MASNet: A Robust Deep Marine Animal Segmentation Network

Zhenqi Fu, Ruizhe Chen, Yue Huang, *Member, IEEE*, En Cheng, Xinghao Ding,
and Kai-Kuang Ma, *Life Fellow, IEEE*

***Abstract*—Marine animal studies are of great importance to human beings and instrumental to many research areas. How to identify such animals through image processing is a challenging task that leads to marine animal segmentation (MAS). Although deep neural networks have been widely applied for object segmentation, few of them consider the complex imaging condition in the water and the camouflage property of marine animals. To this end, a robust deep marine animal segmentation network is proposed in this article. Specifically, we design a new data augmentation strategy to randomly change the degradation and camouflage attributes of the original objects. With the augmentations, a fusion-based deep neural network constructed in a Siamese manner is trained to learn the shared semantic representations. Moreover, we construct a new large-scale real-world MAS data set for conducting extensive experiments. It consists of over 3000 images with various underwater scenes and objects. Each image is annotated with an object-level mask and assigned to a category. Extensive experimental results show that our method significantly outperforms 12 state-of-the-art methods both qualitatively and quantitatively.**

***Index Terms*—Fusion, image degradation, marine animal segmentation (MAS), object camouflage, Siamese network.**

I. Introduction

THE exploration of underwater environments has been an active engagement across a plethora of scientific fields, such as ocean ecology, marine geological sciences, and natural resources discovery. Recently, visually guided underwater robots and intelligent underwater monitoring systems become instrumental tools or equipment to assist these activities effectively, and many image processing algorithms have been developed to serve various purposes. However, existing solutions for conducting marine scene parsing and object segmentation are largely underexplored. In this article, it is our great interest in developing image segmentation algorithms to perform marine animal segmentation (MAS), particularly for a highly ill-conditioned underwater environment.

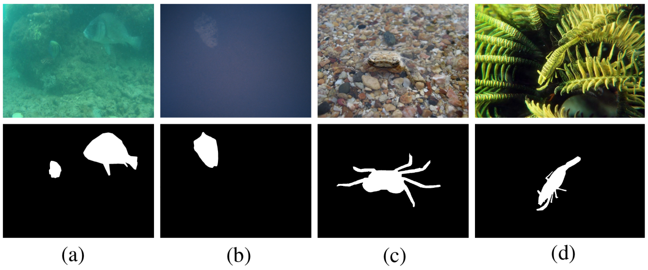

Fig. 1. Four typical marine images for demonstrations. (a) and (b) Degraded into low quality. (c) and (d) Camouflaged objects.

The visual contents of underwater images are usually degraded due to the wavelength-dependent absorption and scattering effects [1], [2], [3]. As a result, images acquired in underwater environments tend to have poor visibility, low contrast, and unwanted color casts. Such severely degraded underwater images are unfavorable and incur obstacles to scene understanding and object segmentation. Challenges can be further compounded if the object of interest is camouflaged and seamlessly blended with the environment. This is often encountered in many marine animals that have been evolving themselves for survival in the underwater ecosystem [4], [5]. The similarity between marine animals and their background further increases the difficulty of accurately segmenting them from underwater scenes. For demonstration, two images with degraded image quality are shown in Fig. 1(a) and (b), and two camouflaged animals are shown in Fig. 1(c) and (d).

Related to marine animal image processing, some methods have been proposed to enhance the image quality [6], [7], [8], [9], [10], [11], [12], [13], [14]. These methods might be indirectly beneficial to the MAS task. On the other hand, camouflaged object segmentation (COS) methods [15], [16], [17] aim to identify objects that are assimilated into their surroundings. However, the current COS approaches might be sensitive to image degradation since they are specifically designed for terrestrial objects, and they do not consider the optical distortions incurred by the water.

In this article, we propose a deep marine animal segmentation network (MASNet) for robust MAS. The core idea of our method is to combine the Siamese architecture with data augmentation techniques to reduce the impact of image degradation and object camouflage. Specifically, the proposed data augmentation strategy is performed at both image and object levels to randomly change the degradation and camouflage attributes of the original image. With the augmented images, we train

Manuscript received 12 August 2022; revised 22 December 2022; accepted 22 February 2023. The work was supported in part by the National Natural Science Foundation of China under Grant 82172033, Grant U19B2031, Grant 61971369, Grant 52105126, Grant 82272071, and Grant 62271430. *(Corresponding author: Xinghao Ding.)*

Associate Editor: R. Cong.

Zhenqi Fu, Ruizhe Chen, Yue Huang, En Cheng, and Xinghao Ding are with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: fuzhenqi@stu.xmu.edu.cn; rzchen1701@stu.xmu.edu.cn; yhuang2010@xmu.edu.cn; chengen@xmu.edu.cn; dxh@xmu.edu.cn).

Kai-Kuang Ma is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ekkma@ntu.edu.sg).

Digital Object Identifier 10.1109/JOE.2023.3252760

MASNet in a Siamese pipeline to force the network to learn the shared semantic features of the two inputs. We elaborately design an attention-guided cross-level fusion module to integrate multilevel features. The fusion-based architecture is helpful to aggregate low-level and high-level features for better local and global perception. We also introduce a modified receptive field block (RFB) [18], [19] to enhance the representations. Two types of loss functions are employed to guide the network training, including a task loss for segmentation and an alignment loss for better representation learning.

To the best of our knowledge, there is only one large-scale object-level labeled data set for MAS [4]. To this end, we construct a new large-scale MAS data set, which contains over 3000 real-world underwater images. We annotate an object-level mask and a related category for each image. This new data set can be used for training and evaluating MAS models.

We summarize the main contributions of this article as follows.

1) To simultaneously tackle the image degradation and the object camouflage problems in MAS, we develop a new data augmentation strategy and train the network in a Siamese fashion to encourage the model to learn shared semantic features.
2) We construct a new MAS data set that contains more than 3000 images with different underwater scenes and objects. Each image is annotated with an object-level mask and a category. This data set can be used for training and testing MAS models.
3) Extensive experiments show that the proposed method achieves state-of-the-art performance in the terms of several objective evaluation metrics on two MAS benchmark data sets.

The rest of this article is organized as follows. In Section II, we introduce related works. In Section III, we detail the proposed approach. In Section IV, we present the experimental results. Finally, Section V concludes this article. The project page of this work is available at https://github.com/zhenqifu/MASNet.

II. Related Work

A. Underwater Image Degradation and Quality Enhancement

Underwater images suffer from quality degradation (e.g., contrast distortion, low visibility, and color shift) due to light absorption and scattering effects. In clean water, red light disappears first at about 5 m water depth. As the depth increases, yellow and green lights disappear subsequently. The blue light with the shortest wavelength travels the furthest distance in the water [20]. As a result, underwater images are usually dominated by green and blue colors. Apart from the wavelength-dependent light absorption, suspended particles in the water affect scene contrast and produce haze-like effects by absorbing lights. Moreover, the diversity of underwater scenes (e.g., turbid water and deep oceanic water) causes different distortion types and levels [21]. Degraded underwater images are detrimental to visual-based applications, such as object identification, detection, and segmentation. To tackle the above issues, underwater image enhancement (UIE) techniques have been proposed and become the essential preprocessing step [2]. Generally, UIE methods can be divided into three categories, i.e., model-free, model-based, and learning-based methods [3]. Model-free approaches enhance underwater images without using the physical imaging model. For example, the authors proposed a fusion-based UIE method in [6] and an improved version in [22]. Model-based methods enhance underwater images under the guidance of the degradation model with manually designed prior constraints. The image dehazing model and its variants are the most used degradation models in existing UIE methods [7], [8], [23]. Researchers also developed various prior features to improve the quality of enhanced/restored underwater images [24], [25], [26], [27]. Recently, deep learning has achieved excellent breakthroughs in various domains [28], [29], [30]. A lot of deep learning-based UIE methods have been proposed [1], [9], [10], [31], [32], [33].

