frontiers in Environmental Science

ORIGINAL RESEARCH
published: 28 February 2022
doi: 10.3389/fenvs.2022.799250

Waterline Extraction for Artificial Coast With Vision Transformers

Le Yang, Xing Wang and Jingsheng Zhai*

School of Marine Science and Technology, Tianjin University, Tianjin, China

Accurate acquisition for the positions of the waterlines plays a critical role in coastline extraction. However, waterline extraction from high-resolution images is a very challenging task because it is easily influenced by the complex background. To fulfill the task, two types of vision transformers, segmentation transformers (SETR) and semantic segmentation transformers (SegFormer), are introduced as an early exploration of the potential of transformers for waterline extraction. To estimate the effects of the two methods, we collect the high-resolution images from the web map services, and the annotations are created manually for training and test. Through extensive experiments, transformer-based approaches achieved state-of-the-art performances for waterline extraction in the artificial coast.

Keywords: coastline extraction, vision transformer, SegFormer, sea–land segmentation, waterline extraction

OPEN ACCESS

Edited by:
Peng Liu,
Institute of Remote Sensing and Digital Earth (CAS), China

Reviewed by:
Zq Gao,
Chinese Academy of Sciences (CAS), China
Min Chen,
Southwest Jiaotong University, China

****Correspondence:***
Xing Wang
xing.wang@tju.edu.cn

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal Frontiers in Environmental Science

Received: *21 October 2021*
Accepted: *10 January 2022*
Published: *28 February 2022*

Citation:
Yang L, Wang X and Zhai J (2022) Waterline Extraction for Artificial Coast With Vision Transformers. Front. Environ. Sci. 10:799250. doi: 10.3389/fenvs.2022.799250

INTRODUCTION

A coastline is the boundary between the dry and wet part in the coastal area when the high tide water is in the mean level Toure et al. (2018). The coastline is a critical geographic information source, and it is of great significance to autonomous navigation, coastal resource management, and protection of the environment Liu et al. (2013). Coastline extraction is a very challenging problem because it is obtained from a region not an instantaneous line. The waterline extraction is the precondition for computing the natural coastline, so the waterline extraction is very important and meaningful. The waterline is the instantaneous boundary between the land and sea. It can be extracted from the high-resolution images without other tools. In the artificial coast, the waterline can be considered as the coastline because the waterline is very slightly influenced by the tides.

With the development of satellite remote sensing technology, it supplies tons of high-resolution images of the coastal area, and they can be used for waterline extraction Roelfsema et al. (2013). Besides buying these remote sensing images directly from the remote sensing image providers, users can obtain many satellite map images freely from the web map services. All these data can be used for the waterline extraction.

The waterline extraction methods mainly include threshold segmentation methods, edge-based methods, object-oriented methods, active contour method, conventional machine learning methods, and deep learning methods. The threshold segmentation methods are intuitive methods that set a threshold value according to the image intensity to segment the land and water. Guo et al. (2016) proposed a method that utilized a normalized difference water index to segment water and land. Chen et al. (2019) used the components of the tasseled cap transformation to extract waterline information. Wernette et al. (2016) presented a threshold-based multi-scale relative relief method to extract the barrier island morphology from high-resolution DEM. These methods are handy and effective for the simple image segmentation task. In these methods, threshold selection is the key and difficult problem. In addition, the methods cannot deal with the images with occlusions or a complex background.

The edge-based methods utilize the distinctive edge feature from the abrupt transition. The common methods including Sobel, Roberts (Yang et al., 2018), Laplacian, and Canny operators (Lin et al., 2013; Paravolidakis et al., 2016; Ao et al., 2017; Widyantara et al., 2017; Paravolidakis et al., 2018) can be adopted to extract the waterline. Wang and Liu (2019) proposed a robust ridge-tracing method utilizing the statistical properties of the pixel intensities in the land and sea to detect the boundary. These methods are easy to detect clear, continuous boundaries. However, in the waterline images with a complex background, they are greatly affected by noise. The continuity of the extracted waterline is hardly guaranteed.

Object-oriented methods no longer use the pixel as the basic processing unit; instead, they use an object composed of homogeneous pixels (Gucluer et al., 2010; Rasuly et al., 2010; Bayram et al., 2017). Ge et al. (2014) presented an object-oriented multi-scale segmentation method using interpretation rule sets for automated waterline extraction from remote sensing imagery. Wu et al. (2018) used the object-oriented classification method to extract the waterline from Landsat images of Shenzhen city. These methods use higher level features to classify images, which can reduce the impact of fine texture characteristics on the results of image classification. However, in face of a large amount of information in high-resolution images, object-oriented classification methods may ignore some of the hidden useful information, and it is difficult to achieve the desired classification accuracy.

The active contour methods can achieve better results for remote sensing images of waterlines with simple backgrounds, strong contrast, and continuous boundaries. Cao et al. (2016) proposed a new geometric active contour model for waterline detection from SAR images, which is adaptive to the speckle noises. Fan et al. (2016) proposed a level set approach with a particle swarm optimization algorithm for waterline automatic detection in SAR images. Elkhateeb et al. (2021) adopted a modified Chan–Vese method for sea–land segmentation, which is initiated by a superpixel-based fuzzy c-means automatically. In the study by Modava and Akbarizadeh, (2017), a waterline extraction method–based active contour for SAR images is proposed, in which the initial contour is obtained from a fuzzy clustering with spatial constraints. In the study by Liu C et al., (2016), the waterline is extracted hierarchically by the level set techniques from single-polarization and four-polarization SAR images. Liu et al. (2017) integrated an edge-based and a region-based active contour model in different scales to fulfill the waterline detection from SAR Images. Due to the characteristics of the active contour model method, the application of this method is feasible for waterline images with a simple background, strong contrast, and continuous boundaries. However, the iterative method inevitably produces a large amount of calculation, which restricts its efficiency.

Conventional machine learning methods distill useful information and hidden knowledge based on a variety of data to extract the waterline. Rigos et al. (2016) and Vos et al. (2019) used a shallow neural network to extract the shoreline from satellite images and video images, separately. Sun et al. (2019) built a superpixel-based conditional random field model to segment the sea and land area. Dewi et al. (2016) presented fuzzy c-means methods to detect positions of the coastline and estimate the uncertainty of the coastline change. Cheng et al. (2016) proposed a graph cut method to segment the sea and land, in which the seed points are achieved by a probabilistic support vector machine. Compared with the traditional waterline extraction method, the shallow neural network, clustering analysis technology, fuzzy logic technology, and support vector machine use intelligent means to find out frequent regular things from a large number of data information effectively. These methods can automatically and efficiently extract regular objects. However, for more complex objects in high-resolution images, the extraction accuracy is unsatisfactory. Some other traditional methods, such as the polarization method (Nunziata et al., 2016), wavelet transform method (Toure et al., 2018), region growing method (Liu Z et al., 2016), and decision tree algorithm (Wang et al., 2020) , are all influenced by noise and cannot process the high-resolution images easily.

