Journal of *Marine Science and Engineering*

MDPI

Article

ADNet: A Real-Time Floating Algae Segmentation Using Distillation Network

Jingjing Xu [1,2] and Lei Wang [3,*]

[1] East China Sea Ecology Center, MNR, Shanghai 201206, China; xujj1979@163.com
[2] Shanghai East Sea Marine Engineering Survey and Design Institute Co., Ltd., Shanghai 200136, China
[3] East Sea Information Center, State Oceanic Administration, Shanghai 200136, China
* Correspondence: wangl@ecs.mnr.gov.cn

Abstract: The segmentation of floating algae is a hot topic in the field of marine environmental research. Given the vastness of coastal areas and complex environments, algae detection models must have both higher performance and lower deployment costs. However, relying solely on a single Convolutional Neural Network (CNN) or transformer structure fails to achieve this objective. In this paper, a novel real-time floating algae segmentation method using a distillation network (ADNet) is proposed, based on the RGB images. ADNet can effectively transfer the performance of the transformer-based teacher network to the CNN-based student model while preserving its lightweight design. Faced with complex marine environments, we introduce a novel Channel Purification Module (CPM) to simultaneously strengthen algae features and purify interference responses. Importantly, the CPM achieves this operation without increasing any learnable parameters. Moreover, considering the huge scale differences among algae targets in surveillance RGB images, we propose a lightweight multi-scale feature fusion network (L-MsFFN) to improve the student's modeling ability across various scales. Additionally, to mitigate interference from low-level noises on higher-level semantics, a novel position purification module (PPM) is proposed. The PPM can achieve more accurate weight attention calculation between different pyramid levels, thereby enhancing the effectiveness of fusion. Compared to CNNs and transformers, our ADNet strikes an optimal balance between performance and speed. Extensive experimental results demonstrate that our ADNet achieves higher application performance in the field of floating algae monitoring tasks.

Keywords: floating algae segmentation; distillation network; transformers; CNNs; marine environmental

Citation: Xu, J.; Wang, L. ADNet: A Real-Time Floating Algae Segmentation Using Distillation Network. *J. Mar. Sci. Eng.* **2024**, *12*, 852. https://doi.org/10.3390/jmse12060852

Academic Editor: Sergei Chernyi

Received: 9 April 2024
Revised: 12 May 2024
Accepted: 16 May 2024
Published: 21 May 2024

1. Introduction

The widespread presence of floating algae (*Ulva prolifera* and *Sargassum*) in the East China Sea poses a serious threat to the marine ecological economy [1–3]. Firstly, the dense growth of green algae obstructs the sunlight, impeding the photosynthesis of submerged organisms and thereby disrupting the marine food chain. Additionally, the extensive decay of floating algae leads to a notable deterioration of seawater quality, resulting in economic losses in coastal aquaculture. Consequently, real-time monitoring and an early warning system for coastal floating algae have become crucial in the development of the circular marine economy [4–7]. The outbreak of green algae, influenced by various factors such as sea breezes and ocean currents, exhibits significant randomness and fluidity in its distribution. Hence, the monitoring system must satisfy stringent requirements in terms of real-time responses, low deployment costs, and high detection accuracy. In comparison to obtaining images from remote-sensing satellites, synthetic-aperture radar, or unmanned aerial vehicles, surveillance cameras installed on ships and shores can offer more stable monitoring, faster responses, and lower application costs. In this paper, we focus on discussing the utility of surveillance RGB images in the algae monitoring task.

Scholars initially employed traditional image processing methods for algae detection. Pan et al. utilized spectral unmixing to estimate green algae area at sub-pixel levels using real-world Geostationary Ocean Color Imager (GOCI) data [8] and synthetic data. Qi et al. [9] addressed the detection and discrimination limits of macroalgae from the perspective of multi-sensor measurements. Xing et al. [10] utilized multi-sensor, multi-scale, and multi-temporal observations to detect large-scale green tide occurrences in specific modes and locations. To address the mixed pixel effect in accurately determining macro-algal coverage from satellite images with moderate and low spatial resolution, Cui et al. proposed the linear pixel unmixing method [11]. To enhance the effectiveness of manual feature design, Yang et al. [12] introduced machine learning methods for floating algae detection. Podlejski et al. [13] built a machine learning model based on spatial features to filter out false cases after the detection process. Although traditional methods offer faster speed and lower deployment costs, they are susceptible to interference from complex marine environments due to artificial features and fixed parameters, leading to poor performance and generalization.

With the successful implementation of artificial intelligence technology across diverse application fields, researchers are focusing on employing deep learning methods for algae detection tasks. The commonly utilized techniques can be categorized into three groups: object detection [14,15], semantic segmentation [16–18], and instance segmentation [19–21]. Object detection, renowned for its swift inference speed and economical deployment costs, has gained widespread adoption in numerous domains. Park et al. [22] proposed to utilize the you only look once (YOLO) [23] model for algal image detection, achieving a balance between inference time and accuracy. Liu et al. [24] introduced an enhanced version of the algae-YOLO approach, using the ShuffleNetV2 [25] as the backbone network to reduce the parameter space.

However, the inherent limitation of these methods is that they cannot provide accurate edge contour information for objects, which is crucial in a monitoring system. To address this issue and meet the requirement for both contour and bounding box information, researchers tend to explore instance segmentation methods. Wang et al. [26] initially proposed the utilization of an instance segmentation framework for floating algae detection, significantly enhancing accuracy through the introduction of the One-Shot Aggregation Version (OSA) [27] and a dual-attention mechanism. Concurrently, Zou et al. proposed AlgaeFiner [28], a high-quality approach designed to achieve precise segmentation of algae targets in complex marine environments. However, a common challenge with these methods is their slower inference speeds and higher deployment costs in online monitoring applications. To address this limitation, scholars have begun exploring the semantic segmentation theory in monitoring systems. Wang et al. [29] proposed an end-to-end method that combines super-pixel algorithms with CNN models for algae segmentation and classification. Furthermore, ERISNet [30], a convolutional and recurrent neural network architecture, has been introduced to detect macroalgae along coastlines using remote sensing data. To extract large-scale green tide information, a deep semantic segmentation network named SRSe-Net [31] has been proposed. Based on the U-Net [32] structure, Liu et al. [33] proposed harmful algal blooms net (HAB-Ne) to effectively capture contextual information of algae targets to improve segmentation accuracy. However, when confronted with complex marine environments, traditional semantic segmentation methods also encounter challenges in achieving an optimal balance between performance and speed.

The methods discussed above provide valuable insights into the challenges inherent in the algae detection task from various angles. Nevertheless, a substantial disparity remains between these methods and their practical implementation in online applications. This paper endeavors to bridge this gap by exploring, through distillation theory, strategies for effectively striking a balance between model detection performance and deployment cost within the context of online algae detection applications.

2. Materials and Methods

2.1. Semantic Segmentation Distillation

The distillation network is composed of a teacher model and a student model. The teacher model typically adopts a transformer-based structure to provide robust performance, while the student network employs a lightweight CNN-based structure to achieve faster speed. During the training phase, the student network utilizes both ground truth labels and teacher results to refine its training. In the deployment phase, only the student network is used to obtain lower application costs. Yang et al. [34] proposed a novel cross-image relational knowledge distillation (KD) method to transfer global pixel correlations from the teacher to the student. Meanwhile, Dong et al. [35] introduced a novel cross-model KD framework to enhance the segmentation performance for high-resolution remote sensing images. A notable advancement in this domain is the single-branch CNN with transformer segmentation network (SCTNet) [36], which utilizes Segformer [37] as a teacher model to guide a lightweight CNN model. SCTNet has demonstrated superior running speeds and performance in the cityscapes dataset, making it suitable for our algae segmentation task. However, the original SCTNet structure still faces challenges in the algae segmentation task regarding the following aspects: (1) The marine environment is intricate and variable, featuring substantial interference. It remains uncertain whether the anti-interference capability of the teacher network can be effectively transmitted to the student network. (2) Algae targets on the sea surface exhibit huge scale differences, necessitating the network to have a robust multi-scale feature learning capability. (3) The construction of the student network requires simultaneously achieving faster speeds and lower deployment costs. To address these challenges, we propose a real-time floating algae segmentation network based on the distillation theory using RGB images, called ADNet, designed to enhance performance while maintaining efficiency.

The RGB images captured from surveillance cameras on ships and shores are particularly susceptible to various interferences, including weather conditions, camera angles, and similar targets, as illustrated in Figure 1. The original SCTNet structure solely relies on convolutional blocks for constructing the student network. Although this approach achieves faster speed and fewer parameters, the inherent limitations of CNN structures in terms of modeling capabilities significantly impede the student;s ability to effectively acquire robust features from the teacher network. To enhance the effectiveness and robustness of the algae distillation network, we introduce a lightweight channel purification module (CPM) in the student structure. The CPM serves a dual purpose by enhancing features related to algae targets while filtering out interference from background elements. It aids the student network in obtaining more accurate semantic information from the teacher network, thereby enhancing the efficacy of distillation. Moreover, the CPM does not introduce any additional learning parameters, thus simultaneously addressing the first and third challenges.

Faced with the second challenge of huge scale disparity in algae targets, we introduce a novel lightweight multi-scale feature fusion network (L-MsFFN). L-MsFFN is designed to enhance the student's ability to model features across different scales. Compared to existing feature fusion methods, L-MsFFN offers superior performance while maintaining minimal computational overhead. Additionally, to mitigate inefficiencies between different scale features in the pyramid, we propose the position purification module (PPM). By substituting the dynamic parameter attention (DPA) [38] mechanism with the PPM during the fusion stage, the PPM can offer accurate control of different pyramids on the final segmentation performance, thereby improving the effectiveness of the distillation process.

The architecture of our proposed ADNet, presented in Figure 2, comprises a CNN-based student network, a transformer-based teacher network, and a multi-scale distillation module. Within the student network, the CPM is integrated at each end of the encoding stage to enhance the modeling effectiveness. Simultaneously, the L-MsFFN is introduced before the decoding stage to strengthen the fusion capability of algae targets. In the L-MsFFN module, the PPM is proposed to address semantic disparities across different

pyramid levels, thereby minimizing the modeling gap between the student and teacher models. The teacher network incorporates the Segformer method to guide the student's learning process. Finally, the feature alignment module and the decoder distillation block from the original SCTNet are retained in our ADNet.

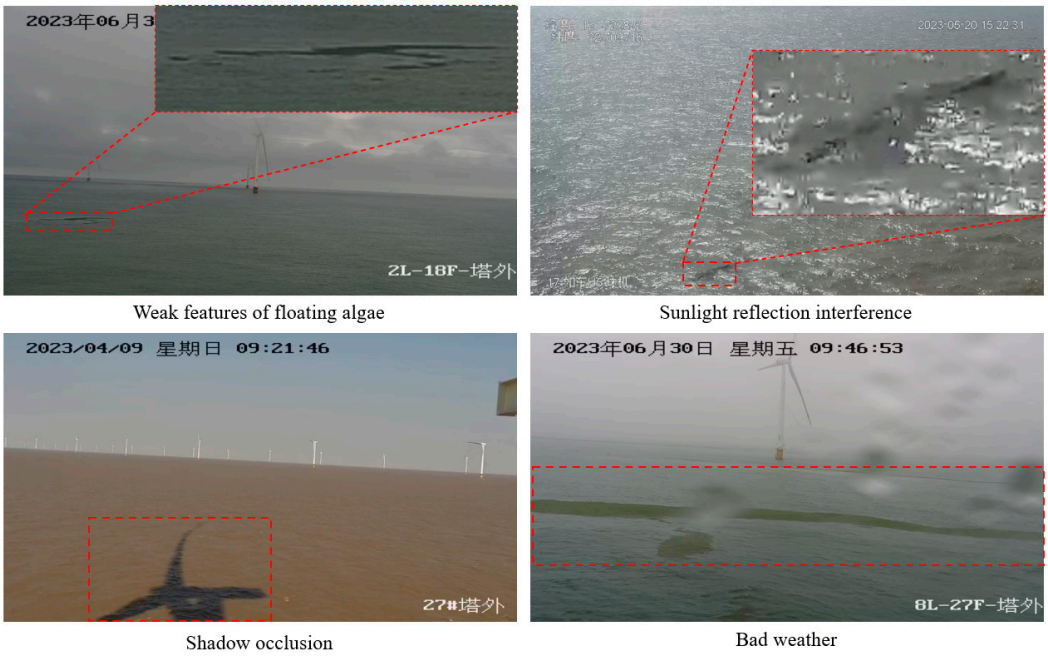

Figure 1. Example of complex marine detection environments in RGB images.

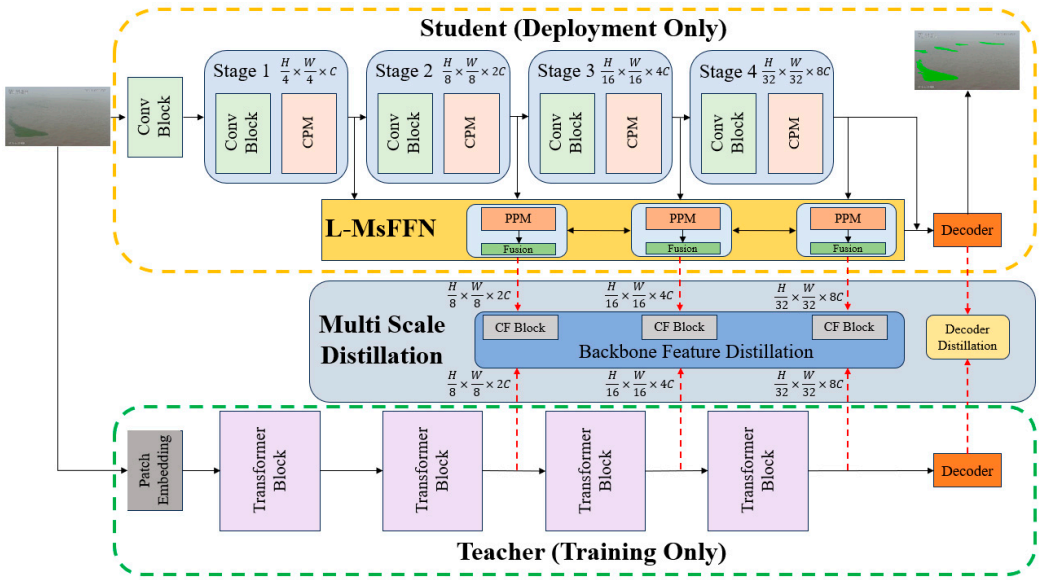

Figure 2. The architecture of ADNet.

2.2. Channel Purification Module

In contrast to the segmentation tasks conducted on public datasets, algae segmentation faces a unique challenge due to the intricate and dynamic environments in which it operates. This challenge requires the network to have strong anti-interference and robustness simultaneously. However, our analysis reveals that the design structure of the original SCTNet lacks specialization in both feature modeling and background anti-interference. Consequently, the teacher network struggles to convey its modeling proficiency to the student, resulting in a notable decline in distillation performance.

In Figure 3, we present some results of the original SCTNet under different interference environments. In the first two cases, influenced by factors such as shadows and fog, the student erroneously identifies wind power facilities and vessels as algae targets, leading to false positives. In the third case, it is obvious that the student fails to segment algae objects situated in shadows, resulting in missed segmentation. Meanwhile, the teacher network did not exhibit similar errors, thus indicating that the transformers possess the ability to model features in complex environments. These results suggest that the feature learning ability of the teacher network has not been effectively transmitted to the student. Despite the SCTNet utilizing a robust transformer-based model to guide the student's learning, the inherent limitation of the CNN structure hampers this distillation process. Consequently, the modeling capability of the student in the feature learning stage is not sufficiently improved even with teacher guidance when faced with strong interference factors. To address this, the CPM is proposed to strengthen the representation ability of the student network; its structure is illustrated in Figure 4.

