remote sensing

MDPI

Article

Sea–Land Segmentation of Remote-Sensing Images with Prompt Mask-Attention

Yingjie Ji [1,2], Weiguo Wu [1], Shiqiang Nie [1], Jinyu Wang [1] and Song Liu [1,*]

[1] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China; jiyingjie@stu.xjtu.edu.cn (Y.J.); wgwu@xjtu.edu.cn (W.W.); shiqiang.nie@xjtu.edu.cn (S.N.); jinyu.wang@xjtu.edu.cn (J.W.)
[2] North China Institute of Computing Technology, Beijing 100083, China
* Correspondence: liusong@xjtu.edu.cn

Abstract: Remote-sensing technology has gradually become one of the most important ways to extract sea–land boundaries due to its large scale, high efficiency, and low cost. However, sea–land segmentation (SLS) is still a challenging problem because of data diversity and inconsistency, "different objects with the same spectrum" or "the same object with different spectra", and noise and interference problems, etc. In this paper, a new sea–land segmentation method (PMFormer) for remote-sensing images is proposed. The contributions are mainly two points. First, based on Mask2Former architecture, we introduce the prompt mask by normalized difference water index (NDWI) of the target image and prompt encoder architecture. The prompt mask provides more reasonable constraints for attention so that the segmentation errors are alleviated in small region boundaries and small branches, which are caused by insufficiency of prior information by large data diversity or inconsistency. Second, for the large intra-class difference problem in the foreground–background segmentation in sea–land scenes, we use deep clustering to simplify the query vectors and make them more suitable for binary segmentation. Then, traditional NDWI and eight other deep-learning methods are thoroughly compared with the proposed PMFormer on three open sea–land datasets. The efficiency of the proposed method is confirmed, after the quantitative analysis, qualitative analysis, time consumption, error distribution, etc. are presented by detailed contrast experiments.

Keywords: image segmentation; prompt pearning; Mask2Former; remote sensing

check for updates

Citation: Ji, Y.; Wu, W.; Nie, S.; Wang, J.; Liu, S. Sea–Land Segmentation of Remote-Sensing Images with Prompt Mask-Attention. *Remote Sens.* **2024**, *16*, 3432. https://doi.org/10.3390/rs16183432

Academic Editors: Masoud Mahdianpari and Peng Liu

Received: 6 August 2024
Revised: 31 August 2024
Accepted: 13 September 2024
Published: 16 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The coastline, fundamentally defined as the delineation between terrestrial and marine environments, serves as a pivotal boundary for various geographical, ecological, and economic processes. This information constitutes the foundational framework for the delineation of land and water resources [1], as well as underpinning the exploration and stewardship of coastal-zone resources. The significance of coastline information extraction and quantification lies in its multifaceted applications, encompassing disaster-risk assessment [2], coastal-engineering planning [3], marine-resource management [4], ecological environmental preservation [5], and coastal-tourism development [6], among others. Furthermore, the extraction of coastlines is indispensable for the analysis of coastline-change mechanisms [7], enabling the examination of various change characteristics, such as coastline length variations, shifting segments, fractal dimensions, and land-use patterns within coastal regions. This, in turn, facilitates the understanding of the intricate interplay between anthropogenic and natural factors influencing such changes. Therefore, the precise extraction and quantification of coastline information is paramount in ensuring the effectiveness of coastal-zone development [8], sustainable utilization, and scientific management strategies. The challenge of rapidly and accurately delineating sea–land boundaries has emerged as a primary research focus in the coastal zone sciences, with profound practical implications for the sustainable future of these vital ecosystems.

The traditional coastline mapping methods mainly include: ground survey [9] (for example, by setting measuring points on the coastline, using the rangefinder, theodolite, and other instruments to measure, and then recording the data for analysis and drawing), or underwater measurement [10] (such as the use of sonar and other underwater measuring equipment, to measure the terrain under the coastline). The complexity and difficulty of these methods depend on factors such as the accuracy requirements of the measurement, the measurement range, and the complexity of the terrain. Ground survey is relatively simple, but requires a lot of manpower and time; underwater measurement requires professional equipment and technology and is affected by factors such as seabed topography and water quality. The traditional method has its own rationality but also faces many problems in practical applications. The investigation of the land–water boundary line is complicated, with a wide range, fast changes, and broken surface features. These traditional detection methods not only have high labor intensity, long working cycles, and low efficiency but also are difficult to use for dynamic monitoring of the coastline. Moreover, limited by geographical environment and other conditions, some survey areas are difficult to reach and difficult to map. Remote-sensing technology is a comprehensive application technology of earth observation based on physical means, geological analysis, and mathematical methods. Its advantages also include high time resolution, high spatial resolution, multi-spectrum and multi-time series. For example, spatiotemporal remote-sensing image fusion such as [11,12] is promising in coastal zone monitoring [13,14]. Furthermore, it is also a large range and less restricted by geographical environment and other conditions. Remote-sensing technology has become an effective means to extract the sea–land boundary line and monitor its changes.

With the fast development of satellite technology, coastline automatic extraction methods based on remote-sensing images have been greatly advanced. How to quickly and accurately automatically extract the land–water boundary from remote-sensing images has been a subject of wide concern of many scholars in recent years. The automatic coastline-extraction methods based on remote-sensing images have been explored and studied [15–17], and a large number of automatic coastline-extraction methods have been proposed. There are many kinds of existing methods, especially for deep-learning-based methods. However, there is no universal coastline-extraction method that can completely and perfectly obtain a good extraction result for all the coastline remote-sensing images. Coastline extraction based on remote sensing still faces three problems. First, data diversity and inconsistency: remote-sensing data come from various sources, including satellite optical remote sensing, microwave remote sensing, Lidar remote sensing, etc. For this study, we focus on optical data. These data have differences in resolution, spectral characteristics, time coverage, and other aspects, which increase the complexity of data processing. Second, noise and interference: remote-sensing images often contain noise and interference, such as atmospheric conditions, cloud cover, ocean reflection, etc., which will affect the accurate extraction of the coastline. Finally, for remote-sensing images, it is not easy to directly introduce non-image geoscience knowledge into the interpretation process due to the phenomenon of "different objects with the same spectrum" or "same object with different spectra" [18], which makes it difficult to control the accuracy of coastline information extraction based on remote sensing.

In this paper, inheriting the meta-architecture of encoder/decoder in the Mask2Former [19], a new sea–land segmentation method PMFormer is proposed. To deal with the insufficiency of prior information caused by data diversity and inconsistency in remote-sensing images, we introduce the prompt mask by normalized difference water index (NDWI) of the target image and prompt encoder architecture. To deal with the problem of large intra-class differences caused by the same object with different spectra, we introduce the query cluster into the block of the prompt encoder in the new architecture. We used three open datasets from different satellites in the experiments. Our proposed method is better at segmenting the sea–land boundary of remote-sensing images with

complex scenes and fine branches. The efficiency of the proposed method is confirmed, after the quantitative analysis and qualitative analysis.

2. Coastline Definition and Classification

In Figure 1, we can observe the seaward boundary, volume-based coastline, instantaneous shoreline, and landward boundary. In this paper, we use the definition sketch of the volume-based coastline position, which is stable against sediment redistribution within a cross-shore zone. The sea–land boundary in this paper mainly refers to the instantaneous shoreline in Figure 1.

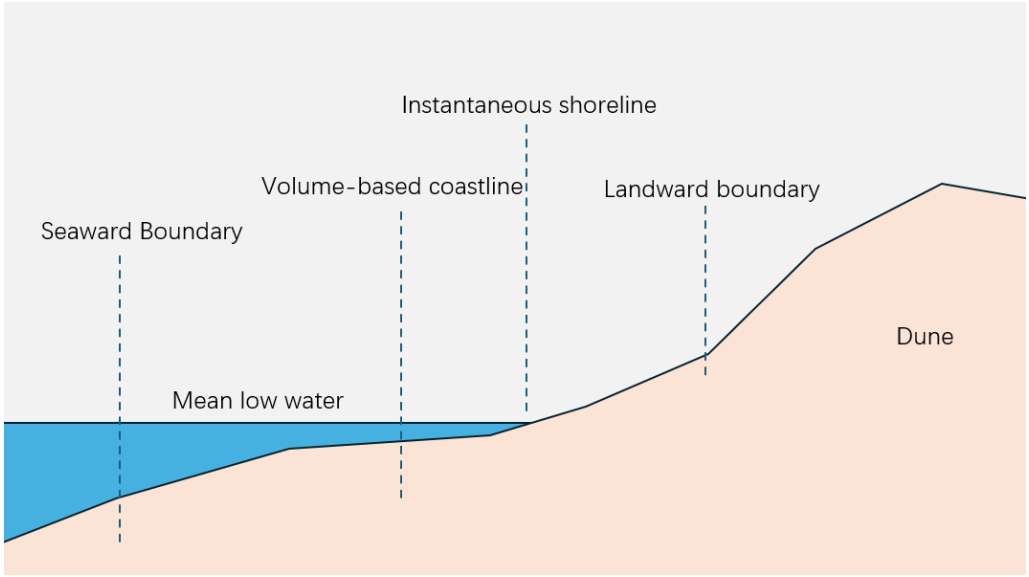


Figure 1. Coastline definition and sea–land edges [20].

The types of shoreline can be divided according to different classification criteria, such as topography, tidal action, sediment type and other factors. Common types of coastlines include: sandy coastlines, rocky coastlines, cliff coastlines, estuarine coastlines, etc. These different types of coastlines have their own characteristics in terms of geology, ecology, and human utilization. Coastlines are often effected in several ways: (1) ocean dynamic effects, (2) climatic factors, (3) sediment movement, and (4) human activities. Considering the above factors, the highly dynamic nature of the coastline is the result of the joint action of natural and human factors. The position of the shoreline is constantly changing, especially because of periodic rising and falling tides. Therefore, the coastline is actually a transitional zone, and has the characteristics of mutation and gradual change at the same time. These characteristics make it difficult to extract the sea–land boundary by remote sensing.

In the next section, we will discuss the basic principles of coastline extraction based on remote sensing, and briefly review the research status of different extraction methods with remote-sensing images, especially those based on deep learning.

3. Related Works for Sea–Land Segmentation

In this section, we mainly review the sea–land segmentation method with remote-sensing images. Satellite remote-sensing images are generated based on the different reactions of different ground objects to electromagnetic waves and the thermal radiation information of the ground objects themselves. Figure 2 is the spectrum of water and other materials. The water of rivers and sea is generally not pure water. The images of water are mainly the result of the interaction between the transmitted light and the chlorophyll, sediment, water depth, and thermal characteristics in the water, so the water bodies are generally green on the remote-sensing images. Based on the above principle, the absorption rate of water in the spectral reaction is generally low in the blue-

green band, and the absorption rate of other bands, especially the infrared band, is high, while the energy absorbed by vegetation and soil in these two bands is small and the reflectivity is high. This makes the water in these two bands obviously different from vegetation and soil. Therefore, in remote sensing, near-infrared band is often used to construct a model to extract water body information. The characteristics of water body reflected in each spectrum of electromagnetic wave are the basis of water body extraction by remote-sensing technology.