In these years, deep learning methods have been rapidly developing with the quickly growing performance of computer hardware. Different from traditional machine learning, it can learn the characteristics of the target more accurately. Some convolutional neural network (CNN) methods are naturally introduced in waterline extraction by segmenting the land and sea. In the study by Liu et al., (2019), a simple CNN with multi-scale features and leaky rectified linear unit (leaky-ReLU) activation function is used for waterline extraction. Liu W et al. (2021) proposed an end-to-end lightweight multitask CNN without downsampling to obtain lakes and shorelines from remote sensing images. Shamsolmoali et al. (2019) adopted a residual dense UNet to facilitate the hierarchical features from the original images for sea–land segmentation. Tsekouras et al. (2018) presented a novel Hermite polynomial neural network to detect the shoreline at a reef-fronted beach. Cheng et al. (2017a) proposed a local smooth regularized deep CNN that can obtain the segmentation and edge results of the sea and land simultaneously. Cheng et al. (2017b) employed a multitasking edge–aware CNN for sea–land segmentation and edge detection simultaneously. Cui et al. (2021) presented a scale-adaptive CNN for sea–land segmentation, which fused multiscale information and emphasized the boundaries' features actively. A sea–land segmentation approach utilizing the fast structured edge network and the waterline database was taken from the study by He et al., (2018). A novel UNet-like CNN was proposed for sea–land segmentation, and the network can be deeper, and the convergence can be faster based on local and global information (Li et al., 2018). Erdem et al. (2021) proposed a majority voting method based on different deep learning architectures to obtain shorelines automatically. Lin et al. (2017) presented a multi-scale end-to-end CNN for sea–land segmentation and ship detection, which can increase the receptive field while maintaining fine details. Even though the CNNs have achieved great performances, the limited receptive field affected the performance because of the structure of the CNN.

Transformers, as the most advanced methods in the semantic segmentation, are migrated to compute vision tasks to solve the problem of long-distance dependence by the self-attention

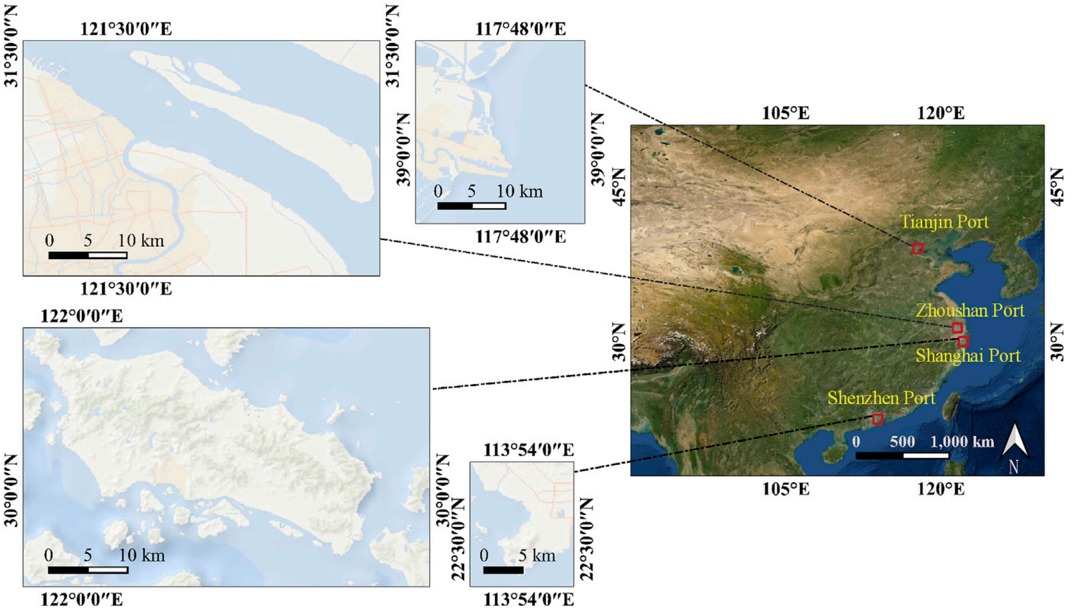

FIGURE 1 | Research area.

mechanism, which is the core of transformers. It determines the global contextual information of each item by capturing its interaction amongst all items. A vision transformer (ViT) is the first work that uses a pure transformer for image classification, which proves that the transformer can achieve the state-of-the-art (Dosovitskiy et al., 2021). It treats each image as a sequence of tokens and then feeds them to multiple transformer layers to make the classification. Subsequently, the dual intent and entity transformer (DeiT) (Touvron et al., 2021) further explores a data-efficient training strategy and a distillation approach for ViTs. The pyramid vision transformer (PVT) is the first work to introduce a pyramid structure in a transformer, demonstrating the potential of a pure transformer backbone compared to CNN counterparts in dense prediction tasks (Wang et al., 2021). After that, methods such as shifted windows (Swin) transformer (Liu Z et al., 2021), convolution transformers (CvT) (Xu et al., 2021), and twin transformer (Chu et al., 2021) enhance the local continuity of features and remove fixed size position embedding to improve the performance of transformers in dense prediction tasks. Segmentation transformers (SETRs) adopt the ViT as a backbone to extract features, achieving impressive performance in segmentation (Zheng et al., 2021). Following it, semantic segmentation transformers (SegFormer) achieved even better results later (Xie et al., 2021).

Therefore, we use the most advanced transformer methods to extract the waterline as an early exploration. This study mainly focuses on the process of extraction of the waterlines for artificial coasts and presents the early research for investigating the potential of transformers in waterline extraction from very high-resolution images.

The rest of the study has the following sections. **Materials and Methods** suggests details about the dataset and methodology. Results reports experimental results with a discussion. Finally, the conclusion section concludes and discusses future research directions.

MATERIALS AND METHODS

Dataset

For this research, we selected Tianjin, Zhoushan, Shanghai, and Shenzhen four ports as research areas, which are shown in **Figure 1**. The waterline images are collected from Mapbox (Mapbox, 2021), Google Maps (Google Maps, 2021), and Bing Maps (Microsoft, 2021) guided by OpenStreetMap (OSM) tiles. The images are in 18 levels in the map, and the initial resolution is 256 × 256. The ground sampling distance (GSD) is about 0.48 m. We combine each neighboring four tiles into a 512 × 512 image. A total of 600 images are chosen as the initial data, and the ground truths of waterlines are created by hand. We also augment it with the random rotation, flip, scale, contrast, brightness, and saturation to 6000 images. Among them, 3600 images are considered as the training set, 1200 images and 1200 images for validation and test, respectively. The images and corresponding annotations are indicated in **Figure 2**. To evaluate the effects of transformers, six CNN segmentation methods are introduced in the experiments.

Methodology

SETR

SETR is an Encoder–Decoder architecture, as seen in **Figure 3**. SETR adopted a high resolution of local features extracted by a CNN and the global information encoded by transformers to segment pixels in an image. Because of quadratic model complexity of the transformer, flattening the whole image as a

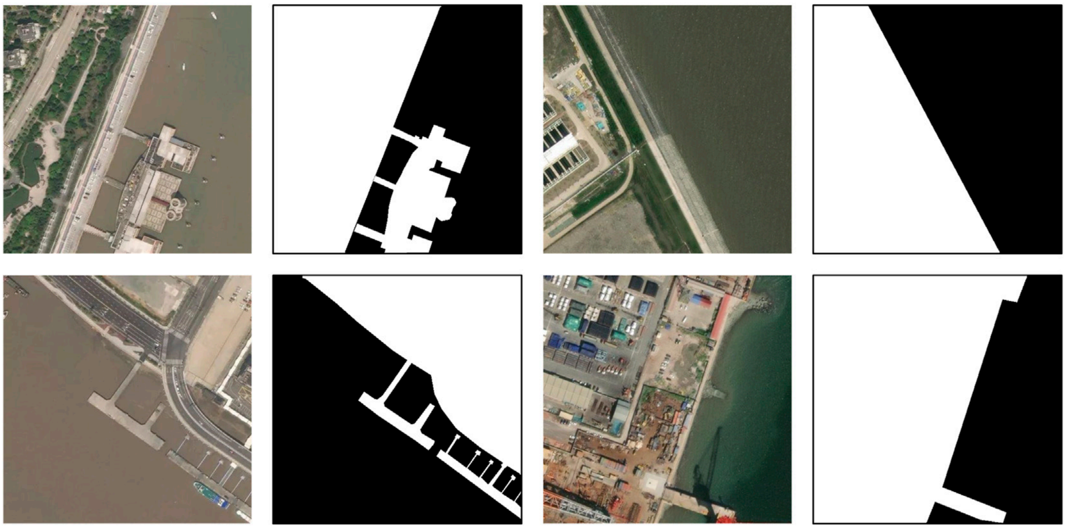

FIGURE 2 | Dataset.

