remote sensing

MDPI

Article

Algal Bed Region Segmentation Based on a ViT Adapter Using Aerial Images for Estimating CO_2 Absorption Capacity

Guang Li [1], Ren Togo [2], Keisuke Maeda [3], Akinori Sako [4], Isao Yamauchi [5], Tetsuya Hayakawa [6], Shigeyuki Nakamae [1], Takahiro Ogawa [2] and Miki Haseyama [2,*]

1 Education and Research Center for Mathematical and Data Science, Hokkaido University, N-12, W-7, Kita-ku, Sapporo 060-0812, Hokkaido, Japan; guang@lmd.ist.hokudai.ac.jp (G.L.); s.nakamae@mdsc.hokudai.ac.jp (S.N.)
2 Faculty of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo 060-0814, Hokkaido, Japan; togo@lmd.ist.hokudai.ac.jp (R.T.); ogawa@lmd.ist.hokudai.ac.jp (T.O.)
3 Data-Driven Interdisciplinary Research Emergence Department, Hokkaido University, N-13, W-10, Kita-ku, Sapporo 060-0813, Hokkaido, Japan; maeda@lmd.ist.hokudai.ac.jp
4 Alpha Hydraulic Engineering Consultants Co., Ltd., Hassamu-14-9, Nishi-ku, Sapporo 063-0829, Hokkaido, Japan; sako@ahec.jp
5 Cold Regions Air and Sea Ports Engineering Research Center, N-11, W-2, Kita-ku, Sapporo 001-0011, Hokkaido, Japan; i_yamauchi@kanchi.or.jp
6 Hokkaido Regional Development Bureau, N-8, W-2, Kita-ku, Sapporo 060-8511, Hokkaido, Japan; hayakawa-t22ac@mlit.go.jp
* Correspondence: mhaseyama@lmd.ist.hokudai.ac.jp

Abstract: In this study, we propose a novel method for algal bed region segmentation using aerial images. Accurately determining the carbon dioxide absorption capacity of coastal algae requires measurements of algal bed regions. However, conventional manual measurement methods are resource-intensive and time-consuming, which hinders the advancement of the field. To solve these problems, we propose a novel method for automatic algal bed region segmentation using aerial images. In our method, we use an advanced semantic segmentation model, a ViT adapter, and adapt it to aerial images for algal bed region segmentation. Our method demonstrates high accuracy in identifying algal bed regions in an aerial image dataset collected from Hokkaido, Japan. The experimental results for five different ecological regions show that the mean intersection over union (mIoU) and mean F-score of our method in the validation set reach 0.787 and 0.870, the IoU and F-score for the background region are 0.957 and 0.978, and the IoU and F-score for the algal bed region are 0.616 and 0.762, respectively. In particular, the mean recognition area compared with the ground truth area annotated manually is 0.861. Our study contributes to the advancement of blue carbon assessment by introducing a novel semantic segmentation-based method for identifying algal bed regions using aerial images.

Keywords: blue carbon; algal beds; carbon dioxide absorption; semantic segmentation; ViT adapter

Citation: Li, G.; Togo, R.; Maeda, K.; Sako, A.; Yamauchi, I.; Hayakawa, T.; Nakamae, S.; Ogawa, T.; Haseyama, M. Algal Bed Region Segmentation Based on a ViT Adapter Using Aerial Images for Estimating CO_2 Absorption Capacity. *Remote Sens.* **2024**, *16*, 1742. https://doi.org/10.3390/rs16101742

Academic Editor: Hossein M. Rizeei

Received: 18 March 2024
Revised: 28 April 2024
Accepted: 13 May 2024
Published: 14 May 2024

1. Introduction

Global warming, predominantly attributed to human-induced factors, has led to a discernible increase in the Earth's average temperature [1,2]. This trend has resulted in various harmful effects, such as the melting of polar ice caps, rising sea levels, and more frequent extreme weather events [3]. In this scenario, developing effective carbon dioxide (CO_2) absorption strategies is vital. "Blue carbon", which refers to CO_2 absorption in oceanic and coastal ecosystems, has emerged as an important element in combating atmospheric levels, with coastal algae playing a critical role because of their high CO_2 absorption capacity [4,5].

Effective use of coastal algae in blue carbon strategies requires accurate measurement of CO_2 absorption ability [6,7]. Accurate determination of the area of the algal bed regions

is a critical step in measuring their CO_2 absorption capacity. The extent of these algal beds is an important metric of their potential for CO_2 absorption; larger algal bed areas generally suggest higher CO_2 absorption ability [8]. This relationship between the size of algal beds and their CO_2 absorption ability shows the importance of accurate area measurements in evaluating their role in blue carbon strategies [9]. Several studies have demonstrated the feasibility of identifying algal beds by analyzing aerial or unmanned aerial vehicle images, even though their primary focus has been on the detection of seaweed presence rather than blue carbon assessment [10,11].

In this study, we tackle the challenge of evaluating the capacity of coastal algae for CO_2 absorption, with a particular emphasis on the automated recognition of algal bed areas [12,13]. Conventional manual methods for surveying these regions, although thorough, are time-consuming and resource-intensive. Such methods typically prove unsuitable for widespread application because of their laborious nature [14]. By leveraging remote sensing and deep learning methods, our method provides a more efficient, scalable, and accurate solution for algal bed region segmentation [15–17]. Our method not only streamlines the process but also improves the feasibility of implementing these strategies on a broader scale.

In our method, we use an advanced semantic segmentation model named the vision transformer adapter (ViT adapter) to effectively identify algal-specific features [18]. The original ViT [19], typically trained on large-scale open datasets, is optimized for specific tasks through the integration of a ViT adapter. This adaptation comprises three components: the spatial prior (SP) module, which captures spatial characteristics from input images; the spatial feature injector (SFI) module, which is designed to infuse these characteristics into ViT; the multiscale feature extractor (MFE), which generates hierarchical features from the ViT output. These modules allow our method to adjust to various conditions, such as different algal bed states, lighting scenarios, and distinct coastal ecosystems, thereby significantly improving the segmentation accuracy. Using the proposed method, we can accurately identify algal bed regions from other coastal and marine elements. We also design a patch-based training strategy to fully utilize aerial image information, which can further improve algal bed region segmentation performance.

Our study findings show that the proposed method can accurately recognize algal bed regions using aerial images to estimate the CO_2 absorption capacity, representing an enhancement over conventional methodologies. The proposed method was trained and tested using an aerial image dataset collected from Hokkaido, Japan. The experimental results show that our method achieves a mIoU of 0.787 and a mean F-score (mF-score) of 0.870. Specifically, in algal bed regions, the IoU and F-score were 0.616 and 0.762, respectively, whereas the background region scored higher with an IoU of 0.957 and an F-score of 0.978. The mean recognition area compared with the ground-truth area annotated manually was 0.861. Furthermore, the scalable nature of our method allows for its effective integration into extensive coastal monitoring systems, thereby enhancing our understanding and management of blue carbon dynamics across vast coastal areas. To the best of our knowledge, this is the first time to explore algal bed region segmentation for estimating CO_2 absorption capacity. We compared with other segmentation methods and proved the superiority of the proposed method.

2. Material

2.1. Study Area

The study area, which extends from Yoichi Town to Shimamaki Village along the Hokkaido coastline, is shown in Figure 1. The varied coastal geography, ranging from rugged cliffs to sandy beaches, offers diverse habitats for algal colonization. The geographical coordinates for this area are defined by the JGD2011 geodetic datum (https://www.gsi.go.jp/sokuchikijun/tky2jgd.html, accessed on 12 May 2024), ensuring accurate location mapping. The eastern and western boundaries are marked by longitudes 140.51 and 139.48, respectively, and the northern and southern boundaries are at latitudes 43.23 and 42.37, respectively.