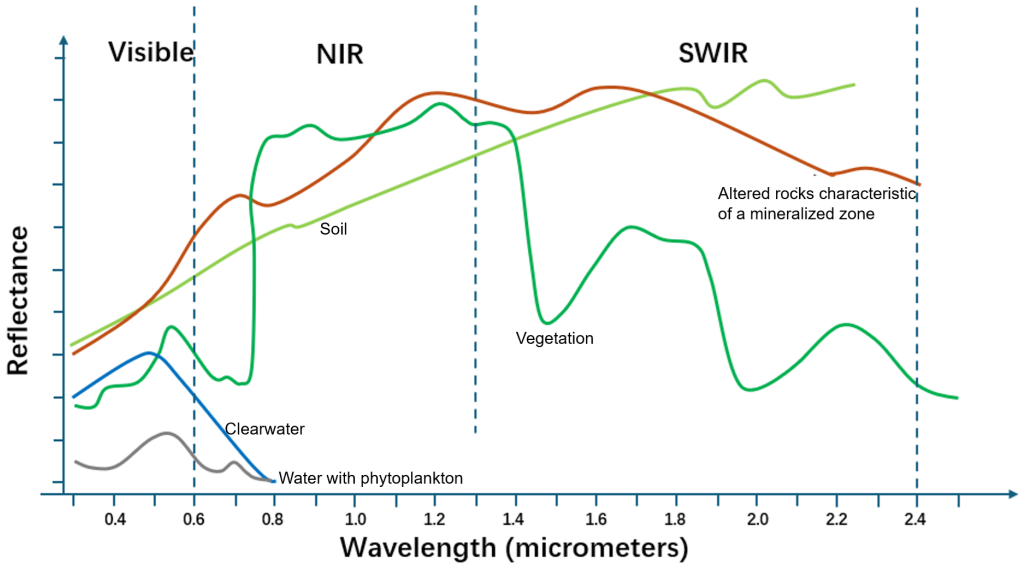


Figure 2. Spectrum of water and other materials [21].

In recent years, a large number of methods for sea–land segmentation based on remote-sensing images have been proposed. They can be mainly divided into: thresholding segmentation methods, traditional machine-learning methods, and deep-learning methods.

The threshold segmentation methods simplify the problem of the land and water boundary to the problem of threshold segmentation. There are various ways to calculate the threshold value. The threshold value can be obtained based on the statistical characteristics of remote-sensing data, or be calculated based on the water index considering spectral characteristics, or by establishing a more complete inversion model of remote sensing, etc. Since remote-sensing images are generally multi-band with relatively rich spectral features, the water index method based on the relationship between bands is very popular in the field of land and water segmentation. For example, normalized difference water index (NDWI) is an index to map waterbodies that employ the green band and NIR band [22]. There are also many improvements for NDWI, such as the second modified normalized difference water index (SMNDWI) [23], the locally adaptive thresholding technique [24], and weighted normalized difference water index (WNDWI) [25], etc. The advantage of thresholding methods is simple and easy to implement. If sea and land are with large spectral differences, the thresholding segmentation methods can usually provide satisfactory results for boundaries of coastal areas. However, the conventional thresholding segmentation methods only utilize the spectral information in remote-sensing images. It is often difficult to accurately discriminate areas or objects with a similar spectrum. More importantly, there is often no optimal thresholding, since the computation of threshold is easily affected by coast types, imaging conditions, sensors, climatic zones, weather, and phenology, etc. Therefore, the application of the thresholding methods are often limited in many sea–land segmentation scenarios.

Traditional machine-learning methods utilize texture-feature extraction to distinguish water bodies from land without a fixed shape in remote-sensing images, which is close to the semantic segmentation in the field of computer vision. A large number of semantic

segmentation methods have emerged. In general, many research achievements in computer vision, including graph cut theory [26,27], the random-walk-based method [28], the active contour-based method [29,30], region merging [31], mean shift [32], etc. are all introduced into sea–land segmentation. These methods mainly consider the similarity in feature spaces but not single spectral information in pixels. These segmentation methods can reduce the interference of the internal information of the pixel, but they belong to the shallow model with manual features. Design of the manual features severely depends on the experiences of the researcher, and their generalizations are limited without large labeled samples [33].

Deep-learning-based methods have shown promising performances in the field of semantic segmentation of images [34]. With the support of big data, semantic segmentation based on deep learning can have more parameters and advanced architectures. Therefore, they can not only learn features beyond the traditional semantic segmentation, but also have faster segmentation speed. In recent years, sea–land segmentation of remote-sensing images based on deep learning has also made great progress. For example, SANet [35] integrates the spectral, textural, and semantic features of ground objects at different scales, ACU-Net [36] replaces the two-layer continuous convolution in the feature extraction part of U-Net with a lightweight ASPP module, SDW-UNet [37] leverages the squeeze-excitation and depth-wise separable convolution to construct a new convolution module, DeepUNet [17] used the residual block on the basis similar to U-Net to extend the depth of the network for sea–land segmentation, RDU-Net [38] used dense connection blocks and combined both downsampling and upsampling paths, and SeNet [39] introduced structured edges into the deep semantic segmentation network.

These methods all show some good segmentation results on certain datasets. However, it is often difficult for them to deal with the crucial problem of foreground–background with a large intra-class difference and insufficiency of prior information in sea–land segmentation. There are still very big challenges in the sea–land segmentation of coasts because of its complexity and uncertainty, which limit the application of many deep-learning methods. In this paper, a new semantic segmentation architecture of transformer based on prompt mask-attention is proposed to solve the problem of prior information insufficiency in remote-sensing images and large intra-class differences in the binary segmentation of sea–land scenes. The methods proposed in this paper will be elaborated in the next section.

4. Method

In this section, an end-to-end method of Prompt Mask2Former (PMFormer) is proposed for semantic segmentation of the remote-sensing images of sea–land boundaries. Its framework is shown in Figure 3. We first introduce the motivation and the pipeline of mask classification that PMFormer is built on. Then, the motivation and pipeline and the major components of the network architecture of PMFormer, such as the multi-scale features, prompt encoding module, and fusion decoding module, will be introduced, respectively, in the following subsections.

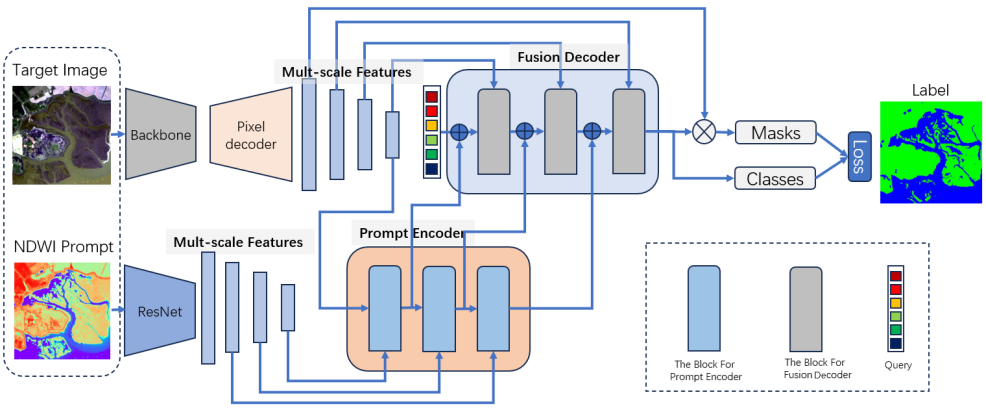


Figure 3. The overview of the PMFormer.

4.1. Motivation and Pipeline

Sea–land segmentation can be taken as a binary classification or a problem similar to foreground–background segmentation. Many methods based on per-pixel segmentation have shown good performances in the field of remote sensing, but it is still difficult to deal with a crucial problem of foreground–background with large intra-class difference and insufficiency of prior information. Such as in sea–land segmentation, there are always more different types of objects on land than in water, and many objects often have partly similar spectra. For example, when we treat sea–land segmentation as a foreground and background problem, the background as one class usually contains many materials such as vegetation, soil, buildings, etc. Other examples, such as the shadow generated by mountains or clouds, swamps or shallow areas, small watershed branches near the estuary, and some soil and vegetation that are easily confused with a water body in terms of spectrum, will produce significant errors in the semantic segmentation process. Therefore, the architecture of networks stills need many improvements before it can adapt to sea–land segmentation.

Mask2Former [19] shows significant advantages in dealing with the foreground–background problem. It converts a multi-class segmentation task into predicting a set of binary masks, which is based on mask-attention and especially suitable for discriminating foreground from background in complex scenes. However, Mask2Former is for a multi-class segmentation task without considering large intra-class differences and its first layer of mask-attention has no input as a prompt. It is a shortcoming because the first layer also needs an initial prompt and the prompt in the first layer will be the important source of all subsequent information.

Inspired by Mask2Former [19], the overall framework of our model is depicted in Figure 3, which is named Prompt Mask2Former (PMFormer). Architecturally, it partly inherits the encoder/decoder meta-architecture in the Mask2Former. Different from Mask2Former, there are two inputs in the proposed PMFormer: one is the target image itself, and the other is the normalized difference water index (NDWI) [22] of the target image.

The target image is first encoded by a backbone and passed into the pixel decoder, and then the first three feature maps from the pixel decoder are connected to the transformer decoder (fusion decoder). The block of the fusion decoder will generate the query and the mask for the next layer. However, the masks are not enough if it is only from the previous layer, especially for the insufficiency of prior information in foreground–background problems. Therefore, a prompt mask is needed.

The NDWI map of the target image is used as a prompt, which can provide the fundamental information to the initial mask-attention. To address the problem of large intra-class differences and insufficiency of prior information in foreground–background head-on, we develop prompt mask branch architecture and introduce a mask fusion mechanism into the transformer decoder to learn the relationships between foreground and background in a new way. In addition, to solve the problem of converting the multi-class problem into binary classification, we construct the new mask-attention block for the prompt branch and propose query cluster mechanisms for the prompt encoder. These two improvements will integrate both target mask-attention and prompt mask-attention into multi-scale features so that the decoding of the sea–land segmentation will be more accurate.

4.2. Feature Pyramid Networks

In the proposed method, PMFormer, both target image and NDWI prompt use feature pyramid networks (FPN).

