Electronic Research Archive

https://www.aimspress.com/journal/ERA

ERA, 32(10): 5615–5632.
DOI: 10.3934/era.2024260
Received: 21 July 2024
Revised: 06 September 2024
Accepted: 23 September 2024
Published: 15 October 2024

Research article

SSMM-DS: A semantic segmentation model for mangroves based on Deeplabv3+ with swin transformer

Zhenhua Wang [1], Jinlong Yang [1], Chuansheng Dong [2], Xi Zhang [3], Congqin Yi [1] and Jiuhu Sun [2,*]

[1] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China
[2] Shandong Land Surveying and Mapping Institute, Jinan 250000, China
[3] Shandong provincial institute of land space data and remote sensing technology, Jinan 250000, China

* **Correspondence:** Email: sunjhgis@126.com; Tel: +8615853135083; Fax: 02161900627.

Abstract: Mangrove wetlands play a crucial role in maintaining species diversity. However, they face threats from habitat degradation, deforestation, pollution, and climate change. Detecting changes in mangrove wetlands is essential for understanding their ecological implications, but it remains a challenging task. In this study, we propose a semantic segmentation model for mangroves based on Deeplabv3+ with Swin Transformer, abbreviated as SSMM-DS. Using Deeplabv3+ as the basic framework, we first constructed a data concatenation module to improve the contrast between mangroves and other vegetation or water. We then employed Swin Transformer as the backbone network, enhancing the capability of global information learning and detail feature extraction. Finally, we optimized the loss function by combining cross-entropy loss and dice loss, addressing the issue of sampling imbalance caused by the small areas of mangroves. Using GF-1 and GF-6 images, taking mean precision (mPrecision), mean intersection over union (mIoU), floating-point operations (FLOPs), and the number of parameters (Params) as evaluation metrics, we evaluate SSMM-DS against state-of-the-art models, including FCN, PSPNet, OCRNet, uPerNet, and SegFormer. The results demonstrate SSMM-DS's superiority in terms of mIoU, mPrecision, and parameter efficiency. SSMM-DS achieves a higher mIoU (95.11%) and mPrecision (97.79%) while using fewer parameters (17.48M) compared to others. Although its FLOPs are slightly higher than SegFormer's (15.11G vs. 9.9G), SSMM-DS offers a balance between performance and efficiency. Experimental results highlight SSMM-DS's effectiveness in extracting mangrove features, making it a valuable tool for monitoring and managing these critical ecosystems.

Keywords: remote sensing image; semantic segmentation; mangrove; swin transformer

1. Introduction

Mangroves are of great ecological significance in shoreline stabilization, reduction of coastal erosion, sediment and nutrient retention, storm protection, flood and flow control, carbon sequestration, and water quality maintenance [1]. However, mangroves are disappearing at an alarming rate each year due to frequent human activities and climate change. The deforestation of mangroves has seriously undermined their capacity for sustained economic value and resource creation for ecology [2]. Monitoring and protecting mangroves is a crucial means of mitigating their disappearance over the years [3]. Due to the wide distribution of mangroves and scattered areas of intensive habitat [4], remote sensing technology has become the primary technical means of observing mangroves. This is attributed to its low cost, high frequency, and ability to provide synchronized observations over large areas [5]. Remote sensing allows for the efficient monitoring of these critical ecosystems, enabling timely interventions and informed decision-making to safeguard mangroves and their ecological functions.

Mangrove monitoring using multi-source remote sensing data has been extensively studied. Vidhya et al. [6] demonstrated the effectiveness of hyperspectral data, support vector machine (SVM) classification, and soil adjusted vegetation Indices (SAVI) for accurately mapping mangrove health and area. Pham et al. [7] explored the potential of Advanced Land Observing Satellite (ALOS） Phased Array L-band Synthetic Aperture Radar (PALSAR) imagery, geographic information system (GIS) data, and logistic model tree (LMT) for monitoring mangrove species in tropical ecosystems. Li et al. [8] analyzed mangrove reserves using the land type transfer matrix, centroid variation, and landscape index to investigate the diffusion characteristics and patterns of Spartina alterniflora in mangrove wetlands. Cao et al. [9] proposed a mangrove classification method integrating UAV hyperspectral imagery, LiDAR data, and the rotation forest (RoF) ensemble learning algorithm, enabling high-resolution monitoring and supporting mangrove restoration and management efforts. The rapid proliferation of remote sensing data has presented a significant challenge to traditional algorithms. Their limitations in processing large-scale data and extracting information quickly are becoming increasingly apparent. Consequently, high-precision calculation models have emerged as crucial elements in the transformation of remote sensing observations from data to information.

The rapid advancement of machine learning and deep learning technology has revolutionized remote sensing image interpretation. These technologies have achieved remarkable results in tasks such as regression, image generation, object detection, and image segmentation. Numerous scholars have successfully applied them in the field of remote sensing [10–12]. Convolutional neural networks (CNNs), with their powerful feature extraction capabilities, have demonstrated exceptional performance in remote sensing image classification [13]. The transformer model, with its ability to model long-distance dependencies, has also opened new opportunities for remote sensing applications [14]. In the specific domain of mangrove monitoring, deep learning methods have shown promising potential. Wan et al. [15] introduced a small patch-based CNN to address the limitations of fixed, large inputs, expanding CNN's applicability to fringe mangroves and achieving superior classification accuracy. Moreno et al. [16] considered spatial, temporal, and polarization dimensions, proposing a method suitable for long-term monitoring of mangrove growth status through time series composition with varying image numbers and sliding window strides. Fan et al. [17] developed a domain adaptation-based remote sensing image segmentation method for mangroves, incorporating a self-attention mechanism to focus on important image channels and combining remote sensing spectral indices to mitigate potential edge information loss. Xu et al. [18] proposed MSNet, a semantic segmentation model that fuses multi-scale features for mangroves. MSNet can simultaneously extract high-level semantic features and learn high-resolution image details, improving segmentation accuracy through the mutual fusion of different scale features.

While deep learning has advanced mangrove segmentation, challenges remain. Mangroves exhibit a high degree of spectral similarity with surrounding vegetation, particularly Spartina alterniflora, making them difficult to distinguish in remote sensing imagery. This spectral similarity poses a significant challenge for segmentation models. Additionally, the scattered distribution and small size of mangroves lead to a highly imbalanced dataset, complicating model training. The multispectral nature of remote sensing imagery, the multi-scale characteristics of mangroves, and the interference of noise further exacerbate the complexity of segmentation tasks. To address these issues, this paper proposes a semantic segmentation model tailored for mangrove segmentation in remote sensing images, building upon the DeepLabv3+ framework.

