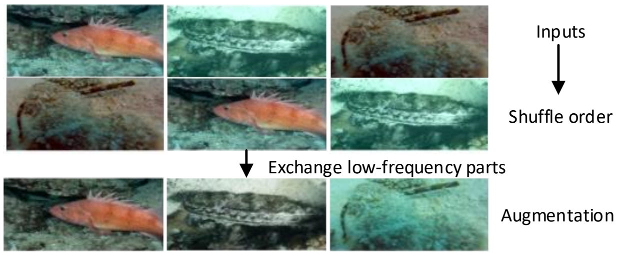

Fig. 3. Example of random style adaption method (RSA). The order of imput images in a batch is randomly shuffled, and the low-frequency part of the images at the same position is exchanged to get an augmentation.

hidden in the local context. Therefore, concerning global context information will be a wise choice for the COD task. In this case, we propose the attention fusion block (AFB) module. As illustrated in Fig.2, the AFB modules accept the output from 1st to 3rd or 3rd to 5th RFB module. High-level features are up-sampled to the same scale as the low-level feature, the three branches are added up and fed into the attention module. We use a simple parameter-free attention module SimAM [10] to calculate attention in both channel and spatial dimensions. The feature fusion process is as follows:

$$X_2^{'} = U(X_2), \quad X_3^{'} = U(X_3). \tag{1}$$

$$w = A(X_1 \oplus X_2^{'} \oplus X_3^{'}). \tag{2}$$

$$F = (w \otimes X_1) \oplus (\frac{1-w}{2} \otimes X_2^{'}) \oplus (\frac{1-w}{2} \otimes X_3^{'}). \tag{3}$$

where A is the attention module, and U is upsampling module. We use the calculated attention w as the weight to fuse the three features via element-wise multiplication $\otimes$ and addition $\oplus$. Fusion of multiple level features can help to form a comprehensive feature expression, therefore obtain global context information.

2.3. Random Style Adaption

There are diverse water degradation styles in underwater images, which may damage the segmentation performance. The network is supposed to focus more on high-frequency edge features in similar texture areas rather than low-frequency degradation style. In order to reduce the impact of water degradation diversity, we propose a random style adaption method (RSA). Different from general data augmentation, we also conduct a siamese structure, adding consistency regularization to the two outputs of a same image at a later stage. As shown in Fig.3, we first take a batch of images, then shuffle the order of images, and exchange the low-frequency part of the images [11] in the same position before and after to get an augmentation, this process can be summarized as follows:

$$A_1 = x, \quad A_1^{'} = \varphi(x). \tag{4}$$

$$A_2 = \mathcal{F}^{-1}(M_\beta \otimes \mathcal{F}(A_1^{'}) \oplus (1 - M_\beta) \otimes \mathcal{F}(A_1)). \tag{5}$$

$$M_\beta = \mathbb{1}_{(h,w)\in[-\beta H:\beta H, -\beta W:\beta W]}. \tag{6}$$

where x is a training batch, $\varphi(x)$ denotes random shuffling of x, $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent Fourier transform and inverse Fourier transform. M_β is a mask whose value is zero except for the center region where $\beta \in (0, 1)$.

2.4. Loss Function

We take BCE-SSIM-IoU loss[3] as our basic loss function, which is defined as follows:

$$l_b = l_{bce} + l_{ssim} + l_{iou}. \tag{7}$$

where l_{bce} indicates BCE loss, l_{ssim} indicates SSIM loss and l_{iou} indicates IoU loss. The above loss supervises the training process at pixel level, patch level and feature map level, the network thus pay more attention to the boundary quality, so as to segment the salient and camouflaged object better. For our network, we calculate loss between the two outputs of network and ground truth, as well as the consistency loss of the two outputs. To be more specific, the consistency loss is also a basic BCE-SSIM-IoU loss, it can make the augmented samples regularize with their original version and improve the ability of the model to deal with water degradation problem. The final loss function is:

$$L = l_b(out_1, GT) + l_b(out_2, GT) + 2 \times l_b(out_1, out_2). \tag{8}$$

3. EXPERIMENTS AND RESULTS

3.1. Dataset

We conduct experiments on a public marine animal segmentation dataset MAS3K [12]. It is the first large-scale dataset for marine animal segmentation, with complex underwater environments and 37 sub-categories of marine animals. It consists of 3,103 images, including 1,588 camouflaged images, 1,322 common images, and 193 background images. The training set has 1,769 images, the testing set has 1,141 images, background images are not included in both sets.

3.2. Implementation details

We train our network using PyTorch. The inputs are resized to 352×352. Adam optimizer [13] is adopted with learning rate 10^{-4}. The network is trained on a NVIDIA TITAN RTX GPU over 250 epochs with a batch size of 16.

3.3. Evaluation Metrics

To comprehensively evaluate our proposed model, we use five commonly used metrics to evaluate the results: mIoU for mean intersection over union, S-measure (S_α) [14] for structure similarities, Weighted F-measure (F_β^w) [15] for overall

