applied sciences

MDPI

Article

Construction of Multi-Scale Fusion Attention Unified Perceptual Parsing Networks for Semantic Segmentation of Mangrove Remote Sensing Images

Xin Wang [1,2,3], Yu Zhang [2,*], Wenquan Xu [3], Hanxi Wang [4], Jingye Cai [1], Qin Qin [3], Qin Wang [5] and Jing Zeng [3]

[1] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China; wxin@guet.edu.cn (X.W.); jycai@uestc.edu.cn (J.C.)
[2] School of Computer Science and Information Security, Guilin University of Electronic Science and Technology, Guilin 541004, China
[3] School of Marine Engineering, Guilin University of Electronic Technology, Beihai 536000, China; xwq@mails.guet.edu.cn (W.X.); qinqin@guet.edu.cn (Q.Q.); zengjing@guet.edu.cn (J.Z.)
[4] School of Geographical Sciences, Harbin Normal University, Harbin 150025, China; wanghanxizs1982@126.com
[5] Basic Teaching Department, Guilin University of Electronic Technology, Beihai 536000, China; wangq@guet.edu.cn
* Correspondence: zy@mails.guet.edu.cn

Abstract: Mangrove forests play a crucial role in coastal ecosystem protection and carbon sequestration processes. However, monitoring remains challenging due to the forests' complex spatial distribution characteristics. This study addresses three key challenges in mangrove monitoring: limited high-quality datasets, the complex spatial characteristics of mangrove distribution, and technical difficulties in high-resolution image processing. To address these challenges, we present two main contributions. (1) Using multi-source high-resolution satellite imagery from China's new generation of Earth observation satellites, we constructed the Mangrove Semantic Segmentation Dataset of Beihai, Guangxi (MSSDBG); (2) We propose a novel Multi-scale Fusion Attention Unified Perceptual Network (MFA-UperNet) for precise mangrove segmentation. This network integrates Cascade Pyramid Fusion Modules, a Multi-scale Selective Kernel Attention Module, and an Auxiliary Edge Neck to process the unique characteristics of mangrove remote sensing images, particularly addressing issues of scale variation, complex backgrounds, and boundary accuracy. The experimental results demonstrate that our approach achieved a mean Intersection over Union (mIoU) of 94.54% and a mean Pixel Accuracy (mPA) of 97.14% on the MSSDBG dataset, significantly outperforming existing methods. This study provides valuable tools and methods for monitoring and protecting mangrove ecosystems, contributing to the preservation of these critical coastal environments.

Keywords: remote sensing image; mangrove monitoring; semantic segmentation; Feature Pyramid Network (FPN); attention mechanism

Academic Editor: Mauro Lo Brutto

Received: 2 November 2024
Revised: 2 January 2025
Accepted: 9 January 2025
Published: 20 January 2025

Citation: Wang, X.; Zhang, Y.; Xu, W.; Wang, H.; Cai, J.; Qin, Q.; Wang, Q.; Zeng, J. Construction of Multi-Scale Fusion Attention Unified Perceptual Parsing Networks for Semantic Segmentation of Mangrove Remote Sensing Images. *Appl. Sci.* **2025**, *15*, 976. https://doi.org/10.3390/app15020976

1. Introduction

1.1. Research Background and Significance

Mangrove forests are wetland woody plant communities, primarily composed of evergreen trees and shrubs, that grow at the land–sea interface. These ecosystems play vital roles in maintaining ecological balance, serving both as marine carbon sinks and providing wind protection and wave attenuation functions. They are renowned as "coastal guardians"

and "marine lungs" [1]. However, these critical ecosystems face multiple threats, including human activities such as aquaculture expansion [2] and river management projects [3], along with natural factors like biological invasion [4] and sea-level rise [5], leading to severe degradation of global mangrove forests.

According to remote sensing data analysis, the global mangrove area decreased by more than 10,000 square kilometers between 1985 and 2020, resulting in significant deterioration of wetland ecological functions [6]. The Global Mangrove Alliance predicts this degradation trend will continue until 2030 [7], further exacerbating global climate change and triggering a vicious cycle of carbon storage loss. To address these challenges, international organizations and national governments have launched multiple initiatives. The International Union for Conservation of Nature and the United Nations Development Programme established the "Mangroves for the Future" program, while the Chinese government formulated the "Special Action Plan for Mangrove Protection and Restoration (2020–2025)". At the local level, the Beihai Municipal Government of Guangxi Zhuang Autonomous Region developed the "Beihai City Mangrove Resource Protection Plan (2020–2030)". These initiatives highlight the importance of effective mangrove monitoring in coastal ecological protection.

1.2. Challenges in Mangrove Monitoring

Remote sensing technology has become an indispensable tool in mangrove monitoring and management due to its advantages of low cost, high efficiency, and broad observation range. It is widely applied in mangrove wetland mapping, community structure analysis, biomass estimation, and disaster warning [8]. However, current monitoring methods face three major challenges.

Firstly, there are significant limitations at the dataset level. Current research primarily relies on medium- to low-resolution satellite data [9,10], with a relatively large spatial resolution that fails to meet the needs of fine-scale monitoring. While drone data [11] and high-resolution satellite data [12] are becoming more available, annotated datasets for deep learning remain scarce. Furthermore, the lack of standardized datasets makes it difficult to directly compare results between different studies, severely constraining algorithm development and validation processes. Additionally, the acquisition of mangrove wetland monitoring data is often limited by weather conditions and satellite revisit cycles, making it challenging to accumulate high-quality training data.

Secondly, compared to horizontal-view images captured by traditional ground cameras, remote sensing images are collected from an aerial perspective, featuring a wide imaging range, high resolution, and high information density [13]. The unique spatial distribution characteristics of mangrove wetlands also pose significant challenges for monitoring. As shown in Figure 1, they primarily grow at the land–sea interface, displaying highly fragmented distribution patterns. (Figure 1a) This spatial discontinuity significantly increases the complexity of target identification. More challengingly, mangrove wetlands exhibit remarkable scale diversity, ranging from discretely distributed small shrub communities to large-scale continuous forest areas [14]. (Figure 1c) These multi-scale characteristics place higher demands on semantic segmentation algorithms. Furthermore, due to their growth in intertidal zones, the image characteristics of mangroves change significantly with tidal variations. (Figure 1b) The same area may present entirely different image features under different tidal conditions, and this dynamically changing background environment further complicates monitoring efforts.

Figure 1. Several challenges in mangrove remote sensing image semantic segmentation.

Thirdly, significant challenges exist in technical processing. With the increasing resolution of remote sensing data, the data volume of single-scene images has increased dramatically [15], placing enormous pressure on computational resources. Processing high-resolution images not only requires more powerful hardware support but also demands algorithm-level optimization to improve processing efficiency. In practical applications, maintaining consistent segmentation accuracy across different scales poses higher requirements for algorithm robustness. How to achieve efficient processing while ensuring segmentation accuracy under limited computational resources remains a significant technical challenge.