B. Camouflage Object Segmentation

Camouflage object segmentation (COS) is an emerging field and attracts increasing attention in recent years. COS aims to segment objects with similar patterns to their surroundings. Traditional works on COS use handcraft visual features, such as color, texture, and gradient [5]. However, the techniques relying on a single feature cannot provide promising performance. Besides, owing to the limitation of hand-crafted features, these methods are only suitable for relatively simple scenarios and may fail in real-world applications. Currently, deep learning has been employed for COS by extracting deep features automatically from extensive training images, which are more generic and effective than hand-crafted features. For example, Fan et al. [5] collected a large-scale COS data set that contains 10 000 images covering camouflaged objects in various natural scenes with over 78 object categories. In addition, the authors proposed a deep-learning-based COS method based on the receptive field and partial decoder component. An improved version of [5] is presented in [34]. The network contains a texture-enhanced module, a neighbor connection decoder, and group-reversal attention to identify objects that are visually embedded in their background. Le et al. [35] constructed a COS data set called CAMO and proposed to leverage both classification and segmentation tasks to promote COS performance. Li et al. [36] leveraged the contradictory information to enhance the detection ability of both salient and camouflaged objects. Yang et al. [37] developed a COS network based on Bayesian learning and Transformer-based reasoning. Lv et al. [15] presented a ranking-based COS network, which simultaneously localized, segmented, and ranked camouflaged objects. Liu et al. [38] designed a confidence-aware network for accurate camouflaged objection. Pang et al. [39] proposed a mixed-scale triplet network to imitate the behavior of human beings when observing camouflaged objects. Other relevant works of COS can be found in [16] and [17].

C. Marine Animal Segmentation

Deep-learning-based object segmentation has made remarkable progress over the past few decades thanks to the advent of

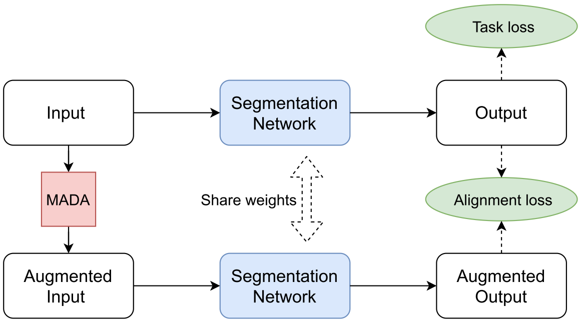

Fig. 2. Overall framework of the proposed method. *In the training phase*, our approach consists of a MADA module, a Siamese segmentation network, a task loss, and an alignment loss. *In the testing phase*, we adopt a single branch to predict the segmentation result from the raw underwater image.

deep models [29], [40] and large-scale annotated data sets [41], [42]. Various methods have been proposed for different visual tasks, such as fully convolutional network [43] for semantic segmentation and U-Net [28] for biomedical image segmentation. Despite the advancements, MAS is considerably less studied. Several important contributions have been made to address the problems of coral reef classification and segmentation [44], [45], fish detection and segmentation [46], [47], underwater semantic segmentation, and saliency prediction [48]. Recently, Li et al. [4] have constructed the first large-scale data set for MAS named MAS3K. This data set contains over 3000 images with different degradation types, such as low illumination, turbid water quality, and photographic distortion. The data set includes both camouflaged and common objects. Each image from the MAS3K data set has rich annotations, including an object-level mask, a category, object attributes, and the camouflage strategy. Based on the data set, the authors proposed a deep MAS network with an interactive feature enhancement module and a cascade decoder. The feature enhancement module aims to refine the features extracted from the backbone network. The cascade decoder receives features and predicts the segmentation result.

III. Method

Fig. 2 outlines the framework of our MASNet, which comprises a marine animal data augmentation (MADA) module, a Siamese segmentation network, a task loss, and an alignment loss. In the training phase, we first generate an augmented instance of each original image. Then, the two images are fed into the Siamese segmentation network yielding two predictions. A task loss is applied to supervise the segmentation. Besides, we align the two outputs to encourage the network to learn the shared semantic information. In the testing phase, we use a single branch of the Siamese segmentation network to predict the result.

A. Data Augmentation for MAS

In this article, we consider the following two obstacles in MAS. The first is the diverse degradation of underwater images.

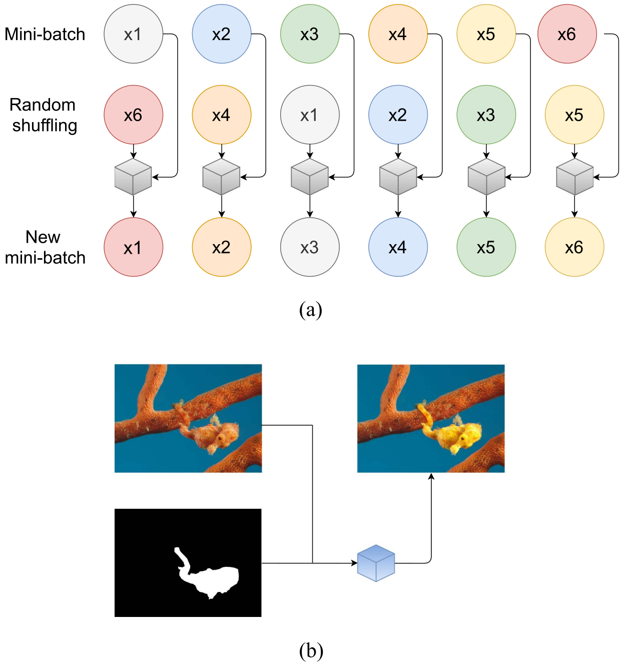

Fig. 3. Overview of the proposed MADA. (a) Image-level random augmentation within a mini-batch. As an example, we set the batch size to 6. Each color denotes an image style and x_i denotes the image content. (b) Object-level random augmentation with the guidance of object masks, where only objects in the image are augmented. Animals may be no longer camouflaged after the object-level augmentation.

Owing to the wavelength-dependent light scattering and absorption, underwater images usually have low quality with different appearances. For example, images captured in coastal water, deep oceanic water, and muddy water show different distortion types. Since we cannot obtain all degradation types and levels, it is challenging to segment marine animals from complex underwater environments with high accuracy and generalization performance. The second is the camouflage properties of marine animals. Body colors, patterns, and other morphological adaptations of marine animals will significantly decrease their probability of being detected and segmented by both humans and machines. In this article, we use data augmentation techniques to reduce the impact of image quality degradation and object camouflage.

To address the degradation problem, we propose an image-level augmentation strategy, which randomly perturbs the style information of each training instance in a mini-batch. Fig. 3(a) shows an example of such augmentation. Let x_1–x_6 be six instances in a mini-batch. Note that the batch size can be arbitrary; we set 6 here as an example. Each color denotes an image style. We first generate a reference batch via random shuffling of the original batch. Then, we replace the low-frequency part of the amplitude of the original image with that of the shuffling image [49]. This results in a new mini-batch, in which each instance has the same semantic content but a different style from the original one. Specifically, for each instance in a mini-batch, we first perform the Fourier transform [50] to obtain the amplitude and phase components. We use $\mathcal{F}^{\mathcal{A}}$ and $\mathcal{F}^{\mathcal{P}}$ to express the amplitude and phase components of an instance, respectively.