Target images first pass through the backbone, which is a ViT [40], as in Figure 3. The backbone is an encoder for target images, but a feature in traditional ViTs is a lack of multi-scale features. Then, high-resolution per-pixel embeddings are generated by a pixel decoder, which up-samples low-resolution features from the output of the backbone, as well as feature pyramid networks (FPN). This stage is similar to a conventional

Mask2Former [19]. What is different is that the lowest resolution features in FPN not only enter into the first block of the fusion decoder to obtain the foreground feature maps but also are inserted into the first block of the prompt encoder.

For NDWI Prompt, we use ResNet50 architecture as the backbone to generate a feature pyramid, since its features are simpler than target images. The prompt is mainly from spectrum information and it can make up for some of the shortcomings caused by the fact that the data-driven method does not make physical sense. Furthermore, it will provide an initial mask for mask-attention in the transformer decoder (fusion decoder). The ResNet50 network is relatively simple, mainly including convolutional layers, followed by batch normalization and rectified linear unit (ReLU) activation functions. For example, the first layer C1: 256 × 256 × 64 indicates that the size of the feature map is 256 × 256, and the number of feature maps is 64. We can refer to reference [41] for more details about ResNet50. Through the ResNet50 network, the feature maps of different stages of the image are obtained, and its C2, C3, and C4 layers are used to establish the pyramid structure of the feature map. These multi-scale features are the prompt information with both water locations and spectrum characteristics. In consequence, they are used as inputs fed into the prompt encoder, which will be addressed in the next subsection.

4.3. Prompt Encoder

The prompt encoder is mainly composed of prompt coding blocks, and the number of blocks is determined by the target image and the multi-layer feature structure of NDWI. We mainly make improvements on two points: first, the coding block is a transformer structure based on the cross-attention mechanism, and its first input is the cross-association calculation of the NDWI layer and the low resolution feature of target images from the pixel decoder; second, the coding block in this paper makes some adjustments to the query, mainly because sea–land segmentation is a binary segmentation problem with large intra-class difference, while the original Mask2Former is more aimed at multi-class segmentation problems.

The architecture of the prompt encoder block is shown in Figure 4. It is composed of multi-head cross-attention, add and norm, and multi-head self-attention. The features from NDWI FPN are used as queries, and the feature from target image features with high resolution is used as key and value. NDWI as a steady feature with physical meaning can provide a good initial estimation for mask-attention, although it is not very precise. Since the queries in traditional Mask2Former are for the multi-class problem, we add query cluster components in the proposed prompt encoder block. The query cluster component is composed of global average pooling (GAP), linear, ReLU, query matrix, FCN, and SoftMax as in Figure 4. Query cluster maps the many queries into few large classes, as well as foreground (sea water) and background (vegetation, soil, buildings, etc.). Because there are many different types of ground objects in the background, the information within the background class is relatively richer. Directly converted into two types of queries, it is easy to generate errors, so the query cluster actually plays a transitional role in pre-classification for the stage of prompt encoding.

Specifically, we use deep cluster networks to realize query clustering. Given a set of data samples $\{\mathbf{x}_i\}_{i=1}^{N}$ (feature samples in this paper), the task of clustering is to group the N data samples into K categories. Similar to K-means, we define that $\mathbf{s}_i$ is the assignment vector of data point i, which has only one non-zero element, $s_{i,j}$ denotes the jth element of $\mathbf{s}_i$, and the kth column of $\mathbf{M}$, i.e., m_k, denotes the centroid of the kth cluster. To prevent trivial low-dimensional representations such as all-zero vectors, we use $g(\cdot)$ to map the feature back to the data domain and require that $g(\cdot)$ and $\mathbf{x}_i$ match each other well under some metric, e.g., mutual information or least squares-based measures. Inspired by DCN [42], the loss for query cluster is:

$$\min \sum_{i=1}^{N} (\|g(f(\mathbf{x}_i)) - \mathbf{x}_i\|_2^2 + \lambda \|f(\mathbf{x}_i) - \mathbf{M}\mathbf{s}_i\|_2^2) \qquad s.t. \quad s_{i,j} \in \{0,1\}, \mathbf{1}^T \mathbf{s}_i = 1, \forall i, j \quad (1)$$

where **1** is a unit vector, $f(\mathbf{x}_i)$ is the encoding architecture, $g(\cdot)$ is the decoding architecture (such as a fusion decoder), $\|f(\mathbf{x}_i) - M\mathbf{s}_i\|_2^2$ is the cluster loss, and $\|g(f(\mathbf{x}_i)) - \mathbf{x}_i\|_2^2$ is the reconstruction loss. However, in our architecture, it is not rigorous reconstruction loss since we perform a image segmentation. Therefore, we use $\|g(f(\mathbf{x}_i)) - \mathbf{l}_i\|_2^2$ at the end of the whole model, where $\mathbf{l}_i$ is the label of pixel $\mathbf{x}_i$. Cluster center $\mathbf{M}$ is the query matrix, where $\mathbf{M} = \{q_1, q_2, \cdots, q_n\}$. Since sea–land segmentation is a binary problem, we need to simplify $\mathbf{M}$ by a full connection network in the end of the query cluster as in the middle of Figure 4. In training, for fixed $\mathbf{M}$ and $\mathbf{s}_i$, SGD is used to update the network parameters; for fixed network parameters and $\mathbf{M}$, the assignment vector of the current sample, i.e., $\mathbf{s}_i$, can be naturally updated in an online fashion.

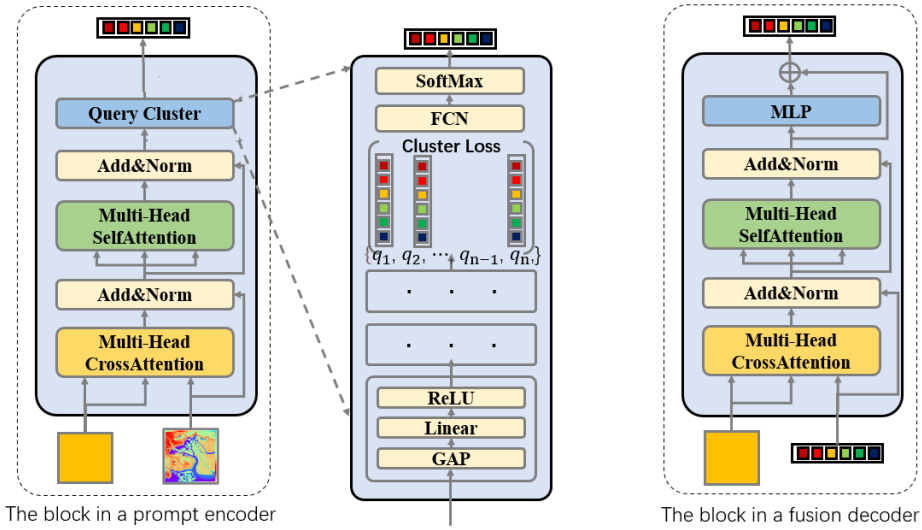


Figure 4. The block for the prompt encoder and the block for the fusion decoder.

The output of each block in the prompt encoder goes two ways: one is used as the part input of the next block, the other goes into the block of the fusion decoder to join into its mask-attention by a fusion way, which will be elaborated in the next subsection.

4.4. Fusion Decoder

The fusion decoder is composed of S blocks, which are similar to a transformer architecture. A standard transformer decoder (STD) mainly consists of a multihead self-attention layer, a cross-attention layer, and a multilayer perceptron (MLP) layer. In STD, the residual connections also are introduced between each layer and sublayers, which is followed by layer normalization. The self-attention layer uses learnable query features as input. More importantly, the cross-attention layer involved both query features and the FPN image features. It is not easy for the cross-attention layer to capture the global context of feature maps, especially in foreground–background problems with large intra-class differences such as sea–land segmentation.

To introduce the proposed fusion decoder, we need to review the mask-attention in Mask2Former. The standard cross-attention function with residual connection is calculated as:

$$X_l = \text{softmax}(Q_l K_l^T) V_l + X_{l-1} \tag{2}$$

where l is the index for layer and X_l is the token in the l layer of attention. $Q_l = f_Q(X_{l-1})$, $K_l = f_K(F_{l-1})$, $K_V = f_V(F_{l-1})$, where $f_Q(\cdot)$, $f_K(\cdot)$ and $f_V(\cdot)$ are linear functions. Q_l is query. K_l, and V_l are key and value for image features F_{l-1}, which are used to update the token. The mask-attentions are computed with Equation (3) as

$$X_l = \text{softmax}(M_{l-1} + Q_l K_l^T) V_l + X_{l-1} \tag{3}$$

Mask2Former switches the order of the self-attention layer and the cross-attention layer and uses mask-attention to reduce the searching area in cascaded transformer structures. Specifically, a query Q_l mask serves as the query of the cross-attention layer of the block, and K_l and V_l are served as the key and value of the cross-attention layer, which are obtained by the transformation of a level of feature map F_{l-1}, respectively. Meanwhile, attention mask M_{l-1} is used to constrain the area of attention operation. The mask-attention in Mask2Former works well on nature images but not enough for sea–land segmentation of remote-sensing images. In this paper, we replace the mask in the cross-attention layer with the proposed fusion mask-attention layer. As a result, mask-attention guiding $\mathcal{M}_{l-1}$ in feature position (x, y) is the value such as Formula (4)

$$\mathcal{M}_{l-1}(x,y) = \begin{cases} 0, & If \quad M_{l-1}(x,y) = 1; \\ -\infty, & otherwise. \end{cases} \tag{4}$$

where $M_{l-1} = M_{l-1}^{pre} + M_{l-1}^{pro}$. M_{l-1}^{pre} is the mask from the previous layer of the fusion decoder block. M_{l-1}^{pro} is the mask form the prompt encoder block. The new M_{l-1} is generated by considering the multi-scale features both in the target image and NDWI prompt.

Since the NDWI prompt mainly comes from the spectrum characteristics of water, its physical meaning will alleviate errors in learning complex foreground backgrounds with data diversity and inconsistency. There is no prompt mask M_{l-1}^{pro} in a conventional Mask2Former architecture. When we recheck the decoder in Mask2Former, there is no mask-attention for the first block in its transformer decoder. It is because its first input is not one layer of FPN but an initial query. However, the first mask is important for the attention, because the decoder needs a proximate relationship between foreground and background at the beginning. Therefore, the proposed PMFormer, provides a first mask as the prompt for the transformer decoder and makes an improvement on the structure of its query.

Now, the motivation and pipeline and the three major components of PMFormer (multi-scale features, prompt encoding module, and fusion decoding module), have been presented, respectively. In the following, we will carry out experiments to verify the proposed method.

5. Experiments and Analysis

5.1. Datasets

To evaluate the proposed segmentation algorithm objectively and comprehensively, we used three different datasets, which are the Sentinel-2 Water Edges Dataset (SWED) [43], Sea-Land-Landsat8 dataset (SLL8) [44], and Sea–Land-GaoFen1 dataset (SLGF) [35].