2. Semantic segmentation model

High-resolution remote sensing images offer rich detail but pose computational challenges for mangrove segmentation. DeepLabV3+, with its ASPP module and dilated convolutions, is well-suited for complex, multi-scale data due to its ability to handle intricate features. However, its Xception backbone is computationally expensive for the relatively simpler task of mangrove segmentation. To address these limitations, we propose a modified DeepLabV3+ architecture incorporating a Swin transformer backbone and a data concatenation module (DCM), as shown in Figure 1.

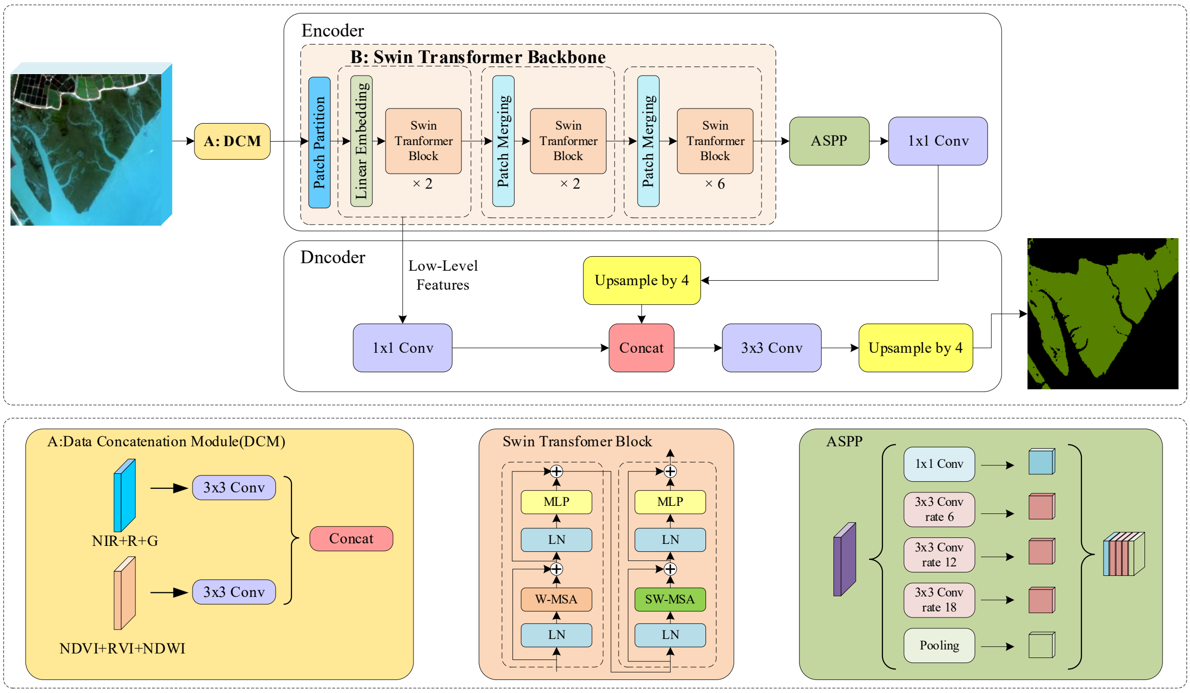


Figure 1. Architecture of the proposed SSMM-DS model. (A) DCM; B) Swin transformer backbone.

- DCM: To better utilize multispectral information and enhance the contrast between mangroves and other vegetation or water, we introduce a DCM (Figure 1(A)). This module expands the dataset and aids in differentiating mangroves from their surroundings.
- Swin transformer backbone: We replace the computationally expensive Xception backbone with a Swin transformer. This efficient vision model preserves global context while processing image patches, improving the model's ability to learn global information and extract detailed features essential for mangrove segmentation (Figure 1(B)).

Furthermore, we employ a combined cross-entropy and Dice loss function to address the sampling imbalance caused by the small areas of mangroves relative to the entire image.

2.1. DCM

Mangroves are unique ecosystems comprised primarily of trees and shrubs. In optical remote sensing images, mangroves can be identified by their distinctive spatial distribution, crown and leaf texture, and spectral characteristics [19]. First, mangroves grow in the intertidal zones of tropical and subtropical coastal environments, forming bands along the coastline [20]. This creates clear boundaries in the image that contrast sharply with inland vegetation and water bodies. Second, as specialized coastal plants, mangroves exhibit unique crown and leaf textures and spectral characteristics that differentiate them from terrestrial forests. These features are crucial for remote sensing identification and monitoring of mangrove forests.

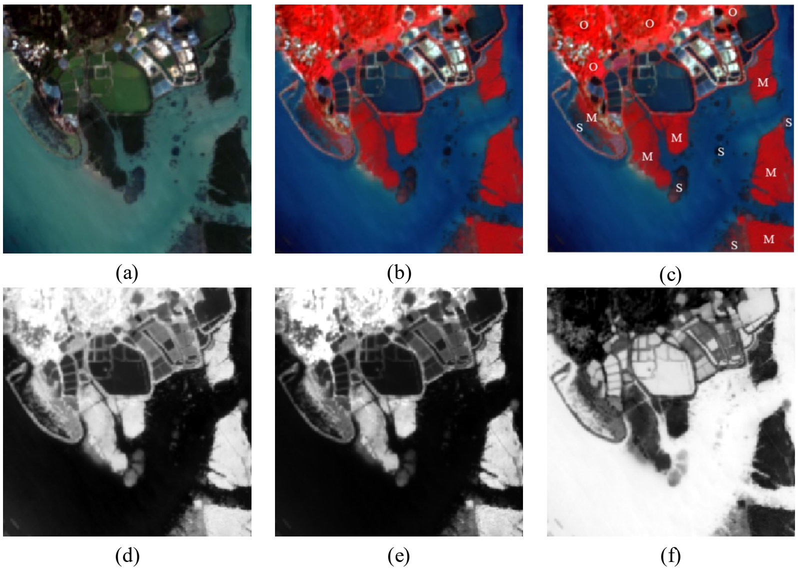
Mangrove exhibits high reflectance in the near-infrared band. Additionally, the reflectivity sharply increases during the transition from red to near-infrared wavelengths [21]. However, since mangroves grow in the intertidal zone, water's influence on the spectrum cannot be ignored. Water strongly absorbs blue light, resulting in low mangrove reflectivity in the blue band. Consequently, the input module of the segmentation model selected three channels from spectral bands near-infrared (NIR), red (R), and green (G) of remote sensing images.