Assuming the input feature is denoted as $f_i \in H \times W \times C$, where H, W, and C represent the height, width, and number of channels of the feature map, respectively. Firstly, we employ adaptive max pooling (AMP) and adaptive average pooling (AAP) operations to model the response of real algae targets (f_m) and interference targets (f_a). Subsequently, the *softmax* operation is applied to calculate the channel response weights on different branches. The resulting weight coefficients are then multiplied with the input features to obtain the corresponding weighted features for algae and other background elements. The calculation processes can be formulated as follows:

$$f_m = AMP(f_i),\ f_m \in 1 \times 1 \times C \tag{1}$$

$$f_a = AAP(f_i), f_a \in 1 \times 1 \times C \tag{2}$$

$$f_m^w = f_i \times softmax(f_m), f_m^w \in H \times W \times C \tag{3}$$

$$f_a^w = f_i \times softmax(f_a), f_a^w \in H \times W \times C \tag{4}$$

In the enhancement branch, the addition operation is employed to obtain the response weight f_m^o, thereby amplifying the attention dedicated to algae targets in the feature map. Conversely, in the purification branch, subtraction operations are utilized to eliminate the background interference response f_a^o from the input features, thereby mitigating the impact of interference on the representation of algae features. Finally, the f_m^o and f_a^o are combined through the weighted add operation (w_m and w_a) to yield the output of the CPM, denoted as f_o. The entire process is expressed as follows:

$$f_m^o = f_i + f_m^w, f_m^o \in H \times W \times C \tag{5}$$

$$f_a^o = f_i - f_a^w, f_a^o \in H \times W \times C \tag{6}$$

$$f_o = w_m \times f_m^o + w_a \times f_a^o, f_o \in H \times W \times C \tag{7}$$

The innovation inherent in our CPM module lies in enhancing the feature response of algae targets while simultaneously diminishing the weight of interfering elements, with the goal of minimizing computational costs. The CPM module operates through a parallel dual-branch design, consisting of the enhancement branch and the purification branch. The purification branch models background information by incorporating average pooling and feature subtraction operations to reduce the response weight of interfering targets. Meanwhile, the enhancement branch employs maximum pooling to extract algae features and utilizes the add operation to amplify these responses throughout the entire encoding stage. Importantly, our CPM structure achieves this operation through computational logic, thereby avoiding the introduction of any additional parameters. In Figure 5, we visualized the attention heatmaps of the enhanced and purified branches during the prediction stage. It can be observed that the enhanced branch has performed the amplification on all potential algae targets in the scene, although there are instances of false enhancement. From the

perspective of interference targets, the purification branch only retains the feature responses confirmed as interest targets and excludes the responses that may be interferences. Finally, by fusing the information from the two branches, the real floating algae targets receive sufficient responses, and the responses of interfering targets have been suppressed.

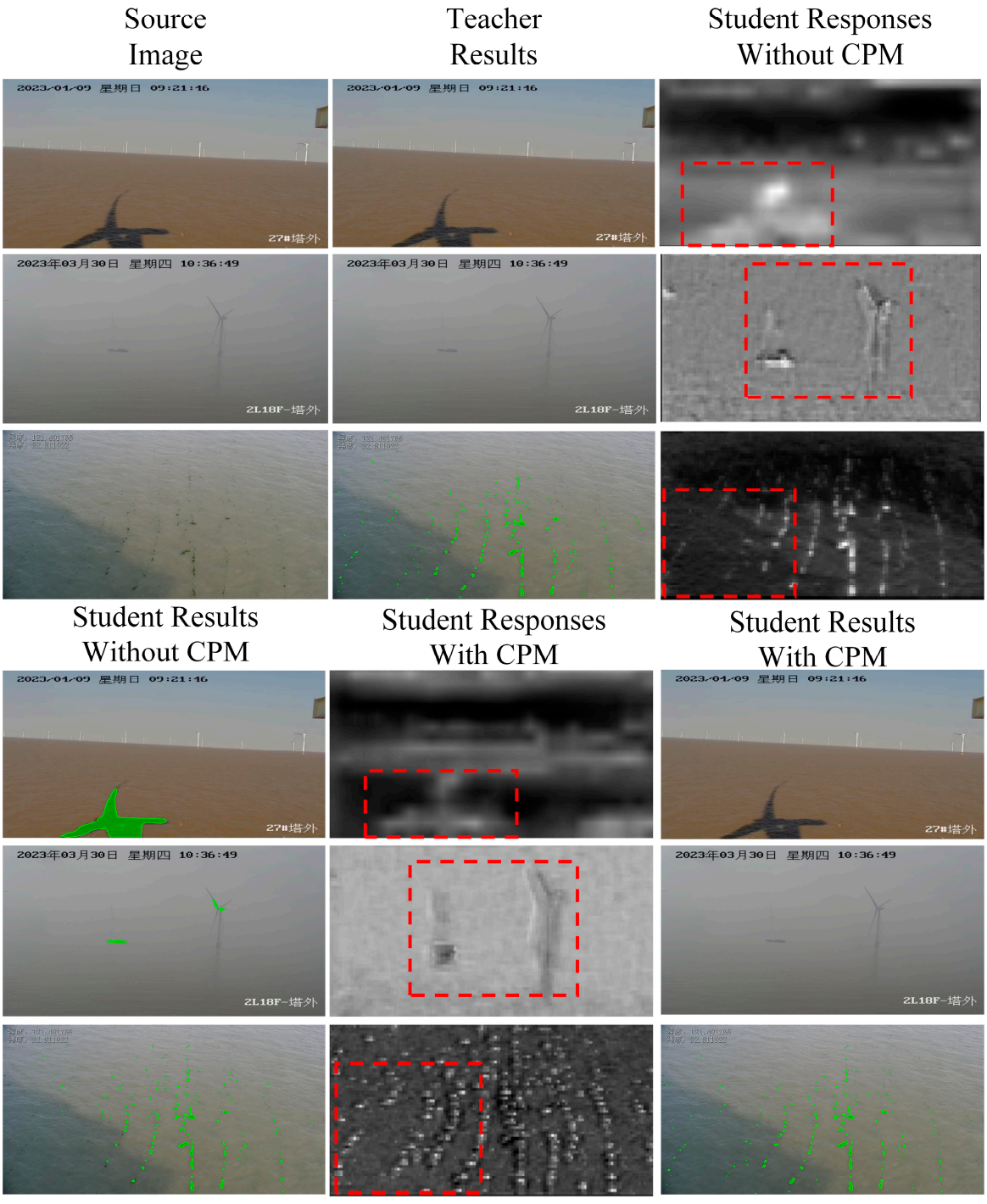

Figure 3. Comparison of CPM in the original SCTNet structure.

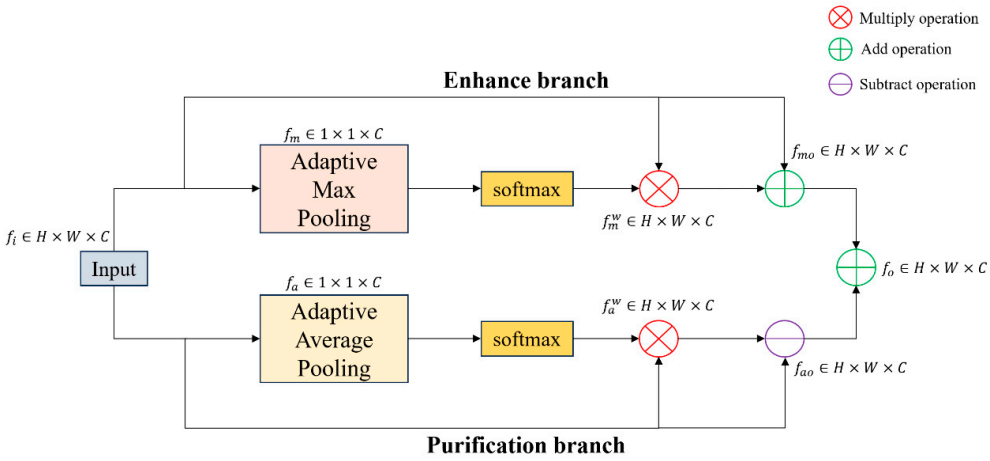

Figure 4. The architecture of CPM.

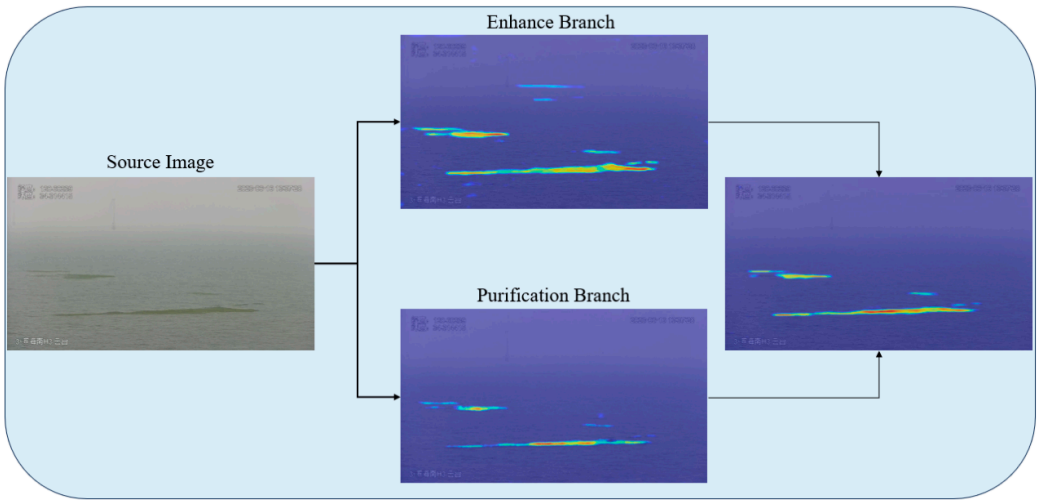

Figure 5. The attention heatmaps of CPM in the prediction stage.

2.3. Multi-Scale Feature Fusion Module

The efficacy of the segmentation model is intricately tied to its ability to handle multi-scale features. However, it is noteworthy that the student structure in the original SCTNet lacks this capability, posing a significant limitation in the algae segmentation task. To underscore this, we conducted a comparative analysis by computing the maximum and minimum sizes of selected categories in both the cityscapes dataset and our floating algae dataset, as depicted in Figure 6. The result reveals that scale differences within the same category are prevalent in different tasks, particularly in our algae category, where the discrepancy spans up to 960 times. Therefore, it is important to strengthen the feature fusion ability across diverse scales in our distillation structure.

In SCTNet, the student network employed the dynamic position attention pyramid pooling module (DAPPM) [36] to simultaneously augment receptive fields and integrate multi-scale contextual information. However, the DAPPM primarily emphasizes the impact of the highest-level feature during the decoding stage, underutilizing the contributions from other pyramid layers. This limitation becomes serious in the results obtained from both the teacher and student networks, as demonstrated in Figure 7. Specifically, in the cityscapes dataset, the student network exhibits challenges in accurately segmenting both large-scale (road category) and small-scale targets (person category). Similarly, in the floating algae dataset, the interference caused by large algae targets in the image leads to significant missed segmentation issues for small-sized algae located nearby.

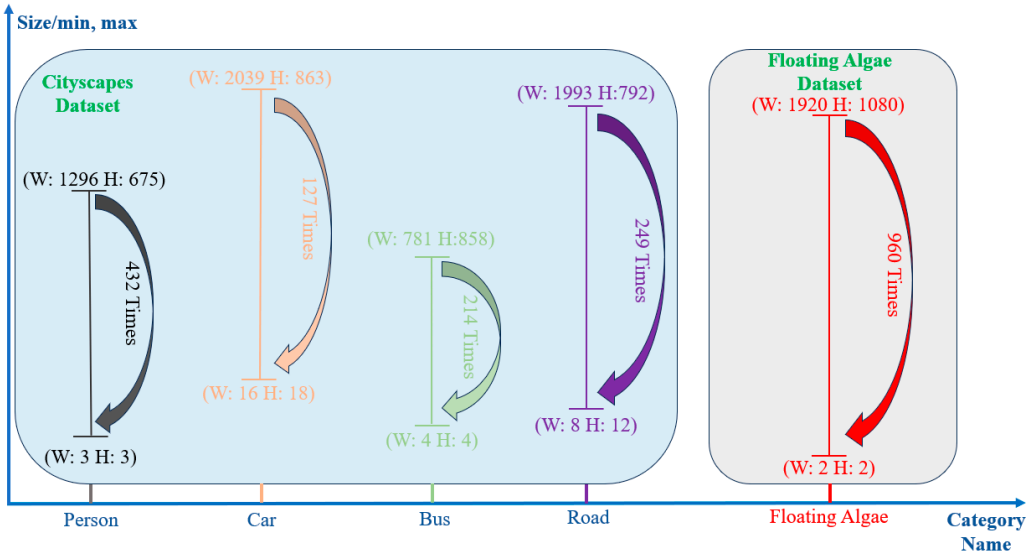

Figure 6. The scale differences in the cityscapes and algae datasets.

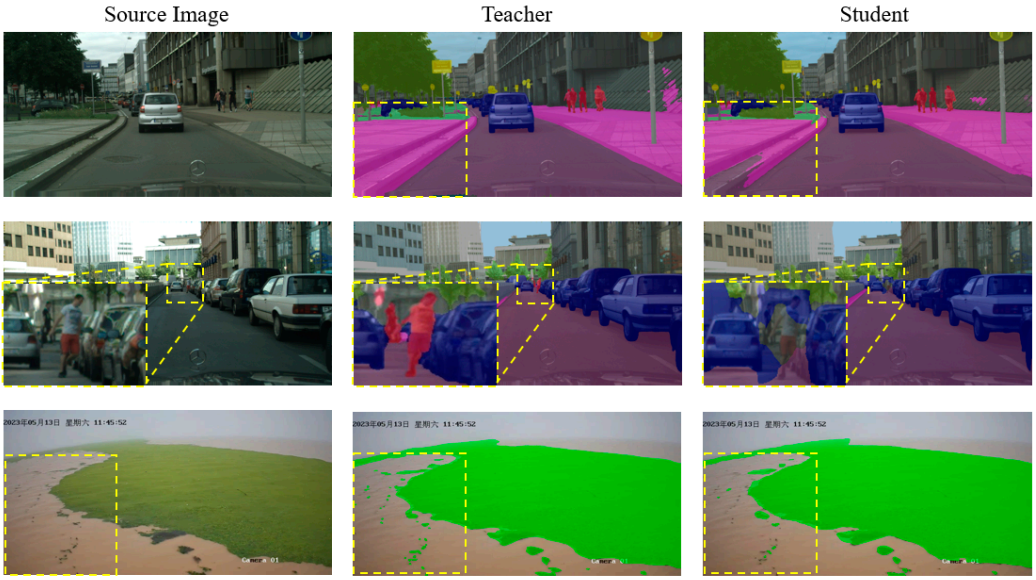

Figure 7. The segmentation results of SCTNet in the Cityscapes and algae datasets.