The Sentinel-2 Water Edges Dataset (SWED) is from reference [43,45]. The authors present a new labeled image dataset of Sentinel-2 [46] SWED for the development and bench-marking of techniques for the automated extraction of coastline morphology data from Sentinel-2 images. Composed of 16 labeled training Sentinel-2 scenes (25,402 samples with a size of 256 × 256), and 98 test label–image pairs, SWED is globally distributed and contains examples of many different coastline types and natural and anthropogenic coastline features.

Sea-Land-Landsat8 dataset (SLL8) is from reference [44]. It is constructed by Landsat-8 OLI [47] images from China's off shore areas, which are from different coastlines. It contains 1950 training patch images with dimensions of 512 × 512 pixels, which are from 17 image samples. Similarly, it also contains 1411 validating and testing images with the same image sizes, which are from 12 image samples. There are two image classes: land and sea. The rivers and lakes in land are treated as land.

The Sea–Land-GaoFen1 dataset (SLGF) is from reference [35]. This dataset was constructed by nine multi-spectral remote-sensing images of GaoFen-1 satellite [48]. The spatial resolution of the multispectral images is 8 m. There are four bands in these images, as well as red, green, blue, and near-infrared bands. The imaging location is in the Lianyungang

coastal zone, Jiangsu Province, China. As illustrated by reference [35], each selected image contains a different type of coastline. Remote-sensing images of coastal areas are cropped and labeled by experts through visual interpretation. The ground truth maps are binary images. The labeled images are divided into 256 × 256 samples by checkerboard segmentation. The training set contains 1544 samples, the validation set contains 178 samples, and the testing set contains 192 samples.

5.2. Experiment Setting

In this paper, the proposed PMFormer method is compared with eight image segmentation methods based on deep learning in experiments. They are FCN [49], UNet [50], PSPNet [51], Deeplabv3 [52], DaNet [53], Segformer [54], PointRend [55], and BiseNet [56]. We also add the NDWI [57] as a typical traditional method for comparison in the experiments.

There are twelve bands in the Sentinel-2 Water Edges Dataset (SWED), and we use all the 12 bands of Sentinel-2 data in segmentation experiments. For Sea–Land-Landsat8 dataset (SLL8), there are 2–6 bands and we use these five bands in experiments. There are four bands in the Sea–Land-GaoFen1 dataset (SLGF) and we use these four bands in the experiments. In the figures, we use false-color images to show the results.

For the parameters, we use the transformer as decoder and encoder with 9 layers and 100 queries. An auxiliary loss is added to every intermediate transformer layer and to the learnable query features. We use the binary cross-entropy loss and the dice loss for our mask loss. The model is trained for 60 epochs with the Adam optimizer. The initial learning rate is 5×10^{-5} and reduces by 10 times in 50-th epoch for both ResNet and Transformer. We use the feature pyramid with resolution 1/32, 1/16, and 1/8 of the original image

For fairness, we utilized the same training configuration and data-augmentation strategy for all models and datasets. All experiments were conducted on an MMSegmentation Development Toolkit [58] with an NVIDIA RTX 3090Ti GPU, utilizing the experimental environment based on PyTorch 1.11.0, CUDA 11.3, CuDNN 11.3, and MMEngine 0.7.3. To ensure a fair and unbiased comparison, these settings were employed to test the performance of all the methods. In this paper, we use Acc (accuracy), mIoU (mean intersection over union), F1 (F1 score) and Pr (precision) as evaluation metrics to evaluate and analyze the experimental results.

5.3. Results and Analysis

5.3.1. Quantitative Comparison

In this subsection, we use the original ratio of the training set, validation set, and testing set for the SWED, SLGF, and SLL8 datasets. All models are trained for 5000 iterations. We validate the models on the validation set and save the current best model weights until the maximum iterations are reached. Finally, we load each best model weight in turn and test the models on the testing set. As shown in Table 1, all models achieve more than 70% mIoU, and more than 90% F1 score.

After comprehensive comparison, we can observe that the performance of PMFormer algorithm is better than other algorithms by considering average indexes in most cases. In Table 1, the main difference in accuracy is shown on the SWED dataset and SLGF dataset, especially the SWED dataset. The mIoU index of the proposed PMFormer algorithm on the SWED dataset can reach 0.9212, which is significantly better than other algorithms. The mIoU of PMFormer on the SLGF dataset can also reach 0.9853. Although it is not the highest, it is better than most other algorithms. On the SLL8 dataset, because all algorithms reach almost 99% mIoU, they are not very differentiated on this dataset. Segformer also show good performances. For many cases, Segformer presents second top index especially for SWED and SLL8, and in some cases it shows relative high index for SLGF dataset. In addition to Segformer, Deeplabv3 also shows good performances on the SLGF and SLL8 datasets. DaNet, PointRend, UNet, etc., also have a high score on a particular dataset, but their comprehensive performances are not good as the proposed PMFormer. Some traditional methods perform reasonably well on average, but are often not as good as new

methods at ranking high on a particular dataset. For example, the mIoU of FCN on the SWED dataset reached 0.84, and the mIoU on the SLGF dataset reached 0.9779, which are relatively good but not outstanding.

Table 1. The results of different methods on different datasets. For all metrics, the first is highlighted with yellow, the second is brown, and the third is green.

Model	SLGF				SWED				SLL8			
	mIoU	F1	Pr	Acc	mIoU	F1	Pr	Acc	mIoU	F1	Pr	Acc
PMFormer	0.9853	0.9938	0.9989	0.9973	0.9212	0.9433	0.9567	0.9487	0.9921	0.9972	0.9966	0.9975
BiseNet	0.9136	0.9549	0.9549	0.9497	0.8455	0.9162	0.9157	0.9178	0.9819	0.9909	0.9916	0.9915
UNet	0.9155	0.9559	0.9569	0.9485	0.8768	0.9343	0.9325	0.9357	0.983	0.9914	0.9925	0.9922
PSPNet	0.9147	0.9555	0.9578	0.9554	0.8597	0.9244	0.9236	0.9259	0.9911	0.9955	0.9953	0.9958
Segformer	0.9725	0.9861	0.9839	0.9861	0.8841	0.9384	0.9483	0.9397	0.9919	0.9959	0.9954	0.9965
Deeplabv3	0.9818	0.9908	0.9909	0.9915	0.7177	0.8356	0.8370	0.8361	0.9916	0.9958	0.9957	0.9959
FCN	0.9779	0.9888	0.9889	0.9858	0.8433	0.9149	0.9136	0.9164	0.9901	0.9951	0.9948	0.9952
DaNet	0.9975	0.9987	0.9988	0.9977	0.7864	0.8803	0.8787	0.8815	0.9898	0.9948	0.9947	0.9954
PointRend	0.9798	0.9898	0.9898	0.9891	0.8119	0.8961	0.8941	0.9002	0.9901	0.9953	0.9950	0.9951
NDWI	0.5968	0.7063	0.6207	0.7186	0.5635	0.6638	0.7801	0.7441	0.4692	0.6624	0.4924	0.8393

We also need to notice that when the datasets are different, they have obvious influences on the coastline segmentation results. For these algorithms, SLL8 almost has the best indicator for all methods. Our approach achieved the most significant improvement on the SWED data. The results are not very stable for SLGF. We believe that this is related both to the model structure and the different data characteristics. For example, DaNet shows some better metrics than the proposed PMFormer method for the SLGF dataset. DaNet as a typical computer vision method is more suitable for natural images. The SLGF dataset with four bands is somehow similar to natural images. In DaNet, the co-channel attention module selectively emphasizes the presence of interdependent channel mappings by integrating the relevant features among all channel mappings. The output of the two attention modules is further improved on in the feature representation, which contributes to more accurate segmentation results. However, the proposed method mainly increased prompt encoding of NDWI and query cluster blocks. Both of these improvements require rich spectral features to work better. The SLGF dataset has only four bands with higher spatial resolution. For SLGF, we speculate that it is difficult to distinguish the spectral differences of ground objects due to insufficient spectral bands, which further leads to the unstable problem of land and water segmentation. Especially for some observation areas with the phenomenon of "different object with the same spectrum" or "same object with different spectra", it will exacerbate the difficulties caused by an insufficient spectrum of SLGF remote-sensing images. However, SLL8 and SWED have more spectral bands, which can provide richer and more stable spectral information, so they are more suitable for the segmentation method based on NDWI as the prompt information proposed in this paper.

Some algorithms are not very stable on different datasets. For example, UNet's mIoU on the SWED dataset is 0.8768, but its performance on the SLGF dataset is only 0.9155. Bisenet behaves similarly to UNet. Others like Deeplabv3 and DaNet both perform well on the SLGF dataset, but poorly on the SWED dataset.

In addition, although the nine algorithms all show high segmentation accuracy on SLL8, the resolution of SLL8 data is relatively low (Landsat-8, 30 m), so the actual information corresponding to the SLL8 high segmentation results may not be as accurate as the information provided by the other two datasets. On the contrary, the accuracy of the SLGF dataset appears to be less than that of the SLL8, because the spatial resolution is 2 or 8 m, which may actually provide more information about the sea–land boundary of the coastal zone.

Finally, although mIoU, Acc, and F1, etc., represent different precision meanings, they still have a certain correlation in general. Overall, for the problem of water and land

segmentation on remote-sensing images, in most cases, the proposed PMFormer method shows the better average performances on different indexes.

5.3.2. Qualitative Comparison

The examples of visual results of different methods for the SWED dataset, SLGF dataset, and SLL8 dataset are shown in Figures 5–7.