To fully leverage the spectral characteristics of remote sensing data and enhance the discrimination ability of mangroves, three indices-NDVI, RVI, and NDWI-were incorporated as additional channels into the segmentation model. These indices effectively highlight mangrove features by calculating the reflectance of light spectra in remote sensing images [22]. The normalized difference vegetation index (NDVI) is commonly used for vegetation assessment, contrasting red and near-infrared light reflectance to highlight vegetation. However, it can saturate under conditions of high biomass [23]. The ratio vegetation index (RVI) offers a measure of vegetation health and density, being sensitive to high-density green vegetation but less sensitive to biomass below 50% [24]. By combining NDVI and RVI, we can better utilize information on both high-density and low-density vegetation. The normalized difference water index (NDWI) effectively distinguishes mangroves from surrounding water bodies by comparing green and near-infrared light reflectance [25]. This index leverages the characteristic coastal distribution of mangroves. Table 1 presents the calculation formulas for these three indices.

Table 1. Three indices calculation formula.

Name	Calculation method
NDVI	(NIR − Red) / (NIR + Red)
RVI	Red / NIR
NDWI	(Green − NIR) / (Green + NIR)

Figure 2 presents a comparative analysis of images using various spectral band combinations and calculated indices. Based on Figure 2(a), it is challenging to differentiate between mangroves and other vegetation. However, Figure 2(b) and (c) reveal distinct characteristics: mangroves appear in dark red or red tones, with clear boundaries, irregular shapes, and a smooth texture. Spartina, on the other hand, presents in red or light red tones, often with a fan-shaped or dot-like appearance and a smooth texture. Other terrestrial vegetation also appears in dark red or red tones but is not located near the coast. Analyzing Figures 2(d) and (e), we observe that water bodies appear as deep black, while mangroves and other vegetation are depicted in light gray and Spartina appears as dark gray. Notably, other vegetation tends to be slightly brighter than mangroves. Figure 2(f) further clarifies the distinctions: Spartina remains dark gray, while water bodies turn bright white. Both mangroves and other vegetation appear dark gray in this image.



(a) (b) (c)

(d) (e) (f)

Figure 2. Comparison of images using various spectral band combinations and calculated indices. (a) RGB composite image (spectral bands R, G, and B); (b) RGB composite image (spectral bands NIR, R, and G); (c) Annotated image (M: mangrove, S: *Spartina*, O: other vegetation); (d) NDVI image; (e) RVI image; (f) NDWI image.

Superior to the raw data module of Deeplabv3+, the data concatenation module expanded the data channels from three (NIR, R, and G bands) to six (NIR, R, G bands, NDVI, RVI, and NDWI), thereby enhancing the features of mangroves (Figure 3). To ensure data consistency, all six channels were normalized using the Z-score method. Subsequently, the three spectral channels (NIR, R, and G bands) were convoluted together, while the additional indices (NDVI, RVI, and NDWI) were convoluted separately. The results of these convolutions were then concatenated and fed into the encoder of the

SSMM-DS model. The data concatenation module improved the model's ability to distinguish mangroves by leveraging the enriched spectral and index information, ultimately leading to more accurate segmentation results.

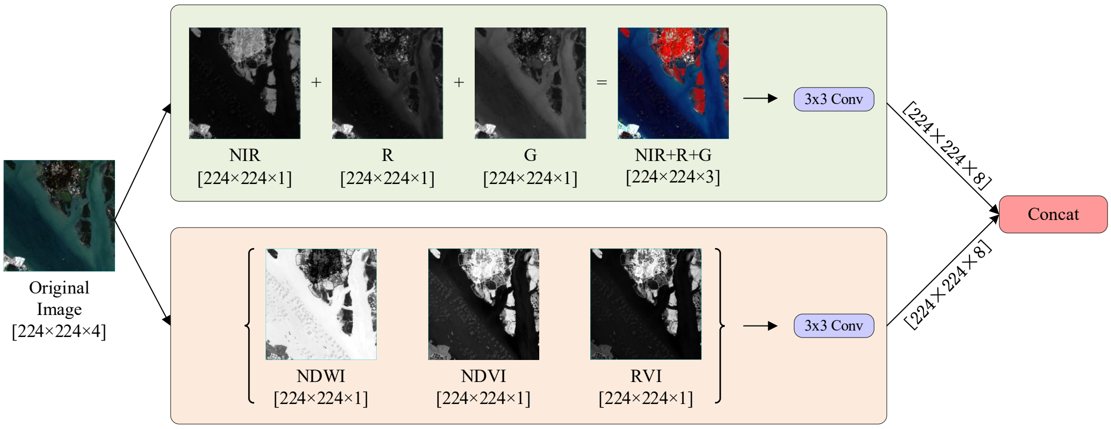


Figure 3. Structure of DCM.

2.2. Swin transformer backbone

The Swin transformer, a computationally efficient vision model, effectively captures multi-scale features within mangroves, from large-scale structures to smaller details. Its hierarchical architecture and shifted window-based self-attention mechanism contribute to its ability to handle high-resolution images.

To enhance performance, we modified the Swin transformer backbone by excluding the original Swin-T Stage 4 component, reducing parameters and computational complexity. Raw data is first processed by the DCM, followed by patch partition and linear embedding. Two Swin-transformer blocks are then applied, yielding low-level features with dimensions reduced to 1/4. These features are further processed through Patch Merging and additional Swin-transformer blocks, resulting in high-level features with dimensions reduced to 1/16, which are fed into the ASPP module. Figure 4 shows the architecture of Swin transformer backbone

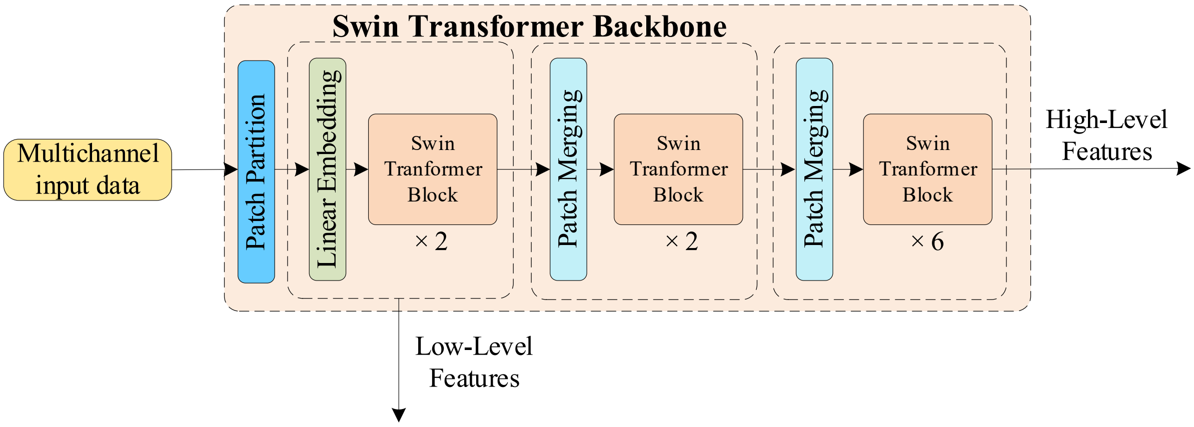


Figure 4. Architecture of Swin transformer backbone.

