Journal of *Marine Science and Engineering*

MDPI

Article

AlgaeMask: An Instance Segmentation Network for Floating Algae Detection

Xiaoliang Wang [1], Lei Wang [1,2,*], Liangyu Chen [1], Feng Zhang [1], Kuo Chen [1], Zhiwei Zhang [1], Yibo Zou [3] and Linlin Zhao [3]

[1] East Sea Information Center, State Oceanic Administration, Shanghai 200136, China
[2] National Marine Data and Information Service, Tianjin 300012, China
[3] The School of Information, Shanghai Ocean University, Shanghai 201308, China
* Correspondence: wangleidett727@163.com

Abstract: Video surveillance on the offshore booster station and around the coast is a effective way to monitor floating macroalgae. Previous studies on floating algae detection are mainly based on traditional image segmentation methods. However, these algorithms cannot effectively solve the problem of extracting *Ulva prolifra* and *Sargassum* at different sizes and views. Recently, instance segmentation methods have achieved great success in computer vision applications. In this paper, based on the CenterMask network, a novel instance segmentation architecture named AlgaeMask is proposed for floating algae detection from the surveillance videos. To address the feature extraction ability of the network in the inter-dependencies for position and channel, we introduce a new OSA-V3 module with the dual-attention block, which consists of a position attention mechanism and channel attention mechanism. Meanwhile, scale-equalizing pyramid convolution is introduced to solve the problem of scale difference. Finally, we introduce the feature decoder module based on FCOS head and segmentation head to obtain the segmentation area of floating algae in each bounding box. The extensive experiment results show that the average precision of our AlgaeMask in the tasks of mask segmentation and box detection can reach 44.22% and 48.13%, respectively, which has 15.09% and 8.24% improvement over CenterMask. In addition, the AlgaeMask can meet the real-time requirements of floating algae detection.

Keywords: floating algae detection; *Ulva prolifera*; *Sargassum*; instance segmentation; video surveillance

Citation: Wang, X.; Wang, L.; Chen, L.; Zhang, F.; Chen, K.; Zhang, Z.; Zou, Y.; Zhao, L. AlgaeMask: An Instance Segmentation Network for Floating Algae Detection. *J. Mar. Sci. Eng.* **2022**, *10*, 1099. https://doi.org//10.3390/jmse10081099

Academic Editor: Marco Cococcioni

Received: 6 July 2022
Accepted: 9 August 2022
Published: 11 August 2022

1. Introduction

In recent years, disaster events involving floating macroalgae have occurred frequently. These events have caused deterioration of the marine ecological environment, as well as serious economic damage to fisheries, marine transportation, and marine tourism in the coastal areas of China [1,2]. The large-scale accumulation of floating algae on the sea surface blocks sunlight and exhausts the oxygen in water during the process of extinction, which seriously affects the survival of marine life. Floating macroalgae disasters in the coastal areas of China mainly include green tides of *Ulva prolifera* and golden tides of *Sargassum*. *Ulva prolifera* is bright or dark green and appears in the Yellow Sea, while *Sargassum* is brownish-yellow or dark-brown and mainly blooms in the East China Sea [3,4].

It is necessary to detect the location and distribution range of these floating algae (*Ulva prolifera* and *Sargassum*) accurately in the sea waters of China across a long period of time. Real-time monitoring of floating algae could provide a reliable basis for the analysis, prevention, and control of disasters to reduce economic and ecological costs. Therefore, a lot of floating algae detection algorithms and methodologies have been researched thus far [5–8].

Satellite remote sensing technology is one of the effective ways to capture the distribution of floating algae in the ocean due to its advantages of broad spatiotemporal

coverage and frequent data acquisition [9–11]. Moderate Resolution Image Spectroradiometer (MODIS) and Geostationary Ocean Color Imager (GOCI) data are commonly used in related research. Xing et al. captured the spatiotemporal features of floating *Sargassum* in the Yellow Sea to calculate their distribution and drifting path via high-spatial-resolution satellite images [4]. Wang et al. proposed a novel method to quantify *Sargassum* distribution and coverage by the MODIS alternative floating algae index (AFAI) over the Central West Atlantic region [12]. Xu et al. conducted the comparison between MODIS survey data and UAV images to verify the detection efficiency and accuracy for the green tides in the Yellow Sea [13]. Shin et al. monitored *Sargassum* distribution on the coast of Jeju Island by GOCI-II imagery captured in 2020 and adopted the GentleBoost model as the detection model [14]. For improving image quality, Cui et al. proposed a super-resolution detection model to reconstruct a high-resolution image of a region from GOCI images in order to distinguish the floating macroalgae patches from the water area more precisely [15]. Liang et al. proposed an extreme learning machine (ELM) method to detect floating macroalgae based on GOCI data, which was insufficiently sensitive to determine the value of threshold for traditional methods [16]. Qiu et al. used multi-layer perceptron (MLP) to monitor floating macroalgae automatically, robust to different environmental conditions, from GOCI imagery in the Yellow Sea [17].

Synthetic aperture radar (SAR) can image the earth in all weather conditions and in high spatial resolution [18]. Shen et al. proposed an unsupervised recognition method for green tide from RADARSAT-2 SAR images, paying attention to the polarimetric characteristics of green macroalgae blooms in both amplitude and phase domains [19]. Ma et al. integrated MODIS with SAR to jointly detect green tide accurately in the Yellow Sea in 2021 and showed the spatiotemporal changes of the green tide in more detail than a single data source [5].

For traditional image processing methods, image transformation and threshold segmentation are adopted to achieve floating algae segmentation effectively. Obviously, the image processing methods have the advantages of simple feature extraction, fast computing speed, and low deployment cost. However, the traditional methods require artificial feature design and predefined templates, and the process of parameter adjustment is very complex. These methods are very sensitive to environment changes and difficult to apply in monitoring floating algae accurately in practice. With the development of deep learning technology in recent years, convolutional neural networks (CNNs) have been successfully applied in the field of object recognition, image segmentation, video analysis, and so on, as they are able to automatically extract useful and rich features. At present, a CNN-based method has become one of the most popular methods in the field of floating algae detection [20]. Valentini et al. proposed a smartphone-camera-based *Sargassum* monitoring system in the French Antilles. The work adopted a pre-trained MobileNet-V2 model for image patch classification and the fully connected CRF to extract semantic segmentation in detail [21]. Arellano-Verdejo et al. designed the ERISNet model based on CNN and RNN to detect floating and accumulated *Sargassum* for MODIS data along the Mexican Caribbean coastline [22]. Wan et al. introduced a novel *Enteromorpha prolifera* (EP) extraction framework from GOCI images. Firstly, a strategy for the sample imbalance between EP and the background was adopted. Then, the network based on 1D-CNN and Bi-LSTM was proposed to make use of the spectral feature and context dependencies of each pixel [23]. For high-resolution aerial images captured by UAV, Wang et al. introduced an *Ulva prolifera* region detection method, using a superpixel segmentation algorithm to generate multi-scale patches and a binary CNN model to determine whether the patches are *Ulva prolifera* or not [24].

Based on CNNs network, U-Net [25] proposes a symmetric structure composed of encoders and decoders to complete the concatenation of low-level and high-level features, and the overall network presents a U-shaped structure. The methods based on U-Net and its related variants have achieved great success in the field of image segmentation [26]. Therefore, a lot of U-Net based methods have been applied to the field of floating algae monitoring. Kim et al. introduced the U-Net framework to detect red tide surrounding

the Korean peninsula, which consists of five U-shaped encoder and decoder layers to capture the spectral features of red tide from GOCI images [27]. Guo et al. constructed an automatic SAR image detection method for green algae in the Yellow Sea based on the deep convolutional U-net architecture [28]. Cui et al. proposed the SRSe-Net to extract large-scale green tides based on U-Net structure and a dense connection mechanism. SRSe-Net has the ability to extract the green tides from the low-resolution MODIS image by the feature mapping learned from the GF1-WFV image domain [29]. Gao et al. proposed the AlgaeNet model based on U-Net to extract floating *Ulva prolifera* from MODIS and SAR images [30].

In the study of computer vision, object detection is the task to locate and classify objects of interest in images. Semantic segmentation is a form of pixel-level prediction to classify each pixel according to the same category, and it only segments targets in different categories and cannot distinguish each individual target in the same category. The instance segmentation methods cannot only locate the corresponding bounding box of target in different categories, but also classify each object at pixel level in the same category. Therefore, the meaning of 'instance' is that the network has the ability to distinguish each individual target in the same category. It is more challenging for instance segmentation as it includes the tasks of object detection and semantic segmentation [31]. The technology of instance segmentation has been widely applied in the fields of autonomous driving, medical image analysis, and video surveillance [32–34]. Mask R-CNN [35] is one of the most widely applied instance segmentation algorithms today, developed from the object detection network, Faster R-CNN [36]. Mask R-CNN adds a semantic segmentation branch for predicting each region of interest (ROI) to the object classification and regression branches, effectively detecting target objects and generating high-quality segmentation masks for each instance. In the Mask R-CNN framework, the final output masks are determined by the object classification branch's highest confidence. However, these predicted masks are not optimal as the correlation between the masks and the confidence is very low. To solve the problem, Mask Scoring R-CNN [37] designed Mask IoU, a mask evaluation strategy, to measure the distance between the real mask and the predicted mask. CenterMask [38] is an anchor-free instance segmentation framework that can simultaneously achieve the target at real-time speed and high accuracy. CenterMask introduced a new spatial attention-guided mask (SAG-Mask) branch to FCOS [39], a one-stage object detection method. SAG-Mask branch could obtain the object bounding boxes to predict segmentation masks on each detected area. The existing floating macroalgae detection and segmentation algorithms have poor portability, and have strict requirements on the observation environment, so it is difficult to apply them in a large range and for a long time. Video surveillance on aboard ships and around the coastline has the advantage of high-definition resolution, real-time image transmission and low cost, so it can be regarded as a useful supplement to remote sensing satellites and SAR, as shown in Figure 1.

In this paper, inspired by the successful application of CenterMask in the field of image recognition and segmentation, we propose a new instance segmentation framework named AlgaeMask for the purpose of floating algae detection, using the surveillance images captured from the on-site imaging such aboard ships and around the coastline. The AlgaeMask integrates the boundingbox detection and edge area segmentation of the floating algae (*Ulva prolifera* and *Sargassum*) simultaneously into a unified architecture, which is applied to practical scenarios effectively.