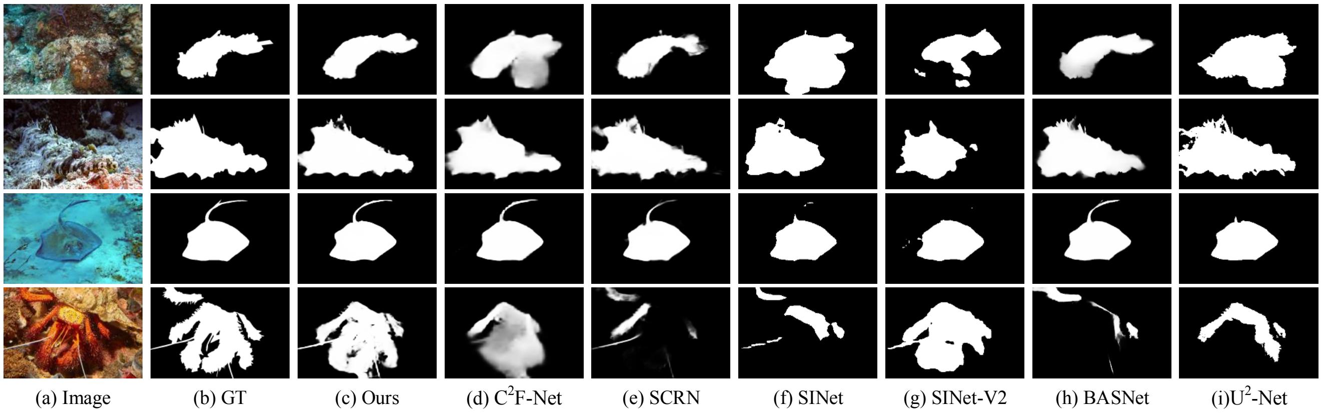
(a) Image (b) GT (c) Ours (d) C^{2}F-Net (e) SCRN (f) SINet (g) SINet-V2 (h) BASNet (i)U^2-Net

Fig. 4. Visualization results of our model and six SOTA SOD/COD methods. The upper two images contain camouflaged objects, and the lower two images contain salient objects. It can be seen from the figure that our method is more accurate than other methods in the segmentation of object boundary, having less uncertainty on the boundary, and is better than other methods in integrity.

Table 1. Performance comparison with SOTA SOD/COD models on MAS3K testing dataset using five metrics(*i.e.*, $mIoU$ ↑, S_α ↑, F_β^w ↑, mE_ϕ ↑ and MAE ↓). "↑"/"↓" indicates that larger or smaller is better. Note that we directly use the results of ECD-Net in [12](marked by "*") since no public code is available for ECD-Net. The best results are highlighted in **Bold** fonts.

Method	MAS3K-Test				
	$mIoU$	S_α	F_β^w	mE_ϕ	MAE
BASNet	0.678	0.820	0.748	0.869	0.044
SCRN	0.690	0.832	0.762	0.884	0.038
SINet	0.657	0.815	0.745	0.885	0.039
U^2-Net	0.651	0.809	0.722	0.851	0.047
ECD-Net*	0.711	0.850	0.766	0.901	0.036
SINet-V2	0.561	0.757	0.648	0.826	0.061
C^{2}F-Net	0.717	0.844	0.781	0.903	0.036
Ours	**0.739**	**0.856**	**0.804**	**0.913**	**0.032**

performance measurement, E-measure (mE_ϕ) [16] for overall and local accuracy, MAE [17] for mean absolute error.

3.4. Results

We compare our model with 7 state-of-the-art SOD/COD methods in MAS3K testing set. Among these models, BASNet [3], SCRN [4] and U^2-Net [5] are SOD models, SINet [6], SINet-V2 [18] and C^{2}F-Net [7] are COD models, ECD-Net [12] is a generic model. The results is shown in Table.1. Fig.4 shows the visualization results. There is no visualization result of ECD-Net since no public code is available for ECD-Net. Our model shows best performance in all mentioned models. Specifically, compared with the best COD model C^{2}F-Net, $mIoU$ increased by 3.07%, S_α increased by 1.42%, F_β^w increased by 2.94%, mE_ϕ increased by 1.11%, and MAE decreased by 1.11%.

Table 2. Ablation study.

Method	MAS3K-Test				
	$mIoU$	S_α	F_β^w	mE_ϕ	MAE
Base	0.691	0.828	0.756	0.886	0.047
Base+RSA	0.708	0.841	0.779	0.910	0.036
Base+RFB	0.709	0.840	0.781	0.889	0.037
Base+AFB	0.720	0.845	0.785	0.906	0.035
Full	**0.739**	**0.856**	**0.804**	**0.913**	**0.032**

3.5. Ablation Study

Hereafter we study the effectiveness of each module, and explore the role of siamese structure. The results are shown in Table.2, note that "Base" denotes that there are only encoders and decoders. The results indicate that the contribution of each module focus on different aspects, and all of the modules are necessary for performance improvement.

4. CONCLUSION

In this paper, we propose a robust underwater object segmentation network WaterSNet, which utilizes multi-scale and multi-level feature information to better adapt to complex underwater environment. We propose an attention fusion module AFB, which fuses the features obtained by the RFB in spatial and channel dimension. We propose a random style adaption method for data augmentation, and siamese structure for regularization. Experimental results show that the proposed method outperforms other state-of-the-art methods.

5. REFERENCES

[1] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[3] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[4] Zhe Wu, Li Su, and Qingming Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7264–7273.

[5] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, pp. 107404, 2020.

[6] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao, "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2777–2787.

[7] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu, "Context-aware cross-level fusion network for camouflaged object detection," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Zhi-Hua Zhou, Ed. 2021, pp. 1025–1031, ijcai.org.

[8] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, 2021.

[9] Songtao Liu, Di Huang, et al., "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 385–400.

[10] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11863–11874.

[11] Yanchao Yang and Stefano Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[12] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen, "Marine animal segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[13] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.

[15] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal, "How to evaluate foreground maps?," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 248–255.

[16] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang, Ed. 2018, pp. 698–704, ijcai.org.

[17] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.

[18] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao, "Concealed object detection," *CoRR*, vol. abs/2102.10274, 2021.