The above observations reveal several significant challenges in the original SCTNet structure. Firstly, even within the same category, the features of large-sized targets in the higher-level pyramid will overshadow the features of small-sized targets, resulting in missed segmentation. The multi-scale feature fusion structure can ensure the student network captures a more comprehensive representation of the floating algae data, thereby enhancing the overall performance. Secondly, the teacher network, which is based on the transformer structure, demonstrates superior performance in multi-scale modeling compared to CNN-based structures. This advantage arises from self-attention's ability to model long-range dependencies. However, this multi-scale fitting capability is difficult for the CNNs to replicate. Finally, relying solely on the highest-level semantics for decoding has been proven inefficient in the algae distillation task. This design structure limits the distillation performance, particularly when dealing with targets that exhibit significant scale variations. The decoding features of the student network ought to incorporate multi-dimensional feature information, rather than solely relying on high-level semantics.

Currently, significant advancements have been achieved in the field of multi-scale fusion research, with multi-scale bidirectional feature pyramid network (Ms-BiFPN) [28]

emerging as a standout performer in these approaches. Based on the bidirectional feature pyramid network (BiFPN) [38] architecture, Ms-BiFPN facilitates efficient and seamless modeling of multi-scale features by integrating the adaptive spatial-fusion block in the instance segmentation task. However, the complex fusion processes inherent in both Ms-BiFPN and BiFPN have restricted their practical application in our monitoring task.

To this end, we propose the L-MsFFN; its structure is depicted in Figure 8. Firstly, we initiate a pruning process on the original BiFPN structure, selectively retaining only essential fusion branches to mitigate unnecessary computational overhead. The original BiFPN structure adopts a top-down, lateral jumping, and bottom-up design to maximize the involvement of each pyramid layer in the feature fusion process, enriching the multi-scale semantics. However, this densely connected nature also introduces redundant computations. Upon analyzing the impact of each fusion link in our algae distillation task, we observe that the lateral jump connections have a negligible effect on performance, rendering them unnecessary. Therefore, we eliminate these lateral skip connections in our L-MsFFN structure. A more detailed discussion about the skip connections will be presented in the experimental section. Additionally, we utilize the PPM as a replacement element for the DPA in the original BiFPN. This substitution aims to narrow the semantic gaps across different pyramid levels, thereby enhancing the effectiveness of the feature fusion process.

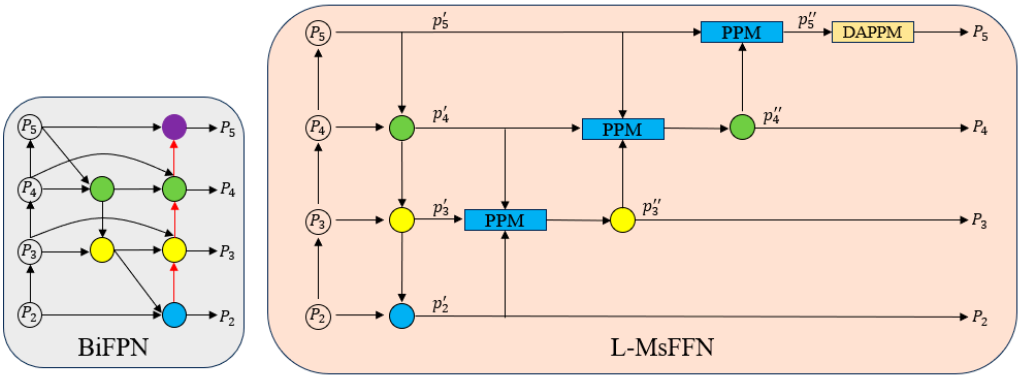

Figure 8. The structure of L-MsFFN.

Assuming the feature maps obtained during the encoding stage are $p_2 \in \frac{H}{4} \times \frac{W}{4} \times C$, $p_3 \in \frac{H}{8} \times \frac{W}{8} \times 2C$, $p_4 \in \frac{H}{16} \times \frac{W}{16} \times 4C$ and $p_5 \in \frac{H}{32} \times \frac{W}{32} \times 8C$. We initiate the top-down fusion process with the following calculations:

$$p'_4 = DownSample(p_5) + p_4 \tag{8}$$

$$p'_3 = DownSample(p'_4) + p_3 \tag{9}$$

$$p'_2 = DownSample(p'_3) + p_2 \tag{10}$$

Here, *DownSample* represents the bilinear interpolation down-sampling operation. In the bottom-up fusion process, each fusion line will be input into the *PPM* to calculate the weight coefficients of each pyramid feature map at the same position. The calculation processes are as follows:

$$p''_3 = PPM(UpSample(p'_2),\ p'_3, DownSample(p'_4)) \in \frac{H}{8} \times \frac{W}{8} \times 2C \tag{11}$$

$$p''_4 = PPM(UpSample(p''_3),\ p'_4, DownSample(p'_5)) \in \frac{H}{16} \times \frac{W}{16} \times 4C \tag{12}$$

$$p''_5 = PPM(UpSample(p''_4),\ p'_5) \in \frac{H}{32} \times \frac{W}{32} \times 8C \tag{13}$$

Here, $UpSample$ represents the bilinear interpolation up-sampling operation. Finally, input p_5'' into the DAPPM to model semantics under different receptive fields.

2.4. Position Purification Module

In feature pyramids, it is essential to consider that lower-level features often contain detailed yet noisy information, whereas high-level features encapsulate the abstract semantics. Directly combining information across different scales can lead to significant mutual interference because the noise present at lower levels may compromise the semantics encoded at higher levels, resulting in the degradation of the network. To mitigate this inconsistency, BiFPN employs the DPA mechanism to dynamically adjust the weight relationship of pyramid features before performing fusion. While this approach is useful, it is inefficient as excessive weight calculations significantly increase the computational overhead at this stage. In contrast to the BiFPN's theory, we recognize that high-level feature maps carry valuable information about the algae targets, while low-level features often contain interfering elements. Therefore, in our L-MsFFN, we adopt a direct fusion approach in the top-down process, bypassing the need for DPA. This approach resembles the original Feature Pyramid Network (FPN) structure, enabling faster fusion speeds. Simultaneously, in the bottom-up process, we introduce the PPM as a replacement for DPA. The structure of PPM is presented in Figure 9.

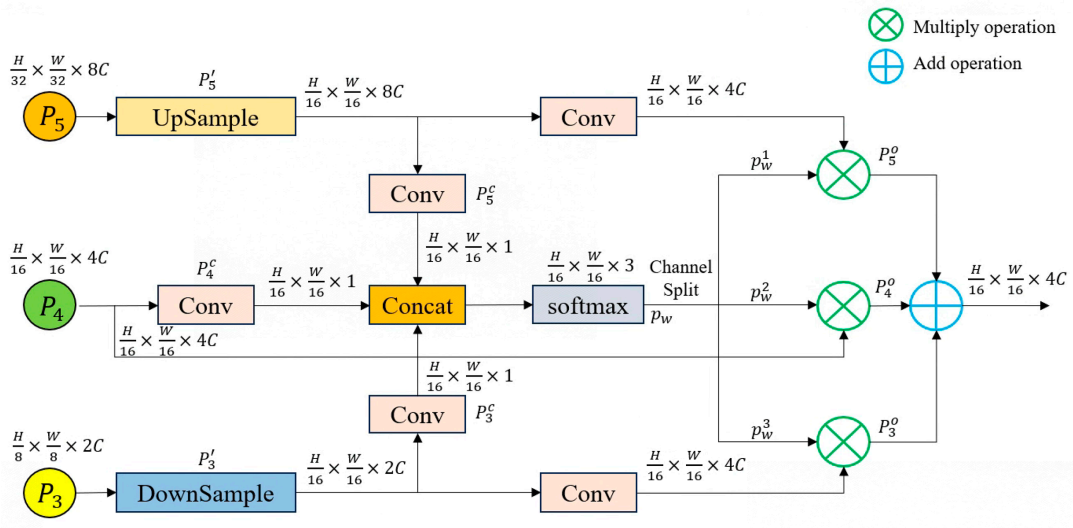

Figure 9. The structure of PPM.

Taking p_3, p_4 and p_5 as example, the calculation process of PPM is as follows: Initially, down-sampling and up-sampling operations are employed to transform the sizes of p_3 and p_5 to $p_3' \in \frac{H}{16} \times \frac{W}{16} \times 2C$ and $p_5' \in \frac{H}{16} \times \frac{W}{16} \times 8C$, respectively. Subsequently, convolution operations are applied to p_3', p_4, and p_5' individually to condense the number of channels to one. Following this convolution, only one channel of information is retained for each pyramid, obtaining the position response feature maps: p_3^c, p_4^c, and p_5^c. The above processes can be calculated as:

$$p_3' = DownSample(p_3) \tag{14}$$

$$p_5' = UpSample(p_5) \tag{15}$$

$$p_3^c, p_4^c, p_5^c = Conv2d(p_3'), Conv2d(p_4), Conv2d(p_5') \tag{16}$$

Here, $Conv2d$ represents the 2D convolutional operation. Afterwards, the p_3^c, p_4^c, and p_5^c are concatenated together, and the $softmax$ operation is applied to calculate the weight coefficients p_w of different pyramid feature maps from the perspective of pixel dimension.

The corresponding weight coefficients p_w^1, p_w^2 and p_w^3 on each pyramid scale can be obtained sequentially through the channel *split* operation.

$$p_w = softmax(concat(p_3^c, p_4^c, p_5^c)) \tag{17}$$

$$p_w^1, p_w^2, p_w^3 = split(p_w) \tag{18}$$

Here, *concat* and *split* represents the concatenation and split operations. Finally, the feature maps at each pyramid scale are sequentially multiplied by the corresponding position weight coefficients p_w^1, p_w^2 and p_w^3 to obtain the position-weighted feature map P_5^o, P_4^o and P_3^o. These are then added up to yield the final fusion result P_4^o of the PPM at the fourth-level scale. The calculations are as follows:

$$P_3^o = p_w^1 \times Conv2d(p_3') \tag{19}$$

$$P_5^o = p_w^3 \times Conv2d(p_5') \tag{20}$$

$$P_4^o = p_w^2 \times p_4 + P_3^o + P_5^o \tag{21}$$

The PPM accomplishes not only the dynamic balance across pyramid levels but also obtains the attention calculation of different pyramid levels at the pixel granularity. Compared to the original BiFPN, which only utilizes two dynamic parameters to represent the entire feature map, PPM allows for more accurate weight coefficient calculation between different scales. Additionally, the PPM exhibits lower computational complexity and higher speed across all fusion methods. Figure 10 delineates the disparity between DPA and PPM.

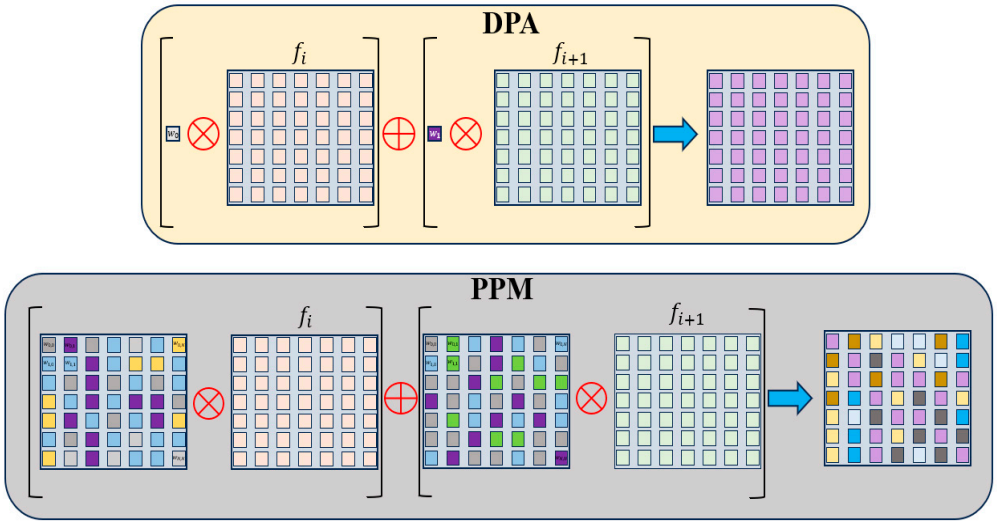

Figure 10. The comparison between PPM and DPA.

2.5. Distillation Loss

During the training stage, the distillation loss serves to quantify dissimilarities between the student and teacher networks. In this study, we adopt the channel-wise distillation loss (CWD) [36] as the training metric. CWD normalizes the activation layer of each channel through a softmax operation, generating a probability map. Subsequently, KL divergence is utilized to minimize the disparity between the teacher and the student network. CWD has demonstrated superior distillation performance in dense prediction tasks, rendering it well-suited for our floating algae distillation task.

The total loss function $loss_{total}$ of our ADNet is divided into the encoding loss $loss_{encoder}$ and the decoding loss $loss_{decoder}$, respectively. The calculation formula is as follows:

$$loss_{total} = loss_{encoder} + loss_{decoder} \tag{22}$$

Specifically, in the encoding stage, feature distillation is individually applied to the different pyramids p_3, p_4, and p_5. In the decoding stage, we directly distill the decoded feature maps from teacher and student networks.

$$\varphi(F) = \frac{\exp(\frac{F,k}{\tau})}{\sum_{k=1}^{W*H}\exp(\frac{F,k}{\tau})} \tag{23}$$

$$loss_{encoder} = \sum_{l=1}^{L=3} w_l \times \frac{\tau^2}{C} \times \sum_{c=1}^{C} \varphi(T_l^e) \times \log(\frac{\varphi(T_l^e)}{\theta(\varphi(S_l^e))}) \tag{24}$$

$$loss_{decoder} = w_d \times \frac{\tau^2}{C} \times \sum_{c=1}^{C} \varphi\left(T^d\right) \times \log(\frac{\varphi(T^d)}{\theta(\varphi(S^d))}) \tag{25}$$

Here, $\varphi(\cdot)$ represents the *softmax* calculation, where exp denotes the natural exponential function. $\theta(\cdot)$ serves to align the feature map size of the student network with that of the teacher network, ensuring compatibility between the two networks. L signifies the number of pyramid layers involved in the distillation process, with a value of 3 specified in our paper. τ, a hyperparameter known as the distillation temperature, facilitates the student network's ability to better approximate the teacher's distributions. In our paper, τ is set to 4. T_l^e and S_l^e represent the feature maps of the l-th pyramid layer in teacher and student, respectively. Conversely, T^d and S^d denote the feature maps derived from the decoding stages of the teacher and student, respectively. w_l and w_d are weight hyperparameters that modulate the impact of different distillation loss layers, ensuring a balanced contribution to the overall distillation process.

2.6. Experiment Setups

In this study, both the training and testing phases were executed on an Intel Xeon Gold 6330 CPU processor operating at 2.0 GHz (Intel, Santa Clara, CA, USA), accompanied by four NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory (NVIDIA, Santa Clara, CA, USA). For the deployment phase, we employed an Intel i7-13700K processor paired with a single NVIDIA GeForce RTX 3090 GPU as the detection server. Table 1 offers a detailed summary of the software utilized in this research, including its names and version numbers.

Table 1. Summary of software names and version numbers.

Software Name	Version Numbers
Ubuntu	18.04
CUDA	12.1
cuDNN	8.9.3
Python	3.8.5
Pytorch	1.13.1
MMCV	2.0.1
MMSegmentation	1.2.2
MMDeploy	1.3.1
Open Neural Network Exchange (ONNX)	1.15.0
ONNX-RunTime-GPU	1.8.1
TensorRT	8.6.1 post1

3. Results

3.1. Datasets

To evaluate the performance of ADNet, we collected RGB-image data from surveillance cameras between April 2023 and August 2023, situated in Nantong and Yancheng in the Jiangsu Sea area. The detailed locations of these surveillance cameras are presented in Table 2. The dataset comprises 4500 labeled RGB images (*Ulva prolifera*) with a resolution of 1920 × 1080. We partitioned 3500 images for training purposes, reserving 1000 for

validation and testing. The distribution of floating algae sizes in the dataset is depicted in Figure 11. In practical applications, we utilize 16 channels of surveillance video data to perform real-time detection and monitoring of floating algae along the coast of Jiangsu.