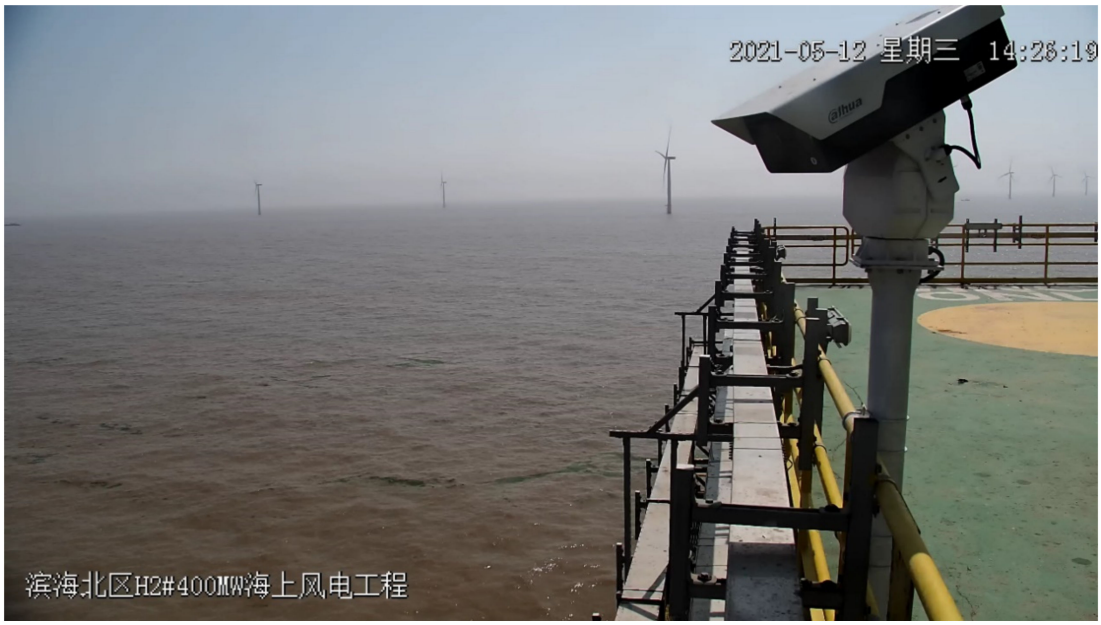

Figure 1. Video surveillance on the ship (These texts not in English are generated from the cameras, which means the date, the camera manufacturer logo and location names of each camera).

The main contributions of our proposed AlgaeMask can be summarized as follows:

(1) A new feature extraction module based on One-Shot Aggregation Version (OSA) and dual-attention mechanism was proposed. By integrating the position attention and channel attention in OSA architecture, the long-range position and contextual information of floating algae can be effectively extracted.
(2) Considering the feature of floating algae at different scales, the multi-scale fusion module is introduced to capture the inter-scale correlation of the feature pyramid, which can effectively capture the invariant features of floating algae.
(3) We evaluate the performance of AlgaeMask and other instance segmentation methods on different scenes. The results show that AlgaeMask can achieve state-of-the-art performance in floating algae detection.

The rest of the paper is organized as follows. Section 2 describes AlgaeMask applied in this paper in detail. Section 3 introduces our experimental results and analysis, including the related dataset, evaluation metrics, qualitative and quantitative performance comparisons, and ablation study. Finally, the conclusions are summarized in Section 4.

2. Methods

As shown in Figure 2, the AlgaeMask consists of a feature extraction module, multi-scale fusion module, and feature decoder module. In the feature extraction module, based on OSA-V2 in CenteMask, the OSA-V3 is proposed to capture the spatial and channel inter-dependencies of floating algae features better by introducing the dual-attention mechanism. In addition, we replace the OSA-V2 block with the original OSA block at Stage 1 and Stage 2. The multi-scale fusion module extracts the scale-invariance features of floating algae by Scale-Equalizing Pyramid Convolution (SEPC) block and Feature Pyramid Network (FPN) block. In the feature decoder module, the Fully Convolutional One-Stage Object Detection (FCOS) head is used to detect the object bounding box at different scales by inputting the output of the multi-scale fusion module. Finally, the segmentation head is performed to obtain the segmentation area of objects in each bounding box.

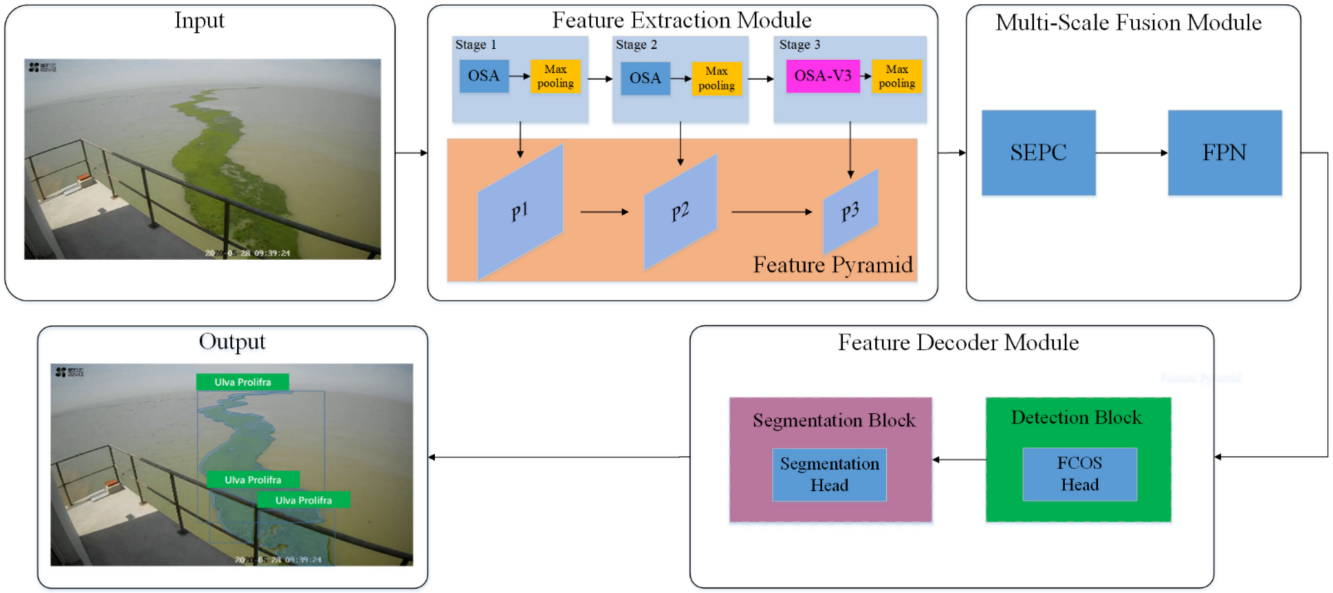

Figure 2. Framework of our proposed AlgaeMask.

2.1. Feature Extraction Module

The environment of floating algae detection is applied is complex and changeable. Therefore, it is necessary for a feature extraction module to have the strong feature extraction and anti-inference. Meanwhile, as the area monitored by one camera is limited, real-time detection on multiple cameras is required for floating algae detection. To deal with this real situation, minimum possible computation costs are desired for our detection model.

Compared to traditional backbone framework such as ResNet, DenseNet, or HRNet, OSA is a computation and energy efficient backbone network, which can capture different receptive fields efficiently. However, due to the lack of attention mechanism, OSA cannot extract long dependencies during the feature extraction phase. In order to enhance the performance of OSA, CenterMask proposes the OSA-V2 block which introduced a channel attention block called effective squeeze-excitation (eSE) [38].

In the floating algae detection, we find that the eSE is only focused on the channel dependencies and ignores the position dependencies between different targets. To handle the insufficiency, we propose a new OSA-V3 block to improve detection accuracy by introducing a dual-attention mechanism, which is composed of channel attention and position attention. The channel attention block can extract the feature interdependency between different channels. The position attention block has the ability to capture the spatial location interdependency under the current scale to help the OSA-V3 block to effectively limit the location of floating algae's regions only above the sea surface and reduce false detection. The architecture of the OSA-V3 block is shown in Figure 3. In the detection of floating algae, it is necessary to input high-resolution images as the number of small targets accounts for the highest proportion. However, the computation cost of the attention mechanism mainly depends on the resolution of images. Therefore, different from the architecture of CenterMask network, we only integrate the OSA-V3 block in Stage 3. In addition, we will also replace the OSA-V2 block with the original OSA block in Stage 1 and Stage 2. We will further discuss how using the origin OSA block in Stage 1 and Stage 2 can not only reduce the model complexity and computation cost, but also demonstrate better performance than the OSA-V2 block.