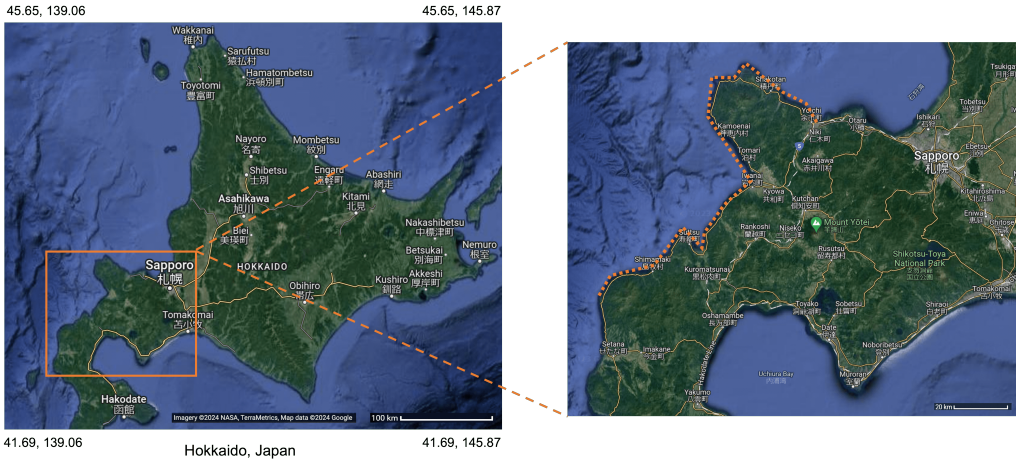


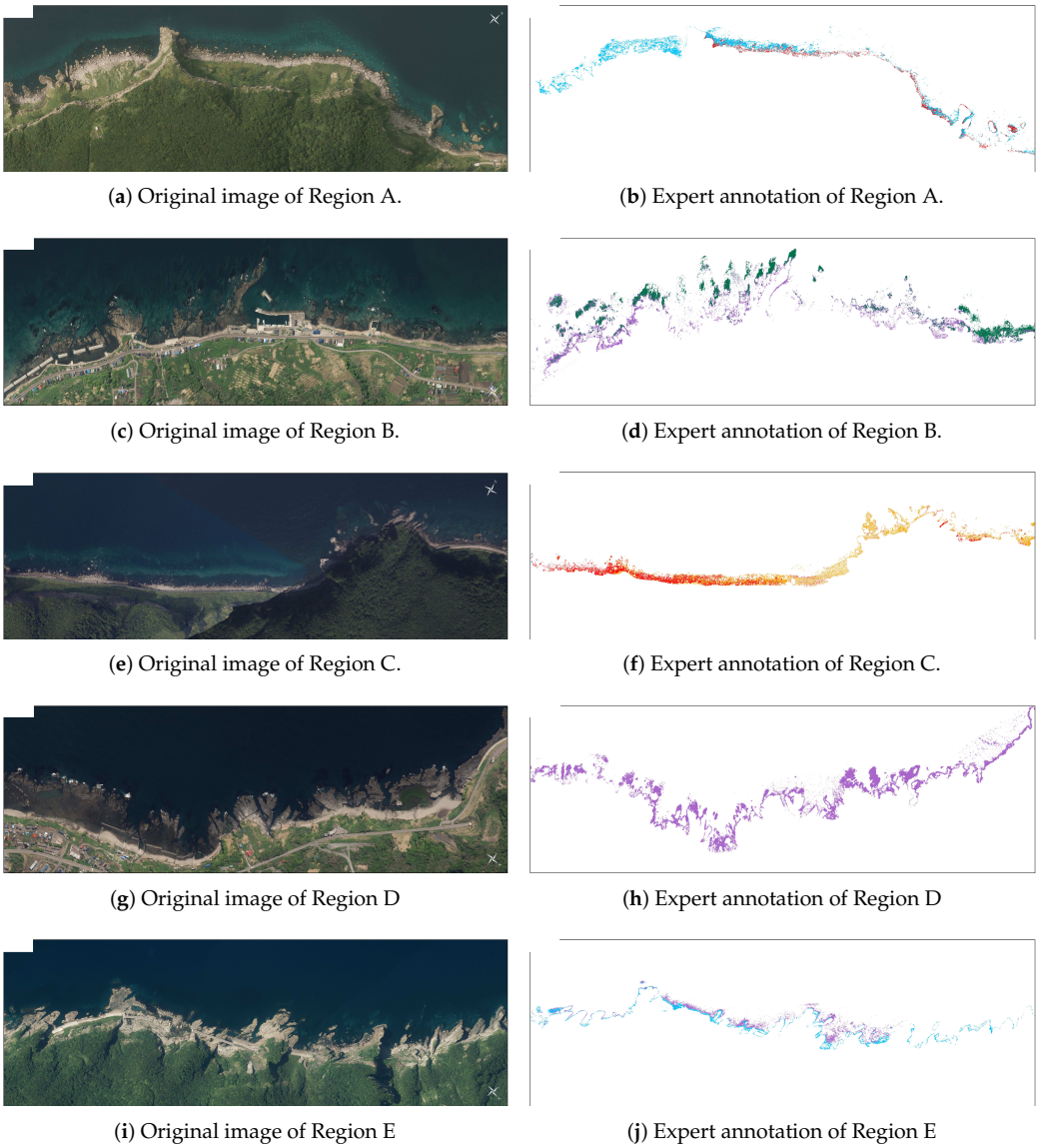
Figure 1. Schematic of the study area.

2.2. Data Collection

The aerial photography process was performed on specific dates in the spring (May) and summer (August and September) of 2015 along the coastline from Yoichi Town to Shimamaki Village, including seven ecological regions. We used an UltraCamEagle digital aerial camera mounted on a Cessna TU-206G aircraft. The camera was set to a ground resolution of 30 cm/pixel, which is aligned with the planned photography altitude and flight paths. Before each photography session, we performed thorough checks of wave conditions, weather, solar altitude, and direction to minimize the halation effect on water surfaces. This involved adjusting the sequence of flight routes to minimize the mirror-like reflection of sunlight.

Furthermore, the aerial photography process maintained an altitude of 4604 m, covering all 12 planned courses. The images were then processed using specialized software to produce high-resolution monochrome and color images. Furthermore, simplified orthophoto images were obtained using numerical elevation models, ensuring an accurate representation of the coastal terrain (https://geospatial.trimble.com/zh-cn/products/software/trimble-inpho, accessed on 12 May 2024). The processed images adhered to public survey standards, offering a reliable and detailed visualization of the coastal region, which is essential for successful mapping and analysis of the distribution of the algal bed region.

Consequently, we acquired 195 aerial images. Expert analysts manually marked the algal bed regions within the images. These expert annotations were used to create binary masks for segmentation. Of the total, five images of different ecological regions were chosen as the validation set, whereas the remaining images were allocated for the training set. Figure 2 shows the original images and expert annotations of the validation set, namely, Shimamaki, Suttu, Iwanai, Tomari, and Kamoenai, which are referred to as Regions A–E. Note that the different colors show different types of algae, and we do not use the information in this study.



(**a**) Original image of Region A. (**b**) Expert annotation of Region A.

(**c**) Original image of Region B. (**d**) Expert annotation of Region B.

(**e**) Original image of Region C. (**f**) Expert annotation of Region C.

(**g**) Original image of Region D (**h**) Expert annotation of Region D

(**i**) Original image of Region E (**j**) Expert annotation of Region E

Figure 2. Original images and expert annotations for Regions A–E.

3. Methodology

Semantic segmentation of algal bed regions from aerial images is a complex process that plays a crucial role in ecological monitoring and environmental management. As shown in Figure 3, the proposed method leverages the ViT adapter, thereby refining the performance of the model in identifying algal-specific features within the images.

3.1. Patch Splitting

The initial step involves processing the raw aerial images using a patch-splitting algorithm. The process segments each image into smaller, overlapping patches to preserve semantic continuity. The patch generation process can be defined as follows:

$$P_i = f_{\text{overlap}}(I, s, o) \quad \forall i \in \{1, 2, \ldots, n\}, \tag{1}$$

where P_i denotes the i-th patch, s represents the patch size, o represents the overlap between patches, and N represents the total number of patches obtained from the aerial image I. The patch-splitting algorithm ensures the preservation of semantic continuity across patches, enabling more detailed and accurate aerial images.