Table 2. Surveillance camera positions in our paper.

Index	Camera Name	Number of Cameras	Longitude (°N)	Latitude (°E)
1	Binhai North District H2#400 MW	2	34.4	120.3
2	SPIC Binhai South H3#300 MW	2	34.3	120.6
3	Jiangsu Rudong H5#	3	32.7	121.7
4	Jiangsu Rudong H14#	3	32.8	121.4
5	Three Gorges New Energy Jiangsu Dafeng 300 MW	2	33.3	121.1
6	Huaneng Jiangsu Dafeng 300 MW	2	33.1	121.4
7	Dafeng Wharf	2	33.2	120.8

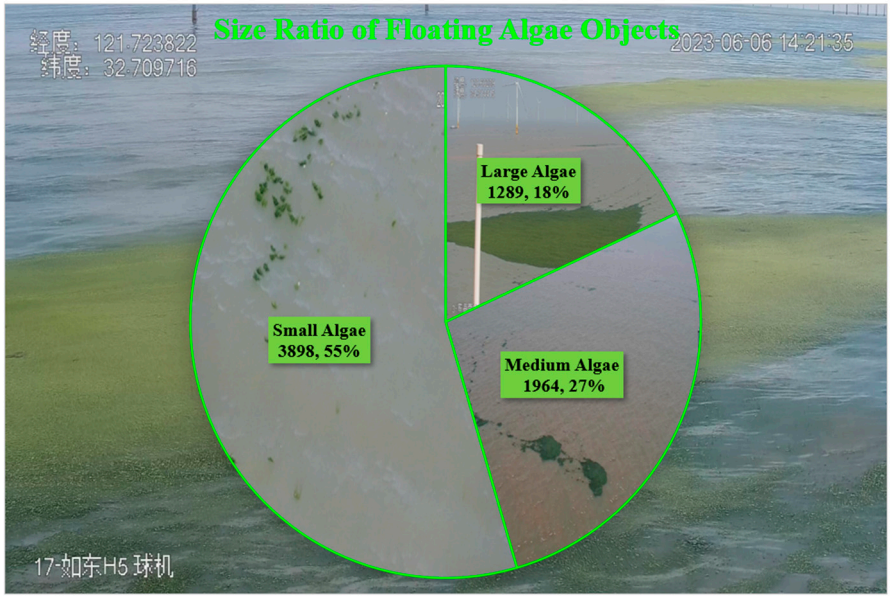

Figure 11. The algae size distribution in our dataset.

3.2. Evaluation of Model Performance

In this paper, we introduce several metrics to assess the performance of our distillation network. These metrics include mean intersection over union (mIoU), inference time, FPS, and the number of learnable parameters (Params). The calculation of mIoU is formulated as follows:

$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{p_{i,\,i}}{\sum_{j=0}^{N}p_{i,j}+\sum_{j=0}^{N}\left(p_{j,\,i}-p_{i,\,i}\right)} = \frac{1}{N+1}\sum_{i=0}^{N}\frac{TP}{FN+FP+TP} \tag{26}$$

where N represents the total number of classes in our algae dataset, with $N+1$ accounting for the inclusion of a background category. The term $p_{i,j}$ represents the count of pixels where the true label is i, but the predicted label is j. Similarly, the definitions of $p_{j,\,i}$ and $p_{i,\,i}$ are analogous to that of $p_{i,j}$. Furthermore, FPS is calculated by dividing the total inference time of the testing images. Specifically, the inference time is determined as the sum of individual inference times for each image, divided by the total number of images K. The calculation of FPS is as follows: $inference\ time = \frac{\sum_{i=0}^{K} I_i}{K}$. Here, I_i is the inference time for the i-th image. In addition, to provide a comprehensive evaluation of our network's performance in actual applications, particularly in terms of inference speed, we employed two widely used model transformation techniques: ONNX and TensorRT (TRT).

To assess the effectiveness of our proposed network, we conducted a comparative analysis with various segmentation methods on the floating algae dataset. The detailed training configurations and hyper-parameters for each model are provided in Table 3, and the detailed comparison results are presented in Table 4. To ensure a fair comparison, all models were evaluated on images with a resolution of $1024 \times 1024 \times 3$. Meanwhile, we applied consistent preprocessing and data augmentation techniques across all models, including image normalization, random cropping, random brightness adjustments, and random size variations. By leveraging these metrics and comparative analysis, we aim to provide a thorough understanding of the strengths and weaknesses of different methods in the field of floating algae monitoring tasks.

Table 3. Training configurations of different methods.

Method Name	Backbone Name	Learning Rate	Iterations	Batch Size
BiSeNetV2	BiSeNetV2	0.01	80,000	16
GCNet [39]	R-50	0.01	120,000	8
	R-101	0.005	150,000	4
OCRNet [40]	HRNet-W18-Small	0.01	40,000	32
	HRNet-W18	0.01	80,000	24
	HRNet-W48	0.005	120,000	8
STDC [41]	STDC1	0.001	40,000	64
	STDC2	0.001	40,000	64
Segformer	MIT-B0	0.01	80,000	16
	MIT-B1	0.01	80,000	16
	MIT-B2	0.01	120,000	12
	MIT-B3	0.0005	150,000	8
	MIT-B4	0.0001	160,000	8
	MIT-B5	0.001	240,000	4
SCTNet	S-Seg50	0.01	80,000	16
	S-Seg75	0.01	80,000	16
	B-Seg50	0.01	120,000	16
	B-Seg75	0.005	120,000	8
	B-Seg100	0.005	120,000	8
ADNet	S-Seg50	0.01	80,000	16
	S-Seg75	0.01	80,000	16
	B-Seg50	0.01	120,000	16
	B-Seg75	0.005	120,000	8
	B-Seg100	0.005	120,000	8

In Table 4, the comparison results across different methods underscore the superiority of transformer-based networks over their CNN-based counterparts. This observation highlights the limitations of CNNs when confronted with complex marine environments. Additionally, results obtained through the distillation method demonstrate higher mIoU performance compared to those trained directly on datasets, providing evidence of the efficacy of this theory. Simultaneously, from the perspective of the speed metric, CNN-based networks exhibit shorter inference times compared to transformers, indicating the computational cost constraints of the latter in practical deployment scenarios. The evaluation of both segmentation performance and inference speed emphasizes the importance of the distillation-based approach in actual application because this approach can achieve an optimal balance between performance and speed. In summary, transformer-based networks excel in performance, while CNN-based networks offer faster speed. The distillation method presents a promising avenue for improving performance while maintaining reasonable speeds. For more intuitive comparison results, refer to Figures 12 and 13.

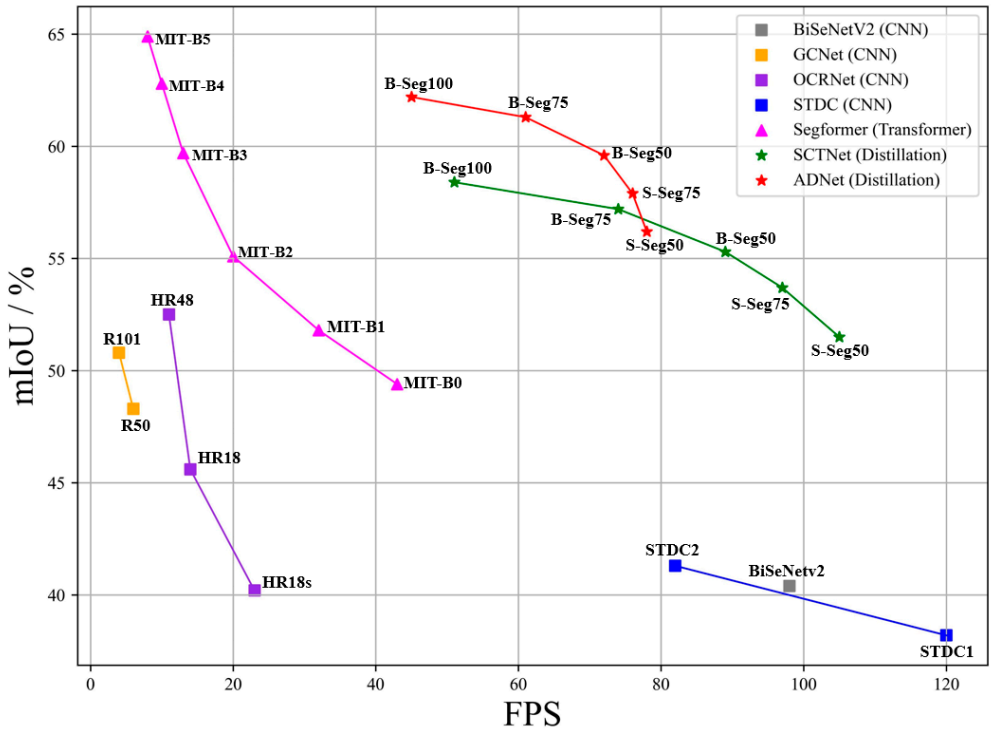

Figure 12. The speed-accuracy performance of different methods.

Table 4. The comparison results of different methods.

Method Name	Backbone Name	Params (MB)	mIoU (%)	mIoU FP16 (%)	FPS (Torch CUDA)	Inference Time (ms)					
						Torch CUDA	TRT CUDA	TRT CUDA FP16	ONNX CUDA	ONNX CUDA FP16	ONNX CPU
BiSeNetV2	BiSeNetV2	14.12	40.4	39.9	98	10.2	8.1	7.2	12.2	10.7	17.1
GCNet	R-50	47.33	48.3	48.1	6	164.4	52.0	51.1	106.7	95.3	112.5
	R-101	65.44	50.8	50.3	4	245.7	74.6	71.2	129.5	125.3	139.6
OCRNet	HRNet-W18-Small	6.1	40.2	39.7	23	43.1	17.7	16.2	30.3	28.5	37.7
	HRNet-W18	11.5	45.6	45.1	14	69.4	23.1	21.5	45.3	48.9	47.6
	HRNet-W48	67.1	52.5	51.9	11	84.4	54.3	51.5	87.2	87.3	89.2
STDC	STDC1	8.2	38.2	35.4	120	8.3	3.9	3.8	17.5	14.9	19.2
	STDC2	12.1	41.3	37.2	82	12.1	5.2	5.2	25.8	19.1	26.3
Segformer	MIT-B0	3.5	49.4	49.0	43	23.2	12.9	12.1	30.9	21.3	35.2
	MIT-B1	13.1	51.8	51.5	32	30.4	17.3	16.5	31.4	20.5	35.7
	MIT-B2	23.6	55.1	54.7	20	49.8	28.6	27.2	47.9	29.5	51.3
	MIT-B3	42.5	59.7	59.6	13	72.3	42.2	41.7	70.1	39.5	79.8
	MIT-B4	58.5	62.8	62.5	10	99.9	59.1	58.6	95.5	51.6	110.7
	MIT-B5	78.2	64.9	64.7	8	118.7	71.8	70.2	115.3	59.8	135.9
SCTNet	S-Seg50	4.6	51.5	51.1	105	9.5	9.2	9.1	19.2	8.7	23.0
	S-Seg75	4.6	53.7	53.2	97	10.3	9.4	9.5	19.3	9.9	22.9
	B-Seg50	17.4	55.3	55.1	89	11.2	10.7	10.5	23.0	13.1	25.2
	B-Seg75	17.4	57.2	56.8	74	13.5	12.3	12.6	24.1	13.8	26.2
	B-Seg100	17.4	58.4	57.9	51	19.6	15.6	15.4	31.7	20.6	35.6
ADNet	S-Seg50	6.5	56.2	55.9	78	12.7	11.6	12.1	21.1	10.7	21.4
	S-Seg75	6.5	57.9	56.8	76	13.1	12.3	12.6	20.5	11.2	22.1
	B-Seg50	19.3	59.6	58.7	72	13.8	13.1	14.3	21.8	14.3	25.7

Table 4. *Cont.*

Method Name	Backbone Name	Params (MB)	mIoU (%)	mIoU FP16 (%)	FPS (Torch CUDA)	Inference Time (ms)					
						Torch CUDA	TRT CUDA	TRT CUDA FP16	ONNX CUDA	ONNX CUDA FP16	ONNX CPU
ADNet	B-Seg75	19.3	61.3	60.6	61	16.3	14.2	14.2	26.4	18.3	29.6
	B-Seg100	19.3	62.2	61.7	45	21.9	16.5	15.4	33.2	23.1	37.5

In the evaluation of the mIoU indices, Segformer, particularly its B5 configuration, demonstrated superior performance, whereas STDC exhibited the lowest. Notably, when guided by the Segformer teacher network, SCTNet-S-Seg50 exhibited a substantial improvement compared to the unguided STDC2, achieving 12.4% improvement. This observation proves the remarkable potential of distillation in enhancing the performance of segmentation networks. Furthermore, within the same network size, our refined ADNet structure exhibited a notable advantage over the original SCTNet, averaging approximately 4% higher in mIoU. The highest index achieved an impressive 62.2%, only trailing the teacher network by 2.7%. This result can be attributed to the contributions of our proposed CPM and L-MsFFN in enhancing the accuracy of floating algae segmentation in complex marine environments. Importantly, ADNet achieved this performance boost without a significant increase in the number of learnable parameters. In terms of speed performance, STDC emerges as the fastest among the evaluated models, closely followed by SCTNet and ADNet. The integration of PPM and L-MsFFN into our ADNet led to a marginal reduction in speed, approximately 3 milliseconds per frame, compared to the baseline SCTNet. Taking the B-Seg100 backbone as an example, despite the ADNet exhibiting a 6 FPS decrease, it significantly improved the mIoU indicators by 3.8%. This result demonstrates the effectiveness of our proposed modules in enhancing performance with minimal computational overhead.

Regarding the module conversion methods, the TRT approach demonstrated superior speed performance compared to the ONNX method. However, this speed advantage comes with a trade-off in flexibility, as ONNX models offer the advantage of being directly executable on the CPU device. Meanwhile, the FP16 model did not significantly enhance the speed when used in the TRT conversion mode. Conversely, in the ONNX mode, adopting the FP16 model can obtain a 30% improvement in speed. These findings provide a valuable insight into the trade-offs between speed and flexibility in different model conversion methods. Employing B-Seg100 within ADNet as an example, the execution time of the FP16 mode decreased by approximately 10 milliseconds compared to the FP32 mode, achieving a notable 33% improvement. In addition, most segmentation methods can maintain their performance in the FP16 mode, underscoring its adaptability in different deployment modes. By capitalizing on both CPU and GPU resources, the mixed deployment mode can offer a substantial cost reduction in applications by leveraging the capabilities of the CPU resources during the inference process. This aspect is crucial in our real-time algae monitoring task, as only one GPU is utilized to process sixteen videos in parallel during the deployment phase.

In scenarios 1 to 3, we focus on assessing the segmentation performance when faced with small algae targets amidst the backdrop of larger algae targets. This evaluation can offer insights into the models' ability to distinguish between different sizes of algae. Moving on, scenarios 4 to 8 discuss the performance in complex environments, including sunlight interference, sea waves, and foggy conditions. By evaluating the models' responses to these varying environmental factors, we can acquire an understanding of their adaptability and robustness in actual applications. The above comparisons not only consider the models' overall accuracy but also evaluate their stability in the presence of strong disturbances.