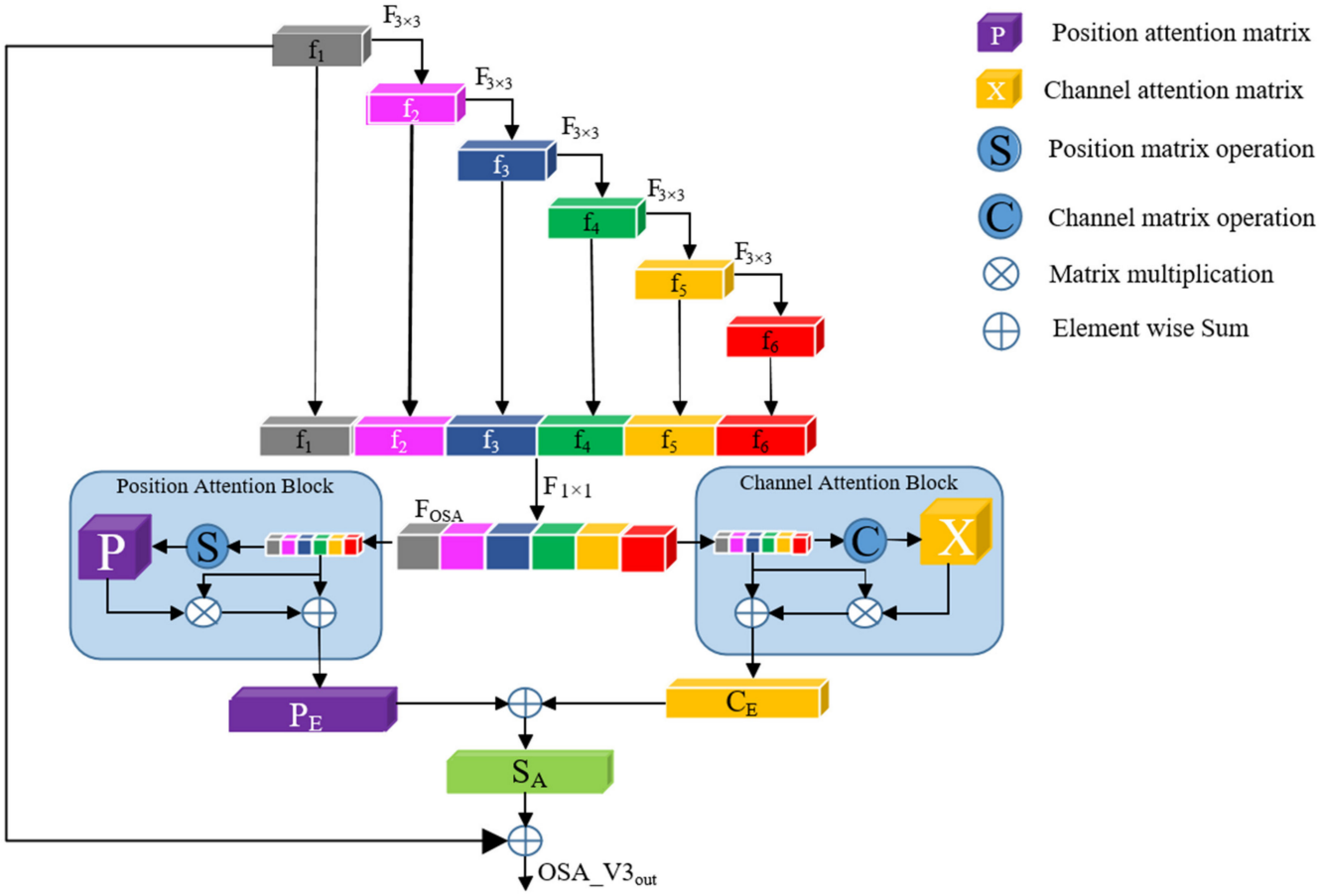

Figure 3. Architecture of OSA-V3 block.

(1) OSA Block

Given the input feature map $f_1 \in R^{C\times H\times W}$, we first use the convolutional operation with kernel size 3 to get $f_2 \sim f_6 \in R^{C\times H\times W}$ in turn. Then we perform the convolutional operation with kernel size 1 on the concatenation result of $f_1 \sim f_6$ to obtain the fusion result of F_{OSA}. The calculation of this process is as follows in (1) and (2):

$$f_i = \text{conv2d}(f_{i-1}),\ i = 2,3,4,5,6 \tag{1}$$

$$F_{OSA} = \text{conv2d}(\text{cat}(f_1, f_2, f_3, f_4, f_5, f_6)) \tag{2}$$

where the conv2d is convolutional operation and cat denotes the concatenate operation on channel dimension.

The output of OSA F_{OSA} is fed into the position attention block and channel attention block in parallel.

(2) Channel Attention Block

In channel attention block, firstly, we reshape the F_{OSA} to $f\_C_{reshape} \in R^{C\times N}$, $N = H \times W$ and transpose $f\_C_{reshape}$ to $f\_C_{transpose} \in R^{N\times C}$.

Secondly, we perform a matrix multiplication between $f_{reshape}$ and $f_{transpose}$ to obtain the channel attention map $X \in R^{C\times C}$.

$$X_{ji} = \frac{\exp\left(f\_C^i_{reshape} \cdot f\_C^j_{transpose}\right)}{\sum_{i=1}^{C} \exp\left(f\_C^i_{reshape} \cdot f\_C^j_{transpose}\right)} \tag{3}$$

where X_{ij} represents the i^{th} channel's impact on the j^{th} channel, exp is exponential operation.

Thirdly, we multiply the attention map X and $f\_C_{reshape}$ and reshape to $R^{C\times H\times W}$ and then multiply with a scale parameter β to obtain the result $f\_C_{att} \in R^{C\times H\times W}$.

Finally, an element-wise sum operation is performed between f_C_{att} and F_{OSA} to obtain the channel attention result C_E.

$$f_{Catt} = \beta \sum_{i=1}^{C} X_{ji} \cdot f_{C_{reshape}}^{i} \tag{4}$$

$$C_E = f\_C_{att} + F_{OSA} \tag{5}$$

(3) Position Attention Block

In position attention block, we first use the convolution operation to generate three new feature maps $f\_P_B \in R^{C\times H\times W}$, $f\_P_C \in R^{C\times H\times W}$, and $f\_P_D \in R^{C\times H\times W}$ and reshape f_P_B, f_P_C, and f_P_D to $R^{C\times N}$, $N = H \times W$, respectively.

Secondly, transpose the f_P_B and $R^{N\times C}$, perform a matrix multiplication between f_P_B and f_P_C, and use softmax operation to calculate the position attention map $P \in R^{N\times N}$.

$$P_{ji} = \frac{\exp\left(f\_P_B^i \cdot f\_P_C^j\right)}{\sum_{i=1}^{C} \exp\left(f\_P_B^i \cdot f\_P_C^j\right)} \tag{6}$$

where P_{ji} represents the i^{th} position's impact on the j^{th} position.

Thirdly, perform a matrix multiplication between f_P_D and P. Then, reshape the result to $R^{C\times H\times W}$ and multiply with a scale parameter α to obtain to obtain the result $f\_P_{att} \in R^{C\times H\times W}$.

Finally, we perform an element-wise sum operation to obtain the final position attention result P_E.

$$f_{Patt} = \alpha \sum_{i=1}^{C} P_{ji} \cdot f_{P_D}^{i} \tag{7}$$

$$P_E = f\_P_{att} + F_{OSA} \tag{8}$$

After the calculation of the position and channel attention, we perform an element-wise sum operation between P_E and C_E to obtain the total attention result S_A. Then, we use the residual connection between the input feature map and S_A to obtain the output of the feature extraction module.

$$S_A = P_E + C_E \tag{9}$$

$$OSA\_V3_{out} = f_1 + S_A \tag{10}$$

2.2. Muti-Scale Fusion Module

By the investigation of floating algae detection applications, the *Ulva prolifera* and *Sargassum* are always displayed in different sizes in the video due to the difference of viewing angles, focal lengths, and distance of the cameras. Therefore, it is very important for our model to be capable of extracting floating algae at different scales, as shown in Figure 4.

Figure 4. Samples of floating algae at different scales.

The feature pyramid network (FPN) is commonly adopted to deal with object detection at different scales by the instance segmentation models such as Mask-RCNN and CenterMask. However, FPN cannot utilize the inter-level correlation in the feature pyramid efficiently. In this paper, we present a multi-scale fusion module (MSF) consisting of the SEPC block and FPN block. The SEPC block can help the MSF module to improve the ability of the scale-invariant feature extraction for floating algae in both spatial and scale dimension. The architecture of MSF module is shown as Figure 5.

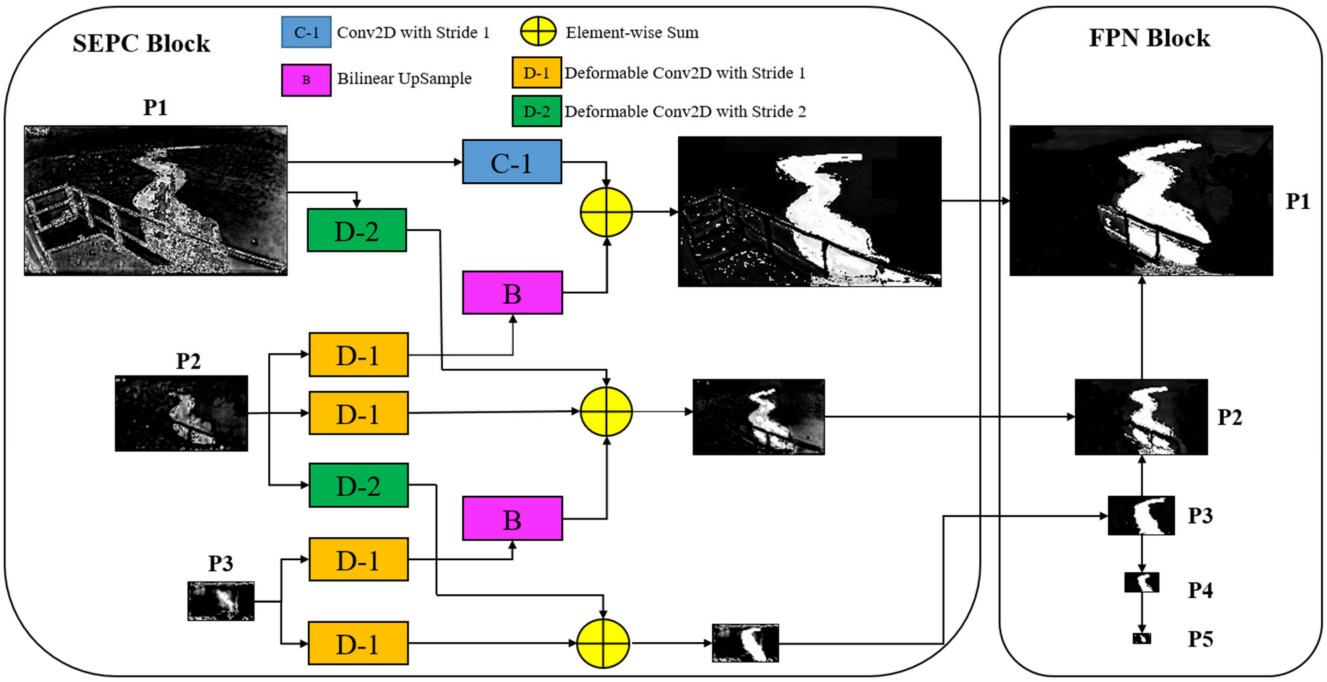

Figure 5. Architecture of the multi-scale fusion module.

Comparing the extracted features in Figure 5, it is obvious that the SEPC block can improve the ability of extracting the robust scale-invariant features of floating algae. By introducing deformable convolution, the SEPC block can compromise the blurring effect of features under different scales.

In the feature extraction module, we can get the feature map with different scales $p_1 \epsilon R^{C_1 \times \frac{H}{8} \times \frac{w}{8}}$, $p_2 \epsilon R^{C_2 \times \frac{H}{16} \times \frac{w}{16}}$, and $p_3 \epsilon R^{C_3 \times \frac{H}{32} \times \frac{w}{32}}$. We will take the p_2 as an example to

illustrate the calculation process of SPEC block. The processes of p_1 and p_3 are same as p_2 and the difference in calculation of p_1 is only that we use the conv2d operation instead of the Deform conv2d.