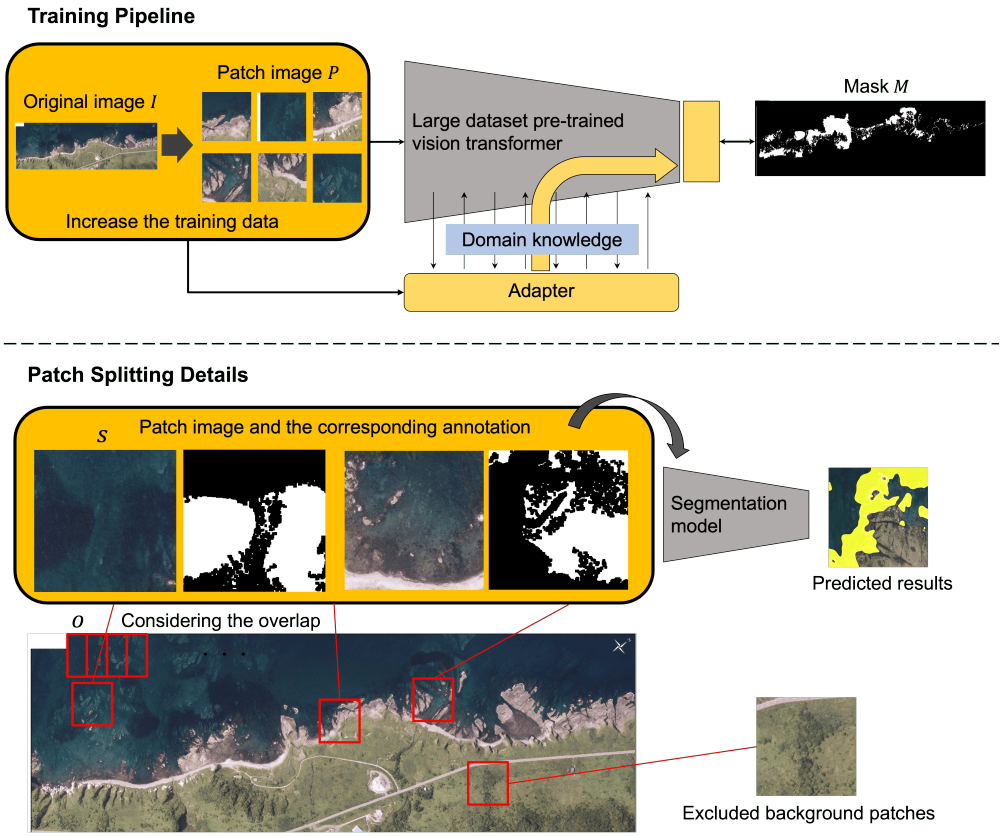


Figure 3. Overview of the proposed method. The upper panel shows the training pipeline, and the lower panel shows the details of the patch-splitting process.

3.2. Vision Transformer with an Adapter

Our model architecture consists of two main components: a ViT and a ViT adapter. The network architecture of the proposed method and the data flow between modules are shown in Figure 4. Following the design principles of Dosovitskiy et al. [19], the ViT starts with a patch-embedding process, where each input patch P is divided into nonoverlapping segments with c pixels. These segments are then flattened (Flatten) and converted into D-dimensional tokens, effectively reducing the feature resolution. The tokens $\mathbf{T}$ are concatenated (Concat) with the corresponding position embeddings $\mathbf{E}$ and then processed through several transformer encoder layers as follows:

$$\mathbf{F}_{\text{vit}}^{j} = \text{Concat}(\text{Flatten}(\mathbf{T}, \mathbf{E})), \tag{2}$$

where $\mathbf{F}_{\text{vit}}^{j}$ denotes the features extracted at the j-th block within ViT from patch P.

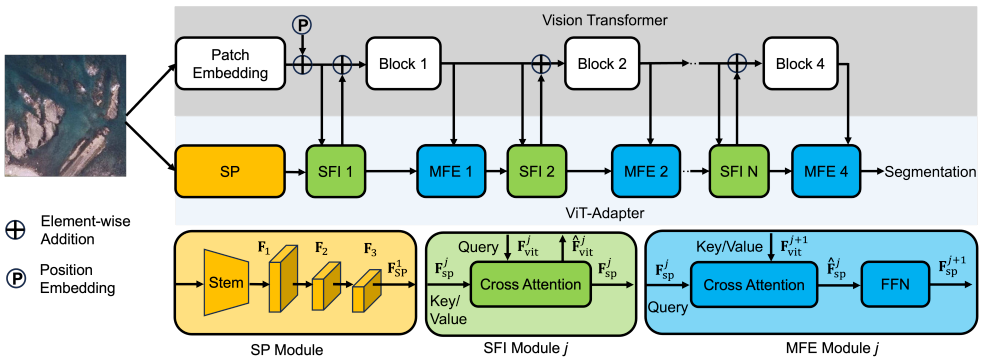


Figure 4. The network architecture of the proposed method and the data flow between modules.

The ViT adapter, designed to enhance the performance of the ViT, includes the following three crucial components: SP, SFI, and MFE modules. The SP module captures spatial features from the input patch. These features, after being flattened and concatenated, are fed into the ViT through the SFI module. The MFE module then processes these features and constructs hierarchical features from the ViT output. This process is repeatedly performed, with the transformer encoders in the ViT segmented into blocks. Each block is involved in enhancing spatial features and hierarchically extracting features. The resulting features form a pyramid structure similar to that used in ResNet [20], enabling the model to efficiently perform various semantic segmentation tasks.

In the ViT adapter, the specialized spatial feature process is critical for capturing detailed spatial information within aerial images. The SP module is designed to process local spatial contexts present in patches. It operates in conjunction with the patch-embedding layer of the ViT and does not alter the original structure of the ViT. The SP module uses a convolutional stem similar to ResNet, consisting of three convolutional layers, followed by a max-pooling layer. This is succeeded by a series of stride-2 3×3 convolutions that increase the channel count while reducing the size of the feature maps. Finally, a set of 1×1 convolutions (Conv) and feature pyramids (FPs) transforms these feature maps into D-dimensional space, encompassing feature maps at resolutions of $1/8$, $1/16$, and $1/32$. The process can be defined as follows:

$$\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3 = \mathrm{FP}(\mathrm{Conv}(P)). \tag{3}$$

Here, $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ represent the tokens at the respective resolutions derived from patch P. Then, we flatten and concatenate these tokens as follows:

$$\mathbf{F}_{\mathrm{sp}}^{j} = \mathrm{Concat}(\mathrm{Flatten}(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)), \tag{4}$$

where $\mathbf{F}_{\mathrm{sp}}^{j}$ denotes the spatial features extracted at the j-th block within the ViT adapter from patch P. The SP module allows for enhanced analysis of spatial features within aerial images, significantly contributing to the overall effectiveness of the model in identifying and interpreting complex visual data.

The plain ViT model typically suffers from suboptimal performance in semantic segmentation tasks because of weak prior assumptions [21,22]. To mitigate this issue, two feature interaction modules, SFI and MFE, based on cross-attention are introduced [23,24]. The SFI module infuses SPs into ViT, whereas the MFE module extracts and processes these features, enhancing the performance of the model in semantic segmentation tasks. The process of the SFI module can be defined as follows:

$$\hat{\mathbf{F}}_{\mathrm{vit}}^{j} = \mathbf{F}_{\mathrm{vit}}^{j} + \mathbf{p}^{j} \cdot \mathrm{SA}(\mathrm{LN}(\mathbf{F}_{\mathrm{vit}}^{j}), \mathrm{LN}(\mathbf{F}_{\mathrm{sp}}^{j})), \tag{5}$$

where $\mathbf{p}^{j} \in \mathbb{R}^{D}$ denotes a learnable parameter that balances the output of SA and the input feature $\mathbf{F}_{\mathrm{vit}}^{j}$, LN denotes layer normalization [25], and SA denotes sparse attention. Sparse attention facilitates a more streamlined and efficient computation, which is especially beneficial in scenarios involving lengthy data sequences [26–28]. Specifically, we used deformable attention [29] in the proposed method. The process of the MFE module is defined as follows:

$$\mathbf{F}_{\mathrm{sp}}^{j+1} = \hat{\mathbf{F}}_{\mathrm{sp}}^{j} + \mathrm{FFN}(\mathrm{LN}(\hat{\mathbf{F}}_{\mathrm{sp}}^{j})), \tag{6}$$

$$\hat{\mathbf{F}}_{\mathrm{sp}}^{j} = \mathbf{F}_{\mathrm{sp}}^{j} + \mathrm{SA}(\mathrm{LN}(\mathbf{F}_{\mathrm{sp}}^{j}), \mathrm{LN}(\mathbf{F}_{\mathrm{vit}}^{j+1})), \tag{7}$$

where we use a feed-forward network (FFN) to extract multiscale features. The spatial feature, denoted as $\mathbf{F}_{\mathrm{sp}}^{j+1}$, which resulted from this injector is then used as the input for the next SFI module. The above processing ensures a comprehensive extraction of image

features, enabling better performance in semantic tasks by effectively using high-resolution aerial images for detailed spatial analysis.

3.3. Semantic Segmentation and Loss Function

In the training phase, the binary mask M is required to perform semantic segmentation. The mask is generated based on detailed annotations to accurately represent the locations of algal beds. The binary mask-creating process can be defined as follows:

$$M(a,b) = \begin{cases} 1, & \text{if pixel}(a,b) \text{ is within an algal bed,} \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where each pixel coordinate (a,b) is labeled as being in the algal bed region or not. Note that the generated binary mask also performs the splitting process corresponding to the aerial image patches, as shown in Section 3.1.