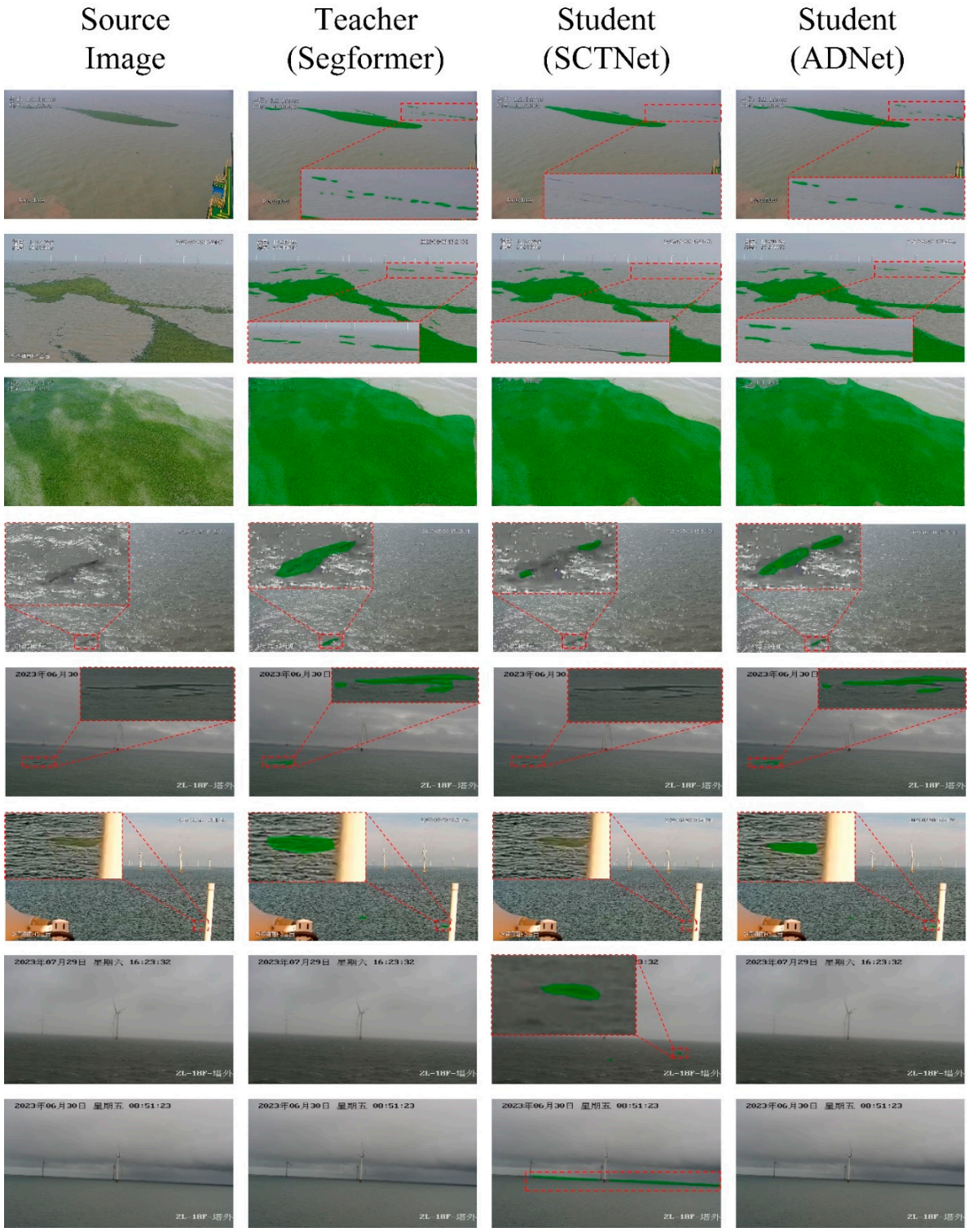

Figure 13. The visualization results of segmentation models.

In the initial set of scenarios, we can clearly observe that all models demonstrate satisfactory performance on large targets, with no instances of missed detections. However, upon closer inspection, SCTNet exhibits limitations for the small algae targets, revealing issues with missing problems. This suggests that the prominent features of large targets overshadow the response of smaller targets during the decoding phase. This limitation becomes serious when dealing with multi-scale targets in a single image. To address this challenge, we introduced the L-MsFFN in our ADNet. This module enables the low-

level features of small targets to resonate with high-level semantics. Additionally, the PPM module is introduced to eliminate the noise features at lower levels, thus preventing them from contaminating the high-level information. Across scenarios 1 to 3, the ADNet significantly enhances the recognition ability for small algae targets while maintaining performance on large targets. This improvement demonstrates the effectiveness of the proposed modifications in enhancing the overall capabilities of the distillation model.

In scenarios 4, 5, and 6, the algae features are overshadowed by strong interference on the sea surface. This poses a significant challenge to networks, requiring them to effectively extract algae features amidst such strong interference. However, despite the Segformer network's impressive performance, the student network of SCTNet still exhibits significant issues. Furthermore, in scenarios 7 and 8, the original SCTNet encounters severe false detection problems due to interference from similar color features on the sea surface. It indicates that the potent feature modeling ability of the teacher network fails to transfer its capabilities to the student network. Consequently, the limited modeling capability of the student hinders its performance in complex marine environments. To address this issue, we introduce the CPM, a module specifically designed to concurrently enhance algae features and reduce interference features. As presented in Figure 14, the student trained using our ADNet structure maintains a high true segmentation rate and a low false rate, even in scenes with strong interference, demonstrating its superiority in handling challenging marine scenarios.

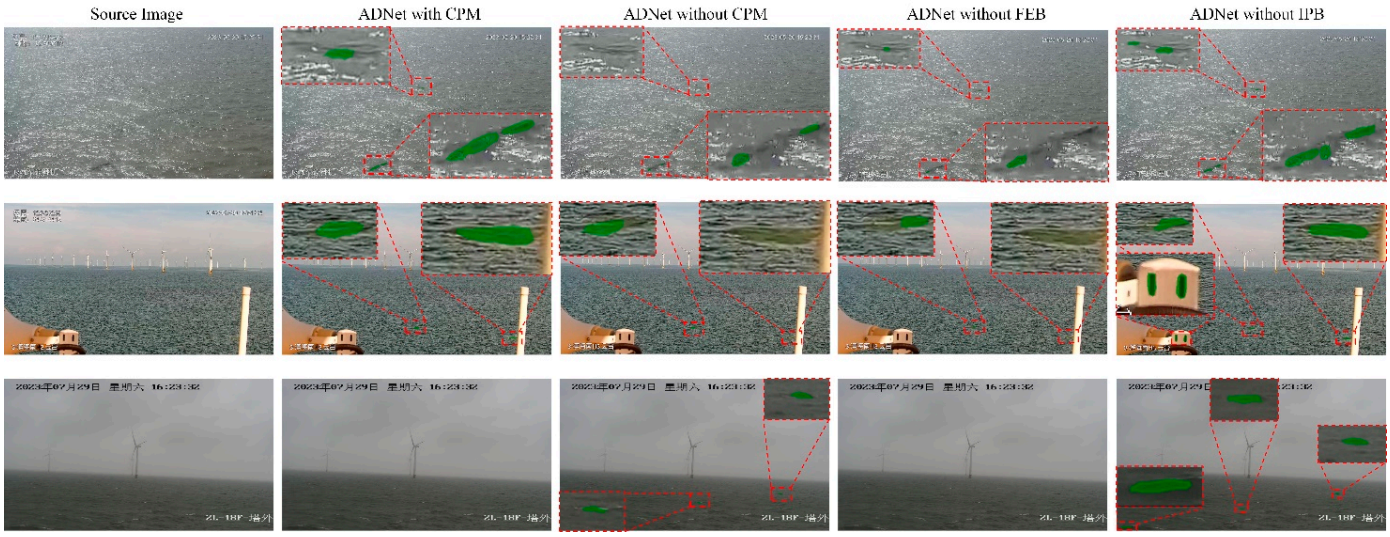

Figure 14. The visualization results of different CPM strategies.

3.3. Ablation Study

3.3.1. The impact of CPM

In this section, we will discuss the impact of CPM on the distillation process for floating algae recognition. To conduct a comprehensive analysis, we implemented the following comparative strategies: First strategy: we completely removed the CPM from the ADNet. Second strategy: we eliminated the feature enhancement branch (FEB) in the CPM. Third strategy: we removed the interference purification branch (IPB) from the CPM. Table 5 summarizes the comparative results obtained through these methods.

The comparison results in Table 5 reveal a significant performance decline when the CPM is totally omitted in our ADNet during the feature encoding phase. Specifically, in the case of the B-Seg100 structure, the absence of CPM leads to a decline of 2.4%. Furthermore, when evaluating the individual contributions of the upper and lower branches in the CPM, it becomes apparent that the IPB plays a more significant role than the FEB. For instance, in the B-Seg100 structure, removing the IPB results in a 1.2% performance loss compared

to removing the FEB. It underscores the importance of deafferent algae and interference features in complex marine environments.

Table 5. The comparison results of different strategies in CPM.

Method Name	Backbone Name	Strategies	mIoU (%)
ADNet	S-Seg50	w/o CPM	52.7
		w/o FEB	54.1
		w/o IPB	53.5
		w/CPM	56.2
	S-Seg75	w/o CPM	55.6
		w/o FEB	56.7
		w/o IPB	55.9
		w/CPM	57.9
	B-Seg50	w/o CPM	57.2
		w/o FEB	58.8
		w/o IPB	57.9
		w/CPM	59.6
	B-Seg75	w/o CPM	58.6
		w/o FEB	60.4
		w/o IPB	59.1
		w/CPM	61.3
	B-Seg100	w/o CPM	59.8
		w/o FEB	61.5
		w/o IPB	60.3
		w/CPM	62.2

Figure 14 offers a visual comparison of segmentation results obtained from different strategies. We can find that the FEB plays a pivotal role in bolstering the capabilities of algae targets, while the IPB excels in mitigating the adverse effects of interference targets. In the first scenario, despite the strong interference caused by sunlight reflection, the model with FEB can successfully identify these algae targets. However, relying solely on feature enhancement can also give rise to potential performance limitations, making the student more prone to false positives. In the second and third scenarios, ADNet equipped with the FEB module exhibited errors in results, which demonstrates that, without the constraints imposed by the IPB, the modeling capabilities for distinguishing between algae and interference features are significantly limited. The removal of interference and invalid background information in the channel dimension can achieve an obvious performance improvement in both the second and third scenes. Consequently, both feature enhancement and purification are deemed crucial in the context of marine environments.

3.3.2. The Impact of Multi-Scale Feature Fusion

The efficacy of the algae segmentation task closely depends on the fusion capability of multi-scale features. In our ADNet framework, the incorporation of the L-MsFFN plays a role in striking an optimal balance between performance and efficiency. This section endeavors to clarify the individual contributions of various fusion methods towards floating algae segmentation performance. Firstly, we present an analysis that examines the disparities between the L-MsFFN and BiFPN. Secondly, we will discuss the functionalities of the PPM modules within the L-MsFFN and BiFPN, respectively. Lastly, we explore the impact of the jumping connections in both the L-MsFFN and BiFPN architectures. Figure 15 illustrates the configurations of the BiFPN (minimal version) and L-MsFFN.

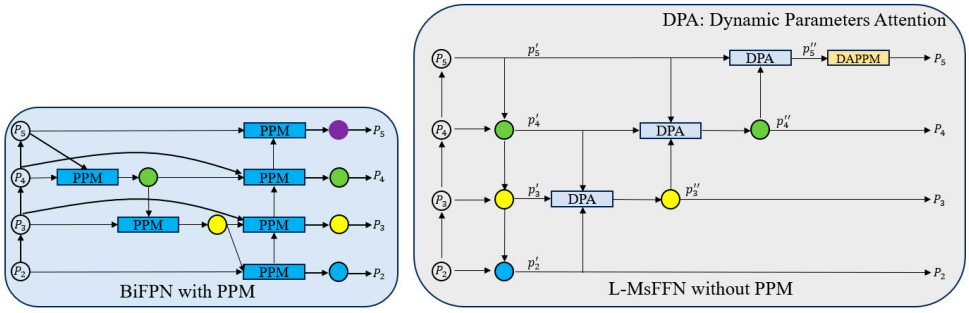

Figure 15. The structure of BiFPN and L-MsFFN w/o PPM.

In addition, to better illustrate the impact of the fusion module on the performance of different algae sizes, we have refined the mIoU metric to incorporate specific size categories: large $mIoU_L$, medium $mIoU_M$, and small $mIoU_S$. The definitions of target sizes align with the reference [29]. The calculations are as follows: $mIoU_L = \frac{1}{N+1}\sum_{i=0}^{N} \frac{p_{i,\,i}}{\sum_{j=0}^{N} p_{i,j} + \sum_{j=0}^{N} \left(p_{j,\,i} - p_{i,\,i}\right)}$, $P \in Algae_L$. Here, $P \in Algae_L$ indicates that the connected domain encompassing the current pixel pertains to a large algae target. The computation procedures for $mIoU_M$ and $mIoU_S$ are analogous to that of $mIoU_L$. The comprehensive results of these comparisons are comprehensively presented in Tables 6 and 7. To facilitate a more intuitive understanding, Figure 16 is referenced, utilizing B-Seg100 as an illustrative example.

Table 6. Comparison of different fusion methods.

Method Name	Parameters (MB)	FLOPs (G)	Inference Time (ms)
SCTNet (Student)	17.4	17.5	6.7
BiFPN	0.14 (0.8%)	7.36 (42.1%)	3.2 (47.8%)
Ms-BiFPN	0.25 (1.4%)	7.42 (42.4%)	3.9 (58.2%)
L-MsFFN	0.50 (2.9%)	1.79 (10.2%)	1.5 (22.4%)

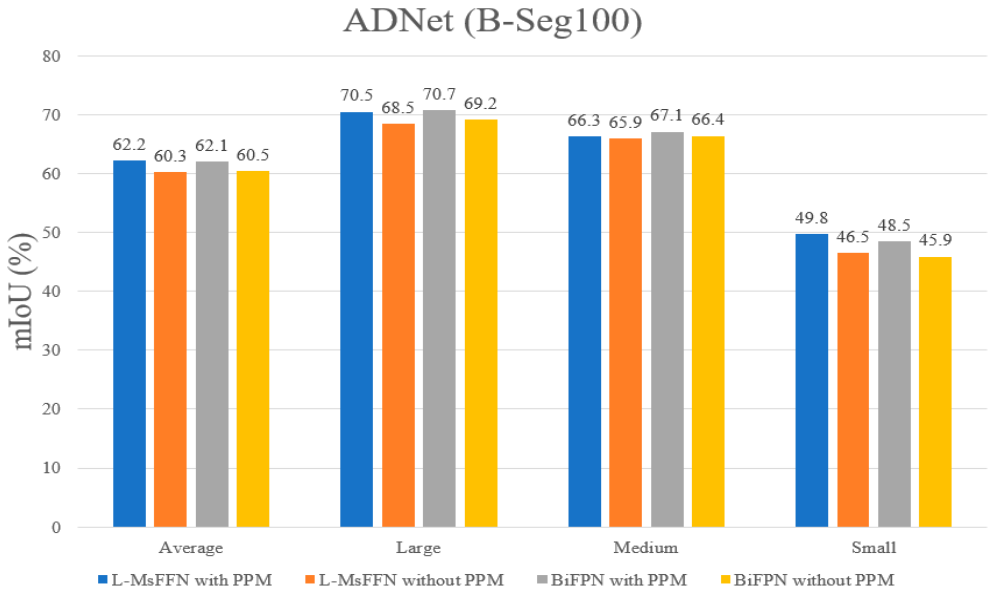

Figure 16. The performance comparison w/o PPM.