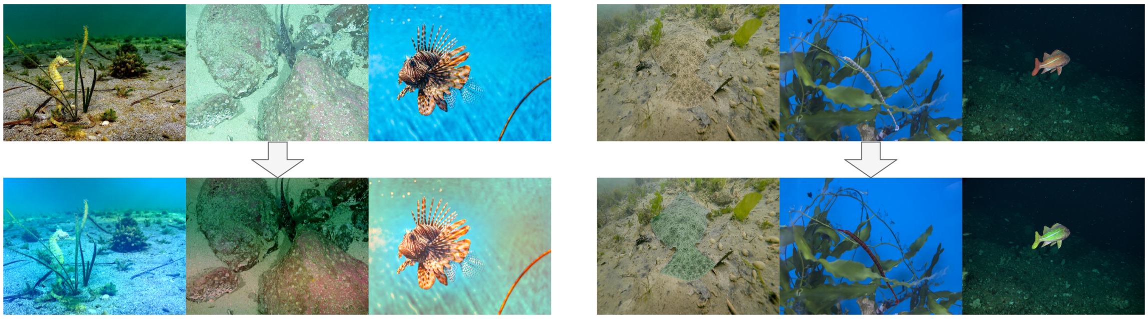

Fig. 4. Examples of MADA results. (a) Image-level augmentation. (b) Object-level augmentation. The original images are presented in the first line and the MADA results are shown in the second line.

Image-level augmentation can be formalized as

$$x^{s\to t} = \mathcal{F}^{-1}(m \circ \mathcal{F}^{\mathcal{A}}(x^t) + (1-m) \circ \mathcal{F}^{\mathcal{A}}(x^s), \mathcal{F}^{\mathcal{P}}(x^s)) \tag{1}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform that maps spectral signals (phase and amplitude) back to the image space. $m \circ \mathcal{F}^{\mathcal{A}}(x^t) + (1-m) \circ \mathcal{F}^{\mathcal{A}}(x^s)$ refers to replacing the low-frequency part of the amplitude of the source image x^s with that of the target image x^t. m is the mask to calculate the low-frequency part, whose value is zero except for the central region with the size of 2×2. Finally, the augmented image $x^{s\to t}$ has same content with s but similar style with t.

Apart from the image-level augmentation, we further propose object-level augmentation to tackle the camouflage issue in MAS. As illustrated in Fig. 3(b), this operation randomly changes the appearance of camouflaged objects with the guidance of their masks. Different from the image-level augmentation that replaces style information in a mini-batch, we directly adjust objects' color, saturation, and contrast, to discriminate them from the background or conceal them in the surroundings. Such objective-level augmentations are implemented following the recommendation of contrastive learning method [51]. Object-level augmentation can be formalized as

$$x^t = M \circ T(x) + (1-M) \circ x \tag{2}$$

where M denotes the object mask. T refers to the transform function of color, saturation, and contrast. Obviously, the augmented image x^t has the same background as the original image but different object styles.

Fig. 4 shows examples of augmented images, where we can notice that image-level augmentation focuses on the image degradation problem, and object-level augmentation concentrates on the camouflage issue. In our method, we simultaneously perform the image-level and object-level augmentation in each iteration to achieve better representation learning.

Since it is impossible to create all degradation types and levels, instead of directly training a network based on the augmented images, we combine the proposed MADA with a Siamese network. As presented in Fig. 2, we additionally employ an alignment loss to encourage the network to learn the shared semantic information. Note that, in the testing phase, we only use one branch of the Siamese network to predict the result. That means our method will not increase the computational and memory costs in the testing phase.

B. *Cross-Level Fusion Network for MAS*

For each branch of the Siamese network, we design a cross-level fusion pipeline to aggregate both low-level and high-level features. Fig. 5 details the overall architecture of the proposed network, which consists of a Res2Net [18] backbone (E1–E5), a feature enhancement module (RFB), and attention-induced cross-level fusion decoder.

1) Backbone and Feature Enhancement Module: We adopt Res2Net-50 [18] as the backbone to extract features at five different scales, denoted as E_i $(i = 1, 2, ..., 5)$. Rather than directly using those features for the segmentation, we leverage an RFB [18] to enhance the extracted features. The RFB is employed to expand the receptive field for capturing richer features in specific layers. We use the same settings as recommended in [18] for implementing the RFB. Fig. 6 shows the details of RFB, which includes five parallel branches b_k $(k = 0, 1, ..., 4)$. In each branch, the first convolutional layer utilizes a 1×1 convolution to reduce the channel size to 64. This is followed by two layers, i.e., a $(2k-1) \times (2k-1)$ convolutional layer and a 3×3 convolutional layer with a specific dilation rate $(2k-1)$ when $k \geq 2$. In our implementation, we factorize the first convolution layer of size $(2k-1) \times (2k-1)$ as a sequence of two steps with $(2k-1) \times 1$ and $1 \times (2k-1)$ kernels, speeding up the inference efficiency without decreasing the representation capabilities. The four branches b_k $(k = 1, ..., 4)$ are then concatenated and their channel size is reduced to 64 using a 1×1 convolutional operation. Finally, the first branch is added, and the whole module is fed to a rectified linear unit (ReLU) activation function to obtain the enhanced feature.

2) Cross-Level Feature Fusion: After extracting feature pyramids from weakly semantic and high-resolution to strongly semantic and low-resolution, we develop an attention-guided cross-level fusion decoder to integrate the diversified features. The reasons we apply such feature pyramids are twofold:

1) There are natural differences in the shape and size among different types of marine animals. The size of similar objects may also vary greatly, due to the observation distance and their relative location to the surroundings.

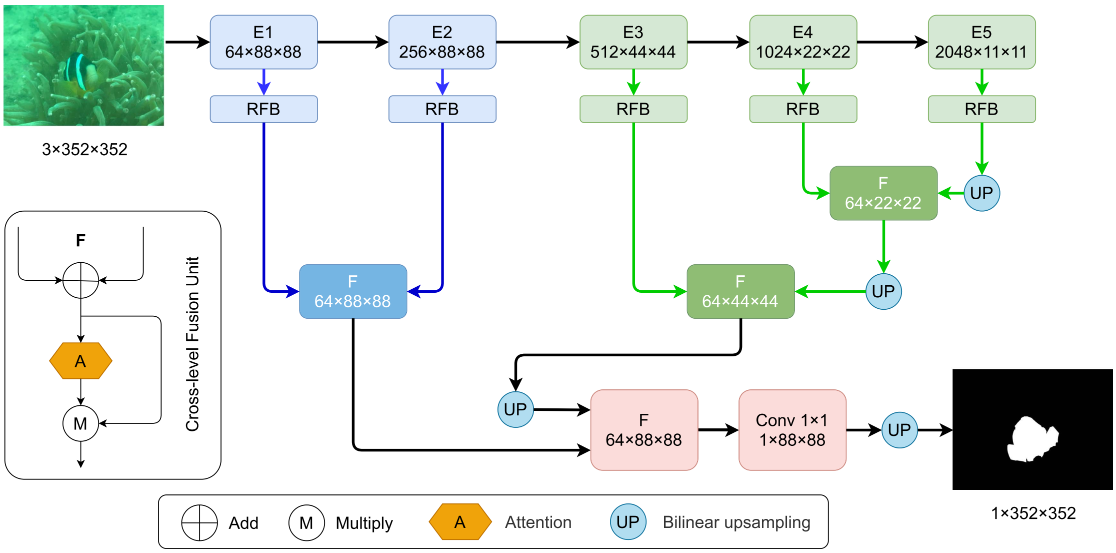

Fig. 5. Architecture of the segmentation network. It consists of a Res2Net backbone (E1–E5), a feature enhancement module, i.e., RFB, and an attention-induced cross-level fusion decoder. The detailed structure of the fusion unit is shown in the left lower part. Specifically, we integrate both low-level (blue line) and high-level (green line) features to achieve better local–global perception.

