SEA-Net: A Severity-Aware Network with Visual Prompt Tuning for Underwater Semantic Segmentation

Jiayong Zhu
Jiangnan University
Wuxi, Jiangsu, China
6223112046@stu.jiangnan.edu.cn

Tao Zhang
Central South University
Changsha, Hunan, China
taozhang@csu.edu.cn

Abstract

Underwater image semantic segmentation is widely used in the recognition and navigation of vision-guided underwater robots. However, due to issues such as insufficient underwater scene illumination and turbidity of the water, the contrast between targets and backgrounds is low. Existing underwater semantic segmentation methods ignore the recognition differences between underwater images, making it challenging for models to extract sufficiently robust visual features from noisy underwater images. To address this problem, we propose SEA-Net, which incorporates a SEA-Adapter and Visual Prompt Tuning. In the framework, we use a severity metric to address various complex noise problems in underwater images. The severity metric classifies all underwater images into High and Low-severity images and combines the Visual Prompt Tuning of two-branch alternating training, this allows the model to learn more robust visual features from different perspectives. On both the SUIM and DeepFish benchmarks, our proposed SEA-Net outperforms state-of-the-art methods in underwater image semantic segmentation tasks.

***Keywords:** underwater semantic segmentation, visual prompt tuning, adapter*

1. Introduction

Underwater semantic segmentation is closely related to industrial information and has wide applications in marine engineering, resource development, and ocean monitoring. Accurately identifying underwater objects and environments enhances operational efficiency, optimizes resource exploration and development processes, and improves environmental monitoring and protection in marine industries. This technology offers precise underwater scene understanding, providing crucial information to support decision-making and operations. More and more researchers are increasingly focusing on the field of underwater vision tasks [21], which includes underwater image target detection [50], underwater image enhancement [1], underwater image semantic segmentation [38]. Underwater image semantic segmentation provides powerful support for underwater robots to explore and exploit resources, and underwater image semantic segmentation models can provide the corresponding semantic information to better help underwater robots identify marine targets.

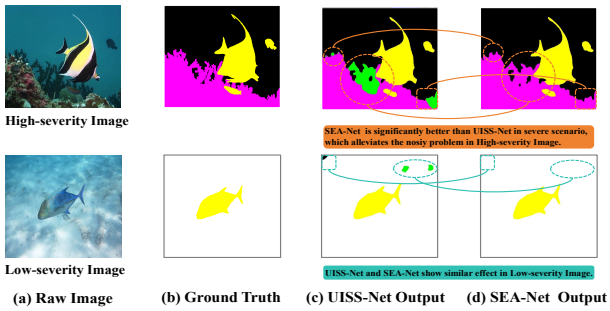


Figure 1. Qualitative results of underwater image semantic segmentation on the SUIM benchmark. For High and Low-severity images, we show the corresponding ground-truth map and the results of UISS-Net and our proposed SEA-Net method. We highlight some improved predictions with white dashed circles. It is evident that the performance of UISS-Net and SEA-Net is similar when processing Low-severity images. However, when processing High-severity images, the SEA-Net outperforms UISS-Net.

However, the quality of underwater images is often unstable due to light absorption and scattering in the water, resulting in an overall blue-green color tone in underwater images. Decreased image quality leads to low contrast between targets and backgrounds, color deviations, blurry edges, and details of objects, as well as uneven brightness [35], which restricts the application of underwater robots in real scenes.

Existing underwater image semantic segmentation methods, such as those proposed by Islam et al. [19], Liu et al. [29], and He et al. [16] ignore the difference in recognition difficulty between various underwater images, resulting in models struggle to extract sufficiently robust visual features from underwater images with noises such as low-

lighting complexity and lack of generalization to images in the presence of other underwater noises. Inspired by Gong et al. [12, 13, 23], we propose SEA-Net to address this problem, which consists of a SEA-Adapter and Visual Prompt Tuning. In the framework of SEA-Net, we provide a novel solution for the underwater image semantic segmentation task, e.g.we use the severity metric to unify the various complex noise problems of underwater images and direct the model to focus on the severity of the image recognition difficulty. The severity metric classifies all underwater images into High and Low-severity images and combines Visual Prompt Tuning with two-branch alternating training, which enables the model to learn more robust visual features from different perspectives. Visual Prompt Tuning, a learnable component within the SEA-Net, activates visual prompts to enhance High and Low-severity images. This mechanism ensures the two branches progressively highlight different severity levels, directing the model's focus toward the severity rather than specific underwater noises of the image. As shown in Fig. 1, the SEA-Net outperforms UISS-Net when processing High-severity images. We also design a SEA-Adapter specifically for underwater scenarios, which consists of a Down-Projection, a ReLU Activation Function, an Up-Projection, a 3x3 Convolution, and a Residual Connection sequentially to optimize the low-level features.

Our contributions are summarized as follows:

- We provide a new perspective on underwater image semantic segmentation tasks by proposing a novel Severity-Aware Network called SEA-Net. The framework employs a severity metric to unify the High-severity and Low-severity images and guides the model to focus on the severity of the recognition difficulty of the underwater images.

- SEA-Net consists of a SEA-Adapter and Visual Prompt Tuning. The SEA-Adapter is designed to be more compatible for underwater scenarios, where both the visual prompts and the SEA-Adapter can automatically update during training, without the need to adjust these hyperparameters manually.