In Table 6, using a 1024 × 1024 × 3 input as an example, we conducted a comparative analysis of the computational performance of various fusion methods. Notably, while there was no significant increase in trainable parameters for both BiFPN and Ms-BiFPN, the adoption of the DPA calculation process in the BiFPN-based structure substantially

heightened the computational demands, resulting in a notable decline in inference speed. Directly integrating either Ms-BiFPN or BiFPN into our distillation network would entail this stage accounting for over half of the entire network's inference time.

Table 7. The comparison results of different feature fusion methods.

Network Name	Backbone Name	Fusion Method	PPM	mIoU (%)	$mIoU_L$ (%)	$mIoU_M$ (%)	$mIoU_S$ (%)	Inference Time (ms)
ADNet	S-Seg50	L-MsFFN	√	56.2	69.3	59.1	40.2	12.7
			×	53.9	66.8	56.1	38.8	11.5
		BiFPN	√	56.1	69.8	60.3	38.2	22.1
			×	54.3	67.6	56.4	38.9	19.2
	S-Seg75	L-MsFFN	√	57.9	69.5	61.1	43.1	13.1
			×	55.3	66.8	59.4	39.7	11.7
		BiFPN	√	57.4	69.4	63.2	39.6	23.4
			×	55.7	67.1	59.9	40.1	19.6
	B-Seg50	L-MsFFN	√	59.6	69.6	64.9	44.3	13.8
			×	58.7	68.1	66.8	41.2	12.4
		BiFPN	√	60.2	70.1	68.1	42.4	24.8
			×	58.6	68.3	66.2	41.3	20.4
	B-Seg75	L-MsFFN	√	61.3	70.2	65.8	47.9	16.3
			×	59.9	69.8	66.7	43.2	14.1
		BiFPN	√	61.7	70.4	68.6	46.1	27.6
			×	59.6	69.9	68.4	40.5	23.4
	B-Seg100	L-MsFFN	√	62.2	70.5	66.3	49.8	21.9
			×	60.3	68.5	65.9	46.5	20.1
		BiFPN	√	62.1	70.7	67.1	48.5	33.1
			×	60.5	69.2	66.4	45.9	29.2

The comparison results presented in Table 7 can yield the following noteworthy observations: Firstly, the PPM module, serving as a versatile component, exhibits remarkable adaptability to both L-MsFFN and BiFPN frameworks. Secondly, replacing the DPA with PPM in the BiFPN structure can achieve superior performance compared to L-MsFFN, albeit with a slight compromise in inference speed. Lastly, without the PPM module, the skip connections in the BiFPN structure demonstrate superior performance over L-MsFFN; however, this advantage is attenuated upon the integration of the PPM module.

In contrast to the BiFPN structure, which utilizes a single parameter to capture attention relationships between different pyramid levels, the PPM structure introduces a pixel-level attention calculation approach for feature fusion. Employing the B-Seg75 structure as a benchmark, the BiFPN with PPM structure exhibits a notable 2.1% performance improvement compared to the DPA approach. Notably, the PPM demonstrates a remarkable 3.6% enhancement in small algae targets. This significant improvement suggests that the PPM structure effectively addresses the challenges associated with small-sized algae targets compared to the original DPA approach.

The utilization of finer-grained attention methods in the fusion stage enables the model to accurately capture the distribution of algae targets. This capability is important in the algae segmentation task because the occurrence of algae objects often presents a combination of large and small targets. By incorporating the PPM module, the student model effectively mitigates the suppression problem of small target responses that can occur due to the dominance of large target features in the decoding stage. The segmentation results presented in Figure 17 further demonstrate the improvements achieved by L-MsFFN and BiFPN with the PPM structure. Conversely, a notable issue of missed segmentation

can be observed when DPA is employed in the original BiFPN structure, highlighting the limitations of the DPA approach in handling complex and diverse algae targets.

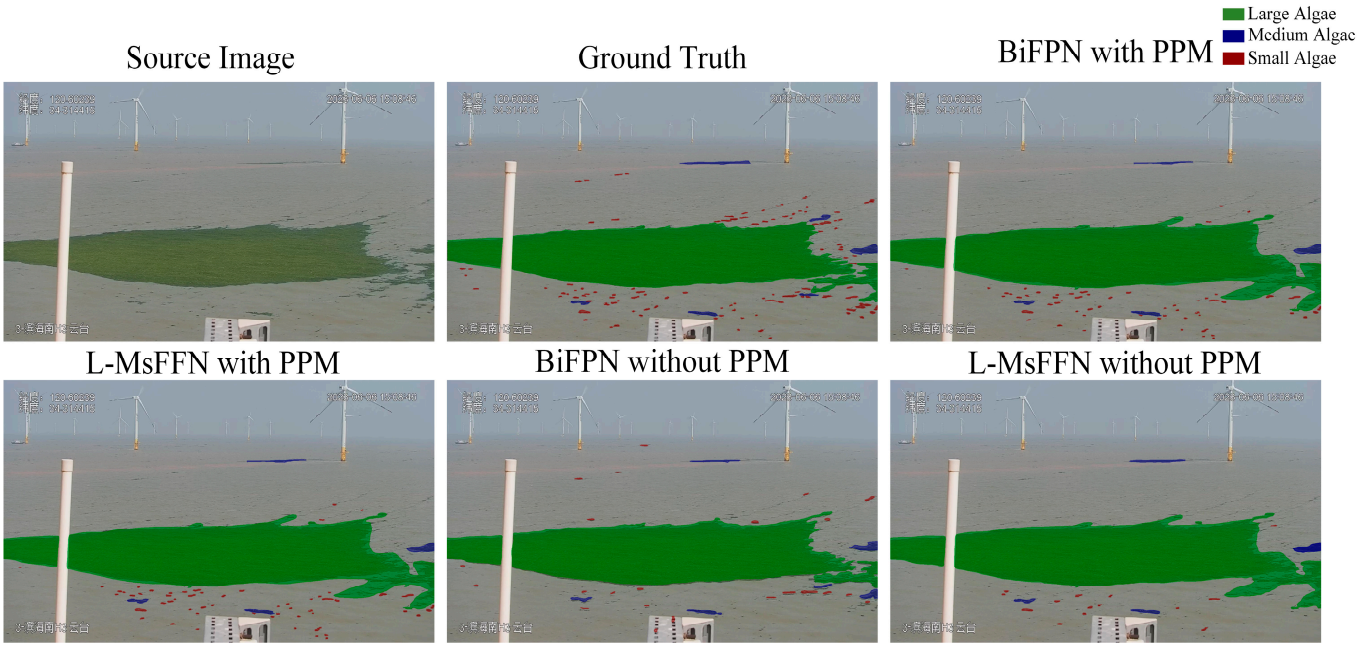

Figure 17. The visualization results of different fusion strategies.

In a horizontal comparison between the L-MsFFN and BiFPN structures, it is evident that L-MsFFN exhibits superior performance in algae segmentation, particularly with its faster inference speed. Using B-Seg100 as an example, L-MsFFN manages to reduce inference speed by a noteworthy 7 ms while achieving a significant 1.7% improvement in the mIoU compared to BiFPN. Additionally, as depicted in Figure 17, a notable issue of missed segmentations arises when L-MsFFN lacks the PPM module. Meanwhile, when comparing L-MsFFN and BiFPN without PPM modules, we found that L-MsFFN exhibited a performance degradation of 0.3%. At the same time, we should note that adding skip connections for small algae targets resulted in a performance decrease of 2.7%. It indicates that skip connections have a beneficial effect on medium and large targets but have a detrimental effect on small ones. Furthermore, skip connections inflate the computational cost of the module, hindering real-time performance. Consequently, we made the decision to eliminate skip connections in our L-MsFFN structure, achieving a balanced trade-off between performance and speed.

3.3.3. The Impact of Distillation Branch Selection

The original SCTNet structure employed feature maps from the top two pyramid layers for knowledge distillation, whereas our proposed ADNet utilized the last three layers. In this section, we will discuss the impact of various pyramid distillation levels on distillation performance. To achieve this, we conducted a comparative analysis involving six distinct distillation structures, as shown in Figure 18, and the results are summarized in Table 8. Through these evaluations, we aim to gain a deeper understanding of the factors that may influence the effectiveness of algae distillation.

In Table 8, it is evident that strategy 2 yields the best performance in both SCTNet and ADNet simultaneously, while strategy 6 exhibits the poorest performance. In strategy 1, all layers in the pyramid participate in the distillation phase. This approach maximizes the involvement of low-level semantic features during the feature decoding phase, thereby enhancing the modeling capability of the student network. The segmentation results for small targets indicate that both SCTNet and ADNet achieve their highest performance in this strategy, reaching 48.2% and 52.6%, respectively. However, this method also encoun-

ters challenges related to erroneous segmentation due to noise features in the low-level pyramids, which block the student network's capacity to model higher-level semantics, ultimately leading to a performance decrease. Therefore, while strategy 1 shows promising results for small targets, the overall performance is not optimal.

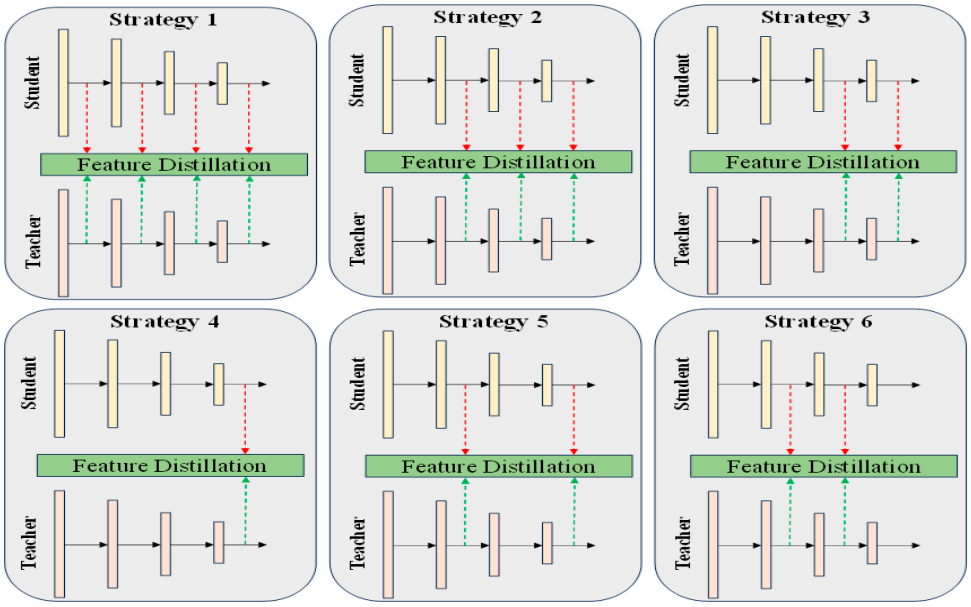

Figure 18. Different algae distillation strategies.

Table 8. The comparison results of different feature distillation strategies.

Backbone Name	Method Name	Distillation Strategy	mIoU (%)	$mIoU_L$ (%)	$mIoU_M$ (%)	$mIoU_S$ (%)
B-Seg100	SCTNet	1	57.3	63.1	60.6	48.2
		2	59.2	67.5	64.5	45.9
		3	58.4	67.8	63.6	43.8
		4	56.1	68.2	59.9	40.2
		5	53.7	65.4	54.2	41.5
		6	52.1	63.9	50.7	41.7
	ADNet	1	60.9	66.3	63.8	52.6
		2	62.2	70.5	66.3	49.8
		3	61.3	71.2	65.5	47.2
		4	58.9	71.6	61.2	43.9
		5	57.6	65.4	64.8	42.6
		6	55.4	63.2	62.5	40.5

To achieve optimal segmentation performance across all algae sizes, it is critical to strike a balance between low-level features and high-level semantics. Building upon strategy 1, strategy 2 opts to omit the distillation of the lowest-level features, enabling the student network to focus on the middle levels of the pyramid. It is widely recognized that both the teacher and student models often contain a large amount of noise and interference information at the lowest level of the pyramid. Distilling these features can potentially damage the high-level valuable semantics, leading to a decrease in performance. Conversely, the second-level features in the pyramid undergo further filtration while still maintaining a responsive representation of small targets. Initiating distillation learning from the second level allows the student network to simultaneously attend to both fine details and abstract semantics across different scales. Illustrating this point, in the case of SCTNet, incorporating features from the second layer enhances the performance from 58.4% to 59.2%, with a noteworthy 2.1% improvement for small algae targets. However, a

reduction in the number of distillation layers, as exemplified in Strategies three and four, leads to a significant performance decline in both SCTNet and ADNet. This underscores the importance of selecting the appropriate distillation layers to maximize performance. Additionally, Figure 11 also underscores the preponderance of small-sized algae targets in our dataset.

In contrast to previous strategies that emphasized distillation of high-level features, Strategies five and six prioritize learning low-level features. These approaches relinquish the guidance from teacher networks on higher-level semantics. However, this shift results in a significant decline in performance, as presented in Table 8. Taking ADNet as an example, strategy 6 exhibits a notable 6.8% decrease in the mIoU compared to strategy 2. Furthermore, despite the distillation network's focus on learning low-level features, the segmentation performance for small-sized algae also suffers a 9.3% decrease in the student network. Figure 19 illustrates numerous false segmentations under Strategies five and six, leading to a significant reduction in the mIoU metrics. This proves that an exclusive focus on low-level pyramids has an adverse impact on the efficiency of algae distillation. It becomes evident that a high-performance distillation strategy should effectively capture both low-level details and high-level semantics in the algae distillation task. Therefore, in our ADNet, we opt for strategy 2 as the implementation method, as it can achieve superior performance.