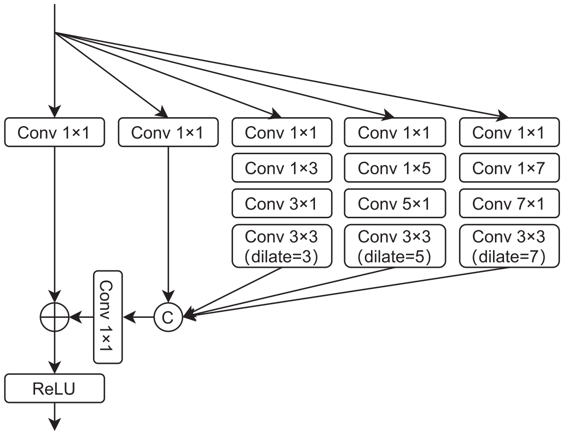

Fig. 6. Architecture of RFB.

2) Features at different levels have different contributions to the MAS task. Fusing features at multiple levels can complement each other to obtain a comprehensive feature expression.

The details of the cross-level fusion decoder can be found in Fig. 5. Concretely, we first add two branch features elementwisely. If the feature size is different, a bilinear upsampling operation will be first conducted. Then, with the guidance of attention, we obtain the fusion weight for each spatial position. Finally, elementwise multiplication is performed to compute integrated features. The cross-level fusion process can be formulated as

$$f_{ab} = (f_a \oplus f_b) \otimes A\left(f_a \oplus f_b\right) \tag{3}$$

where $\oplus$ denotes the elementwise add operation and $\otimes$ represents elementwise multiplication; A indicates the attention function. In this article, we employ SimAM [52] to calculate the fusion weight. It optimizes an energy function to find the importance of each neuron according to some well-known neuroscience theories. In addition, SimAM is parameter-free and computationally efficient. This is important for underwater-related applications that prefer low-complexity and low-power algorithms. Note that designing novel attention methods is not our target, and a more advanced attention operator may further promote our segmentation performance. Different from most of the existing methods that only use high-level features [5], [16], we consider that both low-level and high-level features will benefit the MAS task. High-level features are more relevant to semantic information, while low-level features are location dependent and can capture rich edge information. Therefore, we utilize all five-scale features to calculate the final segmentation prediction.

C. Loss Functions

Two kinds of loss functions are designed in MASNet. The first one is the task loss for learning binary segmentation maps. Specifically, we use the binary cross-entropy (BCE) loss to independently calculate the loss of each pixel to form a pixel restriction on the network. To make up for its deficiency of ignoring the global structure, we also employ intersection over union (IoU) loss [53] to form a global restriction on the network. In summary, the segmentation loss function adopted in our network is defined as

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{BCE}}(P_1, G) + \mathcal{L}_{\text{IoU}}(P_1, G) \tag{4}$$

where P_1 refers to the network prediction and G denotes the ground-truth mask.

The second one is the alignment loss that encourages the network to learn semantic-relevant features, which means the

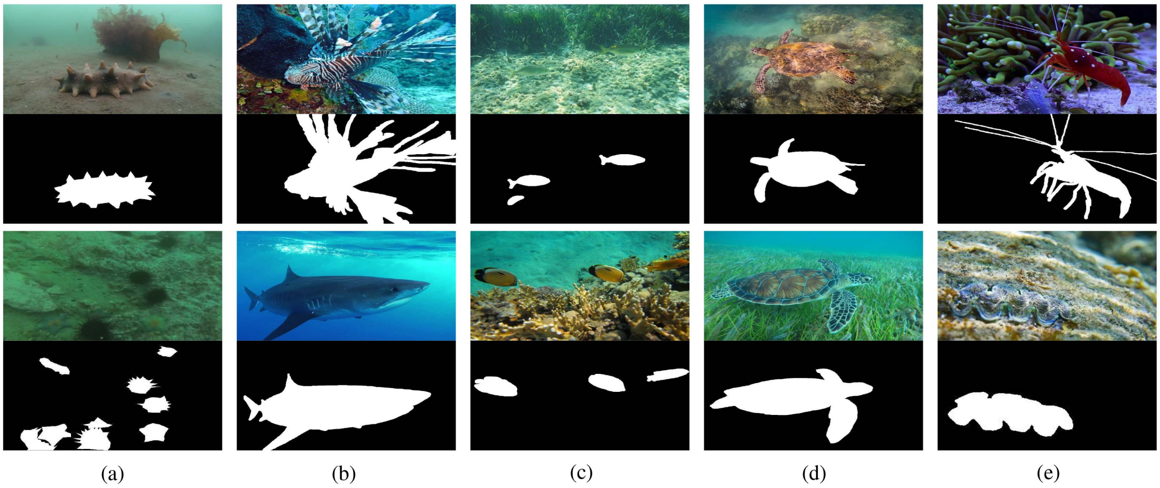

Fig. 7. Examples of images in the RMAS data set. (a) Sea products. (b) Big fish. (c) Small fish. (d) Turtle. (e) Others.

model is expected to output the same predictions of the original and augmented inputs since they share the same objects. Here, we compute the mean square error between two outputs

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{MSE}}(P_1, P_2) \tag{5}$$

where P_1 and P_2 are estimations from original and augmented input, respectively. Finally, the total loss of our method is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Seg}} + \mathcal{L}_{\text{Align}}. \tag{6}$$

D. Real-World Marine Animal Segmentation (RMAS) Data Set

Since there are limited data sets for training and testing MAS models, we construct a new data set named RMAS. The detailed information and properties of the proposed data set are as follows.

1) Image Collection and Annotation: RMAS contains 3014 real-world underwater images with different scenes and degradation types, which are mainly collected from existing underwater image processing data sets including SUIM [48], UFO [54], DeepFish [55], and URPC.[1] In this article, we roughly divide those images into five classes: sea products (including starfish, sea cucumber, sea urchin, and scallop), big fish, small fish, turtle, and others. All images in RMAS are annotated at pixel level and labeled with a class. Fig. 7 shows examples of the collected images and their corresponding annotations. Note that big and small fish are classified according to their relative size in the whole image. The category of others includes shrimp, crab, jellyfish, hippocampus, and so on. Those animals are simply classified as "others" due to the limited quantity.

2) Data Set Features and Statistics: Fig. 8 reports the distribution of each category. We can observe that sea products, big fish, and small fish are the three main categories that account for 92% of the data set. To train and test MAS models, we divide the whole data set into two parts, i.e., the training set and the testing set. Table I lists the number of images in each subset. The original images collected in RMAS have different degradation types with salient or camouflaged objects. The original images are of various spatial resolutions from 221 × 206 to 3840 × 2160. To further explore the ability of each method in handling image degradation and object camouflage, we additionally classify the test images into three categories, i.e., high-quality, low-quality, and camouflage. Table II shows the number of images in each category. Concretely, high-quality and low-quality images are selected according to four measurements, i.e., color deviation, contrast distortion, illumination condition, and the visibility of objects. Note that, if an image has camouflaged objects, it will be divided into "camouflage" regardless of the image quality.