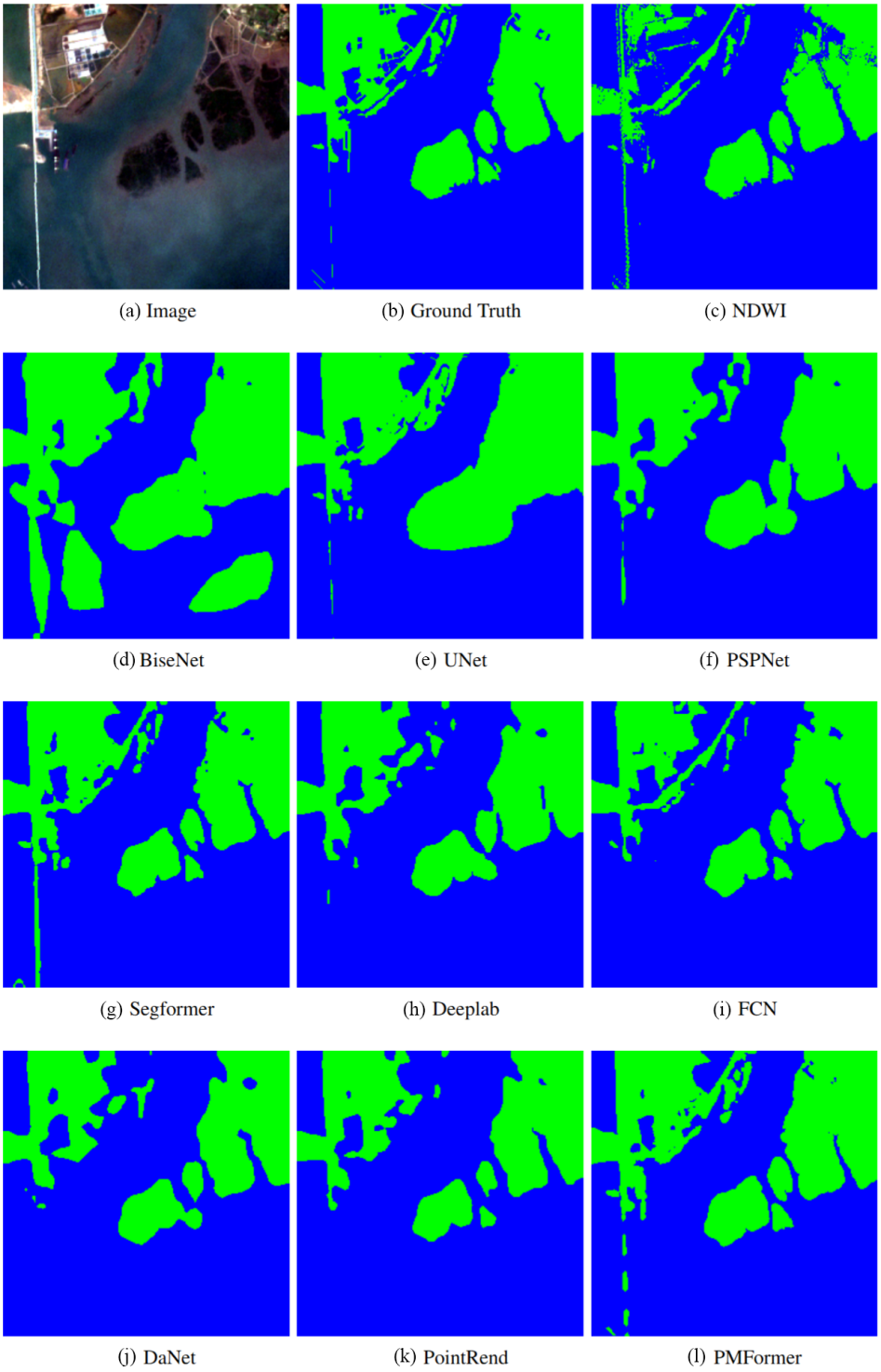


Figure 5. Results of different methods for SWED dataset.

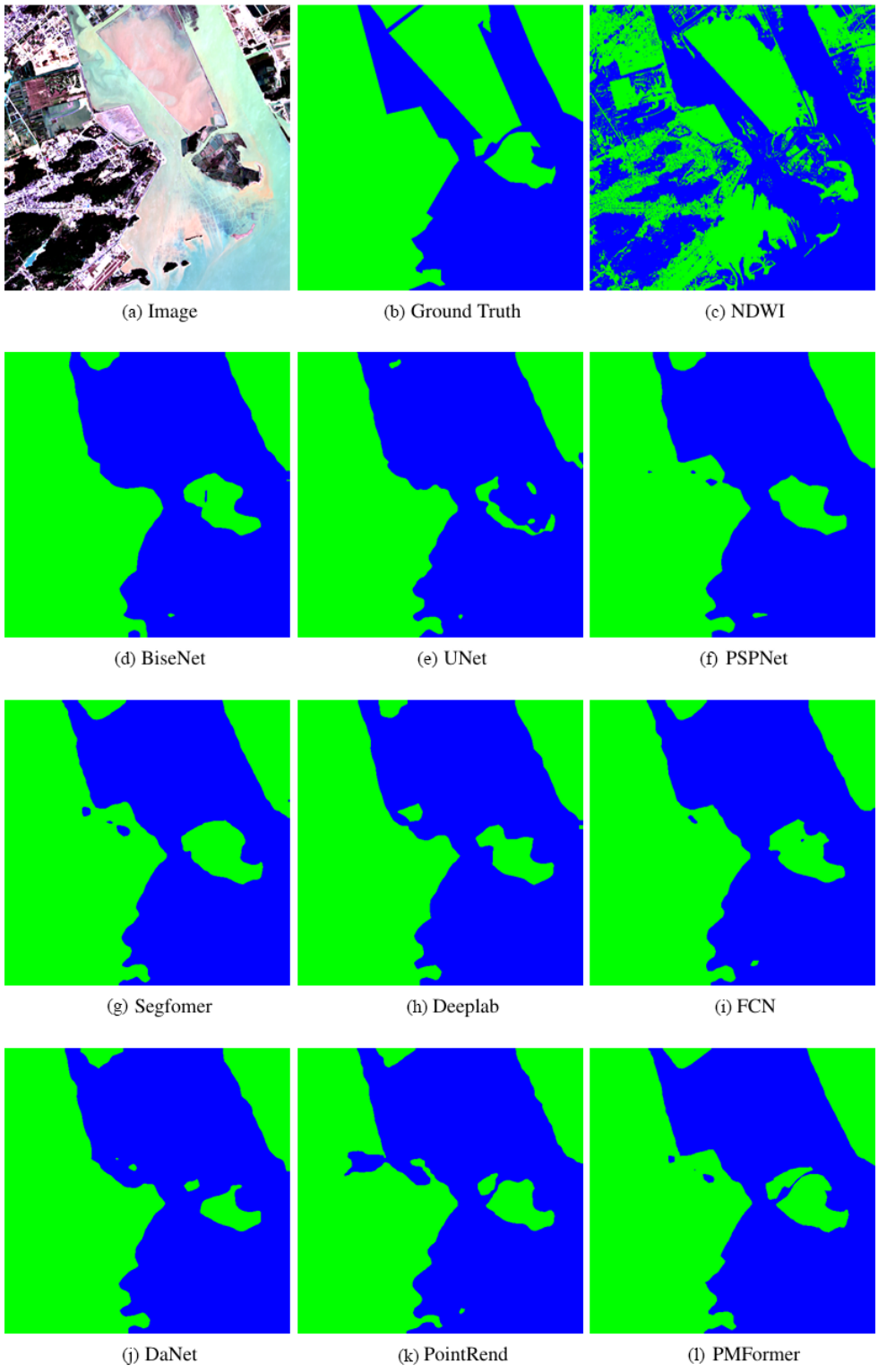


Figure 6. Results of different methods for SLL8 dataset.

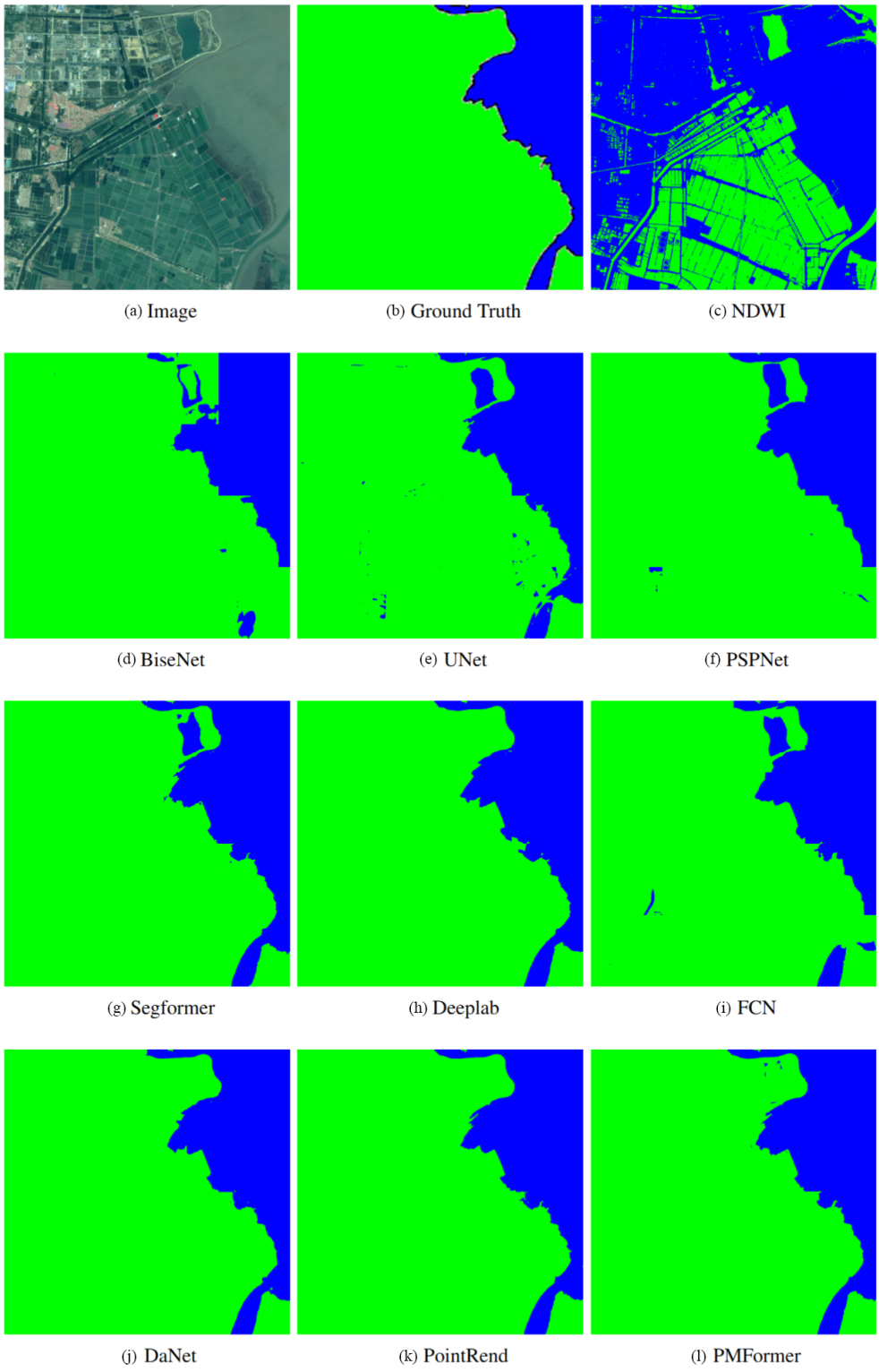


Figure 7. Results of different methods for SLGF dataset.