(1) Patch merging structure

To obtain feature maps at different scales and improve model performance, we incorporated a patch merging structure. This structure divides feature maps into patches, merges them, and down samples them, resulting in feature maps at different resolutions, similar to ResNet [26]. Figure 5 shows the process of patch meagering, where feature maps are divided into four independent sub-maps, concatenates them, and adjusts their dimensions through a 1x1 convolution and down sampling.

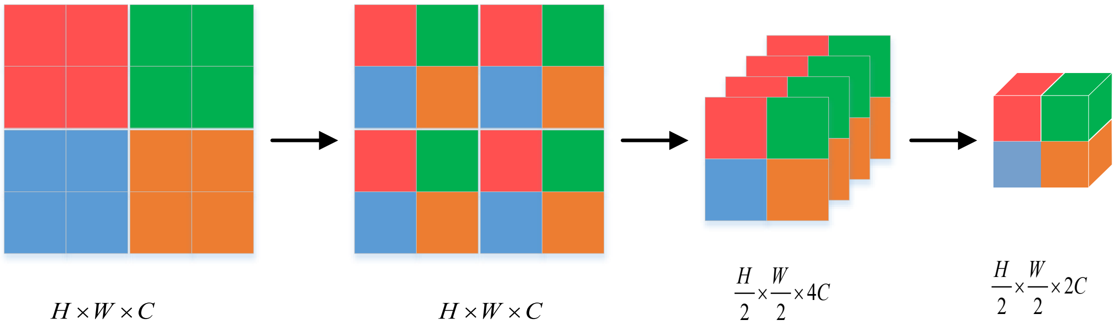

$H \times W \times C$
$H \times W \times C$
$\frac{H}{2} \times \frac{W}{2} \times 4C$
$\frac{H}{2} \times \frac{W}{2} \times 2C$

Figure 5. Patch merging structure.

(2) Swin transformer block

The Swin transformer's hierarchical architecture and shifted window-based self-attention mechanism contribute to its computational efficiency and ability to handle high-resolution images. The hierarchical architecture progressively reduces resolution, increasing the receptive field while decreasing computational load. The shifted window-based self-attention mechanism computes attention locally within non-overlapping windows and periodically shifts them to capture global information.

(3) Computational complexity and feature extraction

While multi-head self-attention (MSA) can improve feature extraction [27], its global-based approach is computationally expensive for large images. Swin transformer addresses this by using window-based MSA (W-MSA), which restricts attention to segmented feature windows. However, W-MSA can limit feature extraction by focusing on local information. To address this, shifted window-based self-attention (SW-MSA) is implemented to connect cross-window features.

The computational complexity of the global-based MSA module and W-MSA are

$$\Omega(MSA) = 4HWC^2 + 2(HW)^2C \tag{1}$$

$$\Omega(W - MSA) = 4HWC^2 + 2M^2HWC \tag{2}$$

2.3. Loss function

The loss function of the SSMM-DS model consists of the cross-entropy loss function and the dice loss function, defined as:

$$loss = \propto Loss_{CE} + (1 - \propto) Loss_{Dice} \tag{3}$$

where $\propto$ is a weight parameter that balances the contributions of the two loss functions.

The cross-entropy loss function is a combination of logarithmic function and softmax function, which can prevent the model from overfitting. Additionally, it has lower time complexity in gradient descent algorithms compared to other loss functions. Therefore, the cross-entropy loss function is often used in semantic segmentation tasks to measure the difference between each predicted pixel value and the true time value. It is defined as

$$Loss_{CE} = -\frac{1}{N}\sum_{I=1}^{N} y_i * log\,\hat{y}_i \tag{4}$$

where *N* is the total number of pixels in the image, $y_i$is the true value of the ith pixel point and $\hat{y}_i$ is the predicted value from the segmentation model.

However, when using the cross-entropy loss function, the gradient becomes very small when the predicted result $\hat{y}_i$ is very close to 1 or 0, significantly impacting the model's learning speed. Additionally, this loss function only focuses on the difference between the predicted value and the true value, failing to consider the distinctions between intermediate states and not achieving the optimal error rate. Furthermore, when the number of negative samples in the dataset greatly exceeds the number of true samples, the model tends to favor learning from the negative samples and ignores the true targets. Therefore, the cross-entropy loss function is not suitable for all training tasks, especially in datasets with imbalanced proportions of positive and negative samples.

The dice loss function effectively addresses class imbalance [28], as every pixel in the image is evaluated and considered. This makes it a better solution for the negative impact caused by the imbalance between the number of target pixels and background pixels in the image. It is commonly used to evaluate the similarity between the number of pixels in a binary image classified as positive classes and the number of pixels that are actually positive. The dice loss function is derived from the Dice coefficient, a metric frequently used in medical imaging to calculate the similarity between two samples.

The dice coefficient is defined as

$$Dice\ coefficient = \frac{2|X \cap Y|}{|X|+|Y|} \tag{5}$$

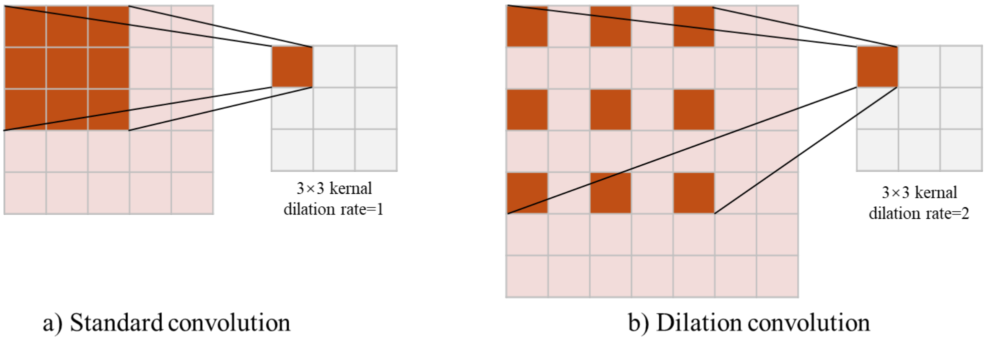
where *X* and *Y* are the set of labeled values and the set of model predictions, respectively. A higher Dice coefficient indicates a greater similarity between the model's predictions and the true results. The dice loss function is defined as

$$Loss_{Dice} = 1 - \frac{2|X \cap Y|}{|X|+|Y|} \tag{6}$$

2.4. Atrous spatial pyramid pooling (ASPP)

In the SSMM-DS model, we integrated the ASPP module [29] to extract multi-scale mangrove features. ASPP is a widely used deep learning module that combines 1 × 1 convolutions, dilated convolutions with rates of 6, 12, and 18, and global pooling. Figure 6 illustrates the difference between standard and dilated convolutions. Compared to standard convolutions, dilated convolutions can capture a larger receptive field. For instance, with a 3 × 3 convolution, the receptive field of a standard convolution is 3 × 3, while a dilated convolution with a rate of 2 expands this to 5 × 5.