- SEA-Net achieves a 3.8% higher mean Intersection over Union (mIoU) compared to the state-of-the-art on the SUIM benchmark for underwater semantic segmentation, with an improvement of 1.1% mIoU in paper. This demonstrates the effectiveness of the proposed method in learning robust features from noisy underwater images, aiding underwater robots in better recognizing marine life.

2. Related Work

2.1. Underwater image semantic segmentation

In recent years, semantic segmentation models based on deep learning methods have made great progress, such as FCN [28], VGG [37], and ResNet [16]. He et al. propose ResNet, which extracts features hierarchically and upsamples the low-dimensional features through a decoder network to generate the final semantic labels. Ronneberger et al. [34] propose U-Net architecture, which significantly improves performance by reusing the output of each encoder layer. Badrinarayanan et al. [2] use the pixel indexes in the encoder that performs maxpool for unpooling, thus eliminating the need for upsampling. Chen et al. [4] propose DeepLabv1, which uses atrous convolution to keep the receptive field consistent based on VGG. Chen et al. [5] introduce ASPP to add multi-scale training and larger receptive fields. Chen et al. [6] improve ASPP and propose both cascade and parallel architectures.

All of the above methods have achieved better results on natural images. However, the quality of underwater images is highly unstable, primarily due to light absorption by water. Therefore, directly transferring the aforementioned semantic segmentation methods to underwater images would be ineffective [41]. In recent years, researchers are committed to presenting underwater segmentation datasets and processing underwater images, Lian et al. [26] have proposed the first underwater image instance segmentation dataset UIIS to facilitate the training and evaluation of underwater instance segmentation models, which provides an in-depth discussion of instance segmentation of underwater scenes. Islam et al. [19] propose the first large-scale semantic segmentation dataset for underwater images, which provides a new benchmark for future research on semantic segmentation of underwater images. Wang et al. [39] use image enhancement based on multi-space transform to improve the quality of the original image. Liu et al. [29] introduce an unsupervised color correction method (UCM) module into the encoder structure of the framework to improve the quality of the images. He et al. [17] use a lightweight network to assist the backbone model and enhance the robustness of the model to better adapt to the images of the underwater scene. SAM, which is the vision foundation model, has achieved excellent results in many application scenarios such as Zhang et al. [47], Li et al. [24], Zhang et al. [48]. Very recently, Xu et al. [42] apply SAM to the underwater foreground segmentation task, achieving better performance on underwater foreground segmentation. O'Byrne et al. [32] propose to use realistic synthetic images to train models.

2.2. Visual Prompt Tuning

Prompt-based learning [20] is initially proposed in Natural Language Processing (NLP), it involves only a few parameters in the input space to fine-tune large pre-trained models for downstream tasks. Subsequently, prompt tuning has tended to be investigated in the field of computer vision (CV). Many approaches [22, 33, 43, 51] conduct some attempts to prompt visual language models (VLMs) in the form of text, and Jia et al. [20] first introduce the concept of "visual prompt" as a learnable vector. Bahng et al. [3] treat prompts as continuous task-specific vectors and individual image perturbations (e.g.soft prompts) are learned by backpropagation with frozen model parameters to demonstrate that prompts are feasible in the CV domain. Gong et al. [13], [12], [23] propose to classify images into High and Low-severity images in the unsupervised domain adaptation semantic segmentation task (UDASS), instructing the model to learn domain-invariant features, but ignoring scene-specific features.

2.3. Adapter

The concept of Adapter is first introduced in NLP [18], which acts as a compact and scalable module for fine-tuning large pre-trained models to each downstream task [7]. In the field of computer vision, Adapters have been recently used in the vision foundation model and vision-language foundation model, such as Li et al. [25] suggest fine-tuning ViT [11] for target detection with minimal modifications. Zhang et al. [46] propose Tip-Adapter, which trains Adapters without any backpropagation, instead of creating weights through a key-value caching model constructed from a small number of training sets, through this non-parametric approach, performance-optimized adapter weights can be obtained without any training. Chen et al. [9] propose ViT-Adapter, which utilizes adapters to enable ordinary ViTs to perform a variety of downstream tasks. Chen et al. [7] propose SAM-Adapter, which is the first method to apply Adapters to the pre-trained segmentation foundation model SAM for camouflage target detection and shadow detection. In this work, we propose the SEA-Adapter based on a vanilla adapter [18], which is designed to adapt to underwater scenes and optimize low-level features.

3. Method

3.1. Architecture

Inspired by U-Net, the overall architecture of our SEA-Net adopts a U-shape structure, using ResNet50 [16] as the backbone, we also follow [17], which uses an auxiliary network [15] for feature extraction in the encoder part to extract richer semantic information. The SEA-Net mainly consists of a SEA-Adapter and Visual Prompt Tuning. Severity-Aware Network uses a severity metric to unify various complex noise problems of underwater images, then all underwater images will be classified into High-severity and Low-severity images based on the severity metric. Visual Prompt Tuning is a learnable component based on the Severity-Aware Network. When the Severity-Aware Network determines the image as a High-severity image, the visual prompts are activated and added to the High-severity images to enhance the High-severity images. Similarly, when the Severity-Aware Network determines the image as a Low-severity image, the visual prompts are activated and added to the Low-severity images to enhance the Low-severity images. By training the dual branches alternately in this way, the model learns more robust visual features from different perspectives. SEA-Adapter, which absorbs prompt information from the visual prompts and optimizes the model's low-level features. The overall structure of SEA-Net is shown in Fig. 2.