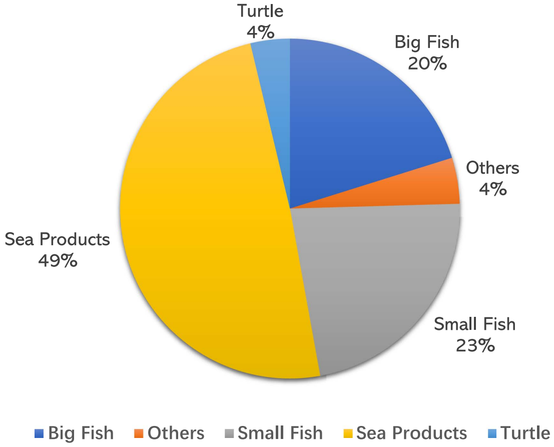

Fig. 8. Distribution of each category.

TABLE I
NUMBER OF IMAGES IN TRAINING AND TESTING DATA SETS

RMAS	Sea products	Big fish	Small fish	Turtle	Others
Training set	1284	501	526	96	107
Testing set	196	107	155	17	25
All	1480	608	681	113	132

[1] [Online]. Available: http://urpc.org.cn/

TABLE II
NUMBER OF IMAGES IN EACH CATEGORY

Category	High-quality	Low-quality	Camouflage
Number	138	328	34

IV. EXPERIMENTS

In this section, we first describe the implementation details, the data set for training and testing, evaluation metrics, and comparison methods. Then, we compare the proposed method with state-of-the-art MAS methods subjectively and objectively. Furthermore, ablation studies are conducted to analyze the effectiveness of each module in MASNet.

A. Implementation Details

We adopt Res2Net50 [18] pretrained on ImageNet as the network backbone. Other layers of MASNet are randomly initialized. In the training phase, the input images are resized to 352×352 and fed into the Siamese network to predict the segmentation mask. We utilize the Adam optimization algorithm to optimize the overall parameters by setting the initial learning rate as $1e^{-4}$. The batch size is set to 16. All the experiments are conducted on a server with Python 3.7, PyTorch 1.7.1, and Nvidia 3080 Ti GPUs. In the testing phase, we use a single branch of the Siamese network to calculate the prediction. Specifically, the image is first resized to 352×352 for model inference, and then, the prediction is resized back to the original size of the input image. Note that both resizing processes use bilinear interpolation.

B. Data Sets

We conduct experiments on two benchmark data sets. The details of each data set are as follows.

1) *MAS3K [4]:* This data set contains a total of 3103 images with seven superclasses, i.e., mammals, reptile, marine fish, arthropod, coelenterate, mollusc, and others. Object-level annotations are provided. According to the original settings in MAS3K, we use 1769 images for training and 1141 images for testing.
2) *RMAS:* The data set consists of 3014 images with five superclasses, i.e., sea products, big fish, small fish, turtle, and others. We divide the whole data set into two parts, i.e., the training set (2514 images) and the testing set (500 images). Each image is annotated with an object-level mask and a category.

C. Competing Methods

We compare the proposed method with 12 state-of-the-art methods, including UNet++ [56], BASNet [53], PFANet [57], SCRN [58], U^2 Net [59], SINet [5], PFNet [17], RankNet [15], C^2 FNet [16],ECDNet [4], OCENet [38], and ZoomNet [39]. Among them, BASNet, PFANet, SCRNet, and U^2 Net are salient object segmentation methods. SINet, PFNet, RankNet, C^2 FNet, OCENet, and ZoomNet are COS approaches. UNet++ is originally designed for medical image segmentation. ECDNet is a recently proposed MAS method. For fair comparisons, we retrain the above methods using default implementations provided by the authors. Note that we directly use the original results of ECDNet for comparison since there are no publicly released implementations.

D. Evaluation Metrics

To comprehensively compare our method with other state-of-the-art approaches, we utilize five popular metrics to evaluate the segmentation performance, i.e., mean intersection over union (mIoU), S-measure (S_α) [60], weighted F-measure (F_β^w) [61], E-measure (mE_ϕ) [62], and mean absolute error (MAE). Among them, mIoU calculates the average of the intersection between a prediction and ground truth divided by their union. S_α measures the object-aware and region-aware structure similarities between a prediction and ground truth. F_β^w calculates the weighted precision and weighted recall to measure the overall performance synthetically. mE_ϕ is based on the human visual perception to evaluate the global and local accuracy. MAE is a widely used metric that evaluates the pixel-level error between the normalized prediction and ground truth. Apart from evaluating the segmentation accuracy, we adopt model parameters and floating-point operations (FLOPs) (i.e., floating-point multiplication-adds [29]) to measure the computational complexity.

E. Performance Comparison

1) Overall Performance: Table III summarizes the quantitative results of different methods on two benchmark data sets. We can observe that the proposed method obtains better performance than previous methods according to five objective metrics. Specifically, compared with the recently proposed MAS method (ECDNet) on MAS3K, our approach significantly improves the segmentation performance. Among all methods, C^2 FNet and ZoomNet achieve relatively good results. The reason may lie in that C^2 FNet and ZoomNet can capture richer global context information from multiscale features and scale integration, respectively.

2) Performance on Each Category: Table III lists the average performance on the whole data set. To further understand the effectiveness of each method, Table IV reports the segmentation performance of each category. Note that, although MASK has seven classes, no image belongs to the "others" category in the testing set. From Table IV, one can find that, on the whole, MASNet and ZoomNet outperform other approaches. The proposed method is good at segmenting marine fish, especially the small fish in RMAS. Besides, no method can obtain the best performance in all categories, and there is a large room to promote the MAS performance.

3) Performance on High-Quality, Low-Quality, and Camouflage Images: The RMAS testing set is divided into three classes to study the effectiveness of each method in handling image degradation and object camouflage. Table V reports the model performance on those three classes. From the table, we can make the following observations.

TABLE III
Quantitative Comparison With State-of-The-Art Methods on MAS3K and RMAS Data Sets According to Five Measurements

Method	Year	MAS3K					RMAS				
		$mIoU \uparrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$mE_\phi \uparrow$	$MAE \downarrow$	$mIoU \uparrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$mE_\phi \uparrow$	$MAE \downarrow$
UNet++ [56]	2018	0.506	0.726	0.552	0.790	0.083	0.558	0.763	0.644	0.835	0.046
BASNet [53]	2019	0.677	0.826	0.724	0.862	0.046	0.707	0.847	0.771	0.907	0.032
PFANet [57]	2019	0.405	0.690	0.471	0.768	0.086	0.556	0.767	0.582	0.810	0.051
SCRN [58]	2019	0.693	0.839	0.730	0.869	0.041	0.695	0.842	0.731	0.878	0.030
U^{2}Net [59]	2020	0.654	0.812	0.711	0.851	0.047	0.676	0.830	0.762	0.904	0.029
SINet [5]	2020	0.658	0.820	0.725	0.884	0.039	0.684	0.835	0.780	0.908	0.025
PFNet [17]	2021	0.695	0.839	0.746	0.890	0.039	0.694	0.843	0.771	0.922	0.026
RankNet [15]	2021	0.658	0.812	0.722	0.867	0.043	0.704	0.846	0.772	**0.927**	0.026
C^{2}FNet [16]	2021	0.717	0.851	0.761	0.894	0.038	0.721	0.858	0.788	0.923	0.026
ECDNet [4]	2021	0.711	0.850	0.766	0.901	0.036	–	–	–	–	–
OCENet [38]	2022	0.667	0.824	0.703	0.868	0.052	0.680	0.836	0.752	0.900	0.030
ZoomNet [39]	2022	0.736	0.862	0.780	0.898	**0.032**	0.728	0.855	0.795	0.915	**0.022**
MASNet	2022	**0.742**	**0.864**	**0.788**	**0.906**	**0.032**	**0.731**	**0.862**	**0.801**	0.920	0.024