By utilizing dilated convolutions with varying dilation rates, ASPP effectively captures features across different receptive fields. Additionally, global pooling is used to capture global context, followed by 1 × 1 convolutions. These features are then combined to enhance the model's ability to represent complex patterns.



a) Standard convolution

b) Dilation convolution

Figure 6. The comparison of Standard convolution and dilated convolution.

3. Experiment data and setup

3.1. Experimental dataset

The image dataset was acquired from GF-1 and GF-6 satellites and collected over Gaoqiao Mangrove Reserve and Shankou Mangrove Ecological Nature Reserve, located along the coast of Lianjiang City in Guangdong Province and Beihai City in Guangxi Province, China. The data was gathered between 2018 and 2021. Figure 7 presents a sample image from the data cube, showcasing spectral bands NIR, R, and G of a GF-6 remote sensing image.

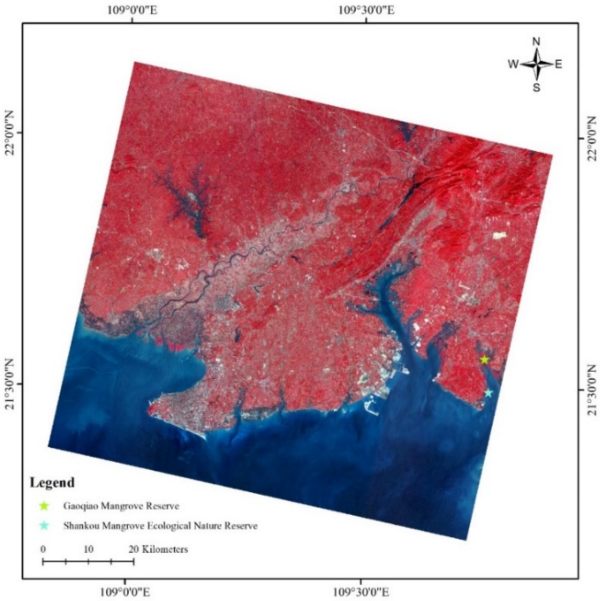


Figure 7. RGB composite image of GF-6 remote sensing imagery (spectral bands NIR, R, and G).

The GF-1 satellite is equipped with two cameras: a 2-meter resolution panchromatic camera and an 8-meter resolution multispectral camera. This configuration combines high spatial resolution, multispectral capabilities, and high temporal resolution, breaking new ground in optical remote sensing technology. Additionally, multi-payload image stitching and fusion techniques are employed. The details of the sensor specifications for GF-1 PMS images are presented in Table 2 below.

Table 2. Characteristics of GF-1 PMS images.

Band	Wavelength (nm)	Spatial resolution (m)	Temporal resolution (Days)	Swath width (km)
B1(Blue)	450–520	8	4	60
B2(Green)	520–590			
B3(Red)	630–690			
B4(Near infrared)	770–890			

GF-6 offers advantages such as high resolution, wide coverage, and rapid high-quality imaging. This significantly enhances the quality and timeliness of remote sensing image acquisition. Its payloads are comparable to those of GF-1. With the network operation of GF-6 and GF-1 satellites, the temporal resolution of remote sensing data acquisition has been reduced from 4 days to 2 days. The details of the sensor specifications for GF-6 PMS images are presented in Table 3.

Table 3. Characteristics of GF-6 PMS images.

Band	Wavelength (nm)	Spatial resolution (m)	Temporal resolution (Days)	Swath width (km)
B1(Blue)	450–520	8	4	60
B2(Green)	520–590			
B3(Red)	630–690			
B4(Near infrared)	770–890			

The metadata was pre-processed with atmospheric correction, maintaining a spatial resolution of 8-meters and including four bands: blue, green, red, and near-infrared. The images were then selectively cropped to a size of 224 × 224 pixels. A total of 843 images were cropped from the 28 remote sensing images and labeled. To enhance the model's generalization ability, we expanded the labeled dataset of 843 images by four times. We achieved this by applying horizontal inversion, rotation, and vertical inversion to the original images, resulting in a total of 3372 labeled remote sensing images. These data augmentation techniques effectively simulate image changes from different angles and directions, improving the model's adaptability to diverse and complex scenarios [30]. Of these images, 80% were randomly selected as the training set, 10% as the validation set, and 10% as the test set.

Since the input GF-1/GF-6 PMS images are in the 16-bit TIFF file format, normalization by dividing by 255 is not applicable. Therefore, Z-score normalization was employed to preprocess the input data, aiming to enhance the convergence speed and stability of the optimization model by standardizing the data distribution. This process standardizes the data to have a mean of 0 and a standard deviation of 1, making it suitable for comparison and analysis across different scales and distributions.

The Z-score can be computed as follows:

$$\mathrm{Z} = \frac{X-\mu}{\sigma} \tag{7}$$

where Z is the pixel value of the output image, X is the raw pixel value to be calculated, μ is the mean value of the pixels in the input image, σ is the standard deviation of the input image.

3.2. Evaluation metrics

Four metrics were calculated to evaluate the performance of the proposed SSMM-DS, including mPrecision, mIoU, FLOPs and Params.

mPrecision is the mean values of precision, where precision indicates the proportion of correctly segmented pixels (i.e., positive segmented pixels). mPrecision and precision are defined by

$$mPrecision = \frac{\sum_{k=1}^{C} Precision(k)}{C} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

where TP, FP, and FN denote the number of the true positive pixels, false positive pixels, and false negative pixels respectively. C denotes the number of pixels in the dataset.

mIoU is a mean value of IoU values, where IoU represents the intersection between the segmentation results and the ground truth. mIoU is defined by

$$mIoU = \frac{\sum_{i=1}^{K+1} \frac{|A_i \cap B_i|}{|A_i \cup B_i|}}{K+1} \tag{10}$$

where A and B represent the ground truth and segmentation results, respectively, and K denotes the number of classification categories.