To optimize the segmentation model, the cross-entropy loss $\mathcal{L}$ is used to measure the discrepancy between the predicted segmentation and ground truth as follows:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{K} \sum_{k=1}^{K} [\mathbf{y}_k \log(\hat{\mathbf{y}}_k) + (1 - \mathbf{y}_k) \log(1 - \hat{\mathbf{y}}_k)]. \tag{9}$$

Here, $\mathbf{y}$ denotes the ground truth, $\hat{\mathbf{y}}$ denotes the predicted results for algal bed pixels, and K represents the total pixel count.

Combining the expansive receptive field of the ViT with the detailed adaptations from the ViT adapter, the proposed method enables refined analysis of aerial images. This integration results in accurate segmentation of algal bed regions, which is essential for effective environmental monitoring.

4. Experiments

4.1. Experimental Settings

In this subsection, we introduce the detailed settings of the proposed method (PM). First, the backbone used in our study is a ViT adapter-L [18], and the segmentation method is UperNet [30]. We leverage the BERT pretraining of image transformers (BEiT-L) [31], which is pretrained on the ImageNet-22K dataset [32] using a masked image modeling approach [33,34]. This refinement significantly alters the original ViT configuration to support an increased input resolution s of 640 pixels, which is facilitated by adjusting the patch size c to 16 pixels. The overlap o was set to 320 pixels. Our architecture has a 24-layer deep network that incorporates 1024-dimensional embeddings and 16 attention heads while using a mixed precision framework to improve memory efficiency during training. The segmentation framework includes a decoder head with four 1024-dimensional embedding channels designed for binary classification problems. An auxiliary head is also integrated to augment segmentation precision and set it up for binary classification.

To improve the performance of the model, data augmentation methods such as resizing, random cropping, flipping, photometric distortion, normalization, padding, and formatting are applied. An AdamW optimizer was used for model training, which was configured with a 2×10^{-5} learning rate and 0.05 weight decay, incorporating a layer-wise learning rate decay strategy across 24 layers at a 0.9 rate for refined parameter optimization. The learning rate schedule was designed according to a polynomial decay model, incorporating a linear warm-up phase over 1500 iterations, strategically balancing exploration and exploitation throughout the training process. The total number of training iterations was set to 50,000. The parameter number of ViT adapter-L was approximately 23.7 M. Table 1 shows the important hyperparameters of the proposed method. For comparative methods, we used a ViT adapter with full image training (CM1), DeepLab V3 (CM2) [35], ViT-L (CM3), and SegFormer (CM4) [36]. All methods use their default settings and train from scratch for fair comparison.

Table 1. Hyperparameters of the proposed method.

Parameter	Value
Training iteration	50,000
Warmup iteration	1500
Image size	640
Overlap size	320
Patch size	16
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.05
Layer-wise decay	0.9
Embedding dimension	1024
Depth	16
Heads number	4
Parameter numbers	23.7 M

In our experiments, we used the PyTorch [37] framework to develop the models and conducted training and testing on an NVIDIA RTX A6000 GPU. The entire training process spanned approximately 1 day. To assess the accuracy of the proposed method, we used the following set of specific metrics.

- Area: A metric that considers the predicted algal bed area against the ground truth.

$$\text{Area} = \frac{P_{\text{pred}}}{P_{\text{gt}}}, \tag{10}$$

  where P_{pred} represents the number of pixels in the predicted algal bed area, and P_{gt} represents the number of pixels in the ground truth algal bed area.
- IoU: A metric that measures the overlap between the predicted and actual algal bed regions.

$$\text{IoU} = \frac{P_{\text{overlap}}}{P_{\text{pred}} + P_{\text{gt}} - P_{\text{overlap}}}, \tag{11}$$

  where P_{overlap} represents the number of overlapping pixels between the predicted algal bed area and ground truth.
- Precision: A metric that indicates the proportion of correctly predicted algal bed areas.

$$\text{Precision} = \frac{P_{\text{overlap}}}{P_{\text{pred}}}. \tag{12}$$

- Recall: A metric that indicates the proportion of actual algal bed areas correctly identified.

$$\text{Recall} = \frac{P_{\text{overlap}}}{P_{\text{gt}}}. \tag{13}$$

- F-score (Fscore): A metric that combines precision and recall into a single measure.

$$\text{Fscore} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{(\text{Precision} + \text{Recall})}, \tag{14}$$

  which can be simplified to:

$$\text{Fscore} = \frac{2 \cdot P_{\text{overlap}}}{P_{\text{pred}} + P_{\text{gt}}}. \tag{15}$$

IoU, Precision, Recall, and F-score metrics are commonly used evaluation metrics in segmentation tasks. The area metric is a newly proposed evaluation metric for the future measurement of CO_2 absorption ability.

4.2. Results and Analysis

This subsection presents the test results of Regions A–E and their subsequent analysis. The effectiveness of the proposed method is quantitatively illustrated in Tables 2–5.

Table 2. Overall semantic segmentation comparison of different methods.

	mIoU	mFscore	mPrecision	mRecall
CM1	0.749	0.831	0.823	0.840
CM2	0.758	0.848	0.882	0.821
CM3	0.763	0.852	0.894	0.820
CM4	0.776	0.863	0.877	0.850
PM	0.787	0.870	0.892	0.851

Table 3. Overall semantic segmentation results of five different algal bed regions.

	mIoU	mFscore	mPrecision	mRecall
Region A	0.826	0.898	0.919	0.880
Region B	0.768	0.859	0.877	0.844
Region C	0.826	0.898	0.880	0.919
Region D	0.769	0.857	0.917	0.815
Region E	0.740	0.831	0.872	0.799
Mean	0.787	0.870	0.892	0.851

Table 4. Background region semantic segmentation results for five different algal bed regions.

	IoU	Fscore	Precision	Recall
Region A	0.971	0.985	0.981	0.990
Region B	0.931	0.964	0.956	0.972
Region C	0.968	0.984	0.988	0.979
Region D	0.951	0.975	0.960	0.990
Region E	0.964	0.982	0.975	0.988
Mean	0.957	0.978	0.972	0.984

Table 5. Algal bed region semantic segmentation results for five different algal bed regions.

	IoU	Fscore	Precision	Recall	Area
Region A	0.681	0.810	0.856	0.769	0.872
Region B	0.606	0.754	0.797	0.716	0.856
Region C	0.685	0.813	0.772	0.859	1.120
Region D	0.640	0.739	0.874	0.640	0.713
Region E	0.515	0.680	0.768	0.610	0.744
Mean	0.616	0.762	0.813	0.718	0.861

As shown in Table 2, the proposed method (PM), which employs a patch-based training strategy with the ViT adapter, exhibits significant advantages in semantic segmentation, achieving a mIoU of 0.787 along with balanced mPrecision and mRecall metrics. This performance surpasses that of other established methods, such as DeepLab V3 (CM2) and SegFormer (CM4), offering more accurate segmentations. Compared to full image training (CM1), PM considerably improves detail resolution, particularly in challenging regions like algal beds. Furthermore, when compared to a standard ViT-L (CM3), PM's enhanced segmentation capabilities stem from the integration of three modules, which collectively contribute to its superior performance.

In Table 3, the mIoU of 0.787 not only suggests high accuracy but also consistent performance across varying marine environments. This consistency is essential for the scalable applications of the proposed method, ensuring reliability in diverse ecological

conditions. Region A, with a mIoU of 0.826, and Region C, with a mIoU of 0.826, are exemplary, indicating the potential of the method to achieve high accuracy in ecologically diverse settings. The mF-score, averaging 0.870, complements this by providing a balanced view of the precision–recall trade-off, with Region A showing a remarkable mF-score of 0.898, attesting to the ability of the method to maintain high recognition performance while minimizing false detections.

Performance differences are evident when comparing the metrics of mean precision (mPrecision) and mean recall (mRecall). An mPrecision of 0.892 shows the ability of the method to correctly identify true algal bed areas, whereas an mRecall of 0.851 reflects the comprehensiveness of algal bed recognition. Region D, despite having a lower mIoU of 0.769, achieves an mPrecision of 0.917, suggesting that although some true algal bed areas may be missed, the areas identified are highly likely correct.

In Table 4, the focus on background region recognition enhances algal bed segmentation, where high scores across IoU, F-score, precision, and recall exceed those shown in Table 3. With mIoU and mF-score values of 0.957 and 0.978, respectively, the method demonstrates its effectiveness in distinguishing background areas, which is an important aspect of accurate algal bed detection. Furthermore, the method exhibits exceptional precision and recall, averaging 0.972 and 0.984, respectively, indicating high accuracy in background pixel classification, particularly in Region A, which exhibits the highest scores.