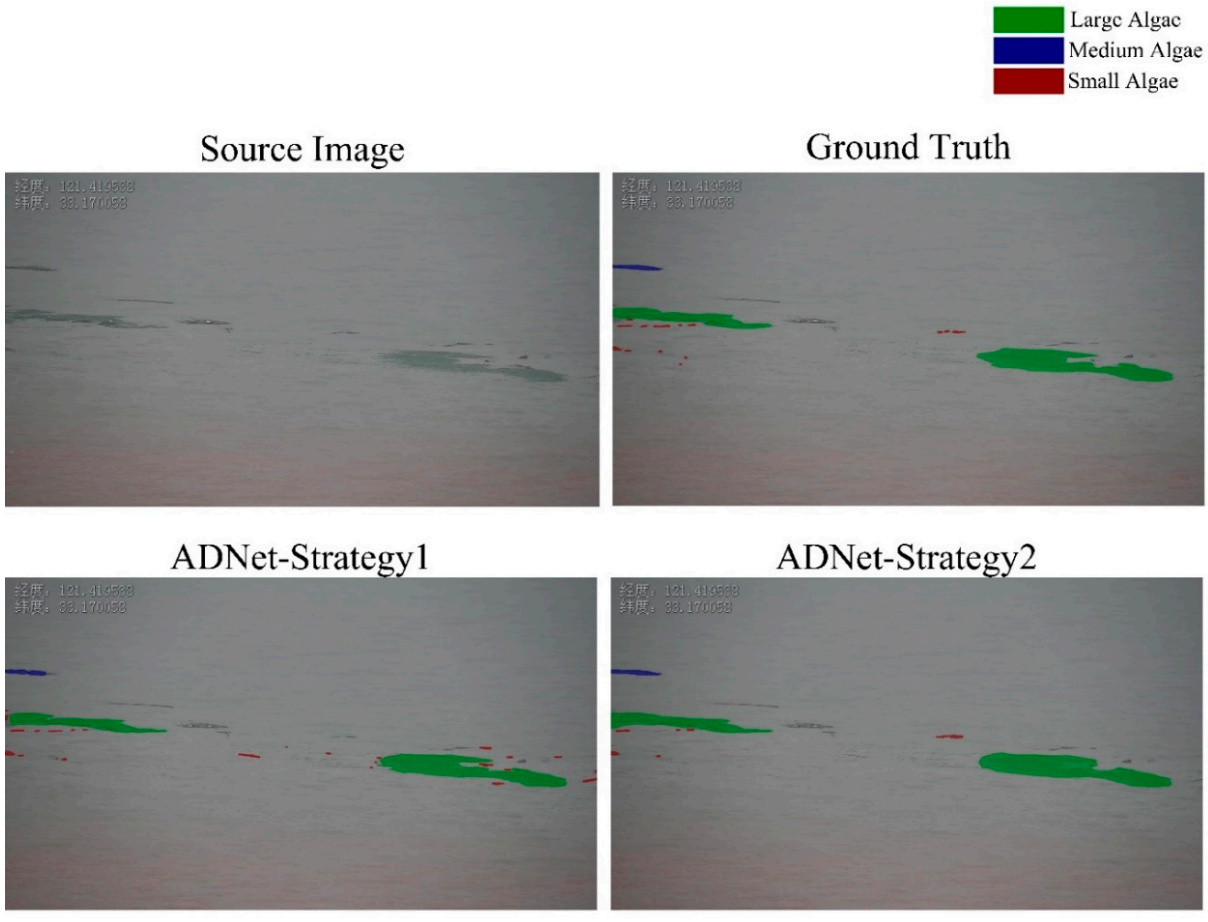

Figure 19. *Cont.*

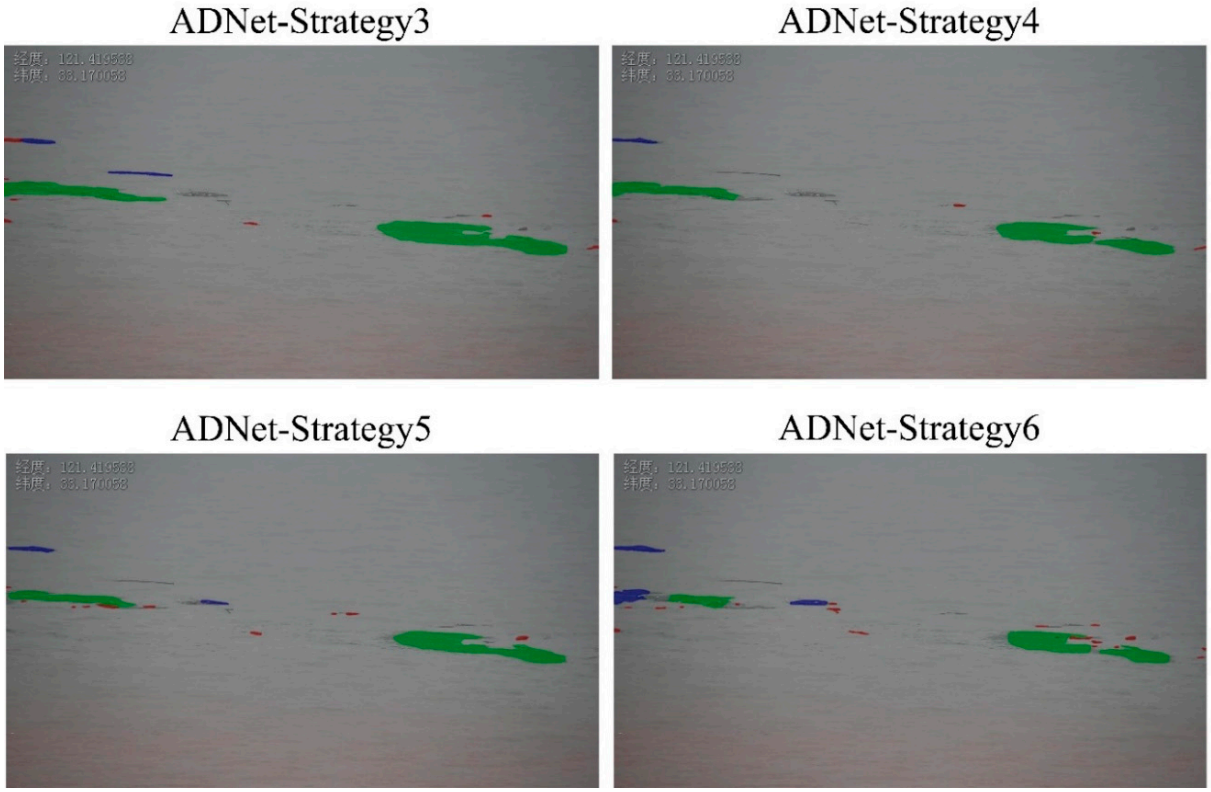

Figure 19. The visualization results of different distillation strategies.

4. Discussion

In previous research, methods such as object detection, semantic segmentation, and instance segmentation have been widely employed to detect algae targets on the sea surface. In Figures 20 and 21, we present a comparison of performance and speed disparities among these methods. For object detection and instance segmentation, we utilize the COCO dataset, while semantic segmentation is evaluated using the cityscapes dataset. The comparison results reveal that instance segmentation methods do not exhibit advantages in terms of either performance or speed compared to object detection methods. Meanwhile, semantic segmentation models, which focus on segmenting targets within the same category, can offer faster speeds and superior performance compared to instance segmentation methods. Meanwhile, the speed of semantic segmentation approaches can achieve approximately 20 frames per second (FPS) faster than instance segmentation theories. Consequently, semantic segmentation techniques have emerged as the preferred choice for the current floating algae detection task.

In Figure 21, CNN-based segmentation methods like STDC and DDRNet have shown notable speed advantages; however, it is important to acknowledge that their performance often falls behind transformer-based approaches like Segformer. Transformers, known for their ability to capture long-range dependencies, demonstrated robust resilience to interference even in complex environments. When confronted with the intricate and dynamic marine conditions, the algae segmentation model must integrate these critical capabilities inherent in transformer structures.

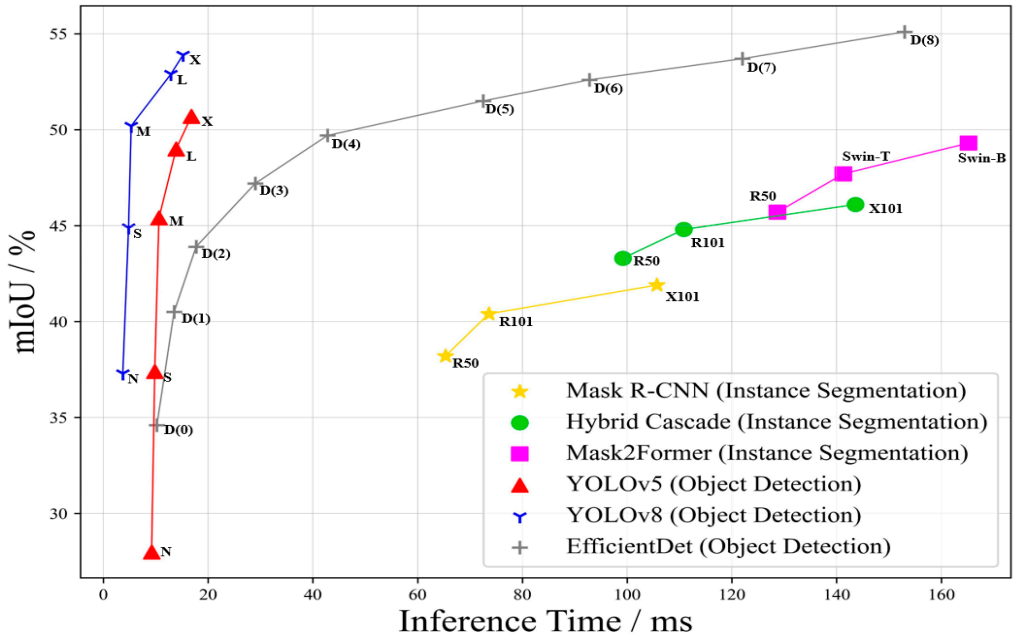

Figure 20. The comparison of different detection methods.

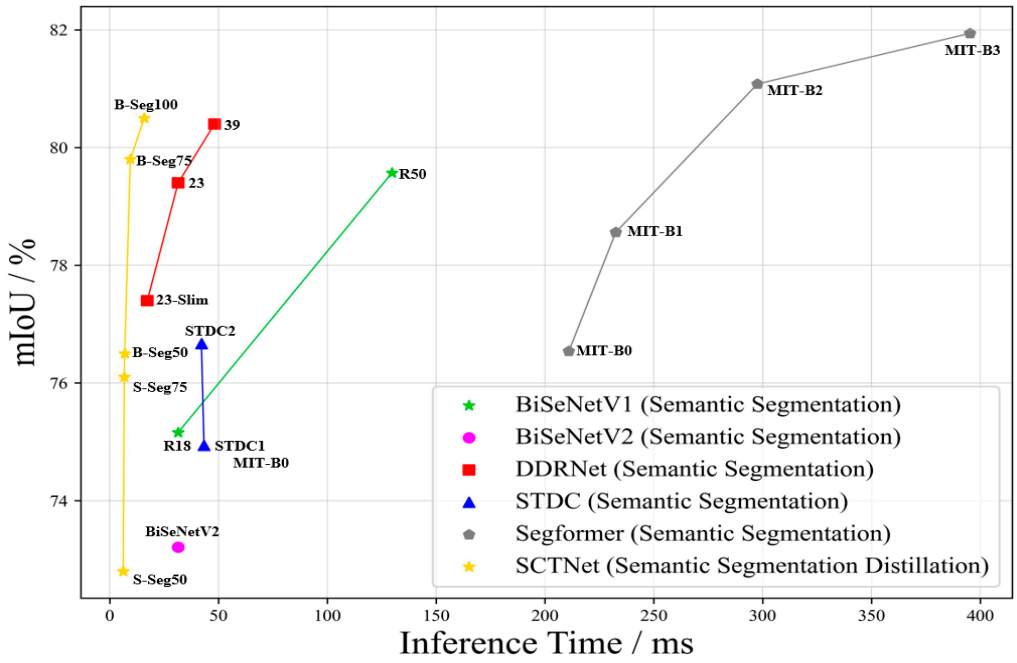

Figure 21. The comparison of different segmentation methods.

To achieve this objective, we introduce the algae distillation method called ADNet in this paper, aiming to attain performance comparable to the transformer-based models on the CNN structures, while preserving its swift inference speed and low deployment cost. In Figure A1 (Appendix A), we present the visualization results of various methods in practical algae monitoring applications. It is evident that ADNet exhibits substantial performance advantages compared to other methods. Moreover, from the perspective of the speed dimension, as we discussed in the Results section, ADNet achieves performance enhancement while maintaining the speed advantage, thereby attaining an optimal balance between performance and speed.

However, our ADNet still faces several challenges, particularly in the following aspects: Firstly, the performance of the teacher network directly influences the student

network's performance due to the distillation structure. Consequently, any incorrect knowledge embedded in the teacher network can be seamlessly transmitted to the student model without any filtering mechanism, ultimately leading to a decrease in the student's performance. As illustrated in Figure 22, we visually compare the segmentation results of ADNet, Segformer, and OCRNet. These comparisons reveal that the teacher network's misidentifications can directly contribute to the errors in ADNet. Conversely, traditional semantic segmentation models do not exhibit similar errors as their training solely relies on the truth-labeled samples. In our ADnet structure, we directly analyze the feature disparities between the teacher and student networks to regulate the student's training behavior. However, this process not only captures the correct features but also incorporates the incorrect, harmful features from the teacher network. Meanwhile, ADNet introduces a purification concept that selectively models the interference and multi-scale features from the perspective of the student network. This purification concept can be further extended to facilitate the selective learning of teacher features during the distillation process. Therefore, we will explore a purification approach in our distillation structure that enables the student to discerningly acquire knowledge from the teacher.

Additionally, although the incorporation of L-MsFFN and PPM in ADNet has achieved improvements from the perspective of multi-scale modeling ability, notable performance gaps persist between CNNs and transformers, particularly in the segmentation of small algae objects. A comparative analysis of the performance between the teacher and the student model on small targets, as presented in Figure 23, reveals that the teacher network, with its transformer structure, exhibits superior detection capabilities. Due to the limitations of CNNs in modeling multi-scale features during the encoding stage, ADNet still struggles with missed detections. Segformer, by integrating high-resolution detailed features with low-resolution abstract features within its attention framework, has achieved remarkable improvements in the detection of small objects. This multi-scale modeling capability in the feature extraction stage remains challenging for our ADNet to replicate because of CNNs' limitations. Therefore, despite ADNet's attempt to leverage the distillation method and multiscale fusion module to forcibly constrain the CNNs to emulate the transformers, it fails to obtain satisfactory performance on small targets. The integration of cost-effective multi-scale modeling techniques with attention-based feature modeling methods is a crucial research topic in the field of floating algae monitoring. Future endeavors should focus on bridging these performance gaps and enhancing the overall accuracy and reliability of our algae monitoring system.