FLOPs and Params are indicative metrics that reveal a model's level of complexity and the computational demands it places on execution resources.

FLOPs measure the total number of floating-point operations required to process an input through the model. This metric is crucial for understanding the computational cost and efficiency of the model, especially when deployed on hardware with limited computational resources. A lower number of FLOPs indicates a more computationally efficient model.

Params refer to the total number of trainable parameters within the model. This metric is an indicator of the model's capacity and complexity. Models with a higher number of parameters generally have a greater capacity to learn from data, but they also require more memory and computational power to train and infer. Managing the number of parameters is essential to balance the trade-off between model performance and resource utilization.

3.3. Experiment setup

The hardware and software configurations used for this study are as follows: an Intel(R) Core(TM) i7-11700 CPU processor and an NVIDIA GeForce RTX 3060 graphics card. The operating system was 64-

bit Windows 11. For deep learning tasks, the PyTorch framework was utilized, and the Matplotlib library was used for visualization.

The parameter settings of the models are as follows: stochastic gradient descent (SGD) with a momentum of 0.9 was selected as the optimization algorithm. The initial learning rate was set to 0.001, the batch size was configured to 4, and the training was conducted for 100 epochs.

4. Results

To evaluate the segmentation performance of the SSMM-DS model, two comparative experiments were conducted. The first experiment was an ablation study, comparing the SSMM-DS model with DeeplabV3+ under different settings. The second experiment was a comparative study, where the segmentation performance of the SSMM-DS model was compared against other models, including FCN [31], PSPNet [32], OCRNet [33], uPerNet [34] and SegFormer [35].

4.1. Ablation experiment

To evaluate the effects of the DCM, the Swin transformer backbone, and the loss function on the segmentation results for mangroves, the proposed SSMM-DS model was compared against the DeepLabV3+ model under different settings. These settings included DeepLabV3+ with the Xception backbone, DeepLabV3+ with the ResNet50 backbone, and DeepLabV3+ with the Swin transformer backbone. The Xception architecture, the primary backbone used in the original DeepLabV3+ implementation, was selected for its efficiency in depthwise separable convolutions, making it highly effective for capturing spatial hierarchies. ResNet50 was chosen due to its deep residual learning framework, which allows for better feature extraction in deeper networks. The Swin transformer backbone was used as the backbone network in this paper, aimed at leveraging long-range dependencies and contextual information within the image. Additionally, tests were conducted to assess the effectiveness of the DCM and the loss function. In validating the impact of the backbone network and loss function on the segmentation model, the experimental data consisted of three spectral bands: green, red, and NIR.

Table 4 presents the results of our ablation studies, which systematically evaluated the impact of different components of the proposed SSMM-DS model. Our findings demonstrate that the Swin Transformer Tiny backbone network significantly improved both mPrecision and mIoU compared to the DeepLabv3 model with the Xception backbone. Specifically, the mIoU increased by 1.25%, and the mPrecision rose by 0.55%. Furthermore, the choice of loss function played a crucial role in segmentation performance. The SSMM-DS model, employing a combination of cross-entropy and Dice loss functions, outperformed models using other loss functions. This combination effectively addressed the class imbalance issue commonly encountered in mangrove segmentation tasks, leading to the highest mIoU (94.72%) and mPrecision (97.65%). To enhance feature representation, we incorporated two spectral vegetation indices and one water index into the original image data, creating multichannel input for the model. This augmentation resulted in a further improvement of 0.39% in mIoU and 0.14% in mPrecision compared to the model using only the original image data. Ultimately, the proposed SSMM-DS model achieved impressive mIoU and mPrecision values of 95.11 and 97.79%, respectively, demonstrating the effectiveness of our approach in leveraging spectral information to improve mangrove segmentation accuracy.

Table 4. Compared results in ablation experiment.

Model	mIoU(%)	mPrecision(%)
Deeplabv3+ with (Xception + CE loss)	92.73	96.84
Deeplabv3+ with (ResNet50 + CE loss)	93.07	96.66
Deeplabv3+ with (Swin transformer tiny + CE loss)	93.98	97.39
Deeplabv3+ with (Swin transformer tiny + DICE loss)	93.69	97.07
Deeplabv3+ with (Swin transformer tiny + proposed loss)	94.72	97.65
Deeplabv3+ with (Swin transformer tiny + proposed loss + DCM)	**95.11**	**97.79**

4.2. Comparison experiment

Table 5 presents a comprehensive comparison of the proposed SSMM-DS model against state-of-the-art segmentation models, including FCN, PSPNet, OCRNet, uPerNet, and SegFormer. Our results demonstrate the superiority of SSMM-DS in terms of both performance and efficiency. The proposed model achieved the highest mIoU (95.11%) and mPrecision (97.79%) while maintaining a significantly lower parameter count (17.48M) compared to other models. Additionally, the SSMM-DS model's FLOPs were lower than FCN, PSPNet, OCRNet, and uPerNet, and comparable to SegFormer. SSMM-DS consistently outperformed models using ResNet50 as the backbone, including FCN, PSPNet, and OCRNet. This highlights the advantages of the Swin transformer architecture in capturing complex relationships within the image data. Even when compared to uPerNet, which also employs the Swin transformer tiny backbone, SSMM-DS achieved higher mIoU and mPrecision while maintaining a significantly lower computational cost. This demonstrates the efficiency of our proposed model design. Finally, in comparison to SegFormer using MiT-B4, SSMM-DS achieved comparable performance, albeit with a slightly higher computational cost. This suggests that our model design effectively balances performance and efficiency.

Figure 8 visually illustrates the segmentation results of different semantic segmentation models applied to mangrove images. The colored areas represent the predicted mangrove regions. Among all models, the SSMM-DS model demonstrated superior performance in extracting mangrove features. Its ability to accurately identify and delineate fine-grained structures within the mangrove ecosystem underscores its potential for practical applications in environmental monitoring and management.

As shown in Table 5 and Figure 8, our proposed SSMM-DS model exhibits exceptional performance in semantic segmentation tasks. It achieved the highest mIoU and mPrecision scores while maintaining a relatively small model size. These results demonstrate that our model can achieve high accuracy while maintaining computational efficiency, making it more suitable for deployment on resource-constrained devices and offering broad application prospects.

Table 5. Performance comparison of different methods.