The calculation process of p_2 can be summarized in (11)~(14).

$$p_1^{down} = \text{Deform2d}(p_1) \in R^{C \times \frac{H}{16} \times \frac{w}{16}} \tag{11}$$

$$p_3^{up} = \text{UpSample}\big(\text{Deform2d}(p_3)\big) \in R^{C \times \frac{H}{16} \times \frac{w}{16}} \tag{12}$$

$$p_2^{deform} = \text{Deform2d}(p_2) \in R^{C \times \frac{H}{16} \times \frac{w}{16}} \tag{13}$$

$$p_2^{SPEC} = p_1^{down} + p_2^{deform} + p_3^{up},\ p_2^{SPEC} \in R^{C \times \frac{H}{16} \times \frac{w}{16}} \tag{14}$$

where Deform2d represents the deformable convolutional operation, p_1^{down} denotes the output of Deform2d with stride 2 on p_1, p_3^{up} represents the output of the up-sample operation on p_3, and the p_2^{deform} is the result of Deform2d on p_2.

Then, the output of SPEC block $p_1 \sim p_3$ is used as the input of FPN to obtain the feature maps with different scales $p_1 \epsilon R^{C_1 \times \frac{H}{8} \times \frac{w}{8}}$, $p_2 \epsilon R^{C_2 \times \frac{H}{16} \times \frac{w}{16}}$, $p_3 \epsilon R^{C_3 \times \frac{H}{32} \times \frac{w}{32}}$, $p_4 \epsilon R^{C_4 \times \frac{H}{64} \times \frac{w}{64}}$, and $p_5 \epsilon R^{C_5 \times \frac{H}{128} \times \frac{w}{128}}$. The calculations of the above process are formulated in (15)~(19).

$$p_3 = \text{conv2d}(p3) \tag{15}$$

$$p_2 = \text{UpSample}(p3) + \text{conv2d}(p2) \tag{16}$$

$$p_1 = \text{UpSample}(p2) + \text{conv2d}(p2) \tag{17}$$

$$p_4 = \text{conv2d}(p3) \tag{18}$$

$$p_5 = \text{conv2d}(p4) \tag{19}$$

2.3. Feature Decoder Module

2.3.1. Detection Block

The traditional detection networks will predict the class category, center point offset, and scaling of width and height of these anchors by the use of the pre-defined anchor boxes. However, the definition of anchor boxes depends on a lot of prior knowledge and may be not reasonable.

In [39], an anchor free framework named fully convolutional one-stage (FCOS) object detection is proposed. By directly predicting the distance up, down, left, and right of each pixel, the FCOS network can not only greatly reduce the complexity of time and space in the training phase, but it can also improve detection accuracy in the testing phase.

We take the outputs of FPN module $P_1 \sim P_5$ as the inputs of FCOS head, consisting of classification head and regression head. These two heads include convolution, group normalization and ReLU operations, as shown in Figure 6.

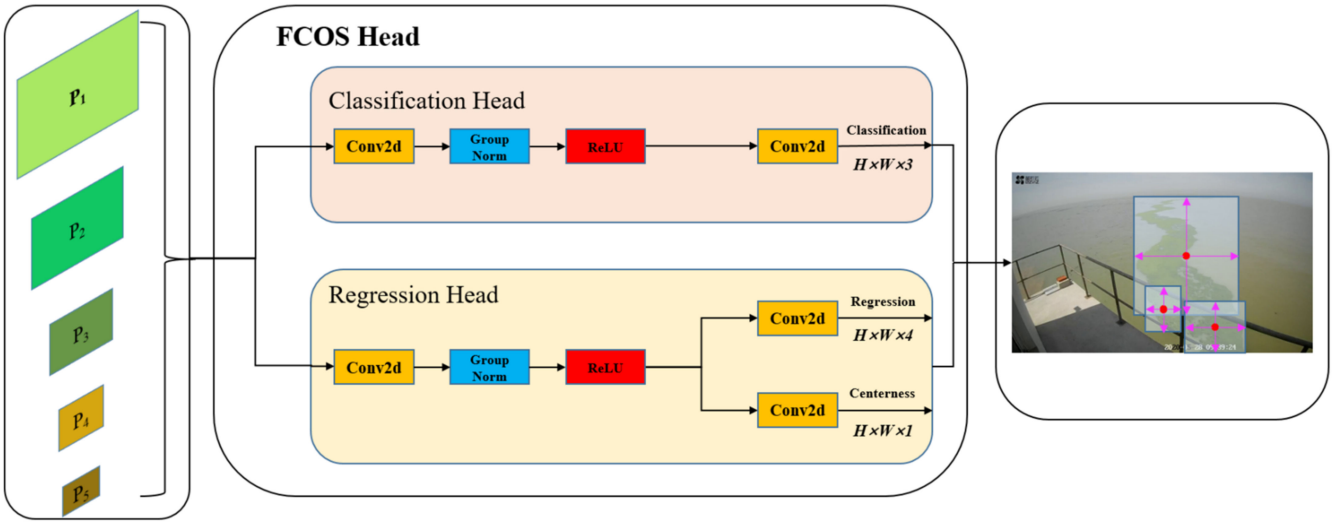

Figure 6. Overview of the detection block based on FCOS head.

In the classification head, the corresponding classification label will be predicted for each position in current feature map. In this paper, our model will predict the following three categories: *Ulva prolifera*, *Sargassum*, and disturbances. In order to reduce the interference of irrelevant objects—such as ships, sea surface, or seabirds—in the environment we define them as a category called 'Disturbances', shown as Figure 7.

Figure 7. Samples of the Disturbances category.

In the regression head, two sub-branches are defined to predict the four boundary distance parameters and one center distance parameter. Assume that the coordinates of a point on the original image is (o_x, o_y), and the scale between the current feature map and the original image is defined as s_i. Then, the relation between the regression branch prediction and the original image position can be summarized in (20)~(23).

$$x\_min_i = o_x - l \times s_i \tag{20}$$

$$y\_min_i = o_y - t \times s_i \tag{21}$$

$$x\_max_i = o_x + r \times s_i \tag{22}$$

$$y\_max_i = o_y + b \times s_i \tag{23}$$

where the x_min_i, y_min_i, x_max_i, and y_max_i are the coordinates of the upper and lower left and right corners of the object bounding box. l, t, r, and b represent the distance to left, upper, right, and bottom of object, respectively.

In the regression head, a parameter named as centerness $\in (0, 1)$ will be predicted that can measure the distance to the object center and the higher value means higher proximity to the object center.

2.3.2. Segmentation Block

The Mask R-CNN network [35] uses the ROIAlign method to realize the alignment of bounding boxes at different scales and it is improved in CenterMask to enhance the detection accuracy of small targets. The calculation of ROIAlign in Mask R-CNN and CenterMask are summarized as follows.

$$\text{MaskRCNN}_{\text{ROIAlign}} = \lfloor k_0 + \log_2(\frac{\sqrt{w \times h}}{224}) \rfloor \tag{24}$$

$$\text{CenterMask}_{\text{ROIAlign}} = \lceil k_{max} - \log_2 \frac{F_{input}}{F_{RoI}} \rceil \tag{25}$$

where the values of k_0 and k_{max} are assigned as 4 and 5. The width and height of each bounding box are denoted as w and h. F_{input} represents the pixel area of input image, and F_{RoI} represents the pixel area of bounding box. Without using the constant value 224 in Mask R-CNN, CenterMask can assign the ROIAlign pooling scale adaptively by the ratio calculation of F_{input}/F_{RoI}, and thus can improve the detection accuracy of floating algaes with different scales.

After the operation of ROIAlign block, we will get the feature maps with same resolution under the inputs at different scales. Then, these feature maps will be fed into the segmentation block to achieve the mask area in the ROI bounding boxes.

As shown in Figure 8, the ROI bounding box in feature maps $p_1 \sim p_5$ with different resolutions will be unified into a fixed size 14×14 after the ROIAlign operation. We fed these ROI features into four convolution layers sequentially.

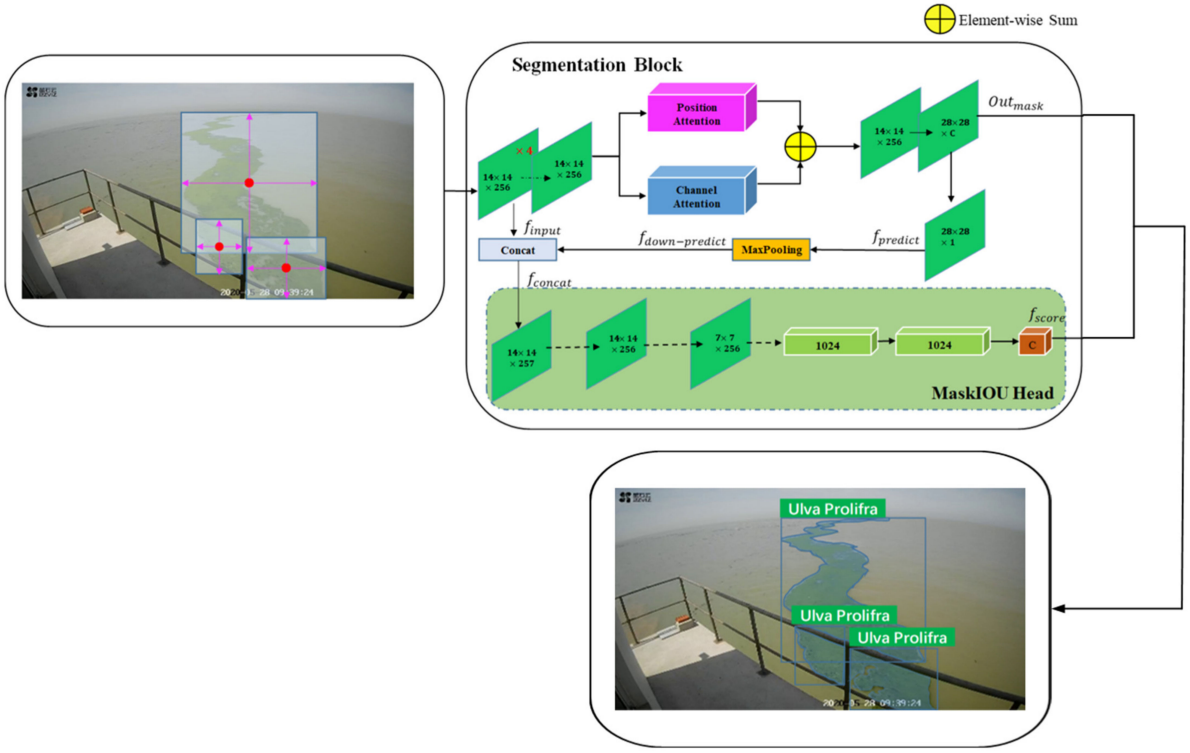

Figure 8. Flowchart of the segmentation block.