3.2. Severity-Aware Network

Severity-Aware Network is primarily designed to classify input images into High-severity and Low-severity images with a severity metric. Visual Prompt Tuning is a learnable component, which is built on top of the Severity-Aware Network to enhance the High and Low-severity images by adding visual prompts.

3.2.1 Severity metric

Severity-Aware Network utilizes a severity metric to address various noise issues in underwater images, treating all noisy pixels in underwater images as severity pixels. Firstly, the input image $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ is converted to a grayscale image $\boldsymbol{X}_g \in \mathbb{R}^{H\times W}$, where H, W, and C denote the height, width, and number of channels of the image respectively. Then, the noisy pixels in the grayscale image with grayscale values lower than the grayscale threshold α are considered as High-severity pixels. As grayscale values between 0 and 50, the image is very dark and details may be difficult to discern, we set the grayscale threshold α to 40. We then calculate the ratio of High-severity pixels in all pixels of the original image. If this ratio is higher than the severity threshold τ, the image is classified as a High-severity image. Otherwise, it is classified as a Low-severity image. As is shown in Fig. 3(b), the darker regions in the image represent High-severity pixels, while the brighter regions represent Low-severity pixels.

$$\text{severity metric} = \begin{cases} \text{High, if } \frac{X_g<\alpha}{H\times W} > \tau \\ \text{Low, else} \end{cases} \tag{1}$$

3.2.2 Visual prompt

We represent the visual prompts as $VPT \in \mathbb{R}^{H_{vpt}\times W_{vpt}\times C}$, the number of visual prompts for

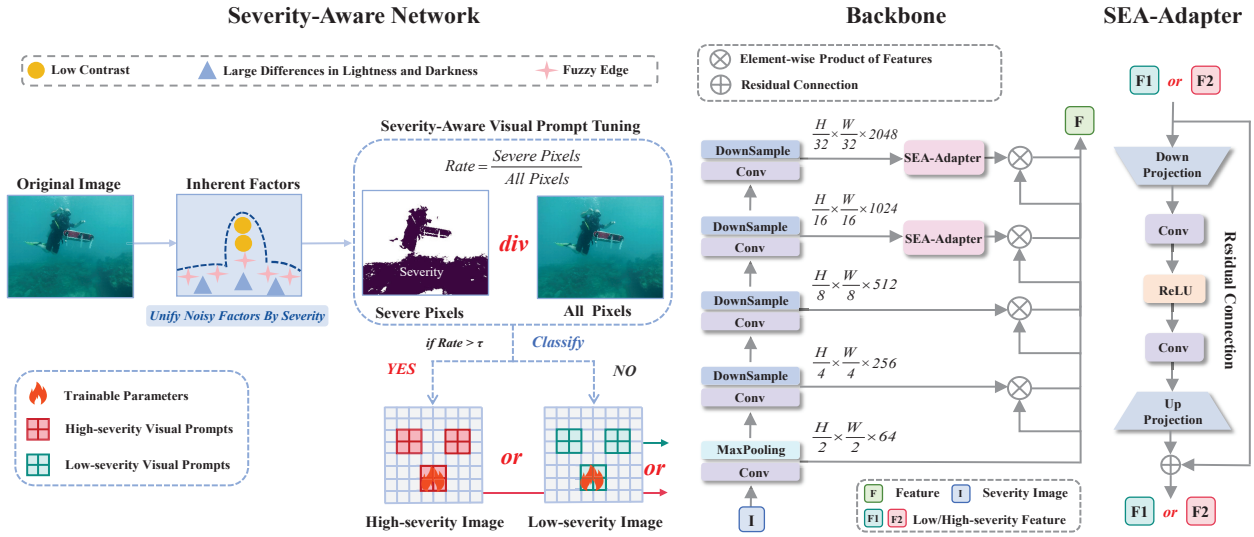


Figure 2. Overview of our proposed SEA-Net framework. It consists of three parts: Visual Prompt Tuning, Backbone, and SEA-Adapter. Blue arrows indicate the flow for data pre-processing, green arrows indicate the flow for Low-severity images, red arrows represent the flow for High-severity images, and gray arrows represent the flow for forward propagation. Given original images, we first pass them through severity metric and Visual Prompt Tuning to obtain High and Low-severity images, then we pass them to Backbone to obtain the final features.

each image is 3, $H_{vpt} = W_{vpt} = 64$. To avoid excessive masking the image by the visual prompts, the three visual prompts are added at the top two corners and the center of the image respectively. As is shown in Fig. 3(c), the equations for adding the visual prompts to an image are as follows:

$$X^H = X^h + 3 \cdot VPT^h \tag{2}$$

$$X^L = X^l + 3 \cdot VPT^l \tag{3}$$

$$\hat{X} = X^H \cup X^L \tag{4}$$

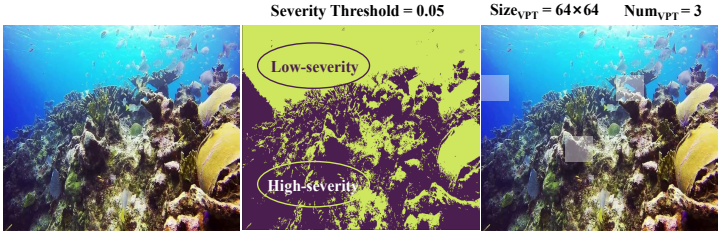
where X^h, X^l denotes a High-severity image and a Low-severity image respectively. VPT^h, VPT^l denotes a High-severity visual prompt and a Low-severity visual prompt respectively. X^H, X^L denotes the High-severity images and Low-severity images after adding the visual prompts, respectively. $\hat{X}$ denotes the enhanced images.