The existence of these challenges indicates the need to develop more advanced data acquisition methods, more effective processing algorithms, and more comprehensive monitoring systems. Particularly in the context of rapidly developing deep learning technology, how to fully utilize new technologies to overcome these limitations and improve the accuracy and efficiency of mangrove monitoring has become a key research direction.

1.3. Current Research Status

In recent years, significant progress has been made in mangrove wetland monitoring with the rapid development of remote sensing and deep learning technologies [16,17]. Early mangrove remote sensing monitoring primarily relied on medium to low-resolution satellite data. Researchers mainly used satellites like Landsat and Sentinel for mangrove distribution monitoring but were limited by low image resolution and complex data preprocessing steps. To improve wetland classification accuracy, research shifted toward using high-resolution drone data and satellite imagery (such as GF, BJ, and SV series). Traditional monitoring methods mainly employed shallow machine learning algorithms like Random Forest (RF) [18], Support Vector Machine (SVM) [19], and object-oriented approaches for mangrove extraction and spatiotemporal analysis based on pixel classification. These methods required manual feature design and intermediate semantic features, lacking flexibility and adaptability when dealing with multi-scale targets and inconsistent category distributions like mangrove wetlands.

With the rapid advancement of deep learning technology [20,21], new research directions for semantic segmentation emerged, providing novel solutions for mangrove monitoring. The introduction of Fully Convolutional Networks (FCNs) [22] pioneered end-to-end, pixel-level classification, using deconvolution to upsample the final feature map to input image size. However, FCNs showed inadequate feature extraction and loss of detailed information during upsampling when processing mangrove images, resulting in

poor performance in mangrove wetland semantic segmentation. Subsequently, numerous semantic segmentation networks were proposed. PSPNet [23] aggregated a context from different regions through Pyramid Pooling Modules and pyramid scene parsing networks, enabling global context understanding. However, its fixed-scale pooling operations struggled to adapt to the varied morphological features of mangrove wetlands. DANet [24] introduced spatial and channel attention mechanisms to balance local features and global dependencies. PSANet [25] explored position-related attention mechanisms by adaptively predicting global attention maps for each position in feature maps through convolutional layers, improving the feature recognition of scattered wetlands but lacking accuracy in mangrove wetland edge regions. DeepLabv3 [26] used Atrous Spatial Pyramid Pooling to collect different-scale receptive fields and effectively expanded filters without increasing parameters and computational complexity but still faced insufficient receptive field issues when processing high-resolution remote sensing images. DeepLabv3+ [27] introduced dilation rates for dilated convolution and improved Atrous Spatial Pyramid Pooling, achieving better segmentation results but remaining inadequate in handling complex mangrove boundary regions. K-Net [28] achieved multi-task unification by assigning different kernels to each task, but its dynamic kernel generation process was computationally expensive. PIDNet [29] implemented semantic segmentation and boundary detection but showed suboptimal detection accuracy for targets with special growth patterns like mangroves. While UperNet [30] could parse visual concepts like cross-scene categories, objects, and textures while performing scene recognition, object detection, and semantic segmentation tasks, its feature pyramid structure showed insufficient information transmission when processing multi-scale mangrove features.

In summary, directly applying these methods to high-resolution mangrove remote sensing image semantic segmentation still faces many challenges. Firstly, the existing models are mostly designed for natural scene images and show inadequate feature extraction when processing remote sensing images with special spectral and texture features, particularly in areas where mangroves mix with other vegetation. While some models adopt multi-scale feature fusion strategies, existing fusion methods still struggle to effectively adapt to the scale variation characteristics exhibited by mangroves under different tidal conditions. In transition areas between mangroves, water bodies, and other vegetation, existing models often show boundary blur and inaccurate segmentation, especially in scattered distribution areas with poor boundary localization accuracy. Additionally, these models show unstable performance across different temporal phases and regions of mangrove images, with limited generalization ability causing it to struggle to adapt to complex and varying monitoring requirements.

To address these model deficiencies in wetland remote sensing recognition, researchers have conducted a series of improvements. Liu et al. [31] comparatively analyzed the performance of different resolution remote sensing datasets in wetland vegetation classification, improving wetland vegetation classification accuracy through enhanced DeepLabv3+ network structure. Addressing the large-scale variations and complex background characteristics of mangrove wetlands, Wang et al. [32] proposed the Swin-UperNet model based on UperNet, successfully achieving high-precision simultaneous segmentation of mangroves and Spartina alterniflora. Li et al. [33] innovatively incorporated prior background knowledge of remote sensing images into network design, proposing the LSKNet Large Selective Kernel network, effectively enhancing the modeling capability of contextual relationships between targets in remote sensing scenes. Wang et al. [34] optimized channel and spatial position relationship modeling by designing adaptive local cross-channel vector aggregation attention modules, achieving significant segmentation accuracy improvements on their constructed MO-CSSSD mangrove dataset. These targeted improvements not

only validate the optimization potential of the existing models but also provide important theoretical and practical references for designing new network architectures.

1.4. Contributions of This Paper

This paper addresses existing problems in mangrove remote sensing monitoring, including data resource scarcity, insufficient feature extraction, inadequate multi-scale adaptability, and low boundary segmentation accuracy. The main contributions are as follows:

(1) The construction of the Mangrove Remote Sensing Image Semantic Segmentation Dataset in Beihai of Guangxi (MRSDBG). This dataset is built upon multi-temporal high-resolution remote sensing images, integrating multiple satellite data sources, and features multi-temporal, multi-scale, and high generalization characteristics. The dataset provides fine-grained, pixel-level annotations, offering crucial support for the development and validation of mangrove remote sensing monitoring algorithms. The construction of this dataset fills the void of high-quality standard datasets in the field of mangrove remote sensing monitoring, providing a reliable data foundation for related research.

(2) The design of an innovative MFA-UperNet network architecture. This network achieves high-precision semantic segmentation of mangrove remote sensing images through the organic combination of three core modules: feature encoder, Auxiliary Edge Neck, and semantic decoder: The feature encoder, based on the ConvNeXt [35] feature extraction backbone network, performs multi-level semantic feature extraction, generating feature maps at different scales, fully utilizing feature information from different levels to establish a solid feature foundation for subsequent precise segmentation. The Auxiliary Edge Neck specifically designs an edge-aware branch, enhancing segmentation accuracy in mangrove boundary regions through fine-grained boundary feature learning. The semantic decoder introduces Cascade Pyramid Fusion Modules and Multi-Scale Selective Kernel Attention Modules, innovatively combining cascade pyramid structure with selective kernel attention mechanisms to achieve efficient multi-scale feature fusion. This not only enhances the model's adaptability to mangrove multi-scale variations but also significantly improves the recognition accuracy of mangrove targets in complex backgrounds.