Model	Backbone	mIoU(%)	mPrecision(%)	FLOPs(G)	Params(M)
FCN	ResNet50	92.08	95.99	37.87	47.12
PSPNet	ResNet50	92.15	95.92	34.23	46.60
OCRNet	ResNet50	91.52	96.11	29.30	36.51
uPerNet	Swin Transformer Tiny	93.08	96.47	44.79	58.94
SegFormer	MiT-B4	94.00	96.95	**9.90**	61.37
Ours	Swin Transformer Tiny	**95.11**	**97.79**	15.11	**17.48**

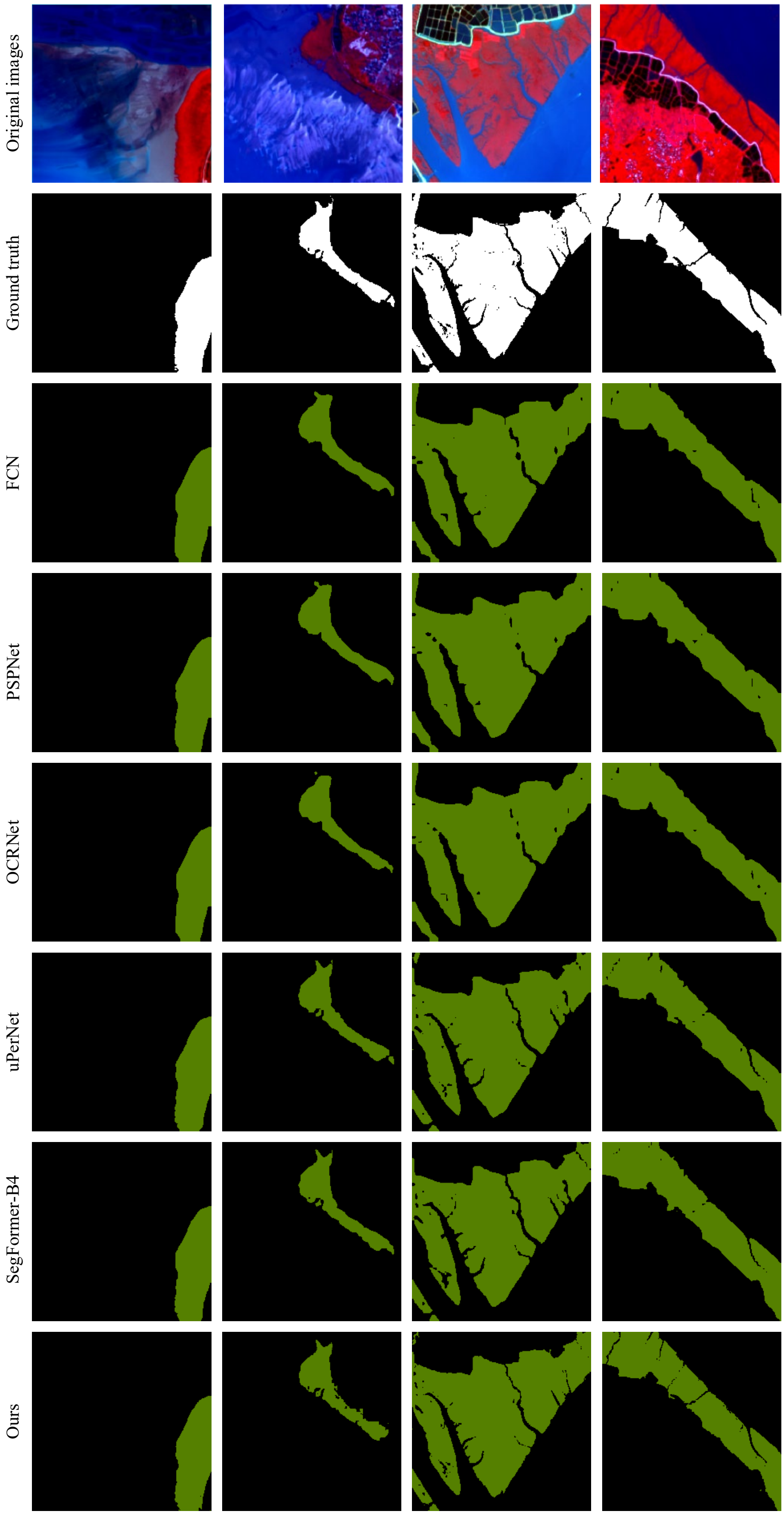


Figure 8. Segmentation results by different models.

5. Conclusions and discussion

Mangrove ecosystems are under severe threat from climate change and human activities. High-resolution remote sensing satellites offer a valuable data source for mangrove monitoring. This paper proposed a novel semantic segmentation model specifically designed for mangroves. To enhance the differentiation between mangroves and other vegetation and water bodies, we constructed multichannel input data by fusing two vegetation spectral indices and one water index from the original data. We employed the Swin transformer as the backbone network to improve the utilization of global features for the segmentation model. To further boost segmentation accuracy for small mangrove areas, we incorporated weighting coefficients into the loss function. Compared to FCN, PSPNet, OCRNet, uPerNet, and SegFormer, our proposed model achieved the highest segmentation accuracy, with mIoU reaching 95.11% and mPrecision attaining 97.79%.

The proposed segmentation model significantly enhances the accuracy of mangrove segmentation. This paper presents a deep learning segmentation model for mangrove segmentation based on GF satellite remote sensing imagery, substantially improving both accuracy and efficiency. However, several areas require further improvement:

(1) Multi-source data fusion: Leveraging multi-source data from satellites like Sentinel, Gaofen, and unmanned aerial vehicles can provide richer information for mangrove monitoring. Future research should focus on systematically fusing these data to enable quasi-real-time, multi-scale monitoring and change detection analysis.

(2) Optimization of high-performance algorithms: The current model faces computational efficiency limitations, hindering its application in large-scale mangrove monitoring. Lightweight network design, model compression, and incorporating prior knowledge of mangrove ecology and geography can effectively improve inference speed and enable real-time monitoring of large-scale mangroves.

(3) Sample imbalance problem: The scarcity of mangrove samples and their similarity to other vegetation can limit the model's generalization ability. Active learning, semi-supervised learning, and transfer learning can help address the insufficient sample problem and improve model robustness.

(4) Intelligent monitoring: High-precision and real-time mangrove monitoring is crucial for coastal zone management and ecological protection. Future research should focus on model interpretability to understand the decision-making process and guide model optimization. Additionally, building an intelligent mangrove monitoring platform can enable real-time monitoring and early warning of mangrove dynamic changes.