3.3. SEA-Adapter

We design a SEA-Adapter for underwater scenarios, which is based on the vanilla adapter. It consists of Down-projection, RELU Activation Function, Up-projection, 3x3 Convolution, and a Residual Connection.

This SEA-Adapter optimizes the low-level features F_1, F_2 among the five features $[F_1, F_2, F_3, F_4, F_5]$ output from Backbone. In this process, the SEA-Adapter aims to leverage beneficial prompts from visual prompts to better optimize the latent features of the model. The pipeline equation of SEA-Adapter is as follows:

$$F_1^{\hat{H/L}} = F_1^{H/L} + \text{Adapter}\left(F_1^{H/L}\right) \tag{5}$$



(a) Original Image (b) Image with severity metric (c) Image with VPT

Figure 3. Visualizations of High and Low-severity pixels and visual prompts on a single image from the SUIM dataset. We set the severity threshold to 0.05 and set the size of visual prompts to 64, and the number of visual prompts to 3.

$$F_2^{\hat{H/L}} = F_2^{H/L} + \text{Adapter}\left(F_2^{H/L}\right) \tag{6}$$

where $F_1^{\hat{H/L}}, F_2^{\hat{H/L}}$ represent the feature maps for High and Low-severity features respectively.

3.4. Overall training flow

The whole training flow is shown in Fig 2. The Severity-Aware Network classifies input images into High-severity images and Low-severity images based on a severity metric. Visual prompts are explicitly added to both types of images to generate enhanced High and Low-severity images. These images are then fed into a Backbone Network to obtain five features at different levels. The SEA-Adapter is used to further optimize two of the lower-level features. Finally, the features from all five layers are upsampled, multiplied, and concatenated to form the final feature map. The entire

training process is supervised by the loss function L_{final}. SEA-Net utilizes both the binary cross-entropy loss [27] and the dice loss functions [31]. While the cross-entropy (CE) loss function excels in multi-classification scenarios by effectively measuring the disparity between model predictions and actual labels, it also promotes faster convergence and enhanced performance. Nonetheless, it demands strong model stability and can be sensitive to uneven sample distributions. On the other hand, the dice loss function is skilled at handling significant imbalances in sample classes, prioritizing foreground region extraction during training. To address dataset irregularities and expedite model convergence, we combine both loss functions. The loss function is concerned with both the model's accurate prediction of segmentation boundaries (Dice Loss) and the model's accurate classification of foreground and background (BCE). The equations for the cross-entropy loss function and dice edge segmentation loss function are as follows:

$$L_{BCE} = -\sum_{i=0}^{N} y_i ln\left(p\left(x_i\right)\right) \tag{7}$$

$$\text{Dice coefficent} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

$$L_{\text{Dice loss}} = 1 - \text{ Dice coefficent} \tag{9}$$

where N is the number of samples, y_i is the label of the sample, and $p\left(x_i\right)$ is the foreground probability of predicting the sample x_i.

$$L_1 = L_{BCE}(GT, \text{ predict }(x)) \tag{10}$$

$$L_2 = L_{\text{Dice loss}}\left(GT_{\text{seg}}, \text{predict}\left(x_{\text{seg}}\right)\right) \tag{11}$$

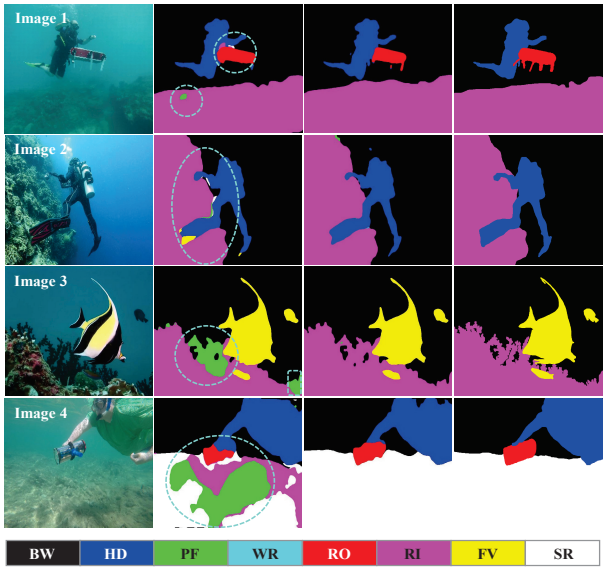
Where L_1 is the main network loss function, GT is the exact semantic labels, predict(x) is the obtained segmentation result, L_{BCE} is the computation process of cross-entropy loss. L_2 is the edge loss, GT_{seg} is the edge label obtained from the real semantic label of the real semantic label, predict (x_{seg}) is extracted from the edge extracted from the segmentation result, $L_{\text{Dice loss}}$ is the calculation process of the dice loss. The loss function of SEA-Net is as follows:

$$L_{\text{final}} = L_1 + L_2 \tag{12}$$

4. Experiments

4.1. Datasets and training setup

SUIM proposed by Islam et al. [19] is a large-scale dataset for underwater image semantic segmentation benchmark. It contains annotations for eight object classes, including Fish and Vertebrates (FV), Coral Reefs and Invertebrates (RI), Aquatic Plants and Sea-grasses (PF), Wrecks or Ruins (WR), Human Divers (HD), Robots (RO), Seafloor and Rocks (SR), and Background (Waterbody) (BW). 1525 RGB images are used for training and validation, while 110 images are used for benchmark evaluation of the model. The images in SUIM have different resolutions including $1906 \times 1080, 1280 \times 720, 640 \times 480, 256 \times 256$. DeepFish proposed by Saleh et al. [36] consists of approximately 40,000 underwater images collected from 20 Australian tropical marine environment habitats. The dataset initially contained only classification labels, and 300 fish semantic segmentation labels were added later. In this paper, the training and test sets are divided according to the ratio of 9:1. The images in both datasets are preprocessed and resized to 512×512. The proposed method is trained for 100 epochs on NVIDIA RTX 3090 GPU with Pytorch version 1.13.0. In addition, we follow the UISS- Net [17], with the initial learning rate set to 0.0004, weight decay set to 0.0001, and momentum set to 0.5.



(a) Original (b) UISS-Net (c) SEA-Net (d) Ground Truth

Figure 4. Quantitative experiments between UISS-Net and SEA-Net on the SUIM dataset. For each target image, we show the corresponding ground truth map and the results of UISS-Net and our proposed SEA-Net. We highlight some improved predictions with blue dashed circles.

4.2. Evaluation criteria

In order to thoroughly assess the model's performance, we have chosen mean Intersection over Union (mIoU), mean Pixel Accuracy (mPA), and Precision (Accuracy) as the evaluation metrics. mIoU is the average of the ratio between the intersection and union of predicted results and ground truth for all classes. mPA is the average ratio of correctly classified pixels for each class. Accuracy is the ratio of pixels that are correctly classified into their respective

Table 1. Comparison between SEA-Net and current mainstream models on the SUIM benchmark. We present the per-class and mean IoU (mIoU). We highlight the best and second-best results in each column in bold and italics, respectively

Model	IoU (%)								mIoU (%)
	BW	HD	PF	WR	RO	RI	FV	SR	
U-Net [34]	79.46	32.25	21.85	33.94	23.65	50.28	38.16	42.16	39.85
U-Net(ResNet)	**90.14**	72.53	2.37	62.65	59.19	69.93	73.13	**69.31**	62.41
U-Net(VGG)	90.03	79.81	4.25	62.23	51.43	*71.23*	74.11	68.27	62.73
SegNet [2]	80.63	45.67	17.45	32.24	55.72	47.62	43.92	51.51	46.85
SUIM-Net [19]	80.64	63.45	23.27	41.25	60.89	53.12	46.02	57.12	53.22
PSPNet [49]	82.51	65.04	28.54	46.56	62.88	55.8	46.78	55.98	55.51
DeepLab [4]	81.82	50.26	17.05	43.33	63.6	57.18	43.56	55.35	51.52
LEDNet [40]	82.96	58.47	18.02	42.86	50.96	58.13	46.13	54.99	51.36
BiseNetv2 [44]	83.67	59.29	18.27	39.58	56.54	58.16	47.33	56.93	52.47
UISS-Net [17]	87.18	**87.03**	*29.48*	**71.27**	**84.11**	70.7	**79.44**	67.54	*72.09*
UISS-Net (test)	89.11	80.88	24.37	68.54	79.22	69.78	76.95	65.96	69.35 (+0.00)
SEA-Net (ours)	*89.75*	*84.69*	**38.45**	*71.11*	*80.49*	**72.68**	*79.11*	*68.92*	**73.15** (+3.80)

categories out of the total number of pixels, assuming that $P_i, (i \in N)$ is the accuracy for each class at pixel i, mIoU, mPA, and Accuracy are defined as:

$$mIoU = \frac{1}{N} \times \sum \left(\frac{TP}{TP + FP + FN} \right) \quad (13)$$

$$Accuracy = \frac{TP + TN}{FP + TP + FN + TN} \quad (14)$$

$$mPA = \frac{1}{N} \times \sum (P_i) \quad (15)$$

where N denotes the number of categories, TP denotes the number of pixels correctly predicted by the model as positive samples, FP denotes the number of pixels that the model correctly predicts as positive samples, TN denotes the number of pixels that the model correctly predicts as negative samples, and FN the number of pixels that the model incorrectly predicts as negative samples.

4.3. State-of-The-Art performance comparison

Table 1 and Table 2 summarize the performance of our method SEA-Net compared with state-of-the-art methods on the SUIM and DeepFish benchmark respectively. Table 1 shows that our proposed SEA-Net method performs much better on average than all state-of-the-art methods on the SUIM benchmark. This advantage is derived from improvements over several classes such as aquatic plants and seaweed (PF), coral reefs, and invertebrates (RI), we achieved the best or second-best performances for all classes. We also present visual comparisons of the proposed SEA-Net with UISS-Net on the SUIM dataset in Fig. 4. Table 2 shows that our proposed SEA-Net method performs much better on average than all state-of-the-art methods on the DeepFish dataset and Fig. 5 presents visual comparisons of the proposed SEA-Net with UISS-Net on the DeepFish dataset. It is evident that, compared with the other methods, our method can not only provide sharper edges of classes in Low-severity images, such as aquatic plants and seaweed, but also improve the detection of classes in High-severity images, such as signs and lights. This demonstrates the effectiveness of the proposed SEA-Net on underwater images.