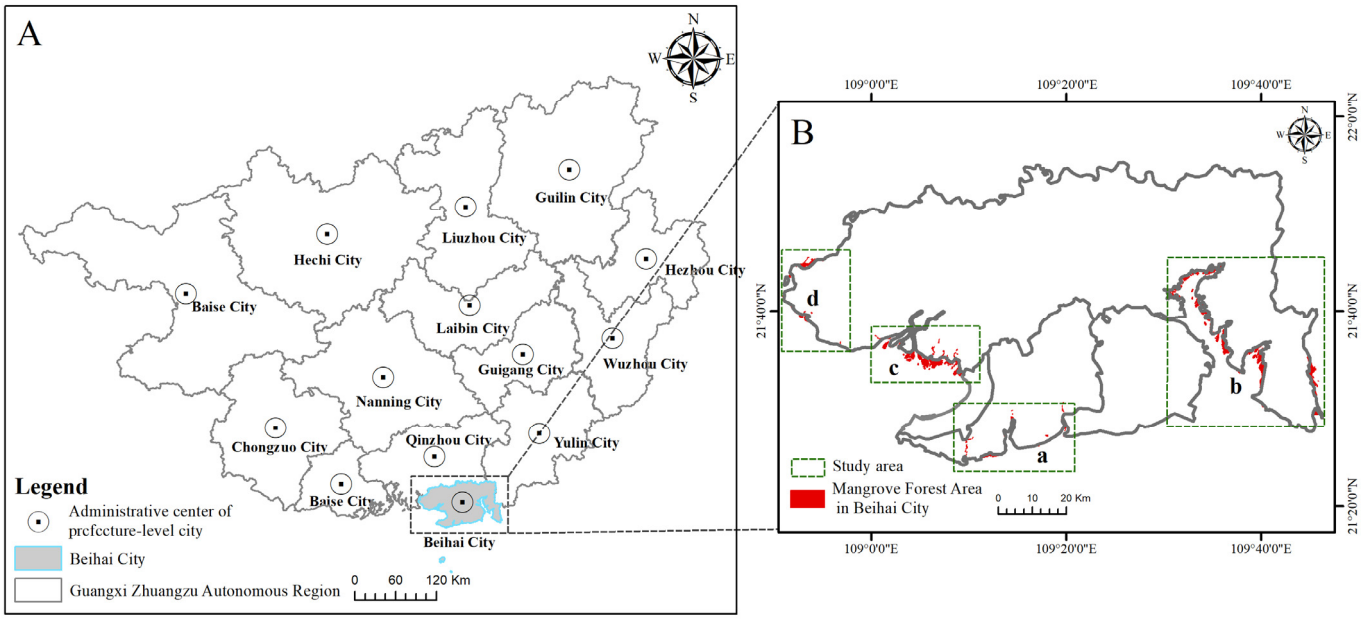
(3) The systematic experimental validation of MFA-UperNet architecture's superior performance in mangrove remote sensing image semantic segmentation tasks. Experimental results demonstrate significant improvements in mangrove segmentation tasks. These improvements provide new technical solutions for enhancing the accuracy and efficiency of mangrove coastal wetland ecological monitoring, offering important practical value for advancing mangrove protection and management.

2. Materials and Methods

2.1. Study Area

The study area is situated in the southern bay (108°84′–109°78′ E and 21°20′–22°00′ N) of Beihai City, Guangxi Zhuang Autonomous Region, China. Beihai City is recognized as one of the 25 major coastal ports in China and serves as the largest distribution area of mangrove resources in the Guangxi Zhuang Autonomous Region. It holds significant importance as part of China's mangrove nature reserve. The area is home to 4210.99 hectares of mangrove forests, which constitute 44.72% of the total mangrove ecosystem area in Guangxi. These forests effectively serve as carbon sinks within marine ecosystems, which contributes to China's pursuit of its "double carbon" goal. The study area comprises two national mangrove nature reserves, one provincial nature reserve, and one national wetland park. The dominant mangrove communities consist of *Rhizophora stylosa*, *Bruguiera gymnoihiza*, *Kandelia obovata*, *Excoecaria agallocha*, *Aegiceras corniculatum*, and *Avicennia marina*. A rare and endangered species *Lumnitzera racemosa* is dispersed within Gongguan Town, Hepu

County. According to the "Beihai Mangrove Resource Protection Plan (2020–2030)" published on the Beihai Municipal Government website, four main mangrove distribution areas were selected as study areas, as shown in Figure 2.



GS (2019) No. 1822

Figure 2. The geographic location of the study area. (**A**) The coastal area of Beihai City, Guangxi, China. (**B**) The study area (a. Coastal National Wetland Park of Beihai, Guangxi, and Hengluoshan area of Yinhai District, Beihai, Guangxi; b. Dugong and Shankou National Nature Reserve area of Hepu, Guangxi; c. the coastal area of Lianzhou Bay of Hepu, Beihai, Guangxi; and d. the coastal area of Maowei Sea and Dafeng River, Guangxi).

2.2. Data Sources

The mangrove label data used in this study primarily come from multiple sources, including multi-source satellite remote sensing data, historical reference data, field survey data, and government planning documents. The satellite data include high-resolution images provided by multiple satellites such as Beijing-2, Beijing-3, Gaofen-7, SuperView-1, and Jilin-1, spanning from 2018 to 2022. The selection of these data sources was based on principles of multi-source nature, timeliness, high resolution, reliability, and complementarity. In practical data experiments, we first screened remote sensing satellite data according to the officially recognized protected area boundaries provided in the "Beihai Mangrove Resource Protection Plan (2020–2030)" and then conducted preliminary verification using the Chinese Mangrove Long Time Series Satellite Remote Sensing Monitoring Dataset (1978–2018) [36] as a historical reference for corresponding years. Finally, cross-validation and refined verification were performed using GPS-positioned field survey data to ensure the accuracy of data labeling. This three-step progressive data processing workflow of "official planning guidance—historical data reference—field verification" not only guarantees the authority of study area selection but also ensures the spatiotemporal continuity and practical applicability of mangrove distribution information. Through the comprehensive utilization of these multi-source, multi-temporal, and multi-scale data, we were able to generate more accurate and reliable mangrove labels, providing high-quality foundational data for subsequent deep learning model training while also ensuring that the annotation results reflect current conditions and possess practical application value.