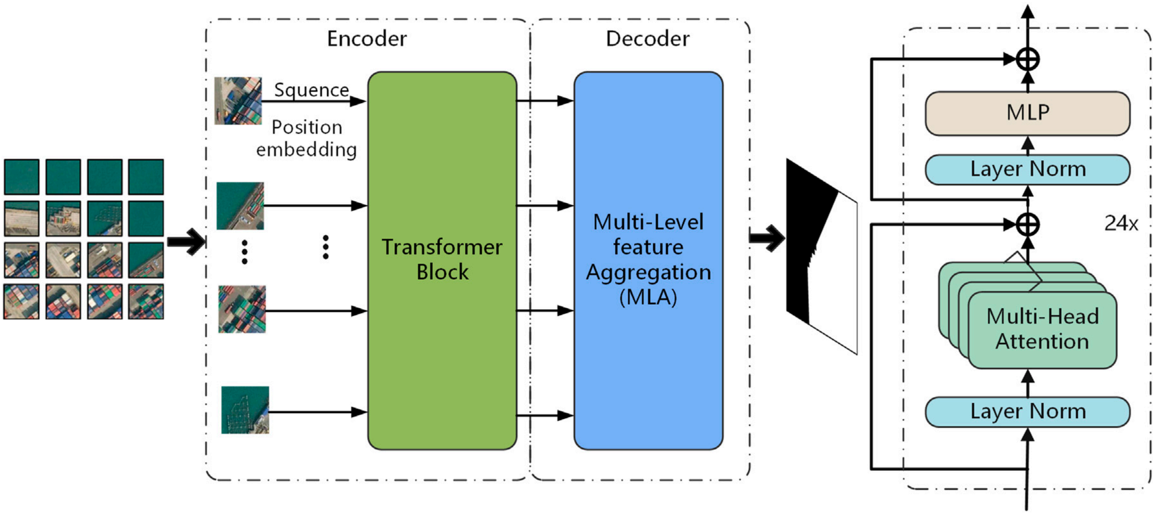

FIGURE 3 | Architecture of the SETR.

sequence makes a huge amount of computation. To speed up, an image is divided into 256 even patches, and then, each patch is flattened into a sequence for input separately.

All the sequences are entered into the pure transformer-based encoder. Therefore, all the transformer layers have a global receptive field, which improve the limited receptive field problem from the CNN. There are 24 layers of transformers in the encoder, in which there are multi-head self-attention (MSA), multilayer perceptron (MLP), and layer normalization blocks residually connected.

The decoder is called the multi-level feature aggregation (MLA). Some feature representations from the transformer layers are first reshaped from 2D to 3D and then aggregated. A 3-layer convolution network downsamples the features at the first and third layer. To enhance the interactions of different levels of features, a top-down aggregation design is introduced. The fused feature is obtained *via* channel-wise concatenation after the third layer. At last, the outputs are upscaled by bilinear operation to the original resolution.

SegFormer

The architecture of SegFormer is depicted in **Figure 4**. The SegFormer consists of two main modules, encoder and decoder. An image as the input is first divided into patches in 4×4. Then, these patches are imported to the hierarchical transformer encoder to obtain multi-level features. These multi-level features are passed to the MLP decoder to predict the segmentation mask at a $H/4 \times W/4 \times N_{cls}$ resolution, where H, W, N_{cls} are the height, width of the image, and the number of categories in the image, respectively.

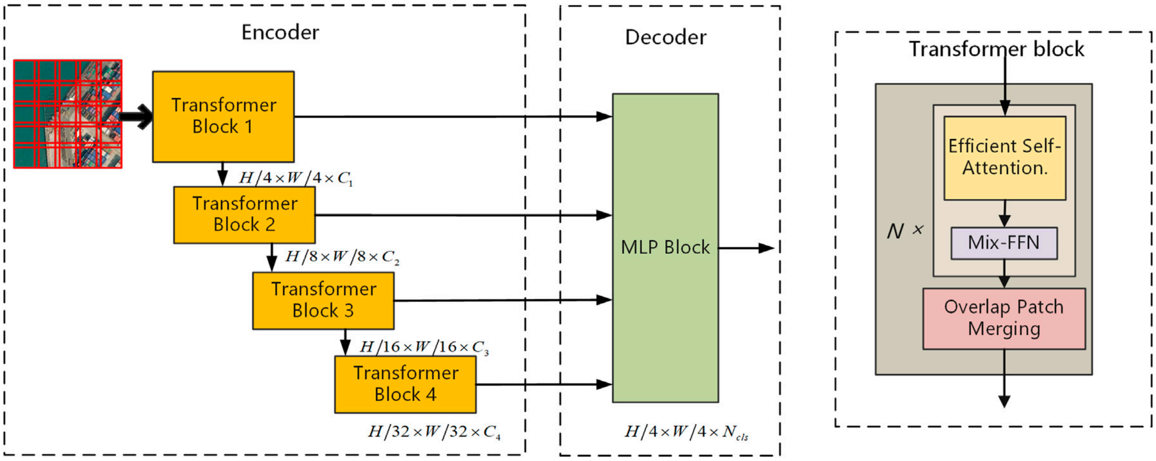
Encoder
Transformer Block 1
$H/4\times W/4\times C_1$
Transformer Block 2
$H/8\times W/8\times C_2$
Transformer Block 3
$H/16\times W/16\times C_3$
Transformer Block 4
$H/32\times W/32\times C_4$
Decoder
MLP Block
$H/4\times W/4\times N_{cls}$
Transformer block
Efficient Self-Attention.
Mix-FFN
N ×
Overlap Patch Merging

FIGURE 4 | Architecture of the SegFormer.

In the encoder, an input image is one with a resolution of $H \times W \times 3$, and C_i is the channel number in the feature map F_i. A hierarchical feature map F_i with a resolution of $H/2^{i+1} \times W/2^{i+1} \times C_i$ is obtained after each transformer block, where $i \in \{1, 2, 3, 4\}$, and C_{i+1} is larger than C_i.

The transformer consists of efficient self-attention, Mix-FFN, and overlap patch merging blocks. Efficient self-attention improves the computational efficiency of the self-attention. In the original multi-head self-attention process, each of the heads has the same dimension $N \times C$, where $N = H \times W$ is the length of the sequence, and C is the channel number. The self-attention is expressed as follows:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_h}}\right)V. \tag{1}$$

In the equation, d_h is the dimension of the head. The computational complexity of this process is $O(N^2)$. To alleviate it, the sequence K is reduced with a reduction ratio R. It is first reshaped into $N/R \times C{\cdot}R$ and then simplified by a fully connected layer. Therefore, the new K has dimensions $N/R \times C$. As a result, the complexity of the self-attention mechanism is reduced from $O(N^2)$ to $O(N^2/R)$.

ViT uses positional encoding (PE) to express the location information. It influences the test accuracy when the image resolution is not the same with that in the training because the positional code needs to be interpolated. To address it, Mix-FFN considers the effect of zero padding to leak location information, and a 3 × 3 convolution is used in the feed-forward network (FFN). Mix-FFN can be formulated as follows:

$$\mathrm{x}_{out} = \mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{MLP}\left(\mathrm{x}_{in}\right)\right)\right)\right) + \mathrm{x}_{in}, \tag{2}$$

where x_{in} is the feature from the self-attention module. Mix-FFN mixes a 3 × 3 convolution and an MLP into each FFN. The Gaussian error linear unit (GELU) (Hendrycks and Gimpel, 2020) is an activation function. x_{out} is the output of the Mix-FFN.