Then, a 2 × 2 de-convolution operation is performed to upsample the feature map to a resolution of 28 × 28. After that, a 1 × 1 convolution is used to predict the class-specific output.

Considering floating algae detection is a multi-class instance segmentation task; however, the score of mask segmentation is shared with the box-level classification result in FCOS head, hardly to measure the mask quality and completeness of instance segmentation.

We introduce the MaskIoU block in our segmentation pipeline to learn a score value for each mask output instead of sharing the box classification confidence. The process of the MaskIoU block can be summarized as follows.

(1) A convolutional operation is performed on the output of mask Out_{mask} to get the prediction mask feature map $f_{predict} \in R^{1\times28\times28}$. $f_{predict}$ is fed into a max-pooling block to get a downsampling result $f_{down-predict} \in R^{1\times14\times14}$.
(2) The input feature map $f_{input} \in R^{256\times14\times14}$ and $f_{down-predict}$ are concatenated to obtain the fusion result $f_{concat} \in R^{257\times14\times14}$.
(3) Four convolution layers (kernel = 3 and stride = 1, and the stride of final convolution is 2 for downsampling the feature map to 7 × 7) and two fully connected layers (outputs with 1024 channels) are performed sequentially on f_{concat} to obtain the result $f_{fc} \in R^{1024\times1\times1}$.
(4) Feed the f_{fc} into task-specific fully connected layers to get the classification score of the current mask $f_{score} \in R^1$.

During the training phase, a binary operation with threshold 0.5 is performed on the predicted mask Out_{mask} to convert the two-dimensional probability image into the binary image f_{binay}. Then, we use the L2 loss between the f_{binay} and the ground truth label image to calculate the mask score loss. During the testing phase, we multiply the classification score in FCOS classification head with the mask classification score in segmentation head as the final object confidence value.

2.3.3. Loss Function

Our loss function consists of the following five parts.

$$loss_{total} = w_1 \times loss_{fcos\_cls} + w_2 \times loss_{fcos\_center} + w_3 \times loss_{fcos\_box} + w_4 \times loss_{seg\_mask} + w_5 \times loss_{seg\_maskIOU} \quad (26)$$

where $loss_{fcos\_cls}$ is the classification loss in FCOS classification head, $loss_{fcos\_center}$ and $loss_{fcos\_box}$ are centerness loss and box regression loss in FCOS regression head, $loss_{seg\_mask}$ is the average binary cross-entropy loss of segmentation mask in segmentation head, $loss_{seg\_maskIOU}$ is the L2 loss in MaskIoU head. The $w_1 \sim w_5$ represent the weight values of each loss, and the values in this paper are 0.5, 1.0, 1.0, 1.0, and 0.5 respectively.

3. Experimental Results and Analysis

3.1. Dataset

According to the location of the floating algae outbreak area in the East China Sea over these years, the data in this paper are collected from the surveillance video captured by seven-way cameras in Nantong and Yancheng of the Jiangsu sea area, from 2020 to 2022. These camera positions are shown in Table 1.

Table 1. Camera position for surveillance adopted in our paper.

Camera Name	Longitude (°N)	Latitude (°E)
Binhai North District H2#400MW	34.490893	120.33805
SPIC Binhai South H3#300MW	34.314415	120.60239
Jiangsu Rudong H5#	32.709716	121.72382
Jiangsu Rudong H14#	32.811922	121.49177
Three Gorges New Energy Jiangsu Dafeng 300MW	33.390532	121.18879
Huaneng Jiangsu Dafeng 300MW	33.170058	121.41954
Dafeng Wharf	33.224975	120.86676

We construct our dataset from 3600 images with the resolution of 1920 × 1080 from the videos mentioned above. The dataset is divided into three categories—*Ulva prolifera*, *Sargassum*, and Disturbances. We adopt 3000 images as a training set and 600 images as a testing set.

3.2. Evaluation Metrics

Considering the AlgaeMask is a type of instance segmentation network, we choose the following metrics to evaluate our model: (1) the mask average precision (AP_{mask}); (2) the box average precision (AP_{box}); (3) the mask average recall (AR_{mask}); (4) the box average recall (AR_{box}). The average precision and average recall can be formulated in (27) and (28).

$$AP = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{27}$$

$$AR = \frac{2}{n}\sum_{i=1}^{n} R_i \tag{28}$$

where the n represents the number of samples, P_i denotes the precision value of the i^{th} sample, and R_i represents the recall value of the i^{th} sample. The calculations of P_i and R_i are as follows

$$P = \frac{TP}{TP + FP} \tag{29}$$

$$R = \frac{TP}{TP + FN} \tag{30}$$

where TP, FP, and FN denote the number of true positives, false positives and false negatives, respectively.

For segmentation evaluation, we compute the TP, FP, and FN by comparing the predicted mask image with the ground truth label image. For bounding box evaluation, we judge whether the IoU value between the predicted box and the label box is greater than the $IoU_{threshold}$ and compute the TP, FP, and FN. The IoU could be calculated in (31).

$$IoU = \frac{area\left(box_{pre} \bigcap box_{gt}\right)}{area\left(box_{pre} \bigcup box_{gt}\right)} \tag{31}$$

where box_{pre} and box_{gt} are corresponding to the predicted box and the ground truth box, $area\left(box_{pre} \bigcap box_{gt}\right)$ represents the area of intersection between box_{pre} and box_{gt}, $area\left(box_{pre} \bigcup box_{gt}\right)$ denotes the area of union between box_{pre} and box_{gt}.

In order to further evaluate the performance of different methods for targets with different sizes, we also provide the following metrics: AP^S, AP^M, AP^L, AR^S, AR^M, and AR^L. We mentioned the larger value of the metrics and the better performance of the network above. The meaning of these metrics are as follows:

- AP^S, AR^S: the average precision or recall for small objects, which the pixel area of object is less than 32^2.

- AP^M, AR^M: the average precision or recall for medium objects, which the pixel area of object is between 32^2 and 96^2.
- AP^L, AR^L: the average precision or recall for large objects, which the pixel area of object is greater than 96^2.

3.3. Experimental Setups

The AlgaeMask is implemented in PyTorch V1.9.1(USA), Detectron2 V0.3(USA), CUDA(USA), and cuDNN V11.4(USA) on four NVIDIA TITAN RTX GPU with 24GB of memory respectively.

In the training phase, in order to enhance the anti-interference ability of the outdoor detection environment, we adopt the data augmentation methods such as random cropping, random brightness, random occlusion, and contrast variation. The batch size and iteration are 8 and 60,000 respectively. The resolution of images captured by surveillance camera is 1920×1080, but we use a resolution of 960×512 for training and testing in this paper.

3.4. Evaluation of Model Performance

In this section, we will compare our proposed AlgaeMask with other instance segmentation methods, including Mask R-CNN, Mask Scoring R-CNN, and CenterMask. The performance comparison is presented in Tables 2 and 3. The results show that our proposed network can reach the best performance on the precision dimension of box detection and mask segmentation. Specifically, compared with Mask R-CNN, Mask Scoring R-CNN, and CenterMask, our AlgaeMask model achieves 28.59%, 22.26%, and 15.13% improvement on AP_{mask} and 43.96%, 37.52%, and 24.24% improvement on AP_{box} in Table 2; 26.15%, 24.35%, and 15.65% improvement on AR_{mask} and 32.59%, 25.68%, and 3.26% improvement on AR_{box} in Table 3, respectively. Additionally, we also visualize the predictions under different scenes in Figure 9.

Table 2. Comparison of experimental results on average precision (AP) by different methods.

Methods	Backbone	AP_{mask}	AP^S_{mask}	AP^M_{mask}	AP^L_{mask}	AP_{box}	AP^S_{box}	AP^M_{box}	AP^L_{box}
Mask R-CNN [35]	ResNet-101	15.63	15.13	39.67	24.87	20.17	18.91	41.23	13.14
Mask Scoring R-CNN [37]	ResNet-101	21.96	19.34	41.34	58.78	26.61	22.94	38.78	60.89
CenterMask [38]	OSA-V2	29.13	15.67	49.67	67.89	39.89	34.74	60.52	64.97
AlgaeMask	OSA+OSA-V3	44.22	36.35	60.54	71.21	48.13	60.27	65.36	68.67

Table 3. Comparison of experimental results on average recall (AR) by different methods.