Beijing-2 satellite (BJ-2) was launched in July 2015 from the Satish Dhawan Space Center Sriharikota in India (Beijing time). For effective mangrove segmentation, five 0.8 m

resolution multispectral images with blue, green, red, and near-infrared bands were selected. Beijing-3 satellite (BJ-3) was launched in June 2021 from the Taiyuan Satellite Launch Center in China (Beijing time), and one 0.3 m resolution multispectral image with blue, green, red, and near-infrared bands was selected. Gaofen-7 satellite (GF-7) was successfully launched in November 2019 from the Taiyuan Satellite Launch Center, and one 0.5 m resolution multispectral image with blue, green, red, and near-infrared bands was selected. SuperView-1 satellite (SV-1) was launched in December 2016 from the Taiyuan Satellite Launch Center, and one 0.5 m resolution multispectral image with blue, green, red, and near-infrared bands was selected. Jilin-1 satellite (JL-1) is China's core commercial high-resolution remote sensing satellite project under construction, encompassing multiple series of high-performance optical remote sensing satellites with high-resolution, wide-swath, video, and multispectral capabilities. One 0.8 m resolution multispectral image with blue, green, red, and near-infrared bands was selected. The relevant remote sensing data information is shown in Table 1.

Table 1. Information on remotely sensed data from the MSSDBG dataset.

Study Area	Satellite	Number of Images	Date
a, c	BJ-2	5	2 October 2018–22 November 2018
b	BJ-3	1	23 December 2022
b	GF-7	1	8 November 2021
b	SV-1	1	12 October 2022
a	JL-1	1	9 April 2022

2.3. Data Preprocessing

This study used the raster processing batch tool in ENVI 5.6 software to complete the preliminary processing steps, including radiometric calibration, atmospheric correction, and image cropping, followed by data annotation work. First, a multi-source data fusion strategy was adopted, combining high-resolution data from multiple satellites, including BJ-2, BJ-3, GF-7, SV-1, and JL-1, covering a time span from 2018 to 2022, with spatial resolutions ranging from 0.3 m to 0.8 m, which greatly reduced potential errors from single data sources. Secondly, regarding the standards for constructing and selecting image samples, we established strict screening criteria: cloud coverage in selected images was controlled below 10%, ensuring high image clarity without stripe noise and systematic distortion. Meanwhile, our study area comprehensively covers mangrove distribution areas in Beihai City, including typical ecosystems such as estuary regions, coastal zones, wetland protected areas, and wetland parks. Special attention was paid to collecting mangrove communities of different density levels, covering areas from sparse to dense growth while also considering various spatial distribution patterns, including continuous and scattered distributions. The implementation of these standards effectively ensured data reliability and accuracy. Finally, experts from geographical science and computer technology fields were brought in for collaborative annotation, combining GIS and remote sensing image interpretation techniques to enhance the professionalism and reliability of the annotations. Additionally, this study referenced historical data, particularly the Chinese Mangrove Long Time Series Satellite Remote Sensing Monitoring Dataset (1978–2018), ensuring the temporal continuity and spatial consistency of annotations. Figure 3 displays examples of the dataset's original remote sensing images and their corresponding annotation results, visually demonstrating the precision and professional level of the annotations. Furthermore, this study employed measures such as iterative optimization, strict quality control processes, and temporal consistency checks to ensure the reliability of the annotation results. This comprehensive approach not only overcomes the limitations of traditional manual marking in terms of

subjectivity, efficiency, and accuracy but also provides valuable experience and reference for large-scale mangrove monitoring and other remote sensing image annotation tasks.

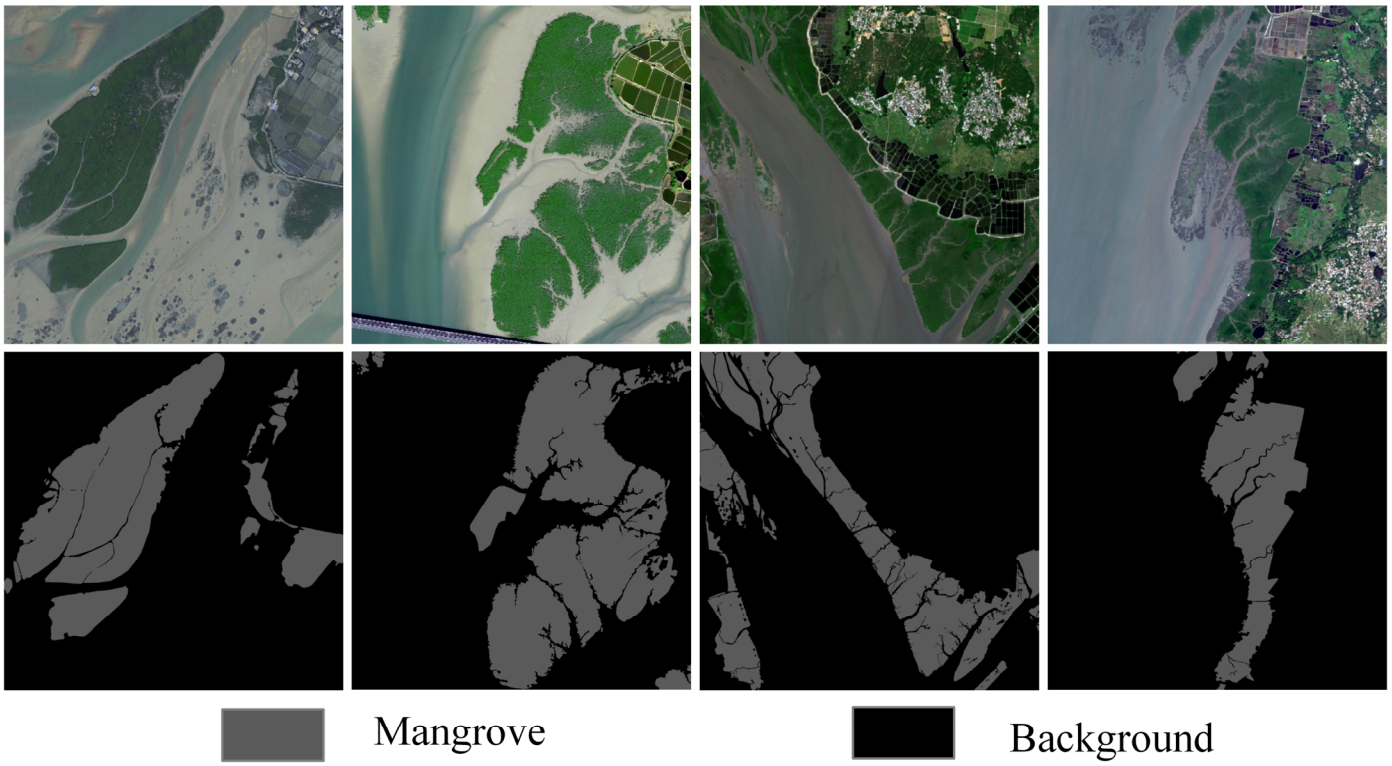


Figure 3. Examples of Original Remote Sensing Images and Their Corresponding Annotation Results.