To preserve the local continuity around those patches, an overlapping patch merging process is used. K is the patch size, S is the stride between two adjacent patches, and P is the padding size. K = 7, S = 4, P= 3, and K = 3, S = 2, P= 1 are set to perform overlapping patch merging to produce features with the same size as the non-overlapping process.

The SegFormer incorporates a lightweight decoder consisting only of MLP layers. The proposed All-MLP decoder consists of four main steps. First, multi-level features from the encoder go through an MLP layer to unify the channel dimension. Then, features are upsampled to 1/4th and concatenated together. Third, an MLP layer is adopted to fuse the concatenated features. Finally, another MLP layer takes the fused feature to predict the segmentation mask.

The Proposed Method

In this section, we demonstrated a method used in the task of waterline extraction. The workflow is shown in **Figure 5**. The transformer first learns the coast features from the training samples. This step is the most time-consuming since most layers of the network are trained in this step. After the learning step, parameters of the model are convergent, and it can infer other new coast images for waterline extraction. Then, a binary mask of the coast is obtained from each input image; the waterline can be extracted from the mask easily. It is worth noting that the contours of the coast at the edges of the image should be excluded because this part is truncated when slicing image tiles.

Metrics

The proposed approaches are evaluated by precision, recall, F1-score, and IoU.

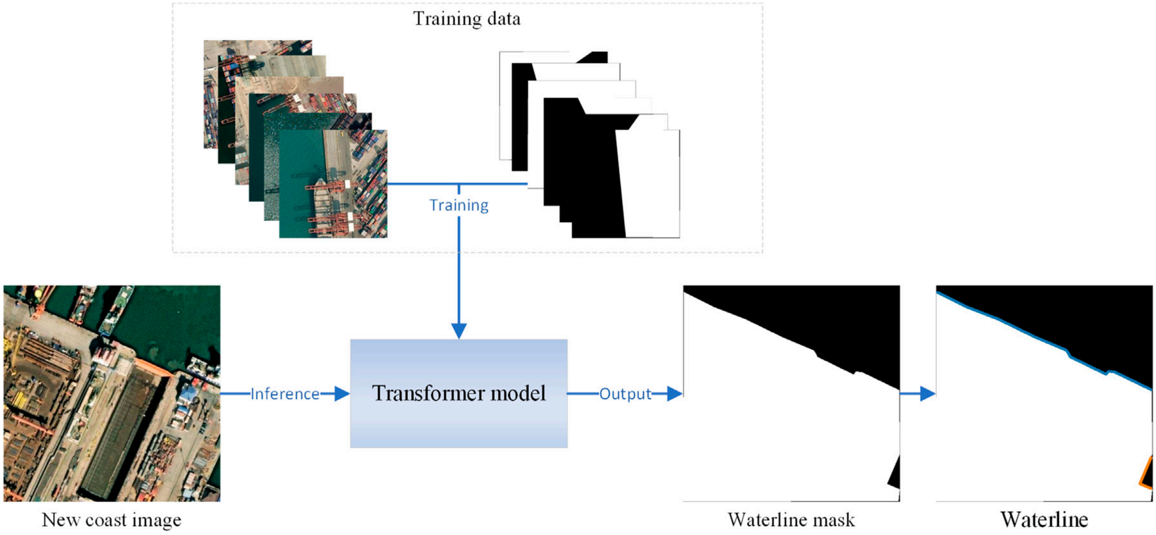

FIGURE 5 | Workflow of the proposed method.

$$P = \frac{T_P}{T_P + F_P}, \tag{3}$$

$$R = \frac{T_P}{T_P + F_N}, \tag{4}$$

$$F_1 = \frac{2 \times P \times R}{P + R}, \tag{5}$$

$$I_U = \frac{T_P}{T_P + F_P + F_N}. \tag{6}$$

Among them, the P and R stand for precision and recall, respectively; the true positive (T_P) stands for the rightly extracted land area; the false positive (F_P) represents the area mistaken as the land; the false negative (F_N) means omitted land pixels. In our study, the reference land images are drawn manually. Precision and recall are contradictory in most cases. To address this, comprehensive metrics F1 (F_1) and IoU (I_U) are employed commonly. Inference time is defined as the average segmentation's time using our test data. The floating-point operations (Flops) represent the computation of the model, and it is a metric for the computational complexity.

RESULTS

Experiment Setting

The proposed transformers were developed under MMsegmentation (MMSegmentation, 2020) by PyTorch (Paszke et al., 2017). Training and testing were performed with eight NVIDIA TITAN Xp GPUs and one NVIDIA TITAN Xp GPU, respectively. In our experimental dataset, there are 3600 images for training, 1200 images for validation, and 1200 images for testing. All the annotations are manually annotated. The resolution of all images is 512 × 512. The SETR uses a learning rate value of 10^{-3}, the number of iterations is 160,000, and the weights are pretrained on ImageNet-21K. The SegFormer was tested using a learning rate value of 10^{-6}, the number of iterations is 40,000, and the weights are pretrained on ImageNet-1K. The other compared methods are all run in 40,000 iterations.

Experimental Results

Qualitative Results

The results of the eight methods are displayed in **Figure 6**. From the results, we can see that PSPNet-UNet, DeepLabV3-UNet, and SETR cannot obtain good results in Image 1, Image 4, and Image 6. A large area in the land is missed, and the fine dock structure is not extracted in all the three methods. For the CNN methods, the methods with the ResNet101 backbone are better than the methods with HRNet, and the methods with the UNet backbone achieve the cheapest results. Only the methods with ResNet101 and HRNet extract the small striped object in Image 6, but no methods can avoid the influence of the ship. In Image 8, only the FCN-ResNet101 and DeepLabV3-ResNet101 gain terrific results. Other CNNs get a lot of false-positive or false-negative parts. For the transformer methods, the SegFormer achieves very nice results in all the images, especially for the fine structures. In contrast, the SETR can also extract the large object effectively in Image 3, but it struggles to the small and thin objects in Image 6 and Image 8. Overall, DeepLabV3-ResNet101, FCN-ResNet101, FPN-ResNet101, and SegFormer are all outstanding, and PSPNet-UNet, DeepLabV3-UNet, and SETR are relatively weak.

We can see in **Figure 7**, the SETR cannot extract the fine objects in Image1, Image 3, and Image 4. The dock and infrastructure are all not complete in the three images. It