TABLE IV
Quantitative Comparison (mIoU) on Each Class

Method	MAS3K						RMAS				
	Mammals	Reptile	Marine fish	Arthropod	Coelenterate	Mollusc	Sea products	Big fish	Small fish	Turtle	Others
UNet++ [56]	0.559	0.584	0.486	0.408	0.797	0.556	0.516	0.674	0.512	0.722	0.555
BASNet [53]	0.788	0.759	0.667	0.497	0.856	0.691	0.629	0.837	0.697	0.889	0.702
PFANet [57]	0.480	0.517	0.386	0.314	0.599	0.431	0.479	0.692	0.528	0.754	0.621
SCRN [58]	0.782	0.801	0.679	0.536	0.847	0.719	0.614	0.838	0.675	0.877	0.715
U^{2}Net [59]	0.731	0.739	0.644	0.493	0.865	0.670	0.590	0.814	0.654	0.864	0.757
SINet [5]	0.764	0.786	0.640	0.465	0.856	0.701	0.594	0.835	0.657	0.876	0.780
PFNet [17]	0.801	0.815	0.677	0.516	**0.892**	0.738	0.612	0.828	0.674	0.881	0.757
RankNet [15]	0.790	0.787	0.642	0.447	0.745	0.704	0.621	0.832	0.697	0.885	0.731
C^{2}FNet [16]	0.783	0.819	0.707	0.563	0.826	0.748	0.631	0.847	0.717	0.892	**0.780**
ECDNet [4]	–	–	–	–	–	–	–	–	–	–	–
OCENet [38]	0.768	0.794	0.645	0.494	0.843	0.729	0.606	0.815	0.655	0.887	0.694
ZoomNet [39]	**0.829**	**0.840**	0.724	**0.570**	0.858	0.765	**0.650**	**0.861**	0.709	**0.914**	0.758
MASNet	0.822	0.825	**0.733**	0.569	0.835	**0.783**	0.649	**0.861**	**0.723**	0.881	0.765

a) High-quality image set achieves comparatively higher performance. This is reasonable since it is easy to segment objects with a clear appearance and texture.

b) Degraded underwater images and camouflaged objects are unfavorable for segmentation due to their poor visibility and low contrast, especially for camouflaged objects that achieve the worst performance.

c) Compared with other methods, MASNet can better handle the degraded image, meanwhile having a competitive performance on high-quality and camouflage sets.

4) Computational Complexity: Table VI reports the model parameters and FLOPs. From the table, we can observe that the proposed method is relatively lightweight compared with the recently proposed deep-learning-based methods, such as ZoomNet and OCENet. Although UNet++ and C^2 FNet obtain the lowest parameters and FLOPs, respectively, their segmentation performance is inferior to our method. Table VI demonstrates that the proposed approach can better balance the computational complexity and segmentation performance.

5) Qualitative Evaluation: Fig. 9 shows the subjective comparison of MASNet with the others. It can be seen that MASNet is capable of accurately segmenting marine animals in diverse degraded underwater scenes. MASNet can achieve better visual results by detecting more accurate and complete objects with rich details. Moreover, MASNet also can infer the camouflaged objects concealed in different environments more accurately. This is mainly because our Siamese network with the MADA module can effectively reduce the impact of image degradation and object camouflage. In contrast, the state-of-the-art methods do not take the degradation and camouflage issues into account simultaneously, which will cause incorrect estimations.

F. Ablation Studies

We further conduct ablation studies to analyze the influence of the data augmentation, cross-level fusion strategy, and the feature enhancement module.

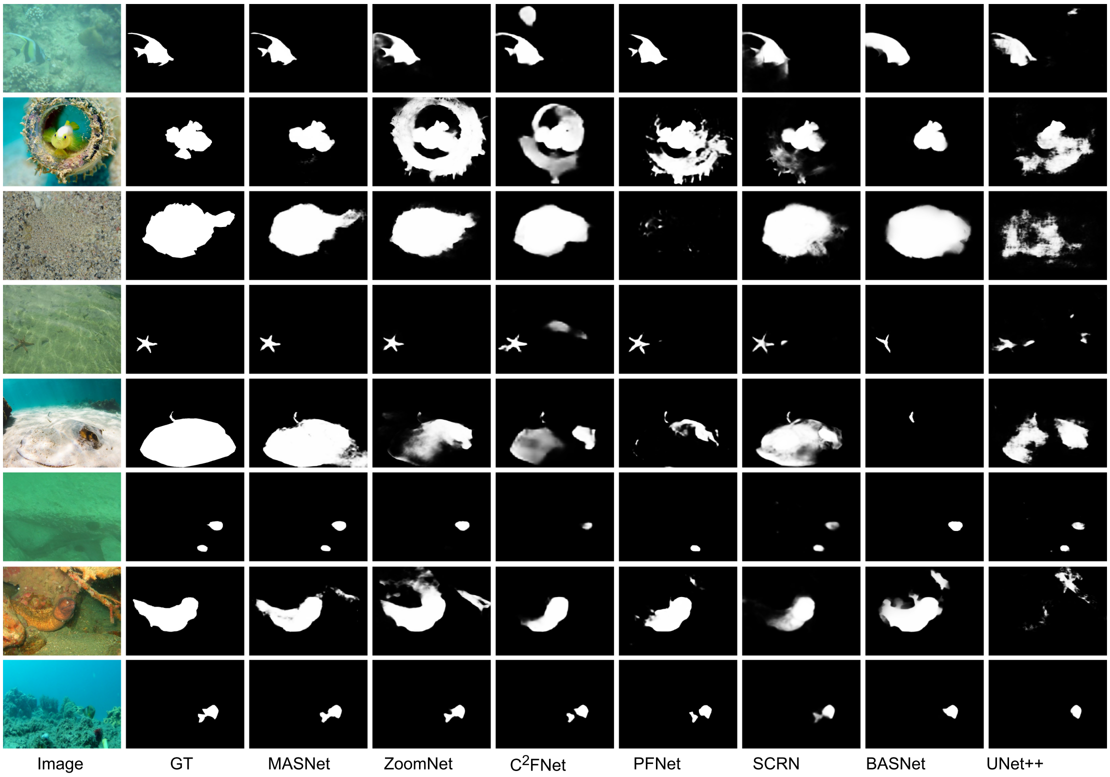

Fig. 9. Visual comparison of the proposed approach with several state-of-the-art methods on the MAS3K and RMAS data sets. The first five images are selected from MAS3K and the last three are from RMAS.