After annotation, this paper used a 512 × 512-pixel sliding window (i.e., moving steps in horizontal and vertical directions) to crop large-scale remote sensing images, as shown in Figure 4, ultimately constructing a dataset containing 2568 remote sensing images. To ensure scientific training and objective evaluation of the model, the entire dataset was divided in a 7:2:1 ratio into mutually independent sets of 1798 training images, 514 validation images, and 256 test images. Firstly, this method satisfies the input requirements of deep learning models (which automatically learn features and patterns in data through multi-layer neural networks), as remote sensing images are typically too large to be directly input into standard memory deep convolutional neural network models. The 512 × 512-pixel window size maintains a sufficient spatial resolution to capture target features while enabling efficient training and inference under existing hardware conditions. Secondly, this cropping method helps optimize computational resources by processing large-scale remote sensing images in batches, reducing resource requirements for single computations and improving computational efficiency and parallel processing capabilities. Furthermore, the 512 × 512 window size preserves local spatial details while covering sufficient ground feature information, helping models better learn and recognize complex ground features such as mangrove distribution and morphology. By cropping different image blocks and combining sliding window overlap settings, more diverse training samples can be generated. Smaller image blocks are also more conducive to data augmentation operations, helping improve model generalization ability and reduce overfitting risk. Setting overlap regions ensures the preservation of boundary details when recombining image blocks, which is particularly important for maintaining spatial continuity and improving overall classification or detection accuracy. Finally, processing smaller image blocks helps better manage computer memory and avoid memory overflow issues. In conclusion, this cropping method not only meets the technical requirements of deep learning models but

also ensures image information integrity while improving processing efficiency and model performance, holding significant importance for large-scale remote sensing image analysis and mangrove monitoring.

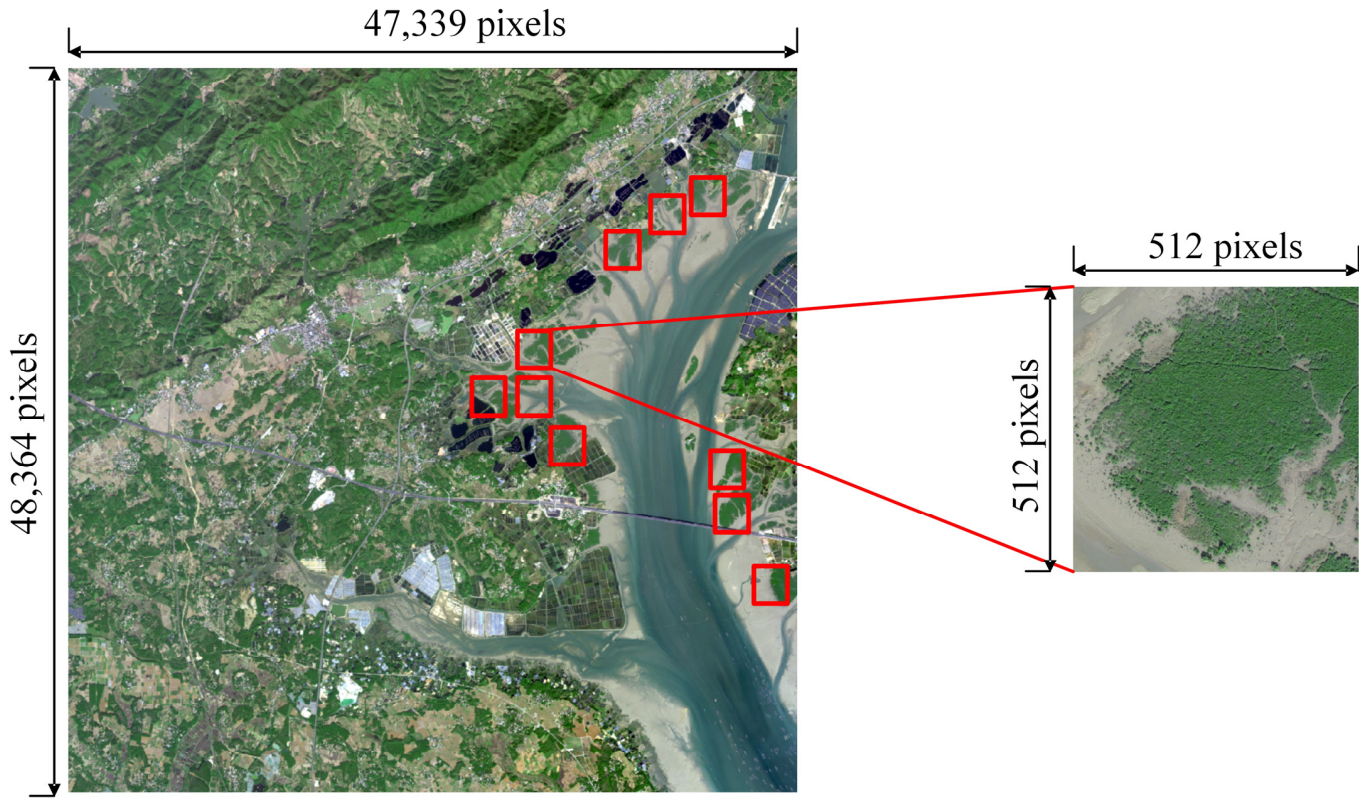


Figure 4. Schematic diagram of the original and cropped images.

After completing the cropping process, this paper performed data augmentation operations, including random image resizing, random cropping and flipping of images and segmentation maps, and photometric distortions. Photometric distortions (such as adjusting brightness, contrast, and saturation) simulated the effects of different lighting conditions and sensor variations on remote sensing images, thereby improving network adaptability across various environments. Random cropping and flipping operations helped the network learn position- and orientation-invariant feature representations, which is crucial for identifying targets (such as mangroves) in different directions and positions within remote sensing images. These techniques significantly increased the diversity of the training data, helping the network learn more robust and generalized feature representations while reducing the risk of overfitting. When training data were limited, data augmentation effectively expanded the training set, improving the network's few-shot learning capability. Additionally, these operations enhanced model robustness to input perturbations and helped balance imbalanced datasets. In summary, through these data augmentation strategies, we were able to fully utilize limited training data to build more reliable and efficient deep learning models, which is significant for remote sensing image analysis and mangrove monitoring.

In conclusion, this study successfully overcame many limitations of traditional manual annotation in mangrove monitoring through multi-source data fusion, expert collaborative annotation, and effective data cropping and augmentation strategies. The proposed comprehensive data preprocessing method not only improved annotation accuracy and efficiency but also significantly enhanced model generalization ability and adaptability, providing a solid technical foundation for large-scale remote sensing image analysis. This

method not only demonstrated significant advantages in mangrove monitoring but also provided a valuable reference for other types of remote sensing image annotation tasks, offering broad application prospects and promotional value.