Table 5 shows the efficacy of the method in the identification of algal bed regions, with the IoU values demonstrating high accuracy in the segmentation task. Notably, an IoU of 0.685 for Region C signifies a considerable alignment between the predicted and actual areas of algal beds. Furthermore, the F-score for Region C is 0.813, which reflects an optimal balance of precision and recall, demonstrating the robustness of the method in accurately identifying algal beds. The area metric relates to the ecological importance of this study. The average ratio of 0.861 across regions for the area metric indicates the ability of the method to accurately estimate the extent of algal bed regions.

The experimental results indicate that the proposed method is not only effective in identifying and delineating algal beds in marine environments but also demonstrates high adaptability and accuracy across different regions. The high precision of our method in background segmentation bolsters its capability in algal bed recognition, providing a reliable tool for environmental monitoring and analysis. The results of this study, along with the promising performance of the method across multiple metrics, show its potential in the field of marine ecology and the broader context of environmental science and climate change mitigation.

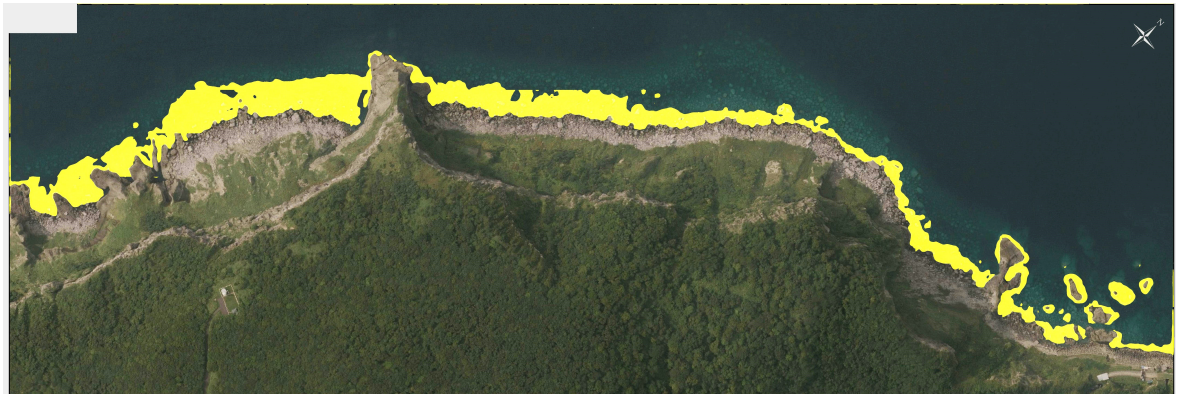
Visualization of the segmentation results is shown in Figures 5–9. It provides a crucial visual complement to the quantitative analysis, offering insight into the practical effectiveness of the proposed method for algal bed region segmentation using aerial images.

The figures present a side-by-side visual comparison of the segmented outputs of the method against manually annotated ground truths for Regions A–E. The visualization results demonstrate the ability of the method to accurately identify and delineate algal bed regions despite common challenges such as varying lighting conditions and water turbidity.

Each figure corresponds to a different coastal region, demonstrating the versatility of our method across diverse geographical areas. This aspect is critical because it shows the robustness and adaptability of the method to different environmental conditions and algal bed densities, which is vital for scalable and reliable environmental monitoring and marine ecosystem management. This visual evidence not only validates the quantitative results but also emphasizes the potential contribution of our method to blue carbon assessment and ecological studies by providing a reliable, efficient, and scalable tool for environmental monitoring.

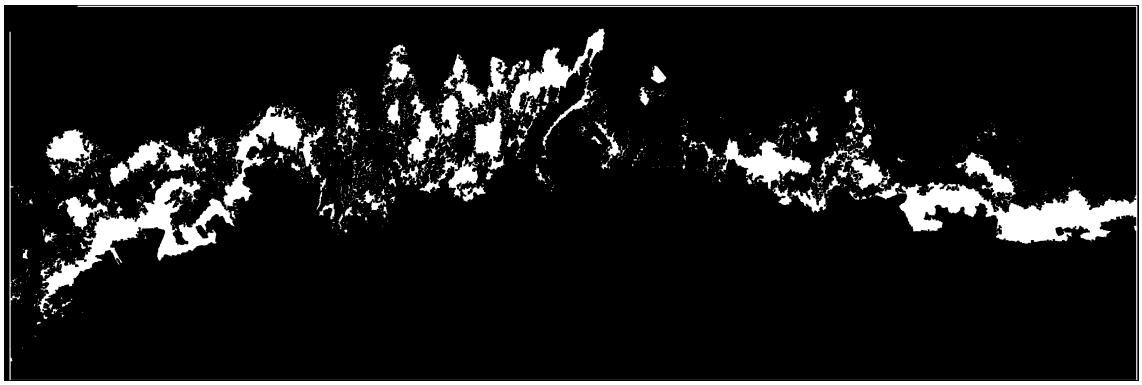
(**a**) Binary mask.

(**b**) Predicted results (CM1).



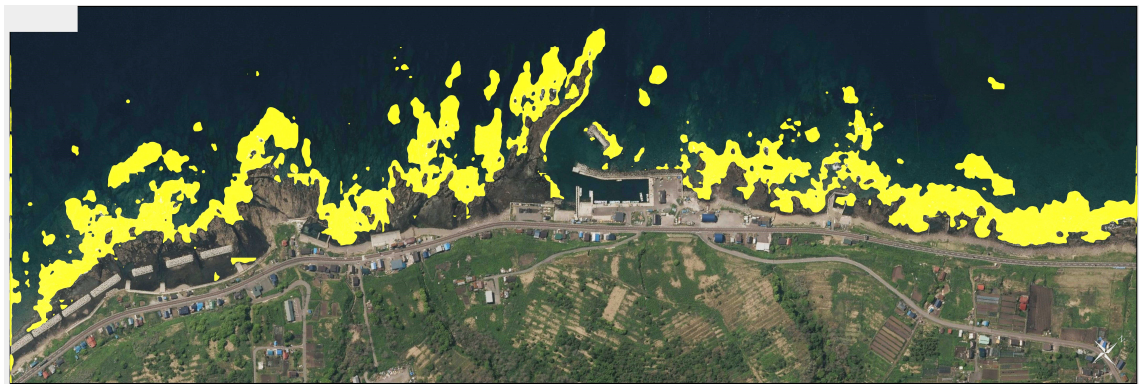
(**c**) Predicted results (PM).

Figure 5. Qualitative evaluation results for Region A.



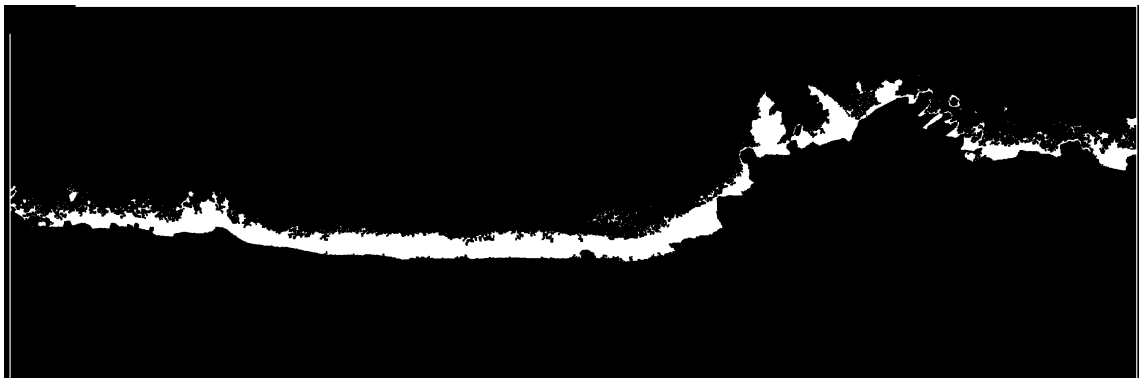
(**a**) Binary mask.

(**b**) Predicted results (CM1).

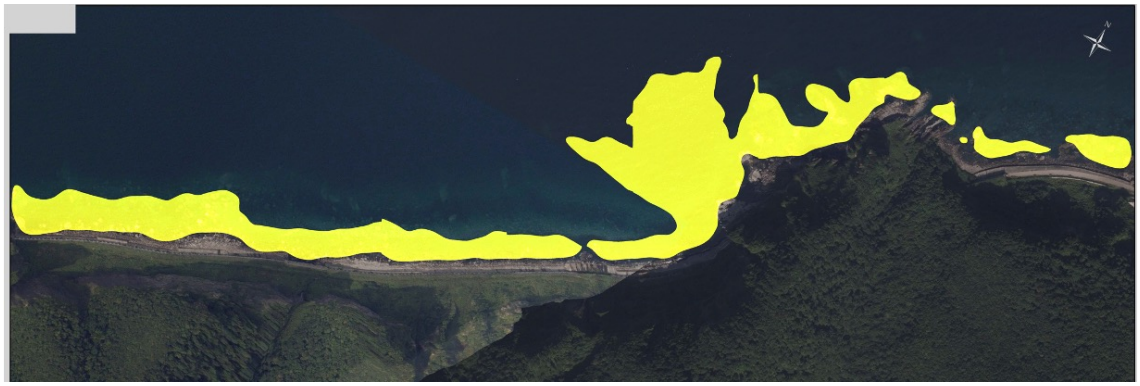


(**c**) Predicted results (PM).