Methods	Backbone	AR_{mask}	AR^S_{mask}	AR^M_{mask}	AR^L_{mask}	AR_{box}	AR^S_{box}	AR^M_{box}	AR^L_{box}
Mask R-CNN	ResNet-101	25.17	21.81	42.39	31.73	37.88	34.82	52.34	35.15
Mask Scoring R-CNN	ResNet-101	26.97	27.68	45.16	57.98	44.79	37.84	49.63	59.82
CenterMask	OSA-V2	35.67	32.71	68.96	65.46	67.21	52.79	54.17	60.12
AlgaeMask	OSA+OSA-V3	51.32	48.90	73.51	68.33	70.47	68.43	63.32	68.91

Figure 9. *Cont.*

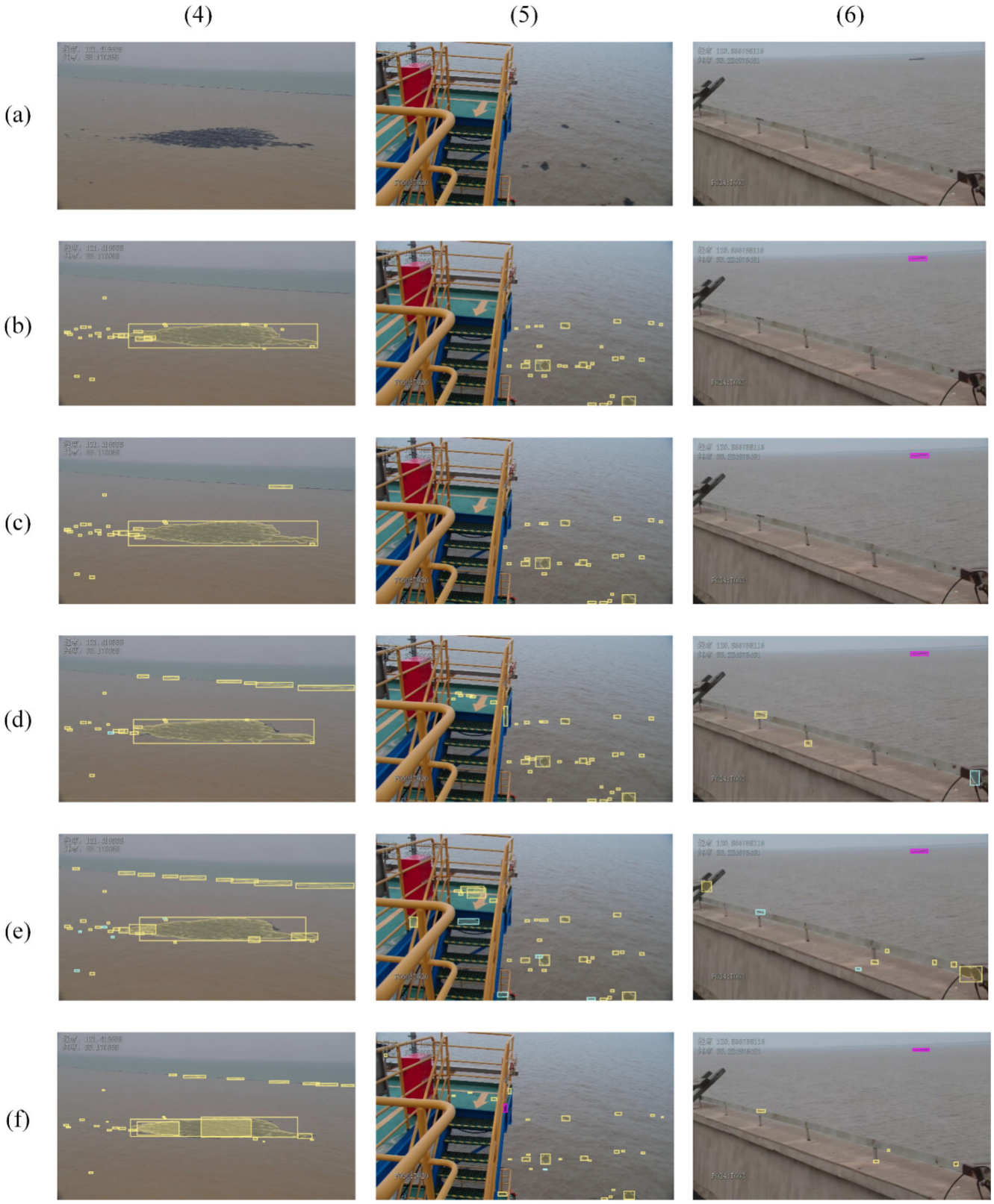

Figure 9. Visualization of different models on testing samples. (**a**) Source image. (**b**) Ground truth. (**c**) AlgaeMask model. (**d**) CenterMask model. (**e**) Mask Scoring R-CNN model. (**f**) Mask R-CNN model. Light green color represents the category of *Ulva prolifera*, the yellow color represents the category of *Sargassum*, and the pink color represents the category of Disturbances.

In testing samples (1), (2), and (3), it is obvious that the networks without the SEPC block—such as Mask R-CNN, Mask Scoring R-CNN, and CenterMask—more readily miss detection. Meanwhile, by introducing the SEPC block to our AlgaeMask network, the success rate has been greatly improved.

In testing samples (4), (5), and (6), we can find out that our proposed model has the minimal false detection rate on these complex scenes. It means that by combing the channel attention and position attention block in our feature extraction phase, the features of floating algae and interference in the environment can be effectively distinguished, which can enhance the feature extraction ability and anti-interference ability of our AlgaeMask network.

Specific to objects in different sizes from Tables 2 and 3, the AlgaeMask can also exhibit better performance on AP and AR in detecting the floating algae. In testing sample (2), a lot of small *Ulva prolifera* are missed detection in CenterMask, Mask Scoring R-CNN and Mask R-CNN. For medium algaes, except for our method, other models demonstrate a large number of false detections in samples (1), (3), and (5). In sample (4), from the perspective of segmentation accuracy and integrity of large algal blooms, our AlgaeMask can achieve the best performance.

Additionally, the performance comparison of different networks for all categories are shown in Tables 4 and 5. Generally, the Disturbances category obtains the best performance in all categories and the *Ulva prolifera* category has the worst performance in all networks. According to testing samples (2), (4), (5), and (6), due to the lack of position attention mechanism and SEPC module, the CenterMask detects a lot of false land objects into *Ulva prolifera* or *Sargassum* and misses a lot of obvious small *Ulva prolifera*, and Mask R-CNN and Mask Scoring R-CNN also have the same problems. Therefore, compared with CenterMask on AP, our proposed network achieves 14.06% and 16.35% improvement for the category of *Ulva prolifera* and *Sargassum* respectively. Meanwhile, on the metric of AR, compared to CenterMask, our proposed AlgaeMask also shows a 13.79% and 6.91% improvement.

Table 4. Comparison of experimental results on AP by different methods for each category.

Methods	Backbone	$AP^{Ulva\ Prolifra}_{mask}$	$AP^{Sargassum}_{mask}$	$AP^{Disturbances}_{mask}$	$AP^{Ulva\ Prolifra}_{box}$	$AP^{Sargassum}_{box}$	$AP^{Disturbances}_{box}$
Mask R-CNN	ResNet-101	9.89	18.71	18.29	11.54	22.19	26.78
Mask Scoring R-CNN	ResNet-101	14.23	22.49	29.16	16.47	33.51	29.85
CenterMask	OSA-V2	17.21	26.43	43.75	24.47	42.56	52.64
AlgaeMask	OSA+OSA-V3	31.27	42.78	58.61	38.26	49.47	56.66

Table 5. Comparison of experimental results on AR by different methods for each category.

Methods	Backbone	$AR^{Ulva\ Prolifra}_{mask}$	$AR^{Sargassum}_{mask}$	$AR^{Disturbances}_{mask}$	$AR^{Ulva\ Prolifra}_{box}$	$AR^{Sargassum}_{box}$	$AR^{Disturbances}_{box}$
Mask R-CNN	ResNet-101	9.14	18.92	47.45	28.62	30.49	54.53
Mask Scoring R-CNN	ResNet-101	12.14	28.31	40.46	22.14	34.67	77.56
CenterMask	OSA-V2	17.37	38.54	51.1	48.29	57.14	96.2
AlgaeMask	OSA+OSA-V3	41.83	43.67	68.46	51.68	63.79	95.94

3.5. Ablation Study

In order to evaluate the performance of the proposed AlgaeMask under different factors and settings, the following ablation studies are conducted.

3.5.1. Impact of Input Resolution

During the training and testing phase, we conduct some experiments on our proposed method with 480 × 256, 960 × 512, and 1920 × 1080 as the input resolution in Table 6.

Table 6. Ablation study on the impact of input resolution on AP and AR.

Resolution	AP_{mask}	AP_{box}	AR_{mask}	AR_{box}	Inference Time (ms)
480 × 256	32.83	35.47	17.34	26.81	96.8
960 × 512	44.22	48.13	51.32	70.47	142.7
1920 × 1080	42.13	42.86	60.39	78.12	380.6

With the increase in input resolution, the value of recall increases obviously because the target floating algaes in small blooms account for the majority ratio in our dataset. When the input resolution is increased from 480 × 256 to 960 × 512, our model could achieve 11.39%, 12.66%, 33.98%, and 43.66% improvement on AP_{mask}, AP_{box}, AR_{mask}, and AR_{box} respectively. It is obvious that the model performance is affected by the resolution of input images greatly. However, the value of precision tends to increase firstly and then decrease with the input resolution increases. When the input resolution is increased from 960 × 512 to 1920 × 1080, the AlgaeMask have 2.09% and 5.27% reduction in AP_{mask} and AP_{box}, and have 9.07% and 7.65% increase in the AR_{mask} and AR_{box}.

This is mainly due to the increase in input resolution helping our model to strengthen its ability to detect small floating algal blooms. However, use of high-resolution images also introduces interference factors in complex environments, which will lead to an increase in false detections. In addition, it will cost more GPU memory resources, and the inference time with an input resolution of 960 × 512 is 2.6× faster than the resolution of 1920 × 1080.

As shown in Table 1, it is necessary for the application of AlgaeMask to process seven channels of video simultaneously. Therefore, the resolution of 960 × 512 is adopted as the input scale of the experiments in this paper, which could meet up the real-time requirements of floating algae detection.

3.5.2. Impact of SEPC Block

In this subsection, we will discuss the impact of SEPC block in our AlgaeMask.

The results are as shown in Figure 10. For the *Sargassum* category, the network with SEPC has 6.56% improvement on AR_{box} and 5.36% improvement on AR_{mask}. For *Ulva prolifera* category, the network with SEPC shows a 3.84% improvement on AP_{box} and 4.35% improvement on AP_{mask}. It is clear that the network with SPEC can attain better performance in all categories.