2.4. Method

2.4.1. Overall Architecture Overview

The overall architecture of the proposed MFA-UperNet network is shown in Figure 5, adopting an encoder–decoder structure design that primarily consists of three core components: feature encoder, Auxiliary Edge Neck, and semantic decoder.

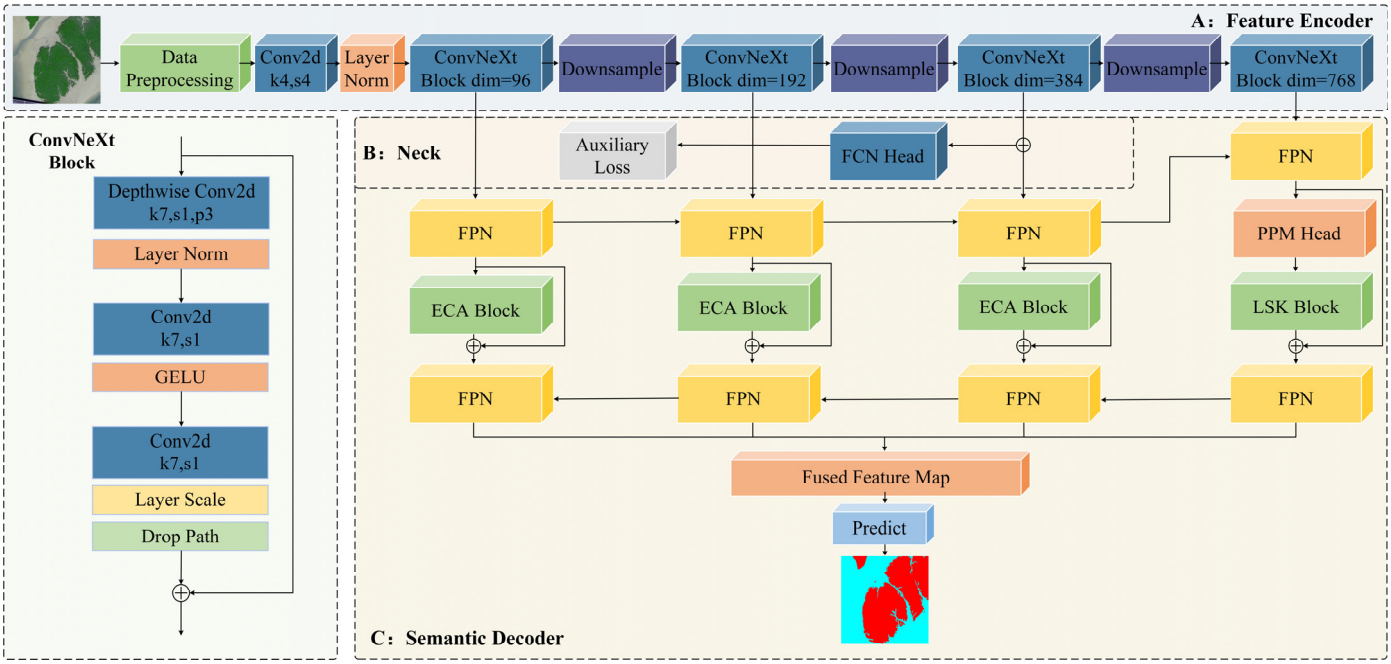


Figure 5. Overall MFA-UperNet Network Structure (**A**. Feature Encoder, **B**. Auxiliary Edge Neck, **C**. Semantic Decoder).

In the feature encoding stage, the network uses ConvNeXt as the backbone, progressively converting high-resolution remote sensing images into multi-scale feature maps through four consecutive feature extraction levels. In the neck region of the network, an innovative auxiliary edge segmentation branch is designed, which improves segmentation accuracy in edge regions while providing additional gradient information for the backbone network, promoting overall model optimization. In the decoding stage, this paper improves the traditional UperNet structure by proposing a Cascade Pyramid Fusion Module (CPFM) and Multi-scale Selective Kernel Attention Module (MSKAM), achieving adaptive feature selection and enhancement, significantly improving the network's ability to recognize mangrove targets at different scales and in complex backgrounds.

2.4.2. Feature Encoder

This study selects ConvNeXt as the feature encoder. As a significant breakthrough in the visual field in recent years, ConvNeXt maintains the efficient computation of convolutional neural networks while integrating the advantageous design elements of Transformer [37]. It possesses powerful feature extraction capabilities and an optimized network architecture, characteristics particularly suitable for processing complex features in mangrove remote sensing images. The specific structure is shown in Figure 6.

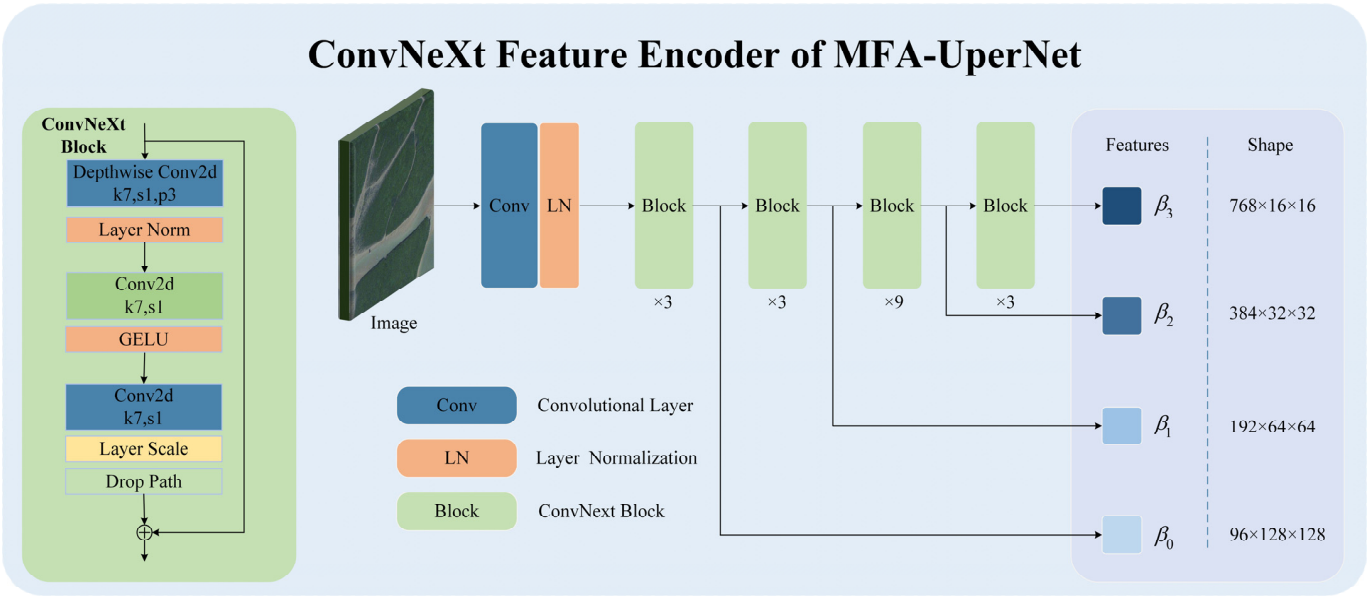


Figure 6. Overall Structure Diagram of ConvNeXt Feature Encoder.