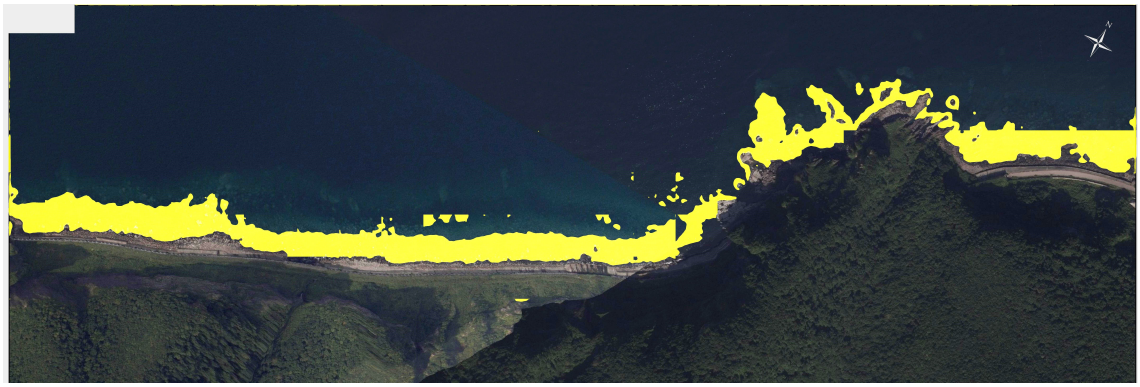
Figure 6. Qualitative evaluation results for Region B.



(**a**) Binary mask.

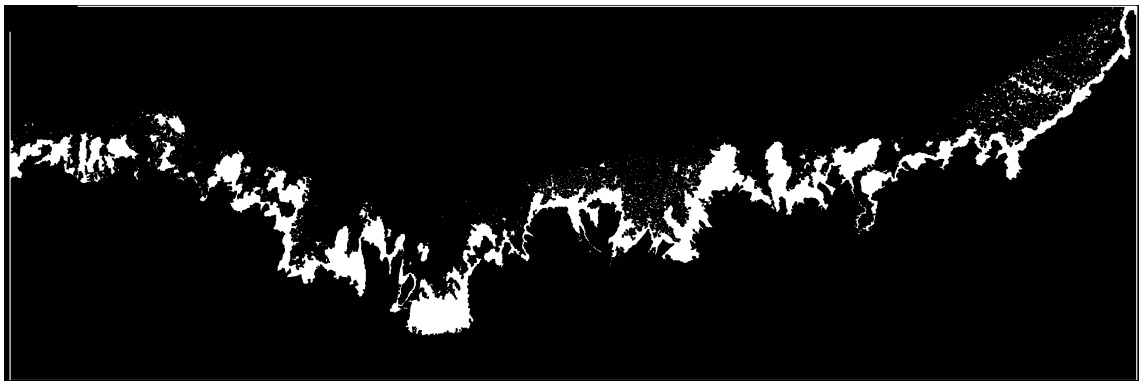


(**b**) Predicted results (CM1).

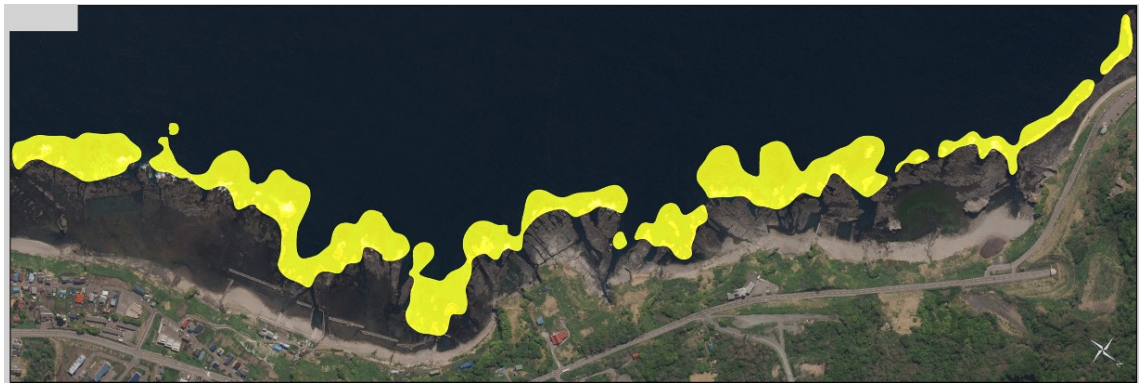


(**c**) Predicted results (PM).

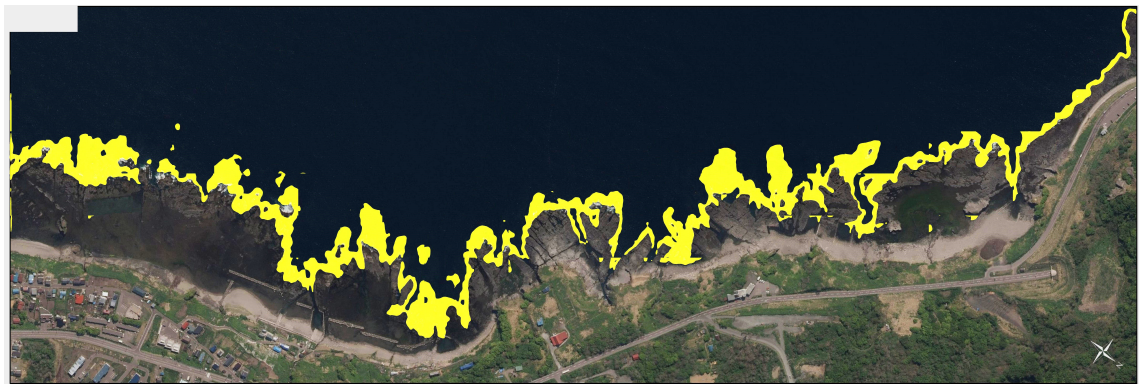
Figure 7. Qualitative evaluation results for Region C.



(**a**) Binary mask.

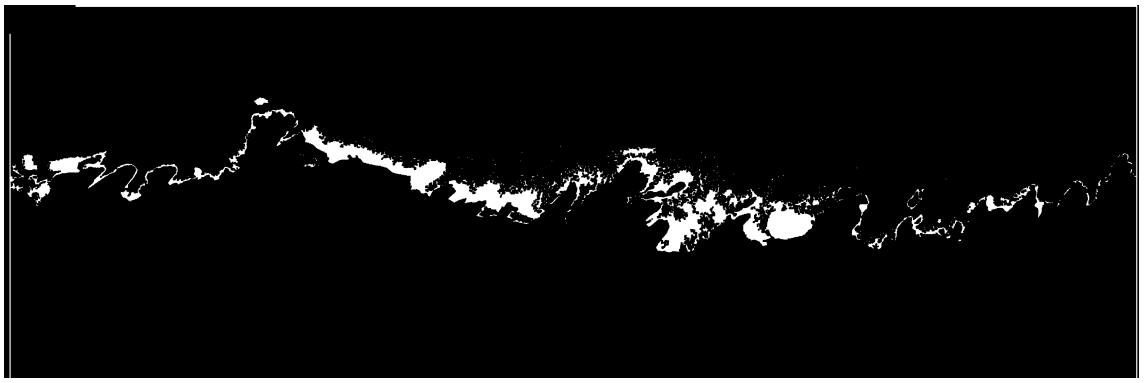


(**b**) Predicted results (CM1).