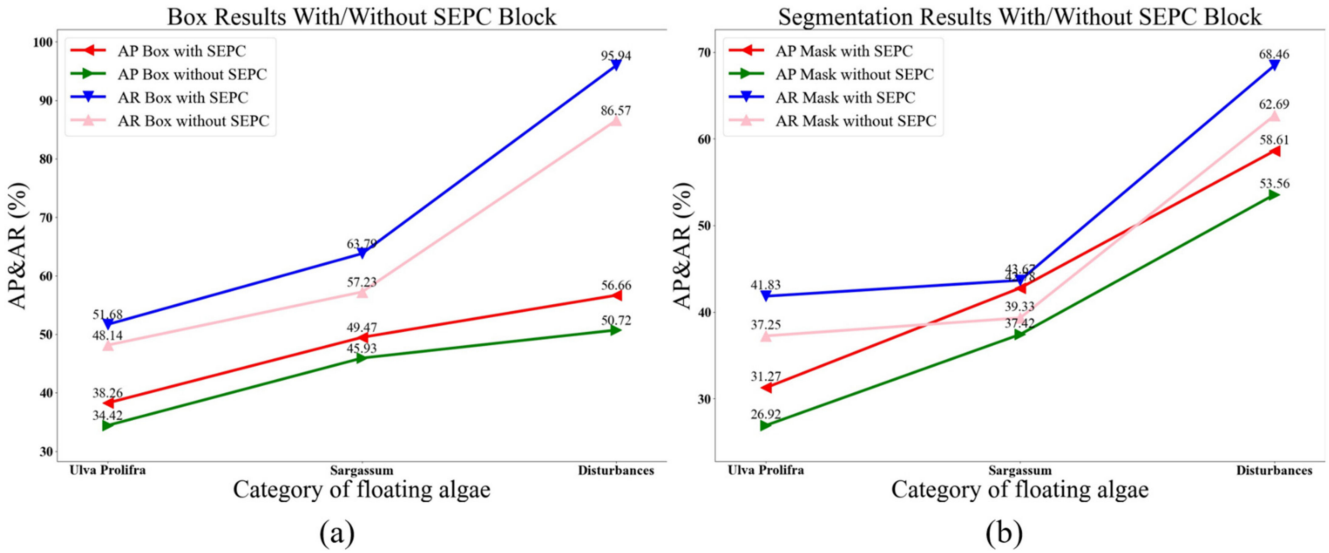

Figure 10. Box and segmentation results with/without SEPC block. (**a**) Box results of AP and AR with/without SEPC block. (**b**) Segmentation results of AP and AR with/without SEPC block.

In practical applications, the camera needs to be installed on the land or other supports, and take pictures from different viewpoints. This leads to a large-scale changes in the image, so that the floating algaes at long distance are usually small and blurry, while the floating algaes at close range are big and clear.

Therefore, the SEPC block—effectively utilizing the invariant features of the floating algaes at different sizes—is very important for our network.

3.5.3. Impact of Dual-Attention Block

In Figure 11, we show the impact of removing the dual-attention block on experimental results. For the *Ulva prolifera* category, the network with the dual-attention block shows a 16.09% and 13.93% improvement on AP_{box} and AP_{mask}. For the *Sargassum* category, the model with the dual-attention block shows a 13.13% improvement on AP_{box} and a 11.52% improvement on AP_{mask}.

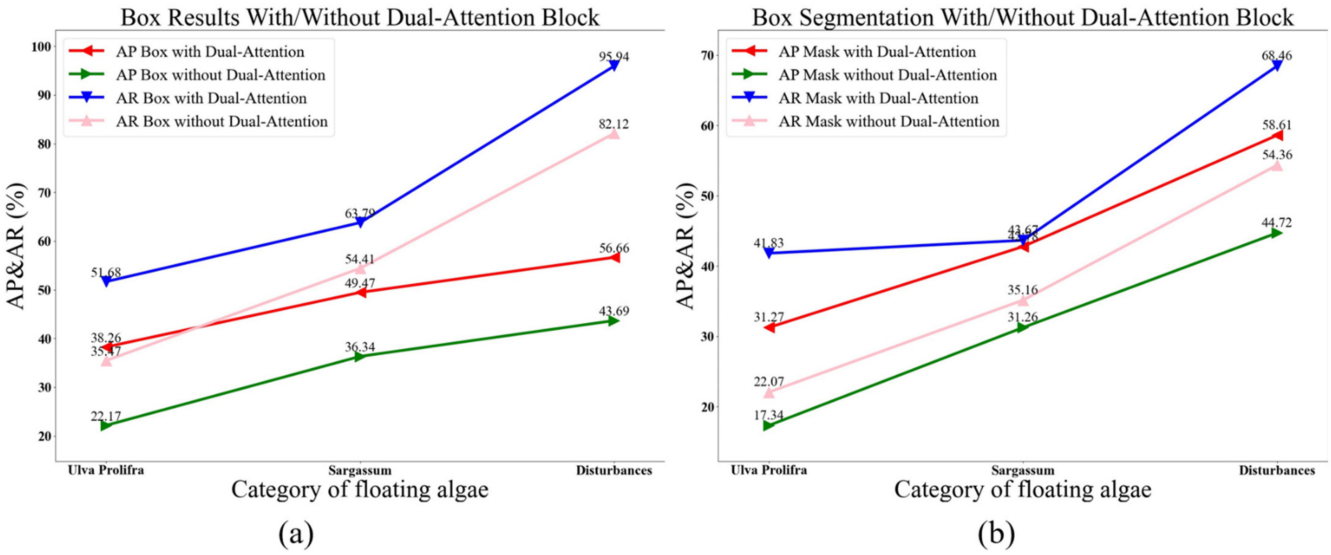

Figure 11. Box and segmentation results with/without dual-attention block. (**a**) Box results of AP and AR with/without dual-attention block. (**b**) Segmentation results of AP and AR with/without dual-attention block.

In contrast with the SEPC block only considering the invariant-features of the floating algae in same category, the dual-attention block can help to learn both the correlation of features at different scales by the channel attention mechanism and the context of features at different spatial positions by position attention mechanism. These abilities of our proposed method are important as the surveillance environment is complex and diverse. By channel attention mechanism, the features of *Ulva prolifera* or *Sargassum* can be extracted effectively in our feature extraction phase. Via position attention mechanism, the similar targets that do not float above the sea surface will be eliminated.

3.5.4. Impact of OSA Block

In this subsection, we will discuss the impact of using OSA and OSA-V2 in Stage 1 and Stage 2 of the feature extraction module.

Based on the CenterMask and AlgaeMask, we replace the OSA-V2 block with original OSA block in Stage 1 and Stage 2 of CenterMask (OSA-V2) to generate the results of CenterMask (OSA). Meanwhile, we replace the OSA block with OSA-V2 block in Stage 1 and Stage 2 of our proposed AlgaeMask (OSA+OSA-V3) to generate the results of AlgaeMask (OSA-V2).

In Table 7, it is obvious that the performance of CenterMask (OSA-V2) and AlgaeMask (OSA-V2) on AP metric are worse than the CenterMask (OSA) and AlgaeMask (OSA+OSA-

V3). Compared with AlgaeMask (OSA-V2) in small targets, AlgaeMask (OSA+OSA-V3) shows a 2.17% and 1.25% increase in AP^S_{mask} and AP^S_{box} respectively.

Table 7. Ablation study on the impact of OSA block on AP.

Methods	Backbone	AP_{mask}	AP^S_{mask}	AP^M_{mask}	AP^L_{mask}	AP_{box}	AP^S_{box}	AP^M_{box}	AP^L_{box}
CenterMask	OSA-V2	29.13	15.67	49.67	67.89	39.89	34.74	60.52	64.97
CenterMask	OSA	30.17	18.23	49.36	67.94	41.22	38.43	59.17	63.85
AlgaeMask	OSA-V2	42.87	34.18	60.19	69.41	47.63	59.02	65.79	67.41
AlgaeMask	OSA+OSA-V3	44.22	36.35	60.54	71.21	48.13	60.27	65.36	68.67

In Figure 12, compared to CenterMask (OSA-V2), the false detection of small targets in CenterMask (OSA) is decreased. Meanwhile, compared to our AlgaeMask (OSA+OSA-V3), the false detection of small targets in AlgaeMask (OSA-V2) are increased.

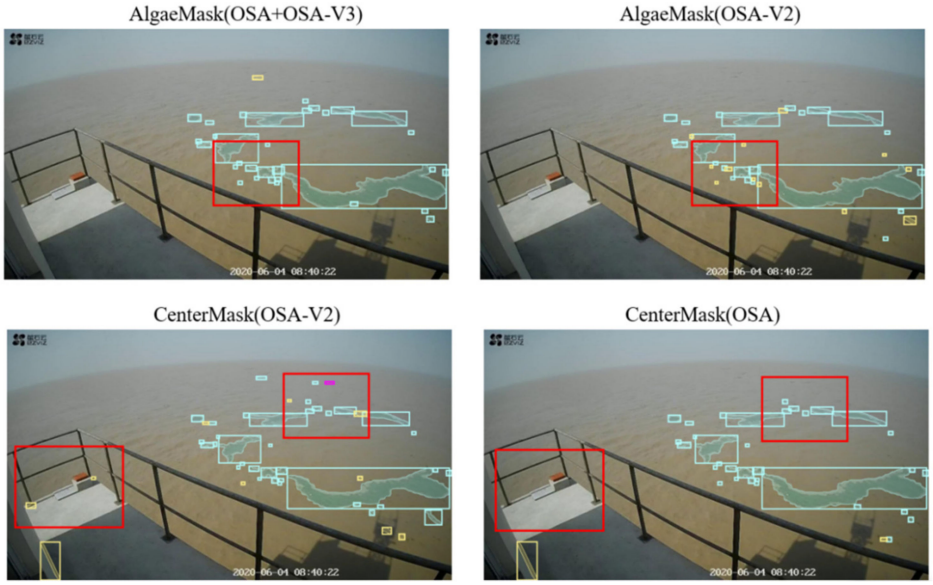

Figure 12. Visualization of AlgaeMask (OSA+OSA-V3), AlgaeMask (OSA-V2), CenterMask (OSA-V2), and CenterMask (OSA). The red solid box represents the main difference between the compared networks.

In floating algae detection, tiny targets have the characteristics of less and simpler feature information. Therefore, these features can be effectively extracted in the first few layers. However, due to the average pooling operation in the eSE block, the features of tiny algaes in different channels may interface with each other, leading to false detection. Meanwhile, some algaes are present at large scale, the features of small algaes may be missed or submerged, resulting in missing detection.

In summary, we choose the original OSA block instead of the OSA-V2 as the backbone of AlgaeMask in Stage 1 and Stage 2 of the feature extraction module.

4. Conclusions

Floating algae detection plays an important role in marine environment monitoring. This study is the first time an instance segmentation method has been applied in floating algae detection. The dataset consisting of multiple marine scenes was built to compare the performance of different instance segmentation networks. The detection precision and time consumption under different input resolutions were also discussed to further verify the actual application capability of our proposed network. In our work, we propose a new instance segmentation framework named AlgaeMask for floating algae detection.

A new feature extraction module based on OSA and a dual-attention mechanism is proposed. The dual-attention block can integrate the position attention and channel attention simultaneously to capture the long-range position and contextual information of floating algae effectively. In floating algae detection, a strong correlation was found between floating algae and interference factors in the environment. Therefore, it is very important to ensure the network can learn the spatial position correlation of targets. The dual-attention mechanism can meet our requirements very well. Meanwhile, to reduce the computation cost of the attention block, we only applied the dual-attention block to the last layer of the feature extraction module.