In Figure 5, we can observe that the best performance for the SWED dataset is the PMFormer algorithm. For small land areas and small line targets, the performance of the PMFormer algorithm is better than that of other methods. The objects on land areas are far more abundant while the characteristics of water are relatively homogeneous. There are both large intra-class differences and insufficiency of prior information. However, these influences from large intra-class differences are less in PMFormer than in other methods. This is attributed to the prompt mask-attention in the architecture of PMFormer. Segformer

is also a very promising method, which in most cases successfully segments small islands and shoal areas. In addition to the Segformer algorithm, the FCN algorithm also performs well. These phenomena are basically consistent with Table 1 data in quantitative analysis. NDWI shows good results on some linear targets, but also has large areas of error because the spectrum was prone to interference. BiseNet, UNet, and PspNet show a lot of errors. BiseNet mistakenly divided large areas of water into land. We can also observe that in UNet and PspNet the localization of edges and details is not accurate, and the shape of the land area in the segmentation results has a large deviation from the ground truth.

In Figure 6, we can observe that there are significant differences in the visual effects of the segmentation results, although the different methods show similarly good results on the SLL8 data simply from Table 1 of the quantitative analysis. The shoal area in the middle of the image visually contains a large amount of sediment, which causes all methods to incorrectly distinguish between shoals and bodies of water. We can observe that the NDWI method shows some good performance on some sediment but it is not stable. In this part of the experiment, large pieces of land were relatively accurate, but small pieces of island shape led to different results. We can observe that the error segmentation of the proposed PMFromer method for the area of sediment is relatively smaller than that of other methods. To some extent, after PMFromer, the segmentation of small islands is also close to the real value for Segformer and FCN methods. UNet, DaNet, and BiseNet do not perform as well as the other methods, which make mistakes in more areas. The quantitative comparison in Table 1 is the average of all validation images. It seems that although the overall error of the SLL8 data is not large, for different algorithms, there are still significantly different results (obvious errors) for specific regions and specific images.

In Figure 7, this is the visualization result of the nine algorithms on the SLGF dataset. The SLGF dataset has a higher resolution, which provides more detail for land and water segmentation, but it also introduces more interference due to more complex texture information. In particular, high-spatial-resolution images generally have a small number of bands in the spectrum, which can make it more difficult for the algorithm to locate the water–land boundary. We can observe that a single NDWI is far inferior to a class of methods based on deep learning because it does not have any supervised information and the spectrum information is not enough in SLGF. For the SLGF dataset, the PMFormer shows slightly better segmentation performances in the shoal region, which is not as significant as in the SWED dataset in Figure 6. Segformer does not perform as prominently as on the other two datasets. In addition, the performances of Deeplabv3, DaNet, and Pointrend are also good and close to PMFormer.

In order to clearly demonstrate the performances of the proposed method. A large image (Sentinel-2) is shown in Figure 8. It is an area that is near the mouth of the Thames River in England, with many beach rivers and small river branches. The boundary between land and water along the way is marked with curves with different color. We can observe that in most areas, the proposed method has achieved good performances.



Figure 8. The result of a large image by PMFormer. The size of the original image is 8000 × 2500. The sea–land boundaries are marked by curve lines with different color.

The method PMFormer proposed in this paper uses three datasets from different satellites, such as Sentinel-2 (SWED), Landsat8 (SLL8), and GaoFen1 (SLGF). The performances

were different. Based on the statistics and analysis of the results, we found that PMFormer is not always the best on the SLGF dataset, but it is almost always the best on the dataset SWED and SLL8. The SLGF dataset has only four bands with higher spatial resolution. For SLGF, we speculate that it is difficult to distinguish the spectral differences of ground objects due to insufficient spectral bands, which further leads to the unstable problem of land and water segmentation. SLL8 and SWED have more spectral bands, which can provide richer and more stable spectral information, so they are more suitable for the segmentation method based on NDWI as the prompt information proposed in this paper. Therefore, we can conclude that the method presented in this paper will have better results for remote-sensing images with rich spectral bands.

5.3.3. Calculation Speed

Table 2 shows the prediction speed of different algorithms on different datasets. In order to better evaluate the speed, each dataset is measured three times. We can observe that the speeds of BiseNet and UNet algorithms are faster than those of other methods. They are followded by PSPNet, Segformer, Deeplabv3, and PointRend. The speed of the proposed PMFormer algorithm is in the middle. The utilization of prompt mask-attention increases the complexity of the network. On the one hand, this architecture improves the accuracy, and on the other hand, it also affects the speed. FCN is obviously the slowest compared to other algorithms. The DaNet algorithm was significantly slowest on the SLL8 dataset, but performed in the middle on the other two datasets. We also listed the trainable params (TPs) and total mult-adds (TMAs) in Table 3. We believe that the different prediction speed for different methods mainly depends on the quantity of parameters (TPs) and the total mult-adds (TMAs) of the architectures.

Table 2. Time consumption of different methods on different datasets (Unit: second).

Model	SLGF			SWED			SLL8		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
PMFormer	4.587	5.148	3.992	8.629	8.013	7.936	32.596	33.215	39.118
BiseNet	1.679	1.678	1.582	3.9875	6.075	4.294	18.693	19.323	29.198
UNet	1.737	1.481	1.357	3.480	3.725	4.700	17.634	17.854	18.346
PSPNet	2.409	2.383	2.351	6.152	5.887	5.202	23.203	24.325	23.128
Segformer	2.386	3.001	2.568	4.874	4.797	6.631	23.462	24.062	34.606
Deeplabv3	2.928	3.479	3.011	6.245	5.254	7.821	23.503	26.579	37.606
FCN	7.689	8.345	7.741	17.968	16.989	15.533	26.135	24.750	26.282
DaNet	2.753	2.447	3.101	5.318	5.546	5.719	61.569	65.453	85.557
PointRend	2.988	2.547	3.026	6.885	6.047	7.549	28.913	30.183	43.410

Table 3. Trainable params (TPs) and total mult-adds (TMAs)

Method	PMFormer	BiseNet	UNet	PSPNet	Segformer	Deeplabv3	FCN	DaNet	PointRend
TPs	7,826,339	13,475,590	29,065,860	48,964,996	3,729,794	49,849,166	68,102,532	9,646,033	28,717,711
TMAs	15.98	18.47	51.49	178.46	6.57	199.19	269.67	18.77	56.11

In Table 2, PMFormer is slower than some of the other methods. For original Mask2Former, there is no prompt encoder modular, so it is fast. However, for the proposed PMFormer, we construct a prompt encoder and make it connect to a fusion decoder. This operation makes the structure more complex. There are more branches of the input and output of the intermediate link in the proposed PMFormer. Furthermore, the multi-scale features have to be computed twice because of the NDWI prompt inputs. Therefore, In Table 2, the timing results show that the proposed PMFormer is slow. Considering the prompt encoder, the proposed PMFormer takes at least half of its computation time to finish this part. We believe that in future work there is still room for improvement in the speed of

this approach if we can simplify the structure appropriately or replace some branches with efficient equivalent structures.

5.4. Ablation Studies

To demonstrate the improvements and advantages of the proposed method compared to the baseline model, we carry out the ablation experiments. There were three ablation experiments: (1) only use NDWI; (2) only use Mask2Former; (3) use PMFormer with both NDWI prompt and mask-attention. The results for the three cases are listed in Table 4. We can observe that the trend of improvement for the segmentation when NDWI provides an initial prompt and mask-attention is added in turns.

Table 4. Ablation experiment on three datasets.

Dataset	Model			mIoU	F1	Pr	Acc
	NDWI	Mask2Former	PMFormer				
SLGF	√	×	×	0.5968	0.7063	0.6207	0.7186
	×	√	×	0.9661	0.9790	0.9808	0.9832
	√	√	√	**0.9853**	**0.9938**	**0.9989**	**0.9973**
SWED	√	×	×	0.5635	0.6638	0.7801	0.7441
	×	√	×	0.8506	0.9058	0.9148	0.9086
	√	√	√	**0.9212**	**0.9433**	**0.9567**	**0.9487**
SLL8	√	×	×	0.4692	0.6624	0.4924	0.8393
	×	√	×	0.9789	0.9712	0.9761	0.9705
	√	√	√	**0.9921**	**0.9972**	**0.9966**	**0.9975**

We can also observe that neither a single NDWI nor a single mask-attention works well. In particular, a single NDWI is far inferior to a class of methods based on deep learning because it only considers spectrum features and does not utilize any supervised information. Mask2Former originally had a good prospect in the field of natural image processing. For sea–land segmentation of remote-sensing images, when we introduced NDWI into Mask2Former in the form of prompt encoding and performed fusion decoding, the overall effects are better on all three datasets than the original model.

Based on the previous network structure, comparative experiment, and ablation experiment, we can summarize three aspects about why our method is superior to others. First, the proposed method is both data driven and model driven. The prompt mask with NDWI can be considered model driven and has physical meaning since it comes from spectrum characteristics of water. However, the Mask2Former is a transformer architecture that is data driven and has excellent generalization ability if it is provided with enough training samples. Combining prompt encoding with Mask2Former, this hybrid model can take advantage of the merits of both sides. Second, we develop prompt mask branch architecture and introduce a mask fusion mechanism into the transformer decoder to learn the relationships between foreground and background in a new way. The NDWI prompt encoding can provide the fundamental information to the initial mask-attention. Especially, it provides the first prompt input for the first layer, which does not exist in the original Mask2Former architecture. Third, when we treat sea–land segmentation as a foreground—background problem, the background as one class usually contains many materials such as vegetation, soil, buildings, etc. To deal with the problem of large intra-class differences in the background, we proposed a query cluster for the block of prompt the encoder. The query cluster plays a transitional role in pre-classification for the stage of prompt encoding and it reduces the errors caused by large intra-class differences. Overall, these improvements allow the new network structure to better adapt to sea–land segmentation of remote-sensing images and obtain more accurate results.

5.5. Discussion

Despite the advantages of deep-learning semantic segmentation for sea–land boundary extraction, such methods still show some error areas. For different semantic segmentation frameworks, the errors of the boundary extraction various examples, such as the shadow generated by mountains or clouds, swamps or shallow areas, small watershed branches near the estuary, and some soil and vegetation that are easily confused with the water body in terms of spectrum, will produce significant errors in the semantic segmentation process.

Due to the characteristics of deep learning itself, many semantic segmentation can learn spectral features, spatial texture features, and even temporal features at the same time, so in fact, when deciding whether a pixel belongs to water or land, the semantic segmentation model generally considers the neighborhood relationship of the current pixel in a large range. This neighborhood relationship can be considered as the receptive field in the field of computer vision. On one hand, it can provide information to the segmentation task more comprehensively, and on the other hand, it also makes the task more complex. This is one of the reasons why the prompt mask-attention was introduced into the model. For example, although GaoFen satellite has only 4 bands, considering only pixel spectrum, in large receptive fields, the segmentation problem is still transformed into a complex high-dimensional problem when neighborhood relations are introduced based on multi-layer convolution.