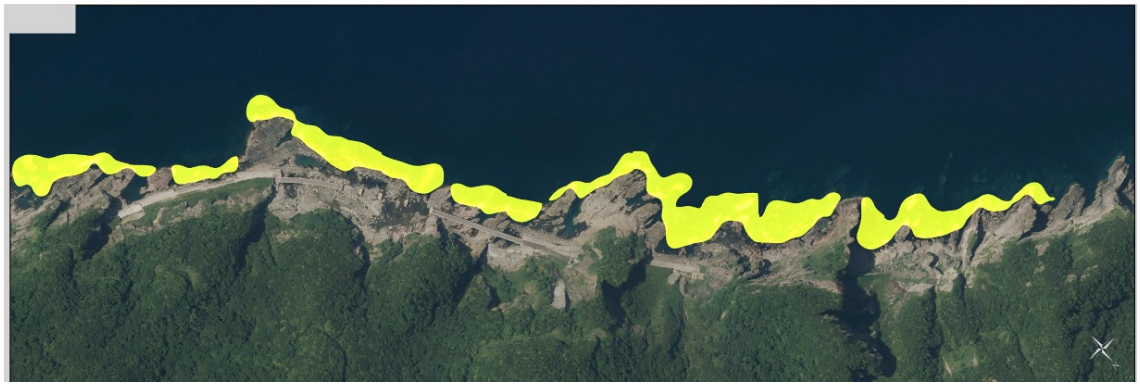


(**c**) Predicted results (PM).

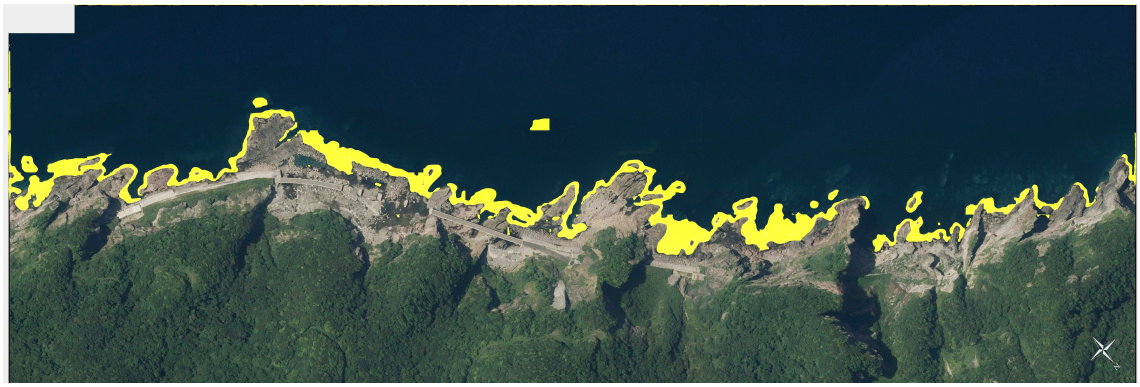
Figure 8. Qualitative evaluation results for Region D.



(**a**) Binary mask.



(**b**) Predicted results (CM1).



(**c**) Predicted results (PM).

Figure 9. Qualitative evaluation results for Region E.

5. Discussion

This study presented a new method for identifying algal beds in coastal regions using a semantic segmentation model based on a ViT adapter. The results indicate high mIoU and mF-score values across various validated regions, demonstrating the effectiveness of the method in identifying algal bed regions. Specifically, the mIoU value of 0.787 and mF-score of 0.870 across the study regions reflect the performance of the method in accurately segmenting algal beds. These performance metrics, particularly mIoU, which is considered a stringent segmentation accuracy metric, indicate the balanced assessment capabilities of the method in terms of both false positives and negatives.

The user-friendly design of the proposed method is particularly advantageous for staff members of local public organizations and fishery cooperatives who play a pivotal role in environmental conservation. The straightforward operation of our method simplifies monitoring tasks. This study is a foundational step in a broader investigative initiative

aimed at enhancing blue carbon evaluation. Using advanced semantic segmentation technology, we intend to improve the accuracy of estimating CO_2 absorption capacity, offering essential assistance to local environmental managers responsible for these significant ecological assessments.

The potential of our method for real-world application is substantial, extending beyond CO_2 absorption to include the early recognition and management of algal beds, which could be used for the early detection and conservation measures of declining seaweed beds. The scalability of the method holds the promise of integration into global monitoring systems, which can revolutionize the management of marine ecosystems in the context of climate change mitigation.

Our method also has some limitations. We present three detailed analyses of the visualization results in Figure 10. In Region C, the overdetection of algal beds can be attributed to edge relationships during image patching. When the image is segmented into smaller patches for analysis, the edges where patches meet can introduce artifacts that our method currently misinterprets as part of the algal beds. This oversegmentation along the patch borders underscores the need for improved edge artifact recognition within our patch-splitting algorithm, which can be addressed by refining the patch-splitting process or by introducing a postprocessing step that considers the continuity of detected features across patch boundaries.

Region D presents a different challenge; our model struggles with the detection of sparsely distributed algal beds. This limitation is partly inherent in semantic segmentation models, which rely on the recognition of patterns and continuity in data. Sparse features without clear, consistent patterns can be easily overlooked or misclassified, leading to underdetection in these areas. Enhancing the sensitivity of the method to less dense and more scattered features without increasing the false positive rate is a delicate balance that requires further research, possibly involving the integration of more sophisticated feature recognition methods.

Region E shows examples of detection errors where our method either fails to identify algal beds or incorrectly marks other regions. This misclassification underlines the need for increased robustness in the model, which can be achieved through more extensive training on a diverse set of images representing a wide range of environmental conditions. Incorporating additional contextual information and using advanced deep learning methods, such as ensemble learning or adversarial training, can also help improve the accuracy and generalizability of the proposed model.

The discrepancies in recognition accuracy, such as a lower IoU in Region E, indicate potential limitations and areas for further development. These variances can be attributed to environmental complexities and varying lighting conditions. The challenge in Region E suggests that the performance of the proposed method can be affected by intricate ecological features or suboptimal imaging conditions, which necessitate further refinement of the algorithm or data preprocessing methods.

Reliance on high-quality aerial images and the computational demands of processing large datasets are factors that can limit the applicability of the method in resource-constrained settings. Furthermore, note that manual annotations are subject to human error and interpretation, which can introduce a degree of variability in the ground truth.

The area metric, which indicates the extent of algal bed coverage identified by our model, has shown promise with high average values in the validation sets. However, the limited sample diversity may not fully represent the complex variability of coastal regions. We recognize this as a constraint on the current applicability of the model. To improve the performance of the model in practical scenarios, we are committed to diversifying our validation datasets. Increasing the number of samples, especially those from underrepresented and ecologically varied regions, will be a critical step toward improving the accuracy and robustness of the method.

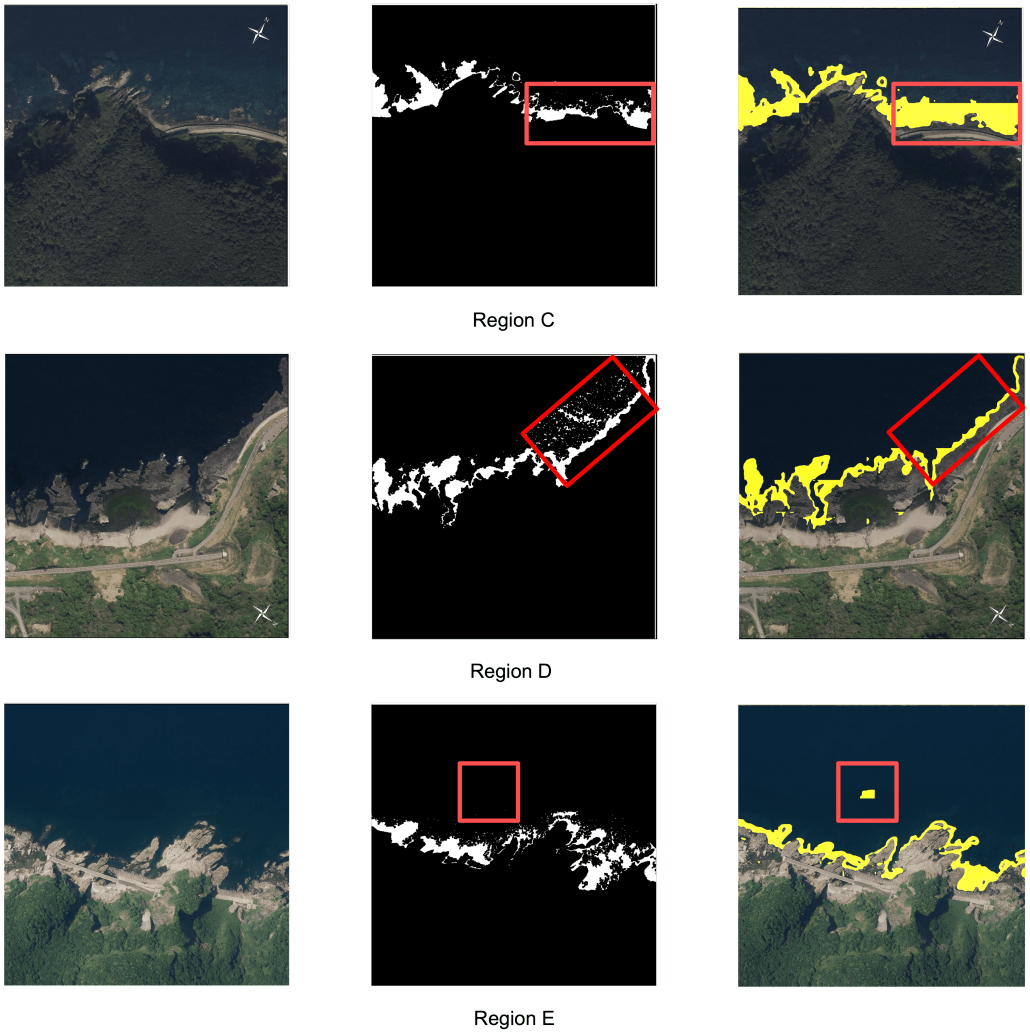


Figure 10. Detailed analyses of the visualization results of the proposed method.