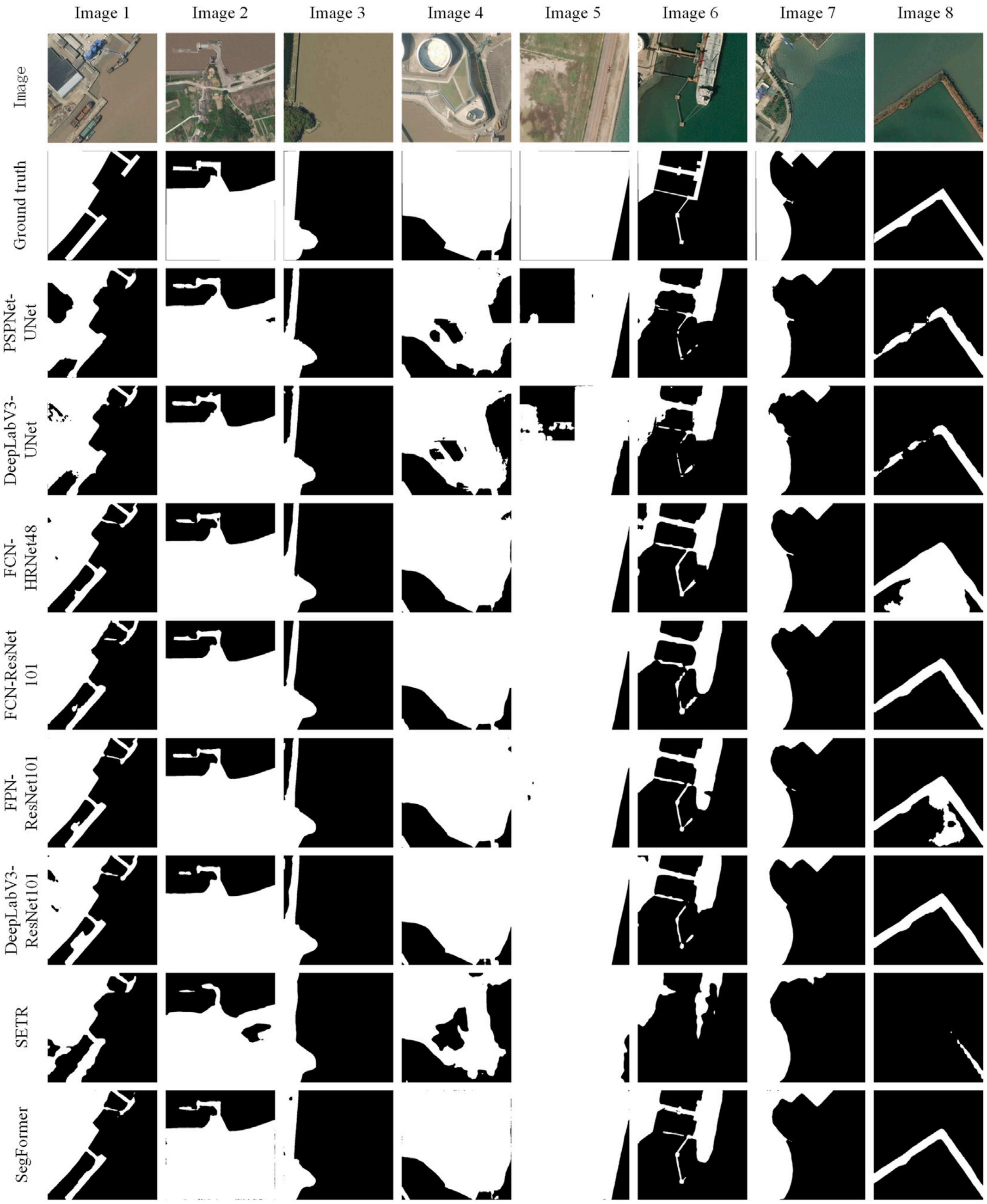

FIGURE 6 | Results of all methods in comparison.

hardly finds the land area near the boundary in Image 2 and Image 3. The connection part in the dock is neglected, and the shape of the harbor is not regular due to the incomplete extraction in Image 2. There are missed land pixels near the frame in Image 3. For large objects, it can perform well, although the edges are not kept fine in Image 1, Image 3, and Image 4.

Figure 8 depicts the results of SegFormer results. We can see that it correctly segments nearly all the pixels. It can even keep the details of objects well, especially in Image 1 and Image 3. The spindly parts in Image 1 and Image 3 are all fine and unmistakable. The integrity and differentiation are impressive. The minor complaints are the small leaks near the edges in Image 1, Image 3, and Image 4 and the small holes in Image 2. The SegFormer can extract the land features so good that the waterline can be extracted completely and accurately.

Quantitative Results

All the experimental methods are reported in **Table 1**. For the CNN methods, the models with ResNet101 achieved best results. Among them, the DeepLabV3 is the best with 0.9056 in precision,

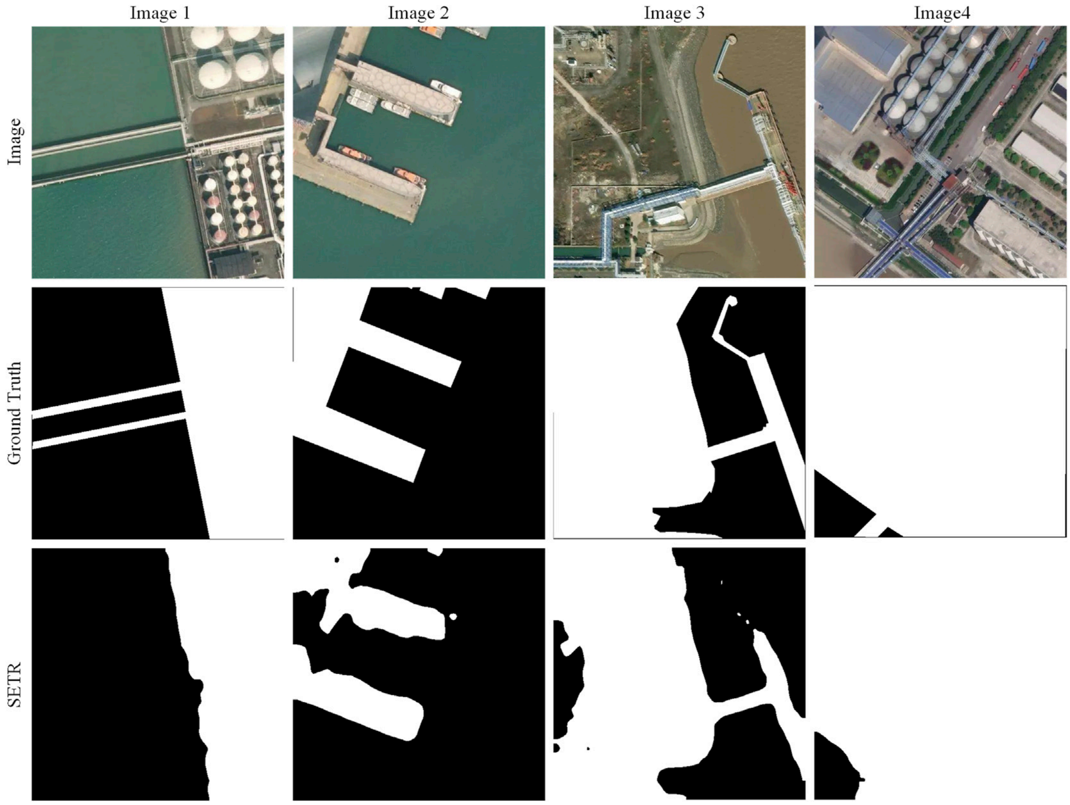

FIGURE 7 | Results of SETR.

0.8814 in recall, 0.8674 in F1, and 0.8169 in IoU. The FCN and FPN with the backbone of ResNet101 also reach or approach 0.9 in precision, 0.88 in recall, 0.86 in F1, and 0.81 in IoU. By comparison, the methods with UNet get lowest scores. PSPNet with UNet has the least scores, with the precision 0.8298, F1 0.8333, and IoU 0.7632. The performances of DeepLabV3 with UNet are slightly higher than that of PSPNet with UNet. The FCN with HRNet48 is moderate, which achieves 0.8964 in precision, 0.8766 in recall, 0.8581 in F1, and 0.8052 in IoU.

For the vision Transformer methods, we can see that the SegFormer reaches the precision 0.9121, recall 0.9104, F1-score 0.8883, and IoU 0.8439, respectively, which prove its accurate and robust performance to segment the land and sea. SETR gets the lowest scores in all metrics. The scores match the results in **Figures 6**, **7**; it cannot acquire the ideal land area, and the shapes are very incomplete.