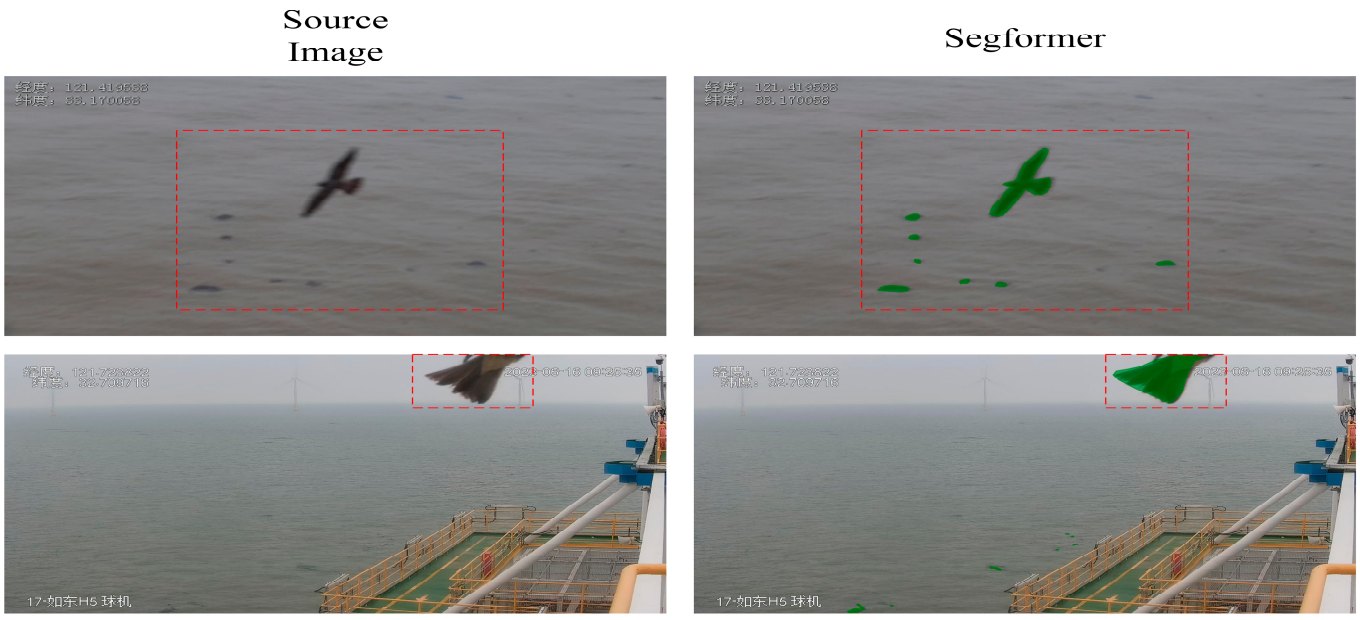

Figure 22. *Cont.*

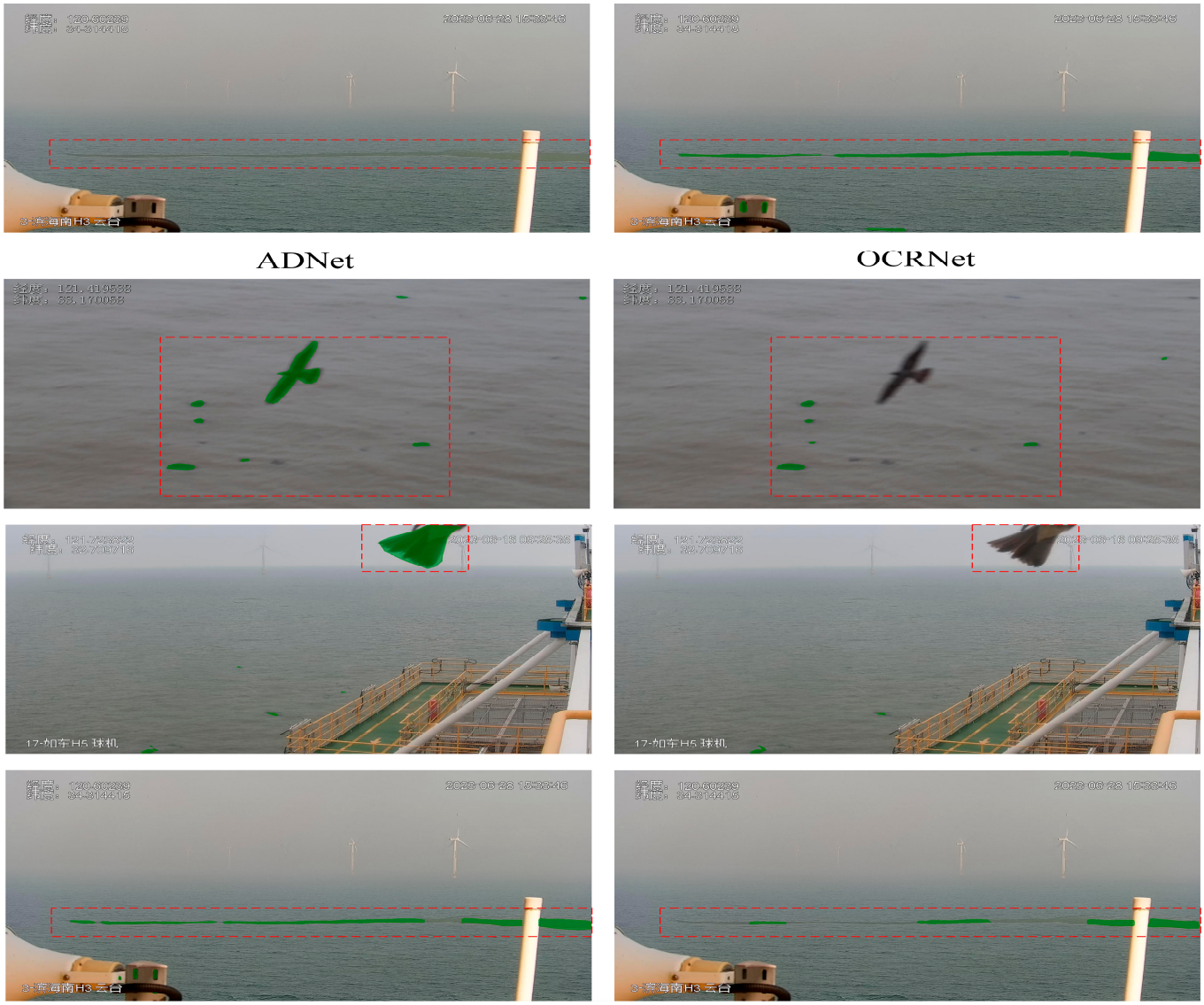

Figure 22. The impact of the teacher on the student during the distillation stage.

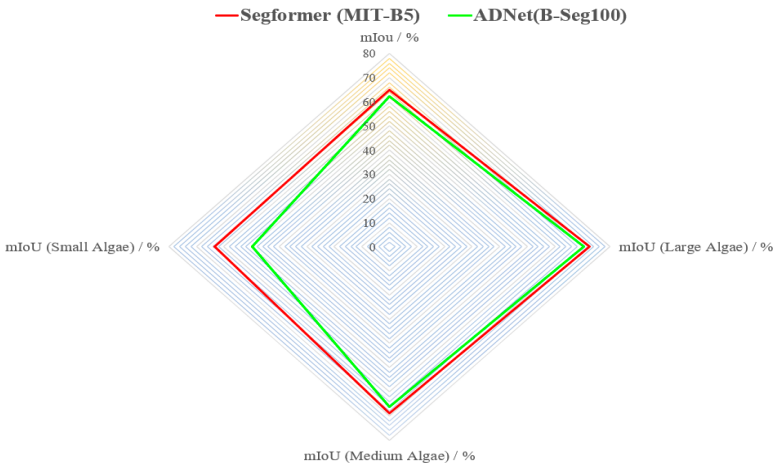

Figure 23. The performance disparities between Segformer and ADNet.

5. Conclusions

In this paper, a novel algae segmentation distillation network named ADNet is proposed. In response to the challenges posed by complex marine environments, we introduce the CPM to the student network. The CPM integrates the FEB and the IPB into a unified framework. The FEB utilizes max-pooling operations to capture foreground algae features against the surrounding background, while the IPB incorporates average-pooling operations to reduce interference responses. The segmentation results demonstrate that the CPM module effectively enhances the response of algae targets while simultaneously diminishing the weight of interfering elements. Furthermore, the CPM achieves this operation without increasing any learnable parameters.

Secondly, considering the huge scale variations of floating algae targets in RGB images due to the cameras' installation angles and positions, we propose the L-MsFFN. L-MsFFN addresses the issue of large target features overshadowing smaller targets during the decoding stage. Specifically, by eliminating skip connections in the original BiFPN structure, the computational complexity of L-MsFFN is effectively reduced. Meanwhile, by replacing the DPA with the PPM, L-MsFFN achieves pixel-level attention control across different pyramid levels. Despite the incorporation of the CPM and L-MsFFN methods into the student network simultaneously, ADNet does not exhibit a notable decrease in speed compared to the original distillation network, demonstrating an optimal balance between performance and speed in the floating algae monitoring system.

In summary, the main contributions of our ADNet are as follows:

(1) We introduce the distillation theory into the floating algae monitoring task in complex marine environments.
(2) A novel channel purification module named CPM is proposed to simultaneously enhance algae semantics while purifying interference features.
(3) We propose a lightweight multi-scale feature fusion network, termed L-MsFFN, to enhance the modeling capability of multi-scale features, reducing the scale-capturing gap between the transformers and CNNs.
(4) A novel position purification module, termed PPM, is introduced to replace the conventional DPA approach during the fusion stage, enhancing the effectiveness and accuracy of L-MsFFN in controlling features across different pyramids.
(5) Extensive experimental results demonstrate that our ADNet can achieve state-of-the-art performance compared to other methods in the floating algae segmentation task.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the editors and the reviewers for their valuable suggestions.

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

Appendix A

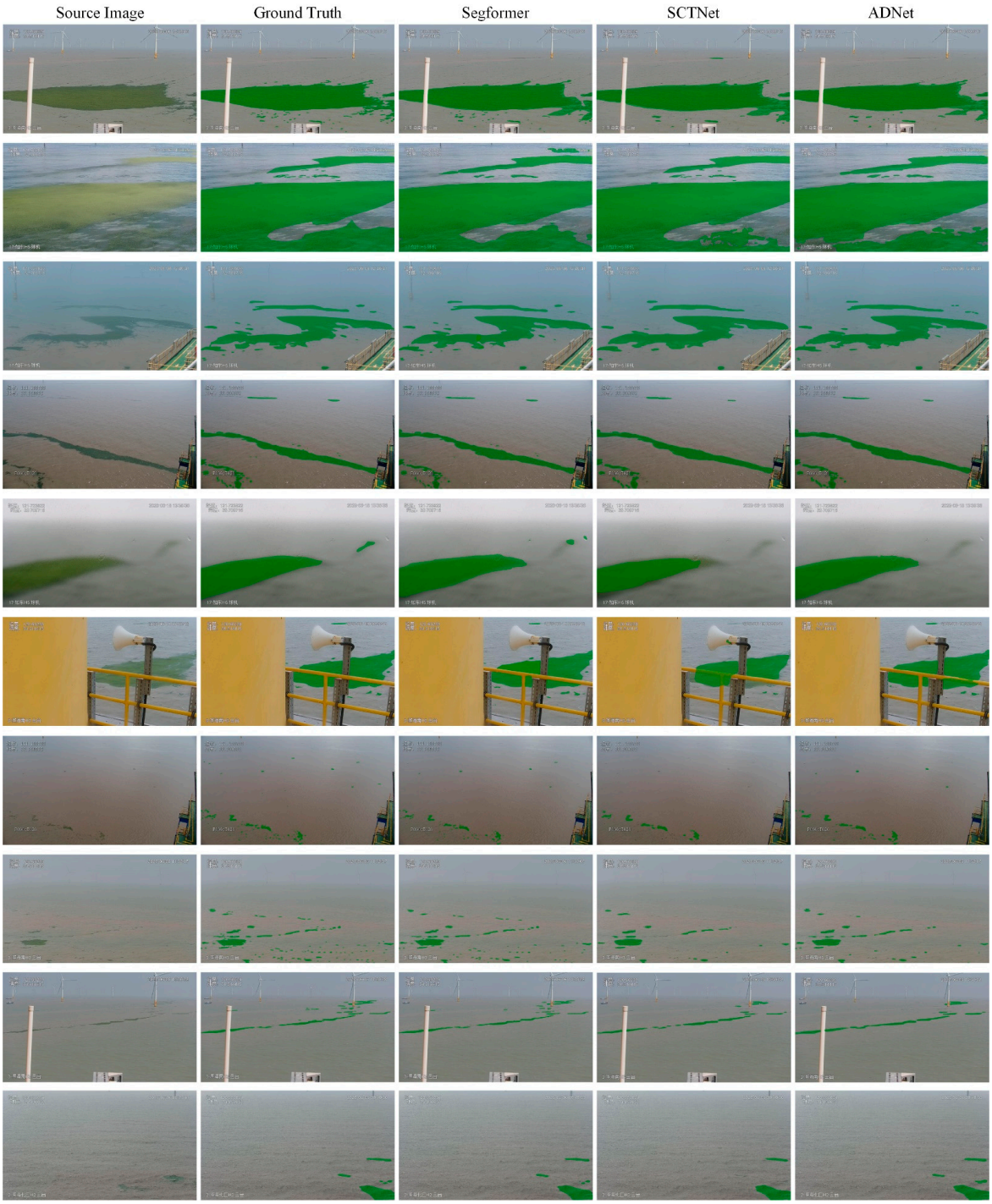

Figure A1. The segmentation results of different distillation models in different marine environments.

References

1. Cuevas, E.; Uribe-Martínez, A.; Liceaga-Correa, M.A. A satellite remote-sensing multi-index approach to discriminate pelagic Sargassum in the waters of the Yucatan Peninsula, Mexico. *Int. J. Remote Sens.* **2018**, *39*, 3608–3627. [CrossRef]
2. 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]
3. Ananias, P.H.M.; Negri, R.G. Anomalous behaviour detection using one-class support vector machine and remote sensing images: A case study of algal bloom occurrence in inland waters. *Int. J. Digit. Earth* **2021**, *14*, 921–942. [CrossRef]

4. Barrientos-Espillco, F.; Gascó, E.; López-González, C.I.; Gómez-Silva, M.J.; Pajares, G. Semantic segmentation based on Deep learning for the detection of Cyanobacterial Harmful Algal Blooms (CyanoHABs) using synthetic images. *Appl. Soft Comput.* **2023**, *141*, 110315. [CrossRef]
5. 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]
6. 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]
7. 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]
8. Pan, B.; Shi, Z.; An, Z.; Jiang, Z.; Ma, Y. A novel spectral-unmixing-based green algae area estimation method for GOCI data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *10*, 437–449. [CrossRef]
9. Qi, L.; Hu, C. To what extent can Ulva and Sargassum be detected and separated in satellite imagery? *Harmful Algae* **2021**, *103*, 102001. [CrossRef]
10. 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]
11. Cui, T.W.; Liang, X.J.; Gong, J.L.; Tong, C.; Xiao, Y.F.; Liu, R.J.; Zhang, X.; Zhang, J. Assessing and refining the satellite-derived massive green macro-algal coverage in the Yellow Sea with high resolution images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *144*, 315–324. [CrossRef]
12. Yang, C.; Tan, Z.; Li, Y.; Shen, M.; Duan, H. A comparative analysis of machine learning methods for algal Bloom detection using remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7953–7967. [CrossRef]
13. Podlejski, W.; Descloitres, J.; Chevalier, C.; Minghelli, A.; Lett, C.; Berline, L. Filtering out false Sargassum detections using context features. *Front. Mar. Sci.* **2022**, *9*, 960939. [CrossRef]
14. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
15. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic detr: End-to-end object detection with dynamic attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2988–2997.
16. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
17. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
18. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
19. 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.
20. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef] [PubMed]
21. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Lin, D. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
22. Park, J.; Baek, J.; Kim, J.; You, K.; Kim, K. Deep learning-based algal detection model development considering field application. *Water* **2022**, *14*, 1275. [CrossRef]
23. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048.
24. Liu, D.; Wang, P.; Cheng, Y.; Bi, H. An improved algae-YOLO model based on deep learning for object detection of ocean microalgae considering aquacultural lightweight deployment. *Front. Mar. Sci.* **2022**, *9*, 1070638. [CrossRef]
25. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 116–131.
26. Wang, X.; Wang, L.; Chen, L.; Zhang, F.; Chen, K.; Zhang, Z.; Zhao, L. AlgaeMask: An instance segmentation network for floating algae detection. *J. Mar. Sci. Eng.* **2022**, *10*, 1099. [CrossRef]
27. Lee, Y.; Hwang, J.W.; Lee, S.; Bae, Y.; Park, J. An energy and GPU-computation efficient backbone network for real-time object detection. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; p. 00103.
28. Zou, Y.; Wang, X.; Wang, L.; Chen, K.; Ge, Y.; Zhao, L. A High-Quality Instance-Segmentation Network for Floating-Algae Detection Using RGB Images. *Remote Sens.* **2022**, *14*, 6247. [CrossRef]
29. 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]
30. 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]

31. 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]
32. 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.
33. Liu, R.; Cui, B.; Dong, W.; Fang, X.; Xiao, Y.; Zhao, X.; Wang, Q. A refined deep-learning-based algorithm for harmful-algal-bloom remote-sensing recognition using Noctiluca scintillans algal bloom as an example. *J. Hazard. Mater.* **2024**, *467*, 133721. [CrossRef]
34. Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; Zhang, Q. Cross-image relational knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 12319–12328.
35. Dong, Z.; Gao, G.; Liu, T.; Gu, Y.; Zhang, X. Distilling Segmenters from CNNs and Transformers for Remote Sensing Images Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5613814. [CrossRef]
36. Xu, Z.; Wu, D.; Yu, C.; Chu, X.; Sang, N.; Gao, C. SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 6378–6386.
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
38. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
39. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; p. 00246.
40. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
41. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9716–9725.