We also need to summarize the main factors of error: (1) Shallow water region: In order to show areas prone to error, we have color-coded the areas that are incorrectly segmented in Figures 9 and 10. We can observe that shoals are the most error-prone area for almost all algorithms. This is mainly because the water in the shoals is relatively shallow and often contains a lot of sediment, and the shoals are not easily correctly distinguished from the visual characteristics or from the spectral characteristics. As shown in Figure 1, it is also true that shoals often have areas of transition between water and land without a particularly clear dividing line. The algorithms with good performances in the previous experiment, such as PMFormer, Segformer, or DeeplabV3, etc., are mainly good in the area of the shoal. Since this paper discusses remote-sensing image segmentation based on deep learning, and deep learning mainly relies on annotated sample data to ensure its accuracy in addition to its own network architecture, labeling as much representative data as possible in the shoal area may provide a more obvious help in improving the segmentation of sea–land boundaries. (2) Small branches of water: As shown in Figures 9 and 10, the small branches of water of the estuary of a river are also difficult to discriminate by some segmentation algorithms. In addition to the large amount of sediment, the estuary of the river is still a line target, which is easily ignored by the segmentation algorithm. For example, BisenNet, PSPNet and FCN algorithms in Figure 10 all produce relatively obvious errors in river estuary (blue areas). However, PMFormer shows promising performances in segmentation of small branches of water. (3) Shadows or clouds: In the process of coastline extraction, cloud and shadow will cause obvious segmentation errors. Especially many ground objects in the shadow area will be closer to the water body in spectral characteristics, so the shadow interference caused by mountains or clouds is difficult to remove based on a single image source. However, if combined with infrared or SAR images, or a longer series of historical images, it is possible to better eliminate shadow interference to a certain extent.

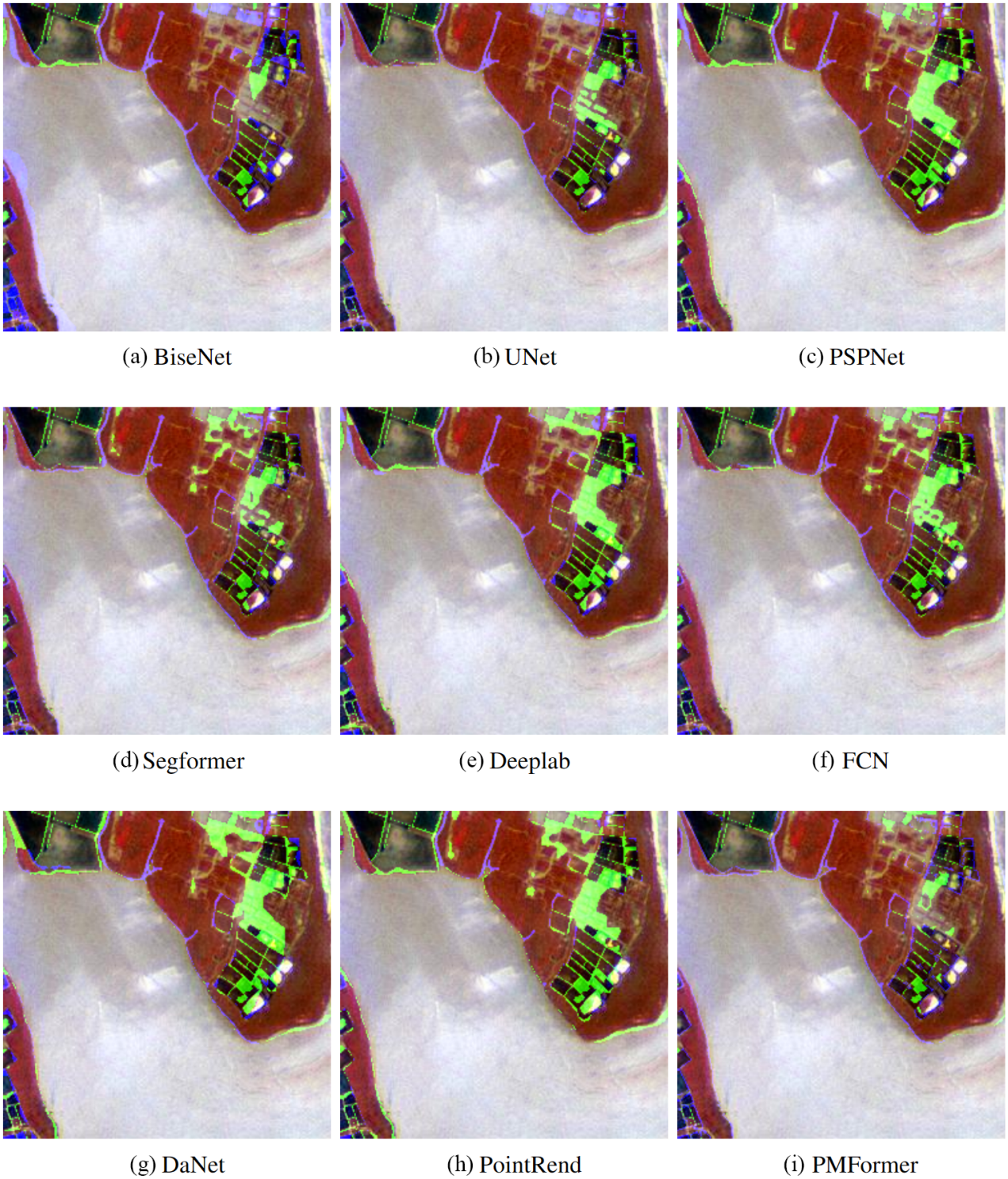


Figure 9. Error areas of different methods for the SWED dataset. The blue areas are where the land is mistakenly divided into water. The green areas are where water is mistakenly divided into land.

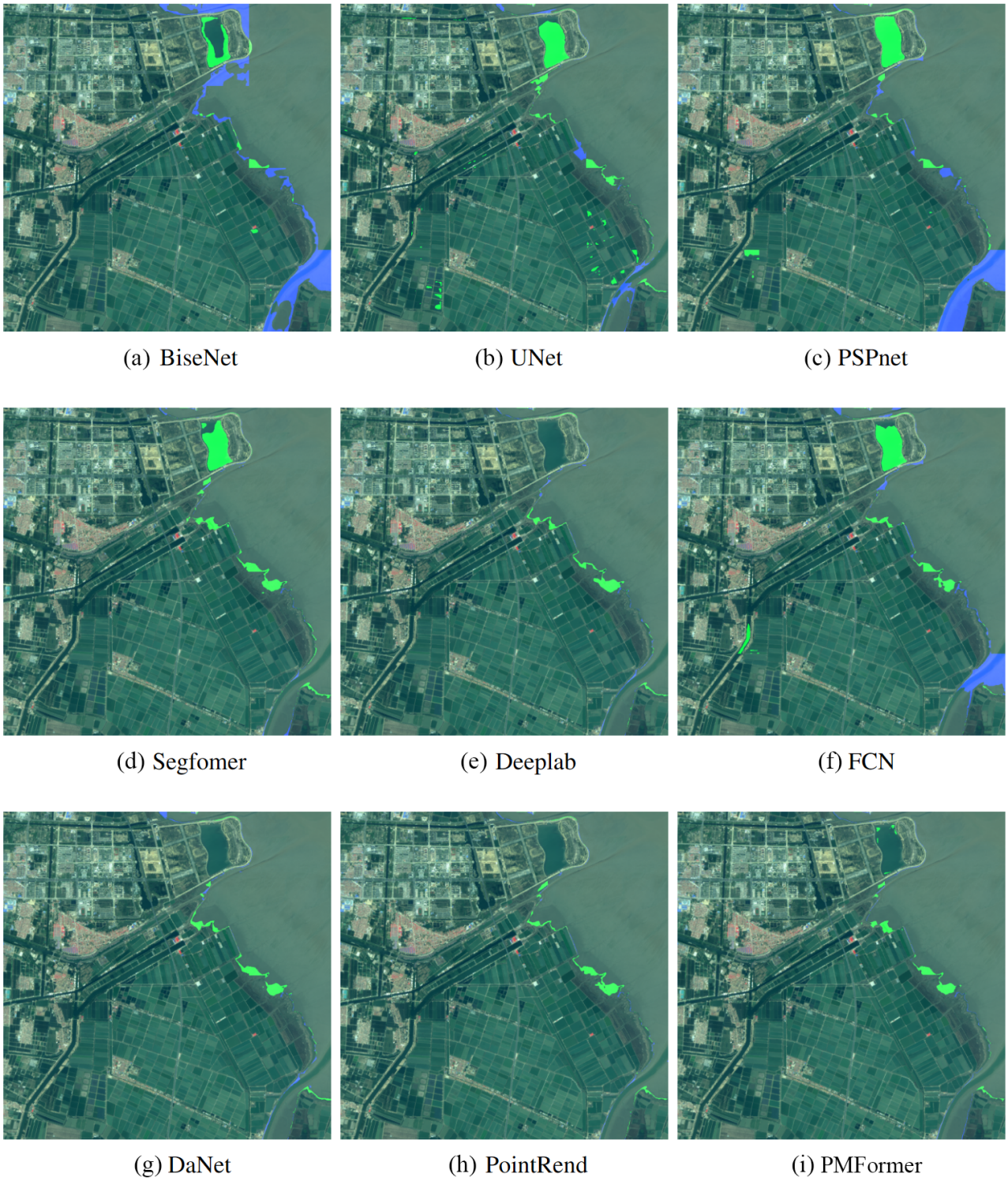


Figure 10. Error areas of different methods for the SLGF dataset. The blue areas are where the land is mistakenly divided into water. The green areas are where water is mistakenly divided into land.

6. Conclusions

In this paper, a new sea–land segmentation method, PMFormer, is proposed. Inheriting the meta-architecture of encoder/decoder in the Mask2Former, we introduce a new branch of encoding as the prompt into its architecture to deal with the problem of insufficiency of prior information and use query cluster to alleviate large intra-class differences in sea–land segmentation. In the experiments, we comprehensively make a comparison of the proposed

method with NDWI and eight other deep-learning sea–land segmentation algorithms. We also used three open datasets from different satellites to make an objective comparison. In terms of accuracy, on the SWED dataset, the proposed PMFormer performs significantly better than other methods; the proposed methods also performed well on SLL8 dataset and SLGF dataset. The results of qualitative analysis are consistent with those of quantitative analysis. The advantages of the proposed method are relatively obvious in the transition area between water and land, and small river branches into the sea or scattered small areas. In addition, when the data are with substantial spectral bands such as the SWED dataset and SLL8 dataset, the proposed method shows stably better abilities for sea–land segmentation than other methods. Overall, the effectiveness of the proposed PMFormer method is confirmed by quantitative comparison, qualitative comparison, ablation experiments, and time-consumption analysis.