TABLE V
COMPARISON (mIoU) ON HIGH-QUALITY, LOW-QUALITY, AND CAMOUFLAGE IMAGES

Method	High-quality	Low-quality	Camouflage
UNet++ [56]	0.635	0.553	0.293
BASNet [53]	0.778	0.697	0.518
PFANet [57]	0.659	0.534	0.357
SCRN [58]	0.776	0.676	0.545
U^{2}Net [59]	0.770	0.650	0.542
SINet [5]	0.775	0.655	0.595
PFNet [17]	0.775	0.672	0.572
RankNet [15]	0.776	0.684	0.601
C^{2}FNet [16]	0.797	0.698	**0.630**
ECDNet [4]	–	–	–
OCENet [38]	0.757	0.661	0.551
ZoomNet [39]	**0.817**	0.703	0.606
MASNet	0.812	**0.710**	0.611

TABLE VI
COMPARISON OF MODEL PARAMETERS AND FLOPS

Method	Parameters (M)↓	FLOPs (G)↓
UNet++ [56]	**9.1**	65.9
BASNet [53]	87.1	240.9
PFANet [57]	16.4	63.4
SCRN [58]	25.2	15.1
U^{2}Net [59]	44.0	71.1
SINet [5]	48.9	19.5
PFNet [17]	46.5	19.0
RankNet [15]	28.7	17.4
C^{2}FNet [16]	26.4	**13.1**
ECDNet [4]	–	–
OCENet [38]	55.0	25.2
ZoomNet [39]	32.4	85.9
MASNet	26.2	17.2

1) Effectiveness of Data Augmentation: MASNet is trained based on the Siamese network with image-level and object-level augmentations. A task loss and an alignment loss are employed to learn degradation and camouflage irrelevant representations. Here, we conduct experiments to investigate the effectiveness of such a training strategy. We have tried the following two variations over the original MASNet.

a) Setting A: Directly training the network with original images (i.e., without MADA and the alignment loss).
b) Setting B: Directly training the network with MADA (i.e., without the alignment loss).

Experimental results are listed in Table VII, we can observe that applying the proposed data augmentation technique and the training strategy can boost the segmentation performance. In the table, we find that Setting B obtains the worst results. The reason may be that directly using MADA for training has a risk

TABLE VII
Ablation Experiments of the Data Augmentation

Method	MAS3K				
	$mIoU \uparrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$mE_\phi \uparrow$	$MAE \downarrow$
Setting A	0.733	0.857	0.780	0.901	0.034
Setting B	0.689	0.832	0.735	0.876	0.044
MASNet	0.742	0.864	0.788	0.906	0.032

TABLE VIII
Ablation Experiments of the Cross-Level Fusion

Method	MAS3K				
	$mIoU \uparrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$mE_\phi \uparrow$	$MAE \downarrow$
E5	0.677	0.835	0.725	0.894	0.041
E5+E4+E3	0.732	0.858	0.776	0.904	0.035
E5+E4+E3+E2+E1	0.742	0.864	0.788	0.906	0.032

TABLE IX
Ablation Experiments of the Feature Enhancement

Method	MAS3K				
	$mIoU \uparrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$mE_\phi \uparrow$	$MAE \downarrow$
MASNet w/o RFB	0.732	0.859	0.778	0.899	0.033
MASNet	0.742	0.864	0.788	0.906	0.032

of overfitting the augmentation operator. For example, objective-level augmentation changes objects' appearance, which allows the network uses shortcuts (e.g., appearance features) to discriminate an object. As a result, important semantic features are ignored. Instead, MASNet adopts a Siamese network to align the two outputs estimated from the original and augmented images, encouraging the network to learn the shared semantic features.

2) Effectiveness of Cross-Level Fusion: Since the proposed network is based on a multiscale fusion structure, it is necessary to conduct an ablation study to understand how each scale features affect the performance. We have tried the following two variations over the original MASNet.

a) E5: Using the features from E5.

b) E5+E4+E3: Using the features from E3, E4, and E5.

We report the test results in Table VIII. From the table, we can observe that the combination of low-level and high-level features (i.e., MASNet) can achieve better performance. The reason lies in that high-level and low-level features have their respective role in characterizing valuable information. For example, the former is more related to semantic features, while the latter can capture rich edge and location-related information.

3) Effectiveness of Feature Enhancement: In MASNet, we use the RFB to enhance features and help establish long-range semantic dependencies. In Table IX, we list the objective results to show the role of RFB. Note that, for the baseline method, we use 1×1 convolutional layers to replace the RFB to meet the specific requirement of channel numbers. From Table IX, we can observe that without RFB, the segmentation performance will drop to a certain degree. This confirms that the RFB is beneficial for the feature learning of the MAS task.

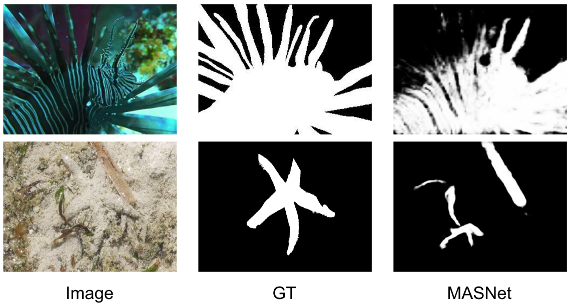

Fig. 10. Failure cases of the proposed method.

G. Limitations and Discussions

In this article, we propose MASNet for MAS. Although our method achieves promising results compared with existing methods, it still has some limitations.

1) Owing to the complex and diverse underwater environments, MASNet fails to handle some specific cases. We show two failure cases in Fig. 10. One can observe that it is difficult to segment objects with high camouflage attributes or animals in poor-quality images with complex shapes. Thus, tremendous efforts are highly demanded, and there exists a large room to promote the accuracy and robustness of MAS.
2) Existing object segmentation methods typically capture rich representations from deep convolutional neural networks (e.g., ResNet [29]) pretrained on large-scale data sets (e.g., ImageNet [63]). Despite the effectiveness, their application is constrained by the model size, computing power, and storage memory, especially for underwater scenarios. Therefore, lightweight models are required to deal with the limited computing and storage resource in specific underwater mobile devices.

V. Conclusion

In this article, we developed a new learning-based method for MAS considering both the image degradation and object camouflage properties. We proposed to combine data augmentation techniques with the Siamese network for the better segmentation of camouflaged and degraded marine objects. For each branch of the Siamese network, we elaborately designed a fusion-based structure to predict a high-quality segmentation map. In addition, we constructed a new data set as a supplement to existing benchmarks. The proposed data set can be used for training and evaluating MAS models. Experimental results on two MAS data sets showed that the proposed method outperforms state-of-the-art approaches significantly. In the future, we plan to develop lightweight models for accurate and real-time MAS.

References

[1] J. Lu, F. Yuan, W. Yang, and E. Cheng, "An imaging information estimation network for underwater image color restoration," *IEEE J. Ocean. Eng.*, vol. 46, no. 4, pp. 1228–1239, Oct. 2021.

[2] C. Li et al., "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2019.

[3] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4861–4875, Dec. 2020.

[4] L. Li, B. Dong, E. Rigall, T. Zhou, J. Dong, and G. Chen, "Marine animal segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2303–2314, Apr. 2022.

[5] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2777–2787.

[6] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 81–88.

[7] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1756–1769, Apr. 2012.

[8] D. Akkaynak and T. Treibitz, "Sea-Thru: A method for removing water from underwater images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1682–1691.

[9] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Trans. Image Process.*, vol. 30, pp. 4985–5000, 2021.

[10] Z. Fu et al., "Unsupervised underwater image restoration: From a homology perspective," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 643–651.

[11] K. Panetta, L. Kezebou, V. Oludare, and S. Agaian, "Comprehensive underwater object tracking benchmark dataset and underwater image enhancement with GAN," *IEEE J. Ocean. Eng.*, vol. 47, no. 1, pp. 59–75, Jan. 2022.

[12] W. Zhang, Y. Wang, and C. Li, "Underwater image enhancement by attenuated color channel correction and detail preserved contrast enhancement," *IEEE J. Ocean. Eng.*, vol. 47, no. 3, pp. 718–735, Jul. 2022.

[13] E. Trucco and A. Olmos-Antillon, "Self-tuning underwater image restoration," *IEEE J. Ocean. Eng.*, vol. 31, no. 2, pp. 511–519, Apr. 2006.