6. Conclusions

This study proposed a novel method for recognizing algal bed regions using aerial images. In our method, we used an advanced semantic segmentation model, a ViT adapter, and adapted it to aerial images for algal bed region segmentation. The proposed method accurately identifies algal bed regions, which is crucial for assessing their CO_2 absorption capacity in the context of blue carbon strategies and climate change mitigation. Experiments on an aerial image dataset collected from Hokkaido, Japan, reveal that the proposed method achieves high accuracy in algal bed region segmentation. This success illustrates its capability to accurately identify algal beds among various coastal and marine features. Such progress promises to improve the monitoring and management of blue carbon ecosystems, thereby playing a crucial role in environmental conservation efforts.

Author Contributions: Conceptualization, G.L., R.T., K.M. and T.O.; methodology, G.L., R.T., K.M. and T.O.; software, G.L.; validation, G.L.; data curation, G.L. and A.S.; writing and original draft preparation, G.L.; writing and review and editing, R.T., K.M., A.S., I.Y., T.H., S.N. and T.O.; visualization, G.L.; funding acquisition, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by JSPS KAKENHI Grant Numbers JP21H03456, JP23K11211, and JP23K11141. This is a collaborative study with Alpha Hydraulic Engineering Consultants Co., Ltd., Cold Regions Air and Sea Ports Engineering Research Center, and the Hokkaido Regional Development Bureau.

Data Availability Statement: The aerial image dataset used in this study is a non-public dataset provided by the Hokkaido Regional Development Bureau, Sapporo, Japan. The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Akinori Sako is employed by the company Alpha Hydraulic Engineering Consultants Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Mitchell, J.F.; Lowe, J.; Wood, R.A.; Vellinga, M. Extreme events due to human-induced climate change. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2006**, *364*, 2117–2133. [CrossRef]
2. Fischer, E.M.; Knutti, R. Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nat. Clim. Chang.* **2015**, *5*, 560–564. [CrossRef]
3. Walsh, J.E.; Ballinger, T.J.; Euskirchen, E.S.; Hanna, E.; Mård, J.; Overland, J.E.; Tangen, H.; Vihma, T. Extreme weather and climate events in northern areas: A review. *Earth-Sci. Rev.* **2020**, *209*, 103324. [CrossRef]
4. Macreadie, P.I.; Anton, A.; Raven, J.A.; Beaumont, N.; Connolly, R.M.; Friess, D.A.; Kelleway, J.J.; Kennedy, H.; Kuwae, T.; Lavery, P.S.; et al. The future of Blue Carbon science. *Nat. Commun.* **2019**, *10*, 3998. [CrossRef]
5. Macreadie, P.I.; Costa, M.D.; Atwood, T.B.; Friess, D.A.; Kelleway, J.J.; Kennedy, H.; Lovelock, C.E.; Serrano, O.; Duarte, C.M. Blue carbon as a natural climate solution. *Nat. Rev. Earth Environ.* **2021**, *2*, 826–839. [CrossRef]
6. Lovelock, C.E.; Duarte, C.M. Dimensions of blue carbon and emerging perspectives. *Biol. Lett.* **2019**, *15*, 20180781. [CrossRef]
7. Bertram, C.; Quaas, M.; Reusch, T.B.; Vafeidis, A.T.; Wolff, C.; Rickels, W. The blue carbon wealth of nations. *Nat. Clim. Chang.* **2021**, *11*, 704–709. [CrossRef]
8. Sondak, C.F.; Ang, P.O.; Beardall, J.; Bellgrove, A.; Boo, S.M.; Gerung, G.S.; Hepburn, C.D.; Hong, D.D.; Hu, Z.; Kawai, H.; et al. Carbon dioxide mitigation potential of seaweed aquaculture beds (SABs). *J. Appl. Phycol.* **2017**, *29*, 2363–2373. [CrossRef]
9. Farrelly, D.J.; Everard, C.D.; Fagan, C.C.; McDonnell, K.P. Carbon sequestration and the role of biological carbon mitigation: A review. *Renew. Sustain. Energy Rev.* **2013**, *21*, 712–727. [CrossRef]
10. Thomasberger, A.; Nielsen, M.M.; Flindt, M.R.; Pawar, S.; Svane, N. Comparative Assessment of Five Machine Learning Algorithms for Supervised Object-Based Classification of Submerged Seagrass Beds Using High-Resolution UAS Imagery. *Remote Sens.* **2023**, *15*, 3600. [CrossRef]
11. Tallam, K.; Nguyen, N.; Ventura, J.; Fricker, A.; Calhoun, S.; O'Leary, J.; Fitzgibbons, M.; Robbins, I.; Walter, R.K. Application of Deep Learning for Classification of Intertidal Eelgrass from Drone-Acquired Imagery. *Remote Sens.* **2023**, *15*, 2321. [CrossRef]
12. Shi, B.; Wideman, G.; Wang, J.H. A new approach of BioCO_2 fixation by thermoplastic processing of microalgae. *J. Polym. Environ.* **2012**, *20*, 124–131. [CrossRef]
13. Ghosh, A.; Kiran, B. Carbon concentration in algae: Reducing CO_2 from exhaust gas. *Trends Biotechnol.* **2017**, *35*, 806–808. [CrossRef]
14. Borja, A.; Bricker, S.B.; Dauer, D.M.; Demetriades, N.T.; Ferreira, J.G.; Forbes, A.T.; Hutchings, P.; Jia, X.; Kenchington, R.; Marques, J.C.; et al. Overview of integrative tools and methods in assessing ecological integrity in estuarine and coastal systems worldwide. *Mar. Pollut. Bull.* **2008**, *56*, 1519–1537. [CrossRef] [PubMed]
15. Bajjouk, T.; Mouquet, P.; Ropert, M.; Quod, J.P.; Hoarau, L.; Bigot, L.; Le Dantec, N.; Delacourt, C.; Populus, J. Detection of changes in shallow coral reefs status: Towards a spatial approach using hyperspectral and multispectral data. *Ecol. Indic.* **2019**, *96*, 174–191. [CrossRef]
16. Lopez-Marcano, S.; Brown, C.J.; Sievers, M.; Connolly, R.M. The slow rise of technology: Computer vision techniques in fish population connectivity. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **2021**, *31*, 210–217. [CrossRef]
17. Chang, Z.; Li, H.; Chen, D.; Liu, Y.; Zou, C.; Chen, J.; Han, W.; Liu, S.; Zhang, N. Crop Type Identification Using High-Resolution Remote Sensing Images Based on an Improved DeepLabV3+ Network. *Remote Sens.* **2023**, *15*, 5088. [CrossRef]
18. Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; Qiao, Y. Vision transformer adapter for dense predictions. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda,1–5 May 2023.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; pp. 9355–9366.
22. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.

23. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
24. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 357–366.
25. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450 .
26. Tay, Y.; Bahri, D.; Yang, L.; Metzler, D.; Juan, D.C. Sparse sinkhorn attention. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 12–18 July 2020; pp. 9438–9447.
27. Roy, A.; Saffar, M.; Vaswani, A.; Grangier, D. Efficient content-based sparse attention with routing transformers. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 53–68. [CrossRef]
28. Gan, Y.; Li, G.; Togo, R.; Maeda, K.; Ogawa, T.; Haseyama, M. Zero-shot traffic sign recognition based on midlevel feature matching. *Sensors* **2023**, *23*, 9607. [CrossRef]
29. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
30. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
31. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2021.
32. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
33. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
34. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. Simmim: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9653–9663.
35. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
36. 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.
37. Imambi, S.; Prakash, K.B.; Kanagachidambaresan, G. PyTorch. In *Programming with TensorFlow: Solution for Edge Computing Applications*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 87–104.