Author Contributions: Conceptualization, Y.J. and S.L.; methodology, Y.J., W.W. and S.N.; software, Y.J. and S.N.; validation, Y.J., S.N. and J.W.; formal analysis, S.N. and J.W.; investigation, W.W. and J.W.; resources, W.W. and S.L.; data curation, S.N. and J.W.; writing—original draft preparation, Y.J. and S.L.; writing—review and editing, Y.J., S.L. and W.W.; visualization, J.W.; supervision, W.W. and S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript

Funding: This work was supported by the National Key R&D Program of China (2022YFB4501604).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: We acknowledge all the reviewers and editors for their constructive suggestions on this paper.

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

References

1. Batista, C.M.; Suárez, A.; Saltarén, C.M.B. Novel method to delimitate and demarcate coastal zone boundaries. *Ocean. Coast. Manag.* **2017**, *144*, 105–119. [CrossRef]
2. Rangel-Buitrago, N.; Neal, W.J.; de Jonge, V.N. Risk assessment as tool for coastal erosion management. *Ocean. Coast. Manag.* **2020**, *186*, 105099. [CrossRef]
3. Slinger, J.; Stive, M.; Luijendijk, A. Nature-Based Solutions for Coastal Engineering and Management. *Water* **2021**, *13*, 976. [CrossRef]
4. Mahrad, B.E.; Newton, A.; Icely, J.D.; Kacimi, I.; Abalansa, S.; Snoussi, M. Contribution of remote sensing technologies to a holistic coastal and marine environmental management framework: A review. *Remote Sens.* **2020**, *12*, 2313. [CrossRef]
5. Jordan, P.; Fröhle, P. Bridging the gap between coastal engineering and nature conservation? A review of coastal ecosystems as nature-based solutions for coastal protection. *J. Coast. Conserv.* **2022**, *26*, 4. [CrossRef]
6. Petrişor, A.I.; Hamma, W.; Nguyen, H.D.; Randazzo, G.; Muzirafuti, A.; Stan, M.I.; Tran, V.T.; Aştefănoaiei, R.; Bui, Q.T.; Vintilă, D.F.; et al. Degradation of coastlines under the pressure of urbanization and tourism: Evidence on the change of land systems from Europe, Asia and Africa. *Land* **2020**, *9*, 275. [CrossRef]
7. Sun, W.; Chen, C.; Liu, W.; Yang, G.; Meng, X.; Wang, L.; Ren, K. Coastline extraction using remote sensing: A review. *Giscience Remote Sens.* **2023**, *60*, 2243671. [CrossRef]
8. Apostolopoulos, D.N.; Nikolakopoulos, K.G. Assessment and quantification of the accuracy of low-and high-resolution remote sensing data for shoreline monitoring. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 391. [CrossRef]
9. Zanutta, A.; Lambertini, A.; Vittuari, L. UAV photogrammetry and ground surveys as a mapping tool for quickly monitoring shoreline and beach changes. *J. Mar. Sci. Eng.* **2020**, *8*, 52. [CrossRef]
10. Domingos, L.C.; Santos, P.E.; Skelton, P.S.; Brinkworth, R.S.; Sammut, K. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. *Sensors* **2022**, *22*, 2181. [CrossRef]
11. Song, B.; Liu, P.; Li, J.; Wang, L.; Zhang, L.; He, G.; Chen, L.; Liu, J. MLFF-GAN: A Multilevel Feature Fusion With GAN for Spatiotemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
12. Liu, P.; Li, J.; Wang, L.; He, G. Remote Sensing Data Fusion With Generative Adversarial Networks: State-of-the-art methods and future research directions. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 295–328. [CrossRef]
13. Chen, D.; Wang, Y.; Shen, Z.; Liao, J.; Chen, J.; Sun, S. Long time-series mapping and change detection of coastal zone land use based on google earth engine and multi-source data fusion. *Remote Sens.* **2021**, *14*, 1. [CrossRef]
14. Cui, J.; Ji, W.; Wang, P.; Zhu, M.; Liu, Y. Spatial–temporal changes in land use and their driving forces in the circum-Bohai coastal zone of China from 2000 to 2020. *Remote Sens.* **2023**, *15*, 2372. [CrossRef]

15. Toure, S.; Diop, O.; Kpalma, K.; Maiga, A.S. Shoreline detection using optical remote sensing: A review. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 75. [CrossRef]
16. Bishop-Taylor, R.; Nanson, R.; Sagar, S.; Lymburner, L. Mapping Australia's dynamic coastline at mean sea level using three decades of Landsat imagery. *Remote Sens. Environ.* **2021**, *267*, 112734. [CrossRef]
17. Li, R.; Liu, W.; Yang, L.; Sun, S.; Hu, W.; Zhang, F.; Li, W. DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3954–3962. [CrossRef]
18. Weng, L.; Gao, J.; Xia, M.; Lin, H. MSNet: Multifunctional Feature-Sharing Network for Land-Cover Segmentation. *Remote Sens.* **2022**, *14*, 5209. [CrossRef]
19. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
20. Dronkers, J. Definition Sketch of the Base Coastline. 2022. Available online: https://www.coastalwiki.org/wiki/File:BasisKustlijn.jpg (accessed on 1 July 2024).
21. Contributors, A. Spectral Signature of Water. 2023. Available online: https://mungfali.com/explore/Spectral-Signature-of-Water (accessed on 1 July 2024).
22. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
23. Li, M.; Zheng, X. A second modified normalized difference water index (SMNDWI) in the case of extracting the shoreline. *Mar. Sci. Bull* **2016**, *18*, 15–27.
24. Liu, H.; Jezek, K. Automated extraction of coastline from satellite imagery by integrating Canny edge detection and locally adaptive thresholding methods. *Int. J. Remote Sens.* **2004**, *25*, 937–958. [CrossRef]
25. Guo, Q.; Pu, R.; Li, J.; Cheng, J. A weighted normalized difference water index for water extraction using Landsat imagery. *Int. J. Remote Sens.* **2017**, *38*, 5430–5445. [CrossRef]
26. Cheng, D.; Meng, G.; Xiang, S.; Pan, C. Efficient sea–land segmentation using seeds learning and edge directed graph cut. *Neurocomputing* **2016**, *207*, 36–47. [CrossRef]
27. Boykov, Y.; Kolmogorov, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1124–1137. [CrossRef] [PubMed]
28. Qin, Y.; Bruzzone, L.; Gao, C.; Li, B. Infrared Small Target Detection Based on Facet Kernel and Random Walker. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7104–7118. [CrossRef]
29. Elkhateeb, E.; Soliman, H.; Atwan, A.; Elmogy, M.; Kwak, K.S.; Mekky, N. A novel coarse-to-Fine Sea-land segmentation technique based on Superpixel fuzzy C-means clustering and modified Chan-Vese model. *IEEE Access* **2021**, *9*, 53902–53919. [CrossRef]
30. Zhu, Z.; Tang, Y.; Hu, J.; An, M. Coastline extraction from high-resolution multispectral images by integrating prior edge information with active contour model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 4099–4109. [CrossRef]
31. He, W.; Song, H.; Yao, Y. An Improved Region Merging Approach for SAR Complex Water Area Segmentation. In Proceedings of the 2019 6th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Xiamen, China, 26–29 November 2019; pp. 1–5. [CrossRef]
32. Jarabo-Amores, P.; Rosa-Zurera, M.; de la Mata-Moya, D.; Vicen-Bueno, R.; Maldonado-Bascon, S. Spatial-Range Mean-Shift Filtering and Segmentation Applied to SAR Images. *IEEE Trans. Instrum. Meas.* **2011**, *60*, 584–597. [CrossRef]
33. Liu, P.; Wang, L.; Ranjan, R.; He, G.; Zhao, L. A Survey on Active Deep Learning: From Model Driven to Data Driven. *ACM Comput. Surv.* **2022**, *54*, 221:1–221:34. [CrossRef]
34. Zhang, Y.; Liu, P.; Chen, L.; Xu, M.; Guo, X.; Zhao, L. A new multi-source remote sensing image sample dataset with high resolution for flood area extraction: GF-FloodNet. *Int. J. Digit. Earth* **2023**, *16*, 2522–2554. [CrossRef]
35. Cui, B.; Jing, W.; Huang, L.; Li, Z.; Lu, Y. SANet: A sea–land segmentation network via adaptive multiscale feature learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 116–126. [CrossRef]
36. Li, J.; Huang, Z.; Wang, Y.; Luo, Q. Sea and Land Segmentation of Optical Remote Sensing Images Based on U-Net Optimization. *Remote Sens.* **2022**, *14*, 4163. [CrossRef]
37. Liu, T.; Liu, P.; Jia, X.; Chen, S.; Ma, Y.; Gao, Q. Sea-Land Segmentation of Remote Sensing Images Based on SDW-UNet. *Comput. Syst. Sci. Eng.* **2023**, *45*, 2. [CrossRef]
38. Shamsolmoali, P.; Zareapoor, M.; Wang, R.; Zhou, H.; Yang, J. A novel deep structure U-Net for sea-land segmentation in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3219–3232. [CrossRef]
39. Cheng, D.; Meng, G.; Cheng, G.; Pan, C. SeNet: Structured edge network for sea–land segmentation. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 247–251. [CrossRef]
40. 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 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Yang, B.; Fu, X.; Sidiropoulos, N.D.; Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 3861–3870.

43. Catherine Seale, Thomas Redfern, P.C. Sentinel-2 Water Edges Dataset. 2022. Available online: https://openmldata.ukho.gov.uk/ (accessed on 1 July 2024).
44. Yang, T.; Jiang, S.; Hong, Z.; Zhang, Y.; Han, Y.; Zhou, R.; Wang, J.; Yang, S.; Tong, X.; Kuc, T. Sea-Land Segmentation Using Deep Learning Techniques for Landsat-8 OLI Imagery. *Mar. Geod.* **2020**, *43*, 105–133. [CrossRef]
45. Seale, C.; Redfern, T.; Chatfield, P.; Luo, C.; Dempsey, K. Coastline detection in satellite imagery: A deep learning approach on new benchmark data. *Remote Sens. Environ.* **2022**, *278*, 113044. [CrossRef]
46. ESA. Sentinel-2. 2024. Available online: https://sentiwiki.copernicus.eu/web/sentinel-2 (accessed on 1 July 2024).
47. NASA. landsat-8. 2024. Available online: https://landsat.gsfc.nasa.gov/satellites/landsat-8/ (accessed on 1 July 2024).
48. Chen, L.; Letu, H.; Fan, M.; Shang, H.; Tao, J.; Wu, L.; Zhang, Y.; Yu, C.; Gu, J.; Zhang, N.; et al. An Introduction to the Chinese High-Resolution Earth Observation System: Gaofen-1 7 Civilian Satellites. *J. Remote Sens.* **2022**, *2022*, 9769536. [CrossRef]
49. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
50. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
51. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
52. 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.
53. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
54. 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.
55. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9799–9808.
56. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
57. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
58. Contributors, M. MMSegmentation Development Toolkit. 2023. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 1 July 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.