Table 2. Comparison between SEA-Net and current mainstream models on the DeepFish benchmark. We present the per-class and mean IoU (mIoU). We highlight the best and second-best results in each column in bold and italics, respectively

Model	IoU (%)		mIoU (%)
	Background	Foreground	
SUIM-Net [19]	99.03	78.4	88.71
SegNet [2]	98.89	68.94	83.91
PSPNet [49]	99.11	71.35	85.23
FCN [28]	99.15	72.61	85.88
DeepLabv3 [6]	99.21	66.3	82.75
HANet [10]	99.25	81.37	90.31
DGCNet [45]	99.21	81.42	90.32
MFAS-Net [14]	99.15	84.86	92.01
DPANet [8]	99.31	82.56	85.88
UISS-Net [17]	*99.55*	*90.55*	*95.05*
UISS-Net (test)	99.49	89.03	94.26 (+0.00)
SEA-Net (ours)	**99.57**	**90.73**	**95.15 (+0.89)**

4.4. Ablation study analysis

4.4.1 Ablation study on different components

To better understand the impact of each component of our SEA-Net, we conducted an ablation study by selectively deactivating each component and measuring the effect on the performance of the underwater semantic segmentation task. Specifically, we defined five nested subset models:

(1) VPT: Using the basic architecture from U-Net with visual prompts in Section 3.2.2 and BCE loss (Eq. (10) and Dice loss (Eq. (11)). We set the number of visual prompts to 3 without severity metric and severity threshold to 0.05.

Table 3. Ablation study of SEA-Net components on the SUIM Dataset. VPT means SEA-Net without the severity metric. SM means SEA-Net with the severity metric. We highlight the best and second-best results in each column in bold and italics, respectively

Input Size	VPT	SM	SEA-Adapter	mIoU (%)	mPA (%)	Accuracy (%)
512 × 512				69.35 (+0.00)	78.26	86.98
512 × 512	✓			70.22 (+0.87)	78.95	87.20
512 × 512	✓		✓	*71.77* (+2.42)	*80.34*	*88.23*
512 × 512			✓	70.87 (+1.52)	79.62	87.23
512 × 512		✓		71.45 (+2.10)	79.90	87.56
512 × 512		✓	✓	**73.15** (+3.80)	**81.52**	**88.01**

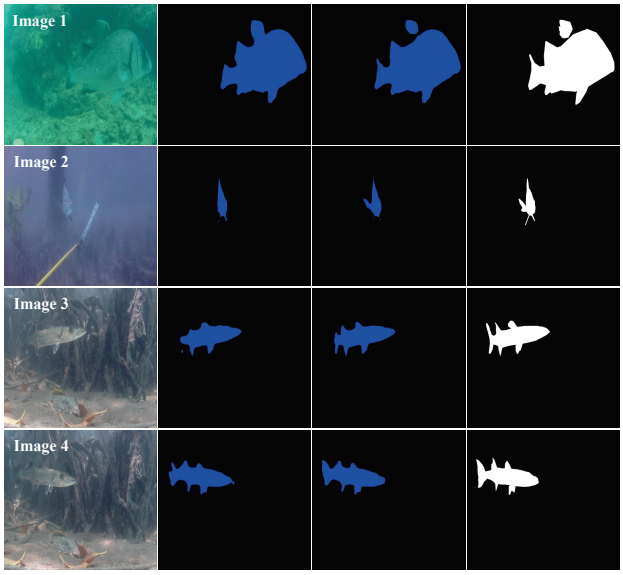
(2) SM: We set the number of visual prompts to 3 with the severity metric and severity threshold to 0.05.

(3) SEA-Adapter: Using the basic architecture from U-Net with SEA-Adapter in Section 3.3.

(4) VPT+SEA-Adapter: Further adding the SEA-Adapter based on VPT.

(5) SM+SEA-Adapter: Further adding the SEA-Adapter based on SM.

The results are presented in Table 3 showing that our overall SEA-Net resulted in a performance gain of 3.8% over the basic architecture in mIoU. The VPT (Visual Prompt Tuning) alone is responsible for a 0.87% improvement, the SM (Visual Prompt Tuning with severity metric produces an additional 1.23% improvement, and the SEA-Adapter provides a 1.52% improvement. This verifies the importance of the Visual Prompt Tuning, severity metric, and SEA-Adapter components of our SEA-Net.



(a) Original **(b) UISS-Net** **(c) SEA-Net** **(d) Ground Truth**

Figure 5. Quantitative experiments between UISS-Net and SEA-Net on the DeepFish dataset. For each target image, we show the corresponding ground truth map and the results of UISS-Net and our proposed SEA-Net.

4.4.2 Ablation study of Visual Prompt Tuning

The size, number, and position of the visual prompts may impact the learning process. To verify this, we set different sizes of visual prompts during training, such as 8×8, 16×16, 32×32, and 64×64. As shown in Table 4, the performance of visual prompts with the size of 64×64 is better than the other sizes of visual prompts, which achieves 71.45% mIoU. However, the performance of visual prompts with the size 8×8 and 16×16 are relatively lower compared to the others. The reason for this could be smaller mask areas constrain the model's representational capacity for specific tasks, preventing it from adequately capturing the features of the images, which causes models to learn worse.