The floating-point operations (Flops) represent the model complexity. **Table 1** shows that the DeepLabV3 models occupy more computing power, followed by SETR and PSPNet. The FCN, FPN, and SegFormer use the least resource in all the methods. Specially, the DeepLabV3-ResNet101 consume the largest computing units, and the SegFormer is the most resource-saving. For the inference time, except the FCN-HRNet48 and SETR with 0.77 and 0.32, other methods are all under 0.3 s. The FCN with ResNet101, FPN with ResNet101, and SegFormer can even infer an image in 0.2 s. FCN-ResNet101 is the fastest method in inference.

DISCUSSION

Performance Analysis of the Superior CNNs

In the six CNN methods, the networks with the backbone of ResNet101 are the best extractors, and they occupy the top three for accuracy. It is followed by the HRNet48, and the UNet is the last. In ResNet101, the convolution layers are very deep, and the features are connected with residual blocks. It can keep more detail features and avoid gradient vanishing by this way. Meanwhile, this ResNet101 uses dilated convolution to increase the receptive field, which makes it more powerful. The HRNet generates high-resolution and low-resolution parallel subnetworks. It can merge the high-resolution features and low-resolution features through the different stages by connecting the multi-resolution parallel subnetworks. Therefore, it can obtain rich high-resolution and low-resolution representations. The best header is DeepLabV3 because it achieves the best scores when different methods use the same backbone of UNet or ResNet101. In DeepLabV3, the atrous spatial pyramid pooling (ASPP) adds a series of atrous convolutions with different dilated rates to increase global contextual information. Global average pooling (GAP) also

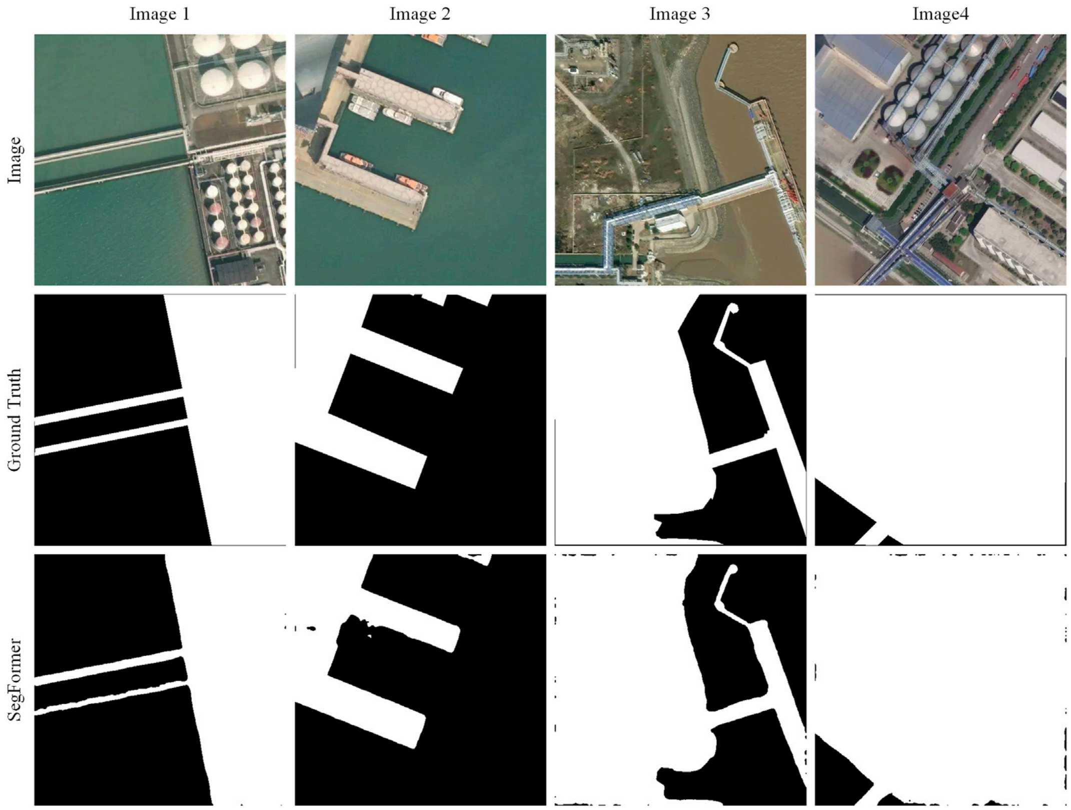

FIGURE 8 | Results of SegFormer.

TABLE 1 | Comparison for all the methods in metrics.

Method	Backbone	Flops (GFLOPs)	Inference Time(s)	Precision	Recall	F1	IoU
PSPNet	UNet	197.76	0.25	0.8298	0.8782	0.8333	0.7632
DeepLabV3	UNet	203.43	0.25	0.8518	0.8703	0.8406	0.7719
FCN	HRNet48	93.38	0.77	0.8964	0.8766	0.8581	0.8052
FCN	ResNet101	76.07	0.08	0.8995	0.8814	0.8636	0.8113
FPN	ResNet101	64.73	0.09	0.9010	0.8793	0.8637	0.8096
DeepLabV3	ResNet101	347.33	0.23	0.9056	0.8814	0.8674	0.8169
SETR	T-Large	212.4	0.32	0.8244	0.8397	0.8018	0.7268
SegFormer	MiT-B5	51.83	0.15	0.9121	0.9104	0.8883	0.8439

combines image-level features. These all make the DeepLabV3 outstanding.

Performance Analysis of the Superior Transformer

In the two vision transformers, the SETR obtain F1 with 0.8018 and IoU with 0.7268, and the SegFormer achieves 0.8883 in F1-score and 0.8439 in IoU. The SegFormer wins the SETR in accuracy completely. It can also be seen from the **Figures 7**, **8**, the SETR cannot extract the integrated and continuous structures in Image 2 and Image 3, and the SegFormer can extract nearly the whole and accurate structures. In SETR, the feature maps after the transformer layers are in the same size, and in the SegFormer, it generates multi-level feature maps. The different scales of feature maps include the high-resolution coarse features and low-resolution fine-grained features, so it can adapt to large and small object extractions. At the same time, the decoder in the SegFormer is made up of only MLP, which is lighter and has a larger effective field than traditional CNN encoders. These all make the SegFormer perform better than the SETR. On the other side, the SETR has more computational complexity with 212.4 GFLOPs than the SegFormer with just 51.83 GFLOPs. The huge amount of computation of the SETR is from the self-attention in the transformer. Because the computational complexity of self-attention is $O(N^2)$, N is the length of the input sequence. The SegFormer uses the efficient self-attention which reduces the computational complexity by some transforms. Therefore, the SegFormer is much easier to compute.

The Most Robust Method

The top CNN DeepLabV3 and the transformer SegFormer are all very competitive. However, the vision transformer SegFormer is superior to DeepLabV3 in precision, recall, F1, and IoU. It also has a smaller complexity and shorter inference time. The limited receptive field in DeepLabV3 requires the ASPP module to enlarge the receptive field, but the model inevitably becomes heavy. The SegFormer benefits from the non-local attention in transformers and enjoys a larger receptive field. The transformer integrates with the MLP decoder an can produce both highly local and non-local attention by adding fewer parameters. These all make the SegFormer more efficient and lighter in waterline extraction.