[14] H.-H. Chang, C.-Y. Cheng, and C.-C. Sung, "Single underwater image restoration based on depth estimation and transmission compensation," *IEEE J. Ocean. Eng.*, vol. 44, no. 4, pp. 1130–1149, Oct. 2019.

[15] Y. Lv et al., "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11591–11601.

[16] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1025–1031.

[17] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8772–8781.

[18] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[19] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[20] Y. Wang, W. Song, G. Fortino, L.-Z. Qi, W. Zhang, and A. Liotta, "An experimental-based review of image enhancement and image restoration methods for underwater imaging," *IEEE Access*, vol. 7, pp. 140233–140 251, 2019.

[21] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541–551, Jul. 2016.

[22] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, Jan. 2018.

[23] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[24] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 132–145, 2015.

[25] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, Dec. 2016.

[26] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1579–1594, Apr. 2017.

[27] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856–2868, Jun. 2018.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. Int. Proc. *Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[31] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robot. Autom. Lett.*, vol. 3, no. 1, pp. 387–394, Jan. 2018.

[32] Y. Guo, H. Li, and P. Zhuang, "Underwater image enhancement using a multiscale dense generative adversarial network," *IEEE J. Ocean. Eng.*, vol. 45, no. 3, pp. 862–870, Jul. 2020.

[33] Z. Fu, X. Fu, Y. Huang, and X. Ding, "Twice mixing: A rank learning based quality assessment approach for underwater image enhancement," *Signal Process.: Image Commun.*, vol. 102, 2022, Art. no. 116622.

[34] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, Oct. 2022.

[35] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Understanding*, vol. 184, pp. 45–56, 2019.

[36] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10071–10081.

[37] F. Yang et al., "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4146–4155.

[38] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1445–1454.

[39] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2160–2170.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.

[41] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[42] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[44] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1170–1177.

[45] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. Murillo, "CoralSeg: Learning coral segmentation from sparse annotations," *J. Field Robot.*, vol. 36, no. 8, pp. 1456–1477, 2019.

[46] M. Ravanbakhsh, M. R. Shortis, F. Shafait, A. Mian, E. S. Harvey, and J. W. Seager, "Automated fish detection in underwater images using shape-based level sets," *Photogrammetric Rec.*, vol. 30, no. 149, pp. 46–62, 2015.

[47] M.-C. Chuang, J.-N. Hwang, K. Williams, and R. Towler, "Automatic fish segmentation via double local thresholding for trawl-based underwater camera systems," in *Proc. IEEE Int. Conf. Image Process.*, 2011, pp. 3145–3148.

[48] M. J. Islam et al., "Semantic segmentation of underwater imagery: Dataset and benchmark," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2020, pp. 1769–1776.

[49] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4085–4095.

[50] M. Frigo and S. G. Johnson, "FFTW: An adaptive software architecture for the FFT," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 1998, vol. 3, pp. 1381–1384.

[51] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[52] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.

[53] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNET: Boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.

[54] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," *in Proc. Robot.: Sci. Syst.*, 2020.

[55] A. Saleh, I. H. Laradji, D. A. Konovalov, M. Bradley, D. Vazquez, and M. Sheaves, "A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[56] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. New York, NY, USA: Springer, 2018, pp. 3–11.

[57] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3085–3094.

[58] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7264–7273.

[59] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested u-structure for salient object detection," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107404.

[60] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.

[61] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps ?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.

[62] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function," *SCIENTIA SINICA Inf.*, vol. 51, 2021, Art. no. 1475.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

Zhenqi Fu received the B.S. degree in electronic information engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016, and the M.S. degree in electronics and communication engineering from Ningbo University, Ningbo, China, in 2019. He is currently working toward the Ph.D. degree in signal and information processing with Xiamen University, Xiamen, China.

From 2021 to 2022, he was a Visiting Student with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include image processing and machine learning.

Ruizhe Chen received the B.S. degree in information engineering from Southeast University, Nanjing, China, in 2020. He is currently working toward the master's degree in artificial intelligence with Xiamen University, Xiamen, China.

His current research interests include computer vision and machine learning.

Yue Huang (Member, IEEE) received the B.S. degree in electrical engineering from the Department of Electrical Engineering, Xiamen University, Xiamen, China, in 2005, and the Ph.D. degree in biomedical engineering from the Department of Biomedical Engineering, Tsinghua University, Beijing, China, in 2010.

She is currently an Associate Professor with the School of Informatics, Xiamen University. Her main research interests include image processing, sparse signal representation, and machine learning.

En Cheng received the Ph.D. degree in communication engineering from the Department of Communication Engineering, Xiamen University, Xiamen, China, in 2006.

He is currently a Professor with the Department Communication Engineering, Xiamen University, where he is also the Director of the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education. His research interests include underwater acoustic communication and networking, spanning from the communication networks, underwater acoustic communication, multimedia signal processing and communication, video/image quality measurement, and embedded system design.

Xinghao Ding was born in Hefei, China, in 1977. He received the B.S. and Ph.D. degrees in precision instruments from the Department of Precision Instruments, Hefei University of Technology, Hefei, China, in 1998 and 2003, respectively.

From 2009 to 2011, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. Since 2011, he has been a Professor with the School of Informatics, Xiamen University, Xiamen, China. His main research interests include machine learning, deep learning, computer vision, and signal processing.

Kai-Kuang Ma (Life Fellow, IEEE) received the Master of Science degree from Duke University, Durham, NC, USA, and the Ph.D. degree from North Carolina State University, Raleigh, NC, in 1992, both in electrical engineering.

From 1984 to 1992, he was with IBM Corporation, Armonk, NY, USA, where he was involved in various digital signal processing and very large scale integration advanced product development. From 1992 to 1995, he was a Member of the Technical Staff with the Institute of Microelectronics, Singapore, working on digital video coding and the MPEG standards. He is currently a Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has coauthored extensively in various international journals and conferences. From 1997 to 2001, he was the Singapore MPEG Chairman and the Head of Delegation. On the MPEG contributions, two fast motion estimations (Diamond Search and MVFAST) produced from his research group have been adopted by the MPEG-4 standard, as the reference core technology for fast motion estimation. His research interests include fundamental image/video processing and applied computer vision.

Dr. Ma was elected a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2008 to 2009. He has served various roles in professional societies, such as the General Co-Chair of 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2017 International Symposium on Intelligent Signal Processing and Communication Systems, 2016 Asian Conference on Computer Vision (ACCV) Workshop, 2013 Visual Communications and Image Processing, and MPEG 2001, the Technical Program Co-Chair of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, 2004 International Conference on Image Processing, 2007 IEEE International Symposium on Intelligent Signal Processing and Communication Systems, 2009 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, and 2010 Pacific-Rim Symposium on Image and Video Technology, and the Area Chair of ACCV 2009 and ACCV 2010. He had extensive editorship contributions in several international journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING as an Associate Editor from 2007 to 2010 and a Senior Area Editor since 2015, from which he received the Merit Award. He was an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2015 to 2018, for IEEE SIGNAL PROCESSING LETTERS from 2014 to 2016, and for IEEE TRANSACTIONS ON MULTIMEDIA from 2002 to 2009, an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS from 1997 to 2012, and an Editorial Board Member for *Journal of Visual Communication and Image Representation* from 2005 to 2014. He has been a Senior Area Editor for IEEE SIGNAL PROCESSING LETTERS since 2018. He was the Chairman of the IEEE Signal Processing Society Singapore Chapter from 2000 to 2002. He is a Fellow of the Academy of Engineering, Singapore.