By harnessing remote sensing technology and advanced algorithms like deep learning, we can obtain real-time and comprehensive monitoring data of mangroves. This empowers us to analyze massive datasets, monitor mangrove growth dynamics, detect potential threats, and predict future trends. These technologies provide valuable decision support for mangrove conservation and management, facilitating the development of more effective conservation strategies.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was generously supported by the grants from scientific research project of Shanghai Oceanic Bureau (grant no. 2023-01).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. N. C. Duke, J. Meynecke, S. Dittmann, A. M. Ellison, K. Anger, U. Berger, et al., A world without mangroves?, *Science*, **317** (2007), 41–42. https://doi.org/10.1126/science.317.5834.41b
2. P. L. Biswas, S. R. Biswas, Mangrove forests: Ecology, management, and threats, *Life Land*, (2019), 1–14. https://doi.org/10.1007/978-3-319-71065-5_26-1
3. C. Giri, Frontiers in global mangrove forest monitoring, *Remote Sens.*, **15** (2023), 3852. https://doi.org/10.3390/rs15153852
4. S. C. Basha, An overview on global mangroves distribution, *Indian J. Geo Mar. Sci.*, **47** (2018), 766–772.
5. Y. Sun, D. Zhao, W. Guo, Y. Gao, X. Su, B. Wei, A review on the application of remote sensing in mangrove ecosystem monitoring, *Acta Ecol. Sin.*, **33** (2013), 4523–4538. https://doi.org/10.5846/stxb201205150715
6. R. Vidhya, D. Vijayasekaran, M. Ahamed Farook, S. Jai, M. Rohini, A. Sinduja, Improved Classification of mangroves health status using hyperspectral remote sensing data, *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, **8** (2014), 667–670. https://doi.org/10.5194/isprsarchives-XL-8-667-2014
7. T. D. Pham, D. T. Bui, K. Yoshino, N. N. Le, Optimized rule-based logistic model tree algorithm for mapping mangrove species using ALOS PALSAR imagery and GIS in the tropical region, *Environ. Earth Sci.*, **77** (2018), 159. https://doi.org/10.1007/s12665-018-7373-y
8. L. Li, W. Liu, Y. Tao, X. Xu, W. Fu, J. Dong, Diffusion dynamics and driving forces of Spartina alterniflora in the Guangxi Shankou Mangrove Reserve, *Acta Ecol. Sin.*, **41** (2021), 6814–6824.
9. J. Cao, K. Liu, L. Zhuo, L. Liu, Y. Zhu, L. Peng, Combining UAV-based hyperspectral and LiDAR data for mangrove species classification using the rotation forest algorithm, *Int. J. Appl. Earth Obs. Geoinformation*, **102** (2021), 102414. https://doi.org/10.1016/j.jag.2021.102414
10. M. Di Cicco, C. Potena, G. Grisetti, A. Pretto, Automatic model based dataset generation for fast and accurate crop and weeds detection, in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2017), 5188–5195. https://doi.org/10.1109/IROS.2017.8206408
11. D. Lee, C. Kim, S. Kim, M. Cho, W. Han, Autoregressive image generation using residual quantization, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 11513–11522. https://doi.org/10.1109/CVPR52688.2022.01123
12. L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, et al., UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, *ISPRS J. Photogramm. Remote Sens.*, **190** (2022), 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008

13. X. Yuan, J. Shi, L. Gu, A review of deep learning methods for semantic segmentation of remote sensing imagery, *Expert Syst. Appl.*, **169** (2021), 114417. https://doi.org/10.1016/j.eswa.2020.114417
14. A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G. Xia, et al., Transformers in Remote Sensing: A Survey, *Remote Sens.*, **15** (2023), 1860. https://doi.org/10.3390/rs15071860
15. L. Wan, H. Zhang, G. Lin, H. Lin, A small-patched convolutional neural network for mangrove mapping at species level using high-resolution remote-sensing image, *Ann. GIS*, **25** (2019), 45–55. https://doi.org/10.1080/19475683.2018.1564791
16. G. M. D. S. Moreno, O. A. D. Carvalho Júnior, O. L. F. D. Carvalho, T. C. Andrade, Deep semantic segmentation of mangroves in Brazil combining spatial, temporal, and polarization data from Sentinel-1 time series, *Ocean Coastal Manage.*, **231** (2023), 106381. https://doi.org/10.1016/j.ocecoaman.2022.106381
17. Y. Fan, Q. Zeng, Z. Mei, W. Hu, Semantic segmentation for mangrove using spectral indices and self-attention mechanism, in *2022 7th International Conference on Signal and Image Processing (ICSIP)*, (2022), 436–441. https://doi.org/10.1109/ICSIP55141.2022.9886553
18. C. Xu, J. Wang, Y. Sang, K. Li, J. Liu, G. Yang, An effective deep learning model for monitoring mangroves: A case study of the Indus Delta, *Remote Sens.*, **15** (2023), 2220. https://doi.org/10.3390/rs15092220
19. C. Kuenzer, A. Bluemel, S. Gebhardt, T. V. Quoc, S. Dech, Remote sensing of mangrove ecosystems: A Review, *Remote Sens.*, **3** (2011), 878–928. https://doi.org/10.3390/rs3050878
20. K. Maurya, S. Mahajan, N. Chaube, Remote sensing techniques: mapping and monitoring of mangrove ecosystem-A review, *Complex Intell. Syst.*, **7** (2021), 2797–2818. https://doi.org/10.1007/s40747-021-00457-z
21. C. Giri, Observation and monitoring of mangrove forests using remote sensing: Opportunities and challenges, *Remote Sens.*, **8** (2016), 783. https://doi.org/10.3390/rs8090783
22. T. V. Tran, R. Reef, X. Zhu, A review of spectral indices for mangrove remote sensing, *Remote Sens.*, **14** (2022), 4868. https://doi.org/10.3390/rs14194868
23. C. Liu, P. Sun, S. Liu, A review of plant spectral reflectance response to water physiological changes, *Chin. J. Plant Ecol.*, **40** (2016), 80–91. https://doi.org/10.17521/cjpe.2015.0267
24. J. Xue, B. Su, Significant remote sensing vegetation indices: A review of developments and applications, *J. Sens.*, (2017). https://doi.org/10.1155/2017/1353691
25. B. Gao, NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space, *Remote Sens. Environ.*, **58** (1996), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3
26. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 6000–6010.
28. X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced NLP tasks, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020), 465–476. https://doi.org/10.18653/v1/2020.acl-main.45

29. L. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Computer Vision-ECCV 2018*, **11211** (2018), 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
30. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data*, **6** (2019). https://doi.org/10.1186/s40537-019-0197-0
31. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
32. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 6230–6239. https://doi.org/10.1109/CVPR.2017.660
33. Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, in *Computer vision-ECCV 2020*, **12351** (2020), 173–190. https://doi.org/10.1007/978-3-030-58539-6_11
34. T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in *Computer Vision-ECCV 2018: 15th European Conference*, (2018), 432–448. https://doi.org/10.1007/978-3-030-01228-1_26
35. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, **34** (2021), 12077–12090.