Specifically, the input mangrove remote sensing image (512 × 512) first passes through a 4 × 4 convolutional layer, followed by a normalization layer. After these two layers of processing, the image resolution becomes one-fourth of the original (128 × 128). The features are then processed through 4 stages of ConvNeXt Blocks, where each ConvNeXt Block includes two-dimensional convolution (Conv2d), layer normalization (LN), GELU activation function, and Layer Scale and Drop Path regularization layers. Each stage contains 3, 3, 9, and 3 blocks, respectively. Through the ConvNeXt feature extractor, four different-sized multi-scale feature information sets are extracted, denoted as (β_0, β_1, β_2, β_3), providing multi-scale feature input for subsequent Auxiliary Edge Neck and semantic decoders.

ConvNeXt plays a crucial role in this study's mangrove semantic segmentation task. First, its powerful feature extraction capability enables the network to effectively capture multi-level information of mangroves from local textures to global structures. Shallow features retain rich spatial details, helping precisely locate mangrove boundaries, while deep features contain rich semantic information, facilitating distinction between different categories. Second, ConvNeXt's optimized architectural design is particularly suitable for processing complex textures and multi-scale features of mangrove remote sensing images, ensuring both inference speed and high feature expression capability. Finally, through deployment, it achieves faster inference speed while maintaining high accuracy, demonstrating the model's efficiency in practical applications. This efficient multi-scale feature extraction mechanism provides a solid feature foundation for MFA-UperNet, significantly improving model performance in mangrove semantic segmentation tasks, showing notable advantages, especially in handling complex backgrounds and ambiguous boundaries.

2.4.3. Auxiliary Edge Neck

In mangrove remote sensing image semantic segmentation tasks, precise edge region identification is particularly crucial. This is because pixels within objects typically have consistent semantic features, while semantic discontinuities mainly occur at boundaries between adjacent objects. Particularly in this paper's mangrove dataset, boundary regions between mangroves and backgrounds often exhibit ambiguity and uncertainty. This phenomenon primarily stems from the special mechanism of wetland vegetation in remote sensing imaging: different types of vegetation may have similar backscattering character-

istics [38], making it difficult to distinguish mangroves from the surrounding vegetation at boundaries.

To address this issue, this study proposes the Auxiliary Edge Neck (AEN). The AEN module directly uses the third layer features (β2) from the feature extractor as input, as this layer retains sufficient spatial details while containing certain semantic information. In terms of network structure, it adopts a lightweight fully convolutional network (FCN) design, processing features through two 3 × 3 convolutional layers. Meanwhile, an auxiliary hybrid loss function is designed (see Section 2.5). The AEN computation process is shown in Formula (1):

$$F = conv_2(conv_1(\beta_2)) \quad (1)$$

where $conv_1$ uses a 3 × 3 convolution operation with stride 1, outputting 256 channels, followed by the BatchNorm2d and ReLU activation function; $conv_2$ is a 3 × 3 convolution that maps feature map channels to the number of target classes (2 in this task, corresponding to edge and non-edge). The AEN structure is lightweight and efficient. During training, AEN loss is weighted and combined with the semantic decoder's loss, with a weight coefficient of 0.4, providing additional gradient information through this auxiliary supervision mechanism to enhance the feature extractor's ability to learn edge features.

2.4.4. Semantic Decoder

The semantic decoder is a key component of MFA-UperNet, primarily consisting of the Cascade Pyramid Fusion Module (CPFM) and Multi-scale Selective Kernel Attention Module (MSKAM). The collaborative work of these two modules achieves an effective fusion of multi-scale features and adaptive enhancement of key information. The overall structure is shown in Figure 7.

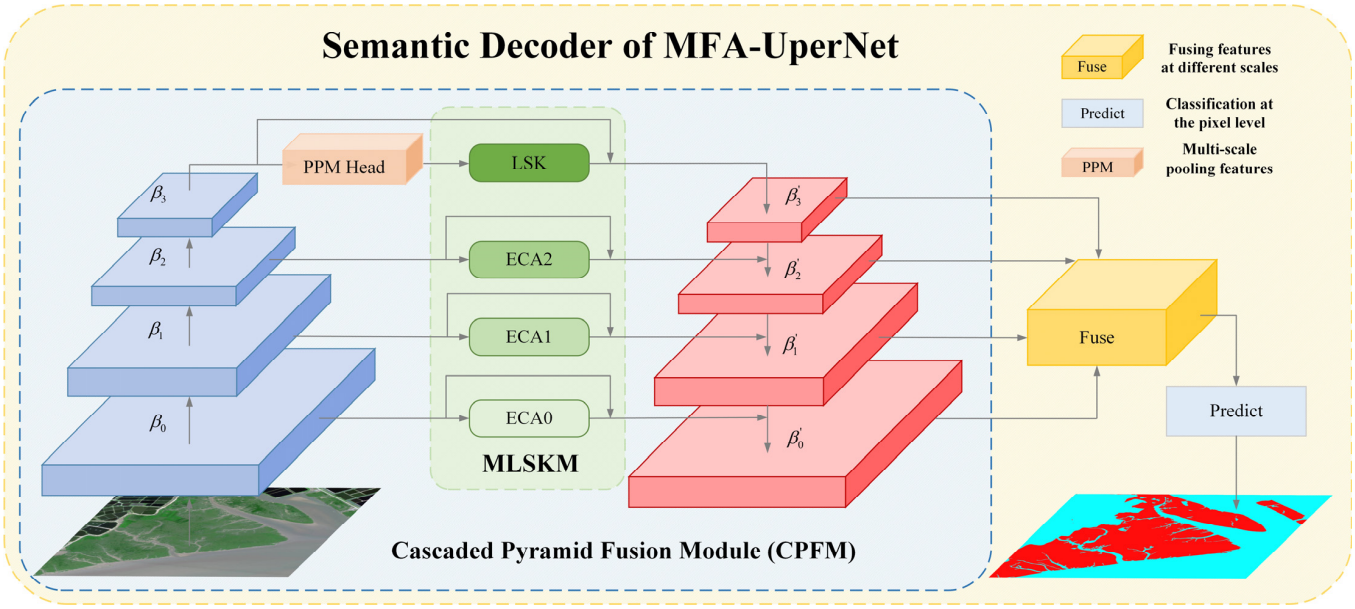


Figure 7. Overall Structure Diagram of MFA-UperNet Semantic Decoder.