CONCLUSION

We propose a new method based on the vision transformers for the waterline extraction by sea–land segmentation. Two transformers, the SegFormer and SETR, are adapted to segment and identify land pixels by a custom dataset from satellite maps. The performances of the two transformers are compared with other state-of-the-art CNN methods, PSPNet, DeepLabV3, FCN, and FPN. The SETR with a pure transformer structure, as an early comer to image segmentation, achieves a nearly equivalent performance compared with the developed CNN methods. More surprisingly, the latter method, the SegFormer outperforms state-of-the-art CNN methods and demonstrates an extraordinary ability to segment land pixels under different conditions. For future work, we hope to improve the method in accuracy and robustness, though it has achieved a fairly good performance.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

LY conceived and designed the analysis, collected the data, performed the analysis, wrote the original draft, and discussed the results. XW verified the analytical methods, discussed the results, and reviewed and edited the manuscript. JZ supervised the work and reviewed and edited the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (Grant nos. 41701480, 42006168, and 41806116) and the Key Research and Development Program of Tianjin (Grant no. 20YFZCSN01040) and the Natural Science Foundation of Tianjin (Grant nos. 20JCQNJC01230). Tianjin Philosophy and Social Science Planning Project of China (Grant no. TJKS20XSX-015).

REFERENCES

Ao, D., Dumitru, O., Schwarz, G., and Datcu, M. P. (2017). "Coastline Detection with Time Series of SAR Images," in Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2017, San Diego, CA, August 6–10, 2017 (SPIE), 70–78. doi:10.1117/12.2278318

Bayram, B., Avsar, O., Seker, D. Z., Kayi, A., Erdogan, M., Eker, O., et al. (2017). The Role of National and International Geospatial Data Sources in Coastal Zone Management. *Fresenius Environ. Bull.* 26, 383–391.

Cao, K., Fan, J., Xinxin Wang, X., Xiang Wang, X., Jianhua Zhao, J., and Fengshou Zhang, F. (2016). "Coastline Automatic Detection Based on High Resolution SAR Images," in 2016 4th International Workshop on Earth Observation and Remote Sensing Applications, Guangzhou, China, July 4–6, 2016 (EORSA), 43–46. doi:10.1109/EORSA.2016.7552763

Chen, C., Fu, J., Zhang, S., and Zhao, X. (2019). Coastline Information Extraction Based on the Tasseled Cap Transformation of Landsat-8 OLI Images. *Estuarine Coastal Shelf Sci.* 217, 281–291. doi:10.1016/j.ecss.2018.10.021

Cheng, D., Meng, G., Xiang, S., and Pan, C. (2016). Efficient Sea-Land Segmentation Using Seeds Learning and Edge Directed Graph Cut. *Neurocomputing* 207, 36–47. doi:10.1016/j.neucom.2016.04.020

Cheng, D., Meng, G., Cheng, G., and Pan, C. (2017a). SeNet: Structured Edge Network for Sea-Land Segmentation. *IEEE Geosci. Remote Sensing Lett.* 14, 247–251. doi:10.1109/LGRS.2016.2637439

Cheng, D., Meng, G., Xiang, S., and Pan, C. (2017b). FusionNet: Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 10, 5769–5783. doi:10.1109/JSTARS.2017.2747599

Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., et al. (2021). Twins: Revisiting the Design of Spatial Attention in Vision Transformers. arXiv [Preprint]. Available at: http://arxiv.org/abs/2104.13840 (Accessed October 20, 2021).

Cui, B., Jing, W., Huang, L., Li, Z., and Lu, Y. (2021). SANet: A Sea-Land Segmentation Network via Adaptive Multiscale Feature Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 14, 116–126. doi:10.1109/JSTARS.2020.3040176

Dewi, R., Bijker, W., Stein, A., and Marfai, M. (2016). Fuzzy Classification for Shoreline Change Monitoring in a Part of the Northern Coastal Area of Java, Indonesia. *Remote Sensing* 8, 190. doi:10.3390/rs8030190

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv [Preprint]. Available at: http://arxiv.org/abs/2010.11929 (Accessed October 20, 2021).

Elkhateeb, E., Soliman, H., Atwan, A., Elmogy, M., Kwak, K.-S., and Mekky, N. (2021). A Novel Coarse-To-Fine Sea-Land Segmentation Technique Based on Superpixel Fuzzy C-Means Clustering and Modified Chan-Vese Model. *IEEE Access* 9, 53902–53919. doi:10.1109/ACCESS.2021.3065246

Erdem, F., Bayram, B., Bakirman, T., Bayrak, O. C., and Akpinar, B. (2021). An Ensemble Deep Learning Based Shoreline Segmentation Approach (WaterNet) from Landsat 8 OLI Images. *Adv. Space Res.* 67, 964–974. doi:10.1016/j.asr.2020.10.043

Fan, J., Cao, K., Zhao, J., Jiang, D., and Tang, X. (2016). "A Hybrid Particle Swarm Optimization Algorithm for Coastline SAR Image Automatic Detection," in 2016 12th World Congress on Intelligent Control and Automation, Guilin, China, June 12–15, 2016 (WCICA), 822–825. doi:10.1109/WCICA.2016.7578256

Ge, X., Sun, X., and Liu, Z. (2014). "Object-oriented Coastline Classification and Extraction from Remote Sensing Imagery," in Remote Sensing of the Environment: 18th National Symposium on Remote Sensing of China, Wuhan, China, October 20–23, 2012 (SPIE), 131–137. doi:10.1117/12.2063845

Google Maps (2021). Available at: https://www.google.com/maps/ (Accessed August 15, 2020).

Gucluer, D., Bayram, B., and Maktav, D. (2010). "Land Cover and Coast Line Change Detection by Using Object Oriented Image Processing in Alacati, Turkey," in Imagin[e,g] Europe, Chania, Greece, 158–165. doi:10.3233/978-1-60750-494-8-158

Guo, Q., Pu, R., Zhang, B., and Gao, L. (2016). "A Comparative Study of Coastline Changes at Tampa Bay and Xiangshan Harbor During the Last 30 Years," in 2016 IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, July 10–15, 2016 (IGARSS), 5185–5188. doi:10.1109/IGARSS.2016.7730351

He, L., Xu, Q., Hu, H., and Zhang, J. (2018). "Fast and Accurate Sea-Land Segmentation Based on Improved SeNet and Coastline Database for Large-Scale Image," in 2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications, Xi'an, China, June 18–20, 2018 (EORSA), 1–5. doi:10.1109/EORSA.2018.8598546

Hendrycks, D., and Gimpel, K. (2020). Gaussian Error Linear Units (GELUs). arXiv [Preprint]. Available at: http://arxiv.org/abs/1606.08415 (Accessed December 6, 2021).

Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., et al. (2018). DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 11, 3954–3962. doi:10.1109/JSTARS.2018.2833382

Lin, L., Pan, Z., Xiao, K., and Ye, N. (2013). "The Coastline Extraction for Fujian Province Based on Long Time Series of Remote Sensing Image," in Proceedings of the 2013 International Conference on Remote Sensing,Environment and Transportation Engineering, Nanjing, China, July 26–28, 2013, 63–66.

Lin, H., Shi, Z., and Zou, Z. (2017). Maritime Semantic Labeling of Optical Remote Sensing Images with Multi-Scale Fully Convolutional Network. *Remote Sensing* 9, 480. doi:10.3390/rs9050480

Liu, Y., Huang, H., Qiu, Z., and Fan, J. (2013). Detecting Coastline Change from Satellite Images Based on beach Slope Estimation in a Tidal Flat. *Int. J. Appl. Earth Obs. Geoinf.* 23, 165–176. doi:10.1016/j.jag.2012.12.005

Liu C, C., Yang, J., Yin, J., and An, W. (2016). Coastline Detection in SAR Images Using a Hierarchical Level Set Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 9, 4908–4920. doi:10.1109/JSTARS.2016.2613279

Liu Z, Z., Li, F., Li, N., Wang, R., and Zhang, H. (2016). A Novel Region-Merging Approach for Coastline Extraction from Sentinel-1A IW Mode SAR Imagery. *IEEE Geosci. Remote Sensing Lett.* 13, 1–5. doi:10.1109/LGRS.2015.2510745

Liu, C., Xiao, Y., and Yang, J. (2017). A Coastline Detection Method in Polarimetric SAR Images Mixing the Region-Based and Edge-Based Active Contour Models. *IEEE Trans. Geosci. Remote Sensing* 55, 3735–3747. doi:10.1109/TGRS.2017.2679112

Liu, X.-Y., Jia, R.-S., Liu, Q.-M., Zhao, C.-Y., and Sun, H.-M. (2019). Coastline Extraction Method Based on Convolutional Neural Networks-A Case Study of Jiaozhou Bay in Qingdao, China. *IEEE Access* 7, 180281–180291. doi:10.1109/ACCESS.2019.2959662

Liu W, W., Chen, X., Ran, J., Liu, L., Wang, Q., Xin, L., et al. (2021). LaeNet: A Novel Lightweight Multitask CNN for Automatically Extracting Lake Area and Shoreline from Remote Sensing Images. *Remote Sensing* 13, 56. doi:10.3390/rs13010056

Liu Z, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. arXiv [Preprint]. Available at: http://arxiv.org/abs/2103.14030 (Accessed October 20, 2021).