In addition, considering the features of floating algae at different distances from the camera, a multi-scale fusion module was introduced to capture the inter-scale correlation of the feature pyramid. In the feature decoder module, the FCOS head and segmentation head were introduced to accurately obtain the segmentation area of the algae in every detection bounding box. The extensive experiment results show that the AlgaeMask can achieve better detection accuracy and at a lower time cost in all compared instance segmentation methods to satisfy the real-time needs of floating algae detection.

Due to the limit amount of marine environment data, our model did not take the interferences of bad weather, reflections, and shadows on the ocean surface into account. For future studies, we will further analyze the performance of deep learning methods under conditions of different complex marine scenes (e.g., rain, fog, and reflections) to enhance the robustness of the floating algae detection network.

Author Contributions: Methodology, L.W.; software, F.Z.; validation, L.Z.; formal analysis, Z.Z.; data curation, K.C.; writing—original draft preparation, L.C.; writing—review and editing, Y.Z.; project administration, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This thesis is supported by the Fund of Technology Innovation Center for Ocean Telemetry, Ministry of Natural Resources 006, Tianjin Enterprise Postdoctoral Innovation Project merit funding project TJQYBSH2018025.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xiao, J.; Wang, Z.; Liu, D.; Fu, M.; Yuan, C.; Yan, T. Harmful macroalgal blooms (HMBs) in China's coastal water: Green and golden tides. *Harmful Algae* **2021**, *107*, 102061. [CrossRef] [PubMed]
2. Qiu, Y.H.; Lu, J.B. Advances in the monitoring of Enteromorpha prolifera using remote sensing. *Acta Ecol. Sin* **2015**, *35*, 4977–4985.
3. Qi, L.; Hu, C.; Shang, S. Long-term trend of Ulva prolifera blooms in the western Yellow Sea. *Harmful Algae* **2016**, *58*, 35–44. [CrossRef]
4. Xing, Q.G.; Guo, R.; Wu, L.; An, D.; Cong, M.; Qin, S.; Li, X. High-resolution satellite observations of a new hazard of golden tides caused by floating Sargassum in winter in the Yellow Sea. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1815–1819. [CrossRef]
5. Ma, Y.F.; Wong, K.P.; Tsou, J.Y.; Zhang, Y.Z. Investigating spatial distribution of green-tide in the Yellow Sea in 2021 using combined optical and SAR images. *J. Mar. Sci. Eng.* **2022**, *10*, 127. [CrossRef]
6. Xing, Q.G.; Wu, L.L.; Tian, L.Q.; Cui, T.W.; Li, L.; Kong, F.Z.; Gao, X.L.; Wu, M.Q. Remote sensing of early-stage green tide in the Yellow Sea for floating-macroalgae collecting campaign. *Mar. Pollut. Bull.* **2018**, *133*, 150–156. [CrossRef] [PubMed]
7. Chen, Y.; Sun, D.; Zhang, H.; Wang, S.; Qiu, Z.; He, Y. Remote-sensing monitoring of green tide and its drifting trajectories in Yellow Sea based on observation data of geostationary ocean color imager. *Acta Opt. Sin* **2020**, *40*, 0301001. [CrossRef]
8. Lu, T.; Lu, Y.C.; Hu, L.B.; Jiao, J.N.; Zhang, M.W.; Liu, Y.X. Uncertainty in the optical remote estimation of the biomass of Ulva prolifera macroalgae using MODIS imagery in the Yellow Sea. *Opt. Express* **2019**, *27*, 18620–18627. [CrossRef]
9. Hu, L.B.; Zheng, K.; Hu, C.; He, M.X. On the remote estimation of Ulva prolifera areal coverage and biomass. *Remote Sens. Environ.* **2019**, *223*, 194–207. [CrossRef]
10. Cao, Y.Z.; Wu, Y.; Fang, Z.; Cui, X.; Liang, J.; Song, X. Spatiotemporal patterns and morphological characteristics of Ulva prolifera distribution in the Yellow Sea, China in 2016–2018. *Remote Sens.* **2019**, *11*, 445. [CrossRef]

11. Xing, Q.G.; An, D.; Zheng, X.; Wei, Z.; Wang, X.; Li, L.; Tian, L.; Chen, J. Monitoring seaweed aquaculture in the Yellow Sea with multiple sensors for managing the disaster of macroalgal blooms. *Remote Sens. Environ.* **2019**, *231*, 111279. [CrossRef]
12. Wang, M.Q.; Hu, C.M. Mapping and quantifying Sargassum distribution and coverage in the Central West Atlantic using MODIS observations. *Remote Sens. Environ.* **2016**, *183*, 350–367. [CrossRef]
13. Xu, F.X.; Gao, Z.Q.; Shang, W.T.; Jiang, X.P.; Zheng, X.Y.; Ning, J.C.; Song, D.B. Validation of MODIS-based monitoring for a green tide in the Yellow Sea with the aid of unmanned aerial vehicle. *J. Appl. Remote Sens.* **2017**, *11*, 012007. [CrossRef]
14. Shin, J.S.; Lee, J.S.; Jiang, L.H.; Lim, J.W.; Khim, B.K.; Jo, Y.H. Sargassum Detection Using Machine Learning Models: A Case Study with the First 6 Months of GOCI-II Imagery. *Remote Sens.* **2021**, *13*, 4844. [CrossRef]
15. Cui, T.W.; Li, F.; Wei, Y.H.; Yang, X.; Xiao, Y.F.; Chen, X.Y.; Liu, R.J.; Ma, Y.; Zhang, J. Super-resolution optical mapping of floating macroalgae from geostationary orbit. *Appl. Opt.* **2020**, *59*, C70–C77. [CrossRef]
16. Liang, X.J.; Qin, P.; Xiao, Y.F. Automatic remote sensing detection of floating macroalgae in the yellow and east china seas using extreme learning machine. *J. Coast. Res.* **2019**, *90*, 272–281. [CrossRef]
17. Qiu, Z.F.; Li, Z.; Bila, M.; Wang, S.; Sun, D.; Chen, Y. Automatic method to monitor floating macroalgae blooms based on multilayer perceptron: Case study of Yellow Sea using GOCI images. *Opt. Express* **2018**, *26*, 26810–26829. [CrossRef]
18. Geng, X.M.; Li, P.X.; Yang, J.; Shi, L.; Li, X.M.; Zhao, J.Q. Ulva prolifera detection with dual-polarization GF-3 SAR data. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *502*, 012026. [CrossRef]
19. Shen, H.; Perrie, W.; Liu, Q.; He, Y. Detection of macroalgae blooms by complex SAR imagery. *Mar. Pollut. Bull.* **2014**, *78*, 190–195. [CrossRef]
20. Li, X.F.; Liu, B.; Zheng, G.; Ren, Y.; Zhang, S.; Liu, Y.; Zhang, B.; Wang, F. Deep-learning-based information mining from ocean remote-sensing imagery. *Natl. Sci. Rev.* **2020**, *7*, 1584–1605. [CrossRef]
21. Valentini, N.; Yann, B. Assessment of a smartphone-based camera system for coastal image segmentation and sargassum monitoring. *J. Mar. Sci. Eng.* **2020**, *8*, 23. [CrossRef]
22. Arellano-Verdejo, J.; Lazcano-Hernandez, H.E.; Cabanillas-Teran, N. ERISNet: Deep neural network for Sargassum detection along the coastline of the Mexican Caribbean. *PeerJ* **2019**, *7*, e6842. [CrossRef] [PubMed]
23. Wan, X.C.; Wan, J.H.; Xu, M.M.; Liu, S.W.; Sheng, H.; Chen, Y.L.; Zhang, X.Y. Enteromorpha coverage information extraction by 1D-CNN and Bi-LSTM networks considering sample balance from GOCI images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9306–9317. [CrossRef]
24. Wang, S.K.; Liu, L.; Yu, C.; Sun, Y.; Gao, F.; Dong, J. Accurate Ulva prolifera regions extraction of UAV images with superpixel and CNNs for ocean environment monitoring. *Neurocomputing* **2019**, *348*, 158–168. [CrossRef]
25. Ronneberger, O.; Philipp, F.; Thomas, B. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
26. Zhou, Z.W.; Rahman, S.M.M.; Tajbakhsh, N.; Liang, J.M. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin, Germany, 2018; pp. 3–11.
27. Kim, S.M.; Shin, J.; Baek, S.; Ryu, J.H. U-Net convolutional neural network model for deep red tide learning using GOCI. *J. Coast. Res.* **2019**, *90*, 302–309. [CrossRef]
28. Guo, Y.; Le, G.; Li, X.F. Distribution Characteristics of Green Algae in Yellow Sea Using an Deep Learning Automatic Detection Procedure. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belguim, 11–16 July 2021; pp. 3499–3501. [CrossRef]
29. Cui, B.G.; Zhang, H.Q.; Jing, W.; Liu, H.F.; Cui, J.M. SRSe-net: Super-resolution-based semantic segmentation network for green tide extraction. *Remote Sens.* **2022**, *14*, 710. [CrossRef]
30. Gao, L.; Li, X.F.; Kong, F.Z.; Yu, R.C.; Guo, Y.; Ren, Y.B. AlgaeNet: A Deep-Learning Framework to Detect Floating Green Algae From Optical and SAR Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2782–2796. [CrossRef]
31. Hafiz, A.M.; Ghulam, M.B. A survey on instance segmentation: State of the art. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 171–189. [CrossRef]
32. Mou, L.C.; Zhu, X.X. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6699–6711. [CrossRef]
33. Maninis, K.K.; Caelles, S.; Chen, Y.; Pont-Tuset, J.; Leal-Taixe, L.; Cremers, D.; Van Gool, L. Video object segmentation without temporal information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1515–1530. [CrossRef]
34. Xu, Y.; Li, Y.; Wang, Y.P.; Liu, M.Y.; Fan, Y.B.; Lai, M.D.; Chang, E.I.-C. Gland instance segmentation using deep multichannel neural networks. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2901–2912. [PubMed]
35. He, K.M.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 22–29 October 2017; pp. 2961–2969.
36. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
37. Huang, Z.J.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.

38. Lee, Y.W.; Park, J.Y. Centermask: Real-time anchor-free instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13906–13915.
39. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9627–9636.