Meanwhile, we also set different numbers of visual prompts such as 1, 3, 4, 5. As shown in Fig. 6(b), the performance of visual prompts with the number 5 is better than other numbers of visual prompts, which achieves 72.13% mIoU. However, the performance of visual prompts with the number 1 is relatively lower compared to the others. The reason for this could be few mask blocks introduce more diverse conditions, which causes models to learn worse.

UISS-Net crops images by placing gray bars on the input image to prevent distortion. However, if visual prompts are placed on the gray bars, the model might learn unrelated features. Therefore, we have also examined how the position of the visual prompts affects performance. Based on the three visual prompts, we arranged them with different starting positions. The first set has center positions at 200 and corner positions at 320. The second set is centered at 200, with corner positions at 400. The third set has a center position of 200 and corner positions at 448. The fourth set has a center position of 200 and corner positions at 496. As shown in Table 5, the visual prompts with center positions at 200, and two corner positions at 400 perform better than the visual prompts with other positions, which achieves 71.05% mIoU.

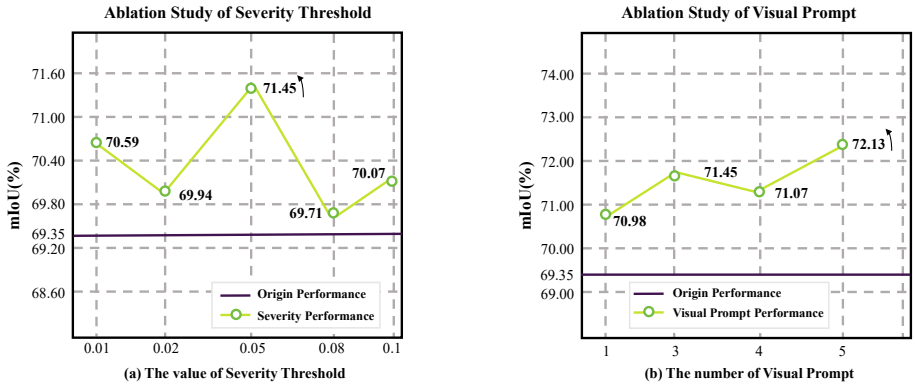


Figure 6. Ablation study of the severity threshold and visual prompt on the SUIM dataset by line graph visualization. The results show that the selection of severity threshold and visual prompt can significantly impact the performance of the model.

Table 4. Sensitivity analysis of Visual Prompt Size on the SUIM dataset. We highlight the best results in each column in bold

Visual Prompt Size	8×8	16×16	32×32	**64×64**
mIoU	69.30	68.53	70.10	**71.45**
mPA	78.40	77.85	79.13	**79.90**
Accuracy	86.28	84.45	87.17	**87.56**

Table 5. Sensitivity analysis of Visual Prompt Position on the SUIM dataset. We highlight the best results in each column in bold

Visual Prompt Position	**320**	400	448
mIoU	**71.45**	71.12	70.31
mPA	**79.90**	79.84	79.37
Accuracy	**87.56**	87.05	86.97

4.4.3 Ablation study of severity threshold

We apply the severity threshold in our method: τ. The setting of the severity threshold also has an effect on the model's performance as well. In order to verify this, we set different severity thresholds during training, such as 0.01, 0.02, 0.05, 0.08, 0.1. As shown in Fig. 6(a), the model performs best at τ = 0.05, which achieves 71.45% mIoU.

Table 6. Sensitivity analysis of SEA-Adapter Position on the SUIM dataset. We highlight the best results in each column in bold

SEA-Adapter Position	mIoU	mPA	Accuracy
$[F_1]$	70.87	79.62	87.23
$[F_2]$	70.62	79.20	88.15
$[\mathbf{F_1}, \mathbf{F_2}]$	**71.77**	**80.34**	**88.23**
$[F_1, F_2, F_3]$	70.01	78.89	85.99
$[F_1, F_2, F_3, F_4]$	66.15	75.50	84.71
$[F_1, F_2, F_3, F_4, F_5]$	69.10	77.50	86.11

4.4.4 Ablation study of SEA-Adapter

We apply the SEA-Adapter in our method. The position of the SEA-Adapter also have an effect on the model's performance as well. In order to verify this, we set different positions of SEA-Adapter by adding SEA-Adapter after each of the five features $[F_1, F_2, F_3, F_4, F_5]$ output from Backbone. As shown in Table 6, the model performs best when adding SEA-Adapter after F_1, F_2, which achieves 71.77% mIoU.

4.5. Limitations and future work

As illustrated by the previous experimental results (Fig. 4 and Fig. 5), the proposed SEA-Net performs well in most situations. However, it still has some limitations, for example, it may fail to detect some underwater fish clusters with very small sizes. Therefore, in future work, we will investigate how to improve the performance by training our model with the help of boundary detection. For example, boundary detection [30] can be helpful for learning a clearer contour for each object. Despite this, differences in biological species and water quality in various marine regions can cause models to perform differently. As a result, future efforts may involve creating specific datasets for different sea areas and utilizing unsupervised domain adaptation techniques for semantic segmentation to enhance the model's generalization across diverse underwater environments.