Mapbox (2021). Available at: https://www.mapbox.com/ (Accessed December 7, 2021).

Microsoft (2021). Bing Maps Imagery API. Available at: https://docs.microsoft.com/en-us/bingmaps/rest-services/imagery (Accessed December 7, 2021).

MMSegmentation (2020). MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. Available at: https://github.com/open-mmlab/mmsegmentation (Accessed December 7, 2021).

Modava, M., and Akbarizadeh, G. (2017). Coastline Extraction from SAR Images Using Spatial Fuzzy Clustering and the Active Contour Method. *Int. J. Remote Sens.* 38, 355–370. doi:10.1080/01431161.2016.1266104

Nunziata, F., Buono, A., Migliaccio, M., and Benassai, G. (2016). Dual-Polarimetric C- and X-Band SAR Data for Coastline Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 9, 4921–4928. doi:10.1109/JSTARS.2016.2560342

Paravolidakis, V., Moirogiorgou, K., Ragia, L., Zervakis, M., and Synolakis, C. (2016). "Coastline Extraction from Aerial Images Based on Edge Detection," in XXIII Congress of International Society for Photogrammetry and Remote Sensing (ISPRS 2016), Prague, Czech Republic, July 12–19, 2016, 153–158. doi:10.5194/isprsannals-III-8-153-2016

Paravolidakis, V., Ragia, L., Moirogiorgou, K., and Zervakis, M. (2018). Automatic Coastline Extraction Using Edge Detection and Optimization Procedures. *Geosciences* 8, 407. doi:10.3390/geosciences8110407

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic Differentiation in PyTorch," in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA.

Rasuly, A., Naghdifar, R., and Rasoli, M. (2010). Monitoring of Caspian Sea Coastline Changes Using Object-Oriented Techniques. *Proced. Environ. Sci.* 2, 416–426. doi:10.1016/j.proenv.2010.10.046

Rigos, A., Tsekouras, G. E., Vousdoukas, M. I., Chatzipavlis, A., and Velegrakis, A. F. (2016). A Chebyshev Polynomial Radial Basis Function Neural Network for Automated Shoreline Extraction from Coastal Imagery. *ICA* 23, 141–160. doi:10.3233/ICA-150507

Roelfsema, C., Kovacs, E. M., Saunders, M. I., Phinn, S., Lyons, M., and Maxwell, P. (2013). Challenges of Remote Sensing for Quantifying Changes in Large Complex Seagrass Environments. *Estuarine Coastal Shelf Sci.* 133, 161–171. doi:10.1016/j.ecss.2013.08.026

Shamsolmoali, P., Zareapoor, M., Wang, R., Zhou, H., and Yang, J. (2019). A Novel Deep Structure U-Net for Sea-Land Segmentation in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 12, 3219–3232. doi:10.1109/JSTARS.2019.2925841

Sun, B., Li, S., and Xie, J. (2019). "Sea-Land Segmentation for Harbour Images with Superpixel CRF," in IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, July 28–August 2, 2019, 3899–3902. doi:10.1109/IGARSS.2019.8899001

Toure, S., Diop, O., Kpalma, K., and Maiga, A. S. (2018). "Coastline Detection Using Fusion of over Segmentation and Distance Regularization Level Set Evolution," in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (Istanbul, Turkey: Copernicus GmbH), 513–518. doi:10.5194/isprs-archives-XLII-3-W4-513-2018

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). "Training Data-Efficient Image Transformers & Distillation Through Attention," in Proceedings of the 38th International Conference on Machine Learning (PMLR), Virtual Event, July 18–24, 2021, 10347–10357. Available at: https://proceedings.mlr.press/v139/touvron21a.html (Accessed October 20, 2021).

Tsekouras, G. E., Trygonis, V., Maniatopoulos, A., Rigos, A., Chatzipavlis, A., Tsimikas, J., et al. (2018). A Hermite Neural Network Incorporating Artificial Bee Colony Optimization to Model Shoreline Realignment at a Reef-Fronted beach. *Neurocomputing* 280, 32–45. doi:10.1016/j.neucom.2017.07.070

Vos, K., Splinter, K. D., Harley, M. D., Simmons, J. A., and Turner, I. L. (2019). CoastSat: A Google Earth Engine-Enabled Python Toolkit to Extract Shorelines from Publicly Available Satellite Imagery. *Environ. Model. Softw.* 122, 104528. doi:10.1016/j.envsoft.2019.104528

Wang, D., and Liu, X. (2019). Coastline Extraction from SAR Images Using Robust Ridge Tracing. *Mar. Geodesy* 42, 286–315. doi:10.1080/01490419.2019.1583147

Wang, C., Yang, J., Li, J., and Chu, J. (2020). Deriving Natural Coastlines Using Multiple Satellite Remote Sensing Images. *J. Coastal Res.* 102, 296–302. doi:10.2112/SI102-036.1

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., et al. (2021). Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. arXiv [Preprint]. Available at: http://arxiv.org/abs/2102.12122 (Accessed October 20, 2021).

Wernette, P., Houser, C., and Bishop, M. P. (2016). An Automated Approach for Extracting Barrier Island Morphology from Digital Elevation Models. *Geomorphology* 262, 1–7. doi:10.1016/j.geomorph.2016.02.024

Widyantara, I. M. O., Wirastuti, N. M. A. E. D., and Asana, I. M. D. P. (2017). "Gamma Correction-Based Image Enhancement and Canny Edge Detection for Shoreline Extraction from Coastal Imagery," in 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Central Java, Indonesia, November 15–16, 2017, 17–22. doi:10.1109/icicos.2017.8276331

Wu, X., Liu, C., and Wu, G. (2018). Spatial-Temporal Analysis and Stability Investigation of Coastline Changes: A Case Study in Shenzhen, China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* 11, 45–56. doi:10.1109/JSTARS.2017.2755444

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv [Preprint]. Available at: http://arxiv.org/abs/2105.15203 (Accessed October 20, 2021).

Xu, W., Xu, Y., Chang, T., and Tu, Z. (2021). Co-Scale Conv-Attentional Image Transformers. arXiv [Preprint]. Available at: http://arxiv.org/abs/2104.06399 (Accessed October 20, 2021).

Yang, C.-S., Park, J.-H., and Rashid, A. H.-A. (2018). An Improved Method of Land Masking for Synthetic Aperture Radar-Based Ship Detection. *J. Navigation* 71, 788–804. doi:10.1017/S037346331800005X

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). Rethinking Semantic Segmentation from a Sequence-To-Sequence Perspective with Transformers. arXiv [Preprint]. Available at: http://arxiv.org/abs/2012.15840 (Accessed August 12, 2021).

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.