5. Conclusions

In this paper, we propose a new approach that uses a severity metric to unify various complex noise problems in underwater images. The proposed SEA-Net consists of a SEA-Adapter and Visual Prompt Tuning. The Severity-Aware Network addresses all the noise problems in underwater images through a severity metric, and Visual Prompt Tuning is a learnable component built on the Severity-Aware Network. This approach achieves state-of-the-art performance on both underwater semantic segmentation datasets SUIM and DeepFish, offering new insights for fu-

ture research. Nonetheless, variations in biological species and water quality across different marine regions may lead to divergent model performance. Thus, future endeavors may entail the creation of specialized datasets for distinct sea areas and the application of unsupervised domain adaptive semantic segmentation methods to enhance model generalization across diverse underwater environments.

References

[1] S. Anwar and C. Li. Diving deeper into underwater image enhancement: A survey. *Signal Processing: Image Communication*, 89:115978, 2020.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] T. Chen, L. Zhu, C. Ding, R. Cao, S. Zhang, Y. Wang, Z. Li, L. Sun, P. Mao, and Y. Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 1(2):5, 2023.

[8] Z. Chen, R. Cong, Q. Xu, and Q. Huang. Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:7012–7024, 2020.

[9] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[10] S. Choi, J. T. Kim, and J. Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9373–9383, 2020.

[11] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Z. Gong, F. Li, Y. Deng, D. Bhattacharjee, X. Zhu, and Z. Ji. Coda: Instructive chain-of-domain adaptation with severity-aware visual prompt tuning. *arXiv preprint arXiv:2403.17369*, 2024.

[13] Z. Gong, F. Li, Y. Deng, W. Shen, X. Ma, Z. Ji, and N. Xia. Train one, generalize to all: Generalizable semantic segmentation from single-scene to all adverse scenes. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2275–2284, 2023.

[14] A. Haider, M. Arsalan, J. Choi, H. Sultan, and K. R. Park. Robust segmentation of underwater fish based on multilevel feature accumulation. *Frontiers in Marine Science*, 9:1010565, 2022.

[15] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Z. He, L. Cao, J. Luo, X. Xu, J. Tang, J. Xu, G. Xu, and Z. Chen. Uiss-net: Underwater image semantic segmentation network for improving boundary segmentation accuracy of underwater images. *Aquaculture International*, pages 1–14, 2024.

[18] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[19] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776. IEEE, 2020.

[20] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[21] M. Jian, X. Liu, H. Luo, X. Lu, H. Yu, and J. Dong. Underwater image processing and analysis: A review. *Signal Processing: Image Communication*, 91:116088, 2021.

[22] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022.

[23] F. Li, Z. Gong, Y. Deng, X. Ma, R. Zhang, Z. Ji, X. Zhu, and H. Zhang. Parsing all adverse scenes: Severity-aware semantic segmentation with mask-enhanced cross-domain consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13483–13491, 2024.

[24] S. Li, J. Cao, P. Ye, Y. Ding, C. Tu, and T. Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*, 2024.

[25] Y. Li, H. Mao, R. Girshick, and K. He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[26] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, and S. Kwong. Watermask: Instance segmentation for underwater imagery.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1305–1315, 2023.

[27] T. Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[29] F. Liu and M. Fang. Semantic segmentation of underwater images based on improved deeplab. *Journal of Marine Science and Engineering*, 8(3):188, 2020.

[30] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018.

[31] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[32] M. O'Byrne, V. Pakrashi, F. Schoefs, and B. Ghosh. Semantic segmentation of underwater imagery using deep networks trained on synthetic imagery. *Journal of Marine Science and Engineering*, 6(3):93, 2018.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[35] D. K. Rout, B. N. Subudhi, T. Veerakumar, and S. Chaudhury. Walsh–hadamard-kernel-based features in particle filter framework for underwater object tracking. *IEEE Transactions on Industrial Informatics*, 16(9):5712–5722, 2019.

[36] 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. *Scientific Reports*, 10(1):14671, 2020.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] R. Vohra, F. Senjaliya, M. Cote, A. Dash, A. B. Albu, J. Chawarski, S. Pearce, and K. Ersahin. Detecting underwater discrete scatterers in echograms with deep learning-based semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–384, 2023.

[39] J. Wang, X. He, F. Shao, G. Lu, R. Hu, and Q. Jiang. Semantic segmentation method of underwater images based on encoder-decoder architecture. *Plos one*, 17(8):e0272666, 2022.

[40] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In *2019 IEEE international conference on image processing (ICIP)*, pages 1860–1864. IEEE, 2019.

[41] M. Waszak, A. Cardaillac, B. Elvesæter, F. Rødølen, and M. Ludvigsen. Semantic segmentation in underwater ship inspections: Benchmark and data set. *IEEE Journal of Oceanic Engineering*, 48(2):462–473, 2022.

[42] M. Xu, J. Su, and Y. Liu. Aquasam: Underwater image foreground segmentation. *arXiv preprint arXiv:2308.04218*, 2023.

[43] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024.

[44] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[45] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. Torr. Dual graph convolutional network for semantic segmentation. *arXiv preprint arXiv:1909.06121*, 2019.

[46] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[47] X. Zhang, Y. Liu, Y. Lin, Q. Liao, and Y. Li. Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22520–22528, 2024.

[48] Y. Zhang, Z. Shen, and R. Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, page 108238, 2024.

[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[50] J. Zhou, L. Pang, D. Zhang, and W. Zhang. Underwater image enhancement method via multi-interval subhistogram perspective equalization. *IEEE Journal of Oceanic Engineering*, 2023.

[51] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.