Cascade Pyramid Fusion Module (CPFM)

Mangrove remote sensing images are characterized by complex backgrounds and variable scales, posing severe challenges for semantic segmentation. Features at different levels contain complementary information: low-level features preserve detailed textures, aiding precise localization, while high-level features contain rich semantics crucial for category recognition. To fully utilize this multi-level information, this paper designed the Cascade Pyramid Fusion Module (CPFM) based on UperNet network and Feature Pyramid

Network (FPN) [39]. Its core concept is to achieve the effective fusion of multi-scale features through FPN and enhance it by introducing the Pyramid Pooling Module (PPM) [23] and Multi-Scale Selective Kernel Attention Module (MSKAM).

The CPFM module workflow is as follows. First, it constructs a feature pyramid, using bilinear interpolation for upsampling and 1×1 convolution to fuse top-down and bottom-up features, followed by element-wise addition, achieving simple and efficient segmentation without any complex refinement modules. Second, it introduces a Pyramid Pooling Module (PPM) at the top layer to aggregate global context information, better performing multi-scale context aggregation and global information collection to assist in remote sensing image scene parsing. Finally, it applies MSKAM (detailed in Section Multi-scale Selective Kernel Attention Module (MSKAM)) for the adaptive enhancement of features at various levels.

Specifically, after the multi-scale features (β_0, β_1, β_2, β_3) generated by the feature encoder are processed through CPF, four different-sized feature maps (β'_0, β'_1, β'_2, β'_3) are obtained, as shown in Formula (2):

$$\begin{cases} \beta'_3 = LSK(PPM(\beta_3)) \\ \beta'_2 = ECA3(\beta_2) + f_1(f_2(\beta'_3)) \\ \beta'_1 = ECA2(\beta_1) + f_1(f_2(\beta'_2) + f_2(\beta'_3)) \\ \beta'_0 = ECA1(\beta_0) + f_1\left(f_2(\beta'_1) + f_2(\beta'_2) + f_2(\beta'_3)\right) \end{cases} \tag{2}$$

where $f_1(\cdot)$ represents a series of standard feature extraction and transformation operations performed in the order of 1×1 convolution, batch normalization, and ReLU activation function. $f_2(\cdot)$ represents upsampling operations using bilinear interpolation, aimed at adjusting the size of high-level features so that the processed features can be aligned and fused with the target feature layer. $ECAi(\cdot)$ and $LSK(\cdot)$ represent feature maps after ECA and LSK in MSKAM (detailed in Section Multi-scale Selective Kernel Attention Module (MSKAM)), and $PPM(\cdot)$ represents feature maps after the Pyramid Pooling Module, which obtains hierarchical information from global to local through different scale pooling. The final output is a composite feature map that fuses multiple scales, thereby achieving the purpose of considering both global mangrove semantic information and local mangrove detail information.

Multi-Scale Selective Kernel Attention Module (MSKAM)

The complex and variable nature of mangrove backgrounds makes structural and semantic information in remote sensing images less distinct. To enhance the representation capability for multi-shaped targets and effectively process scale variations and shape differences of targets in remote sensing images, after using the aforementioned CPFM, it is necessary to consider the localization of key channels [40] during pyramid fusion. For this purpose, MSKAM was designed, consisting of 3 sets of Efficient Channel Attention (ECA) modules and 1 set of Large Selective Kernel (LSK) modules, which can adaptively assess channel value and adjust the spatial receptive field of deep targets to more comprehensively capture channel and spatial information of mangroves. ECA focuses more on important features at lower levels and enhances shallow features' representation capability for targets. Therefore, this paper added 3 ECA modules to feature pyramid modules at different scales to enhance the importance of useful channels in the first three layers of pyramid fusion modules, achieving quick and accurate localization of useful channels. The specific structure of ECA is shown in Figure 8.

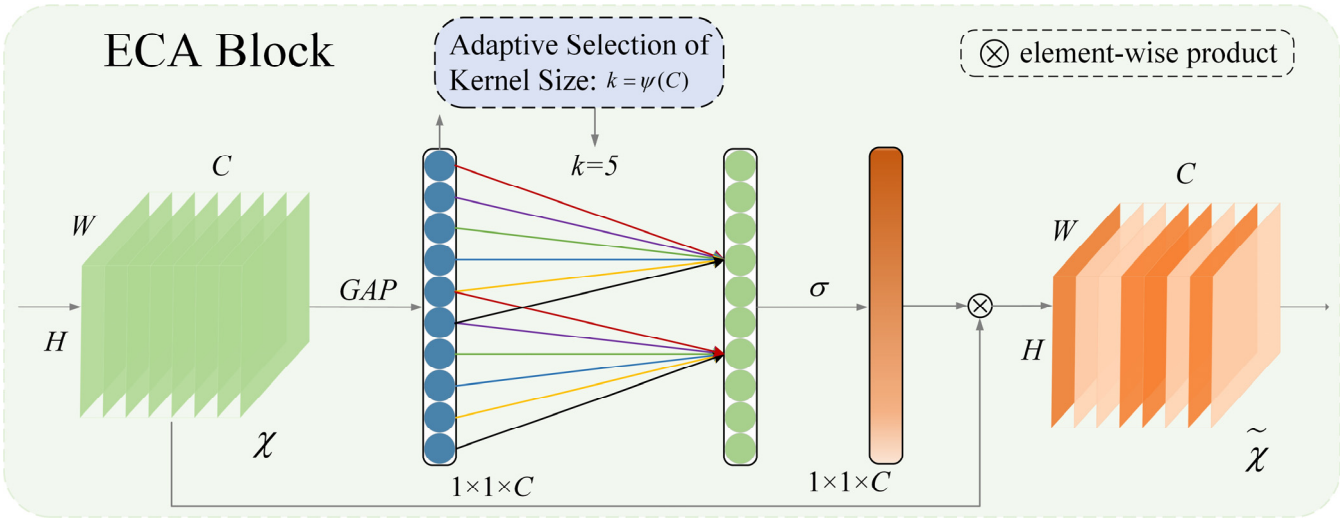


Figure 8. ECA Channel Attention Module Structure.

Specifically, for an input feature map χ, with dimensions C × H × W (channels × height × width), the ECA attention mechanism processing flow is as follows. First, through Global Average Pooling (GAP), the features are compressed into a 1 × 1 × C vector, achieving global spatial information aggregation for each channel. Then, the module adaptively determines the size of the one-dimensional convolution kernel k = ψ(C) (5 in this paper) based on the number of channels C, used to control the range of local cross-channel interactions. Channel-wise local information interaction is achieved through one-dimensional convolution operations, a design that obtains channel dependencies without dimensionality reduction. Finally, channel attention weights are generated and interacted with the original feature map, with the calculation process shown in Formula (3):

$$\widetilde{\chi} = \chi \cdot \sigma(\mathrm{C1D}_k(\mathrm{GAP}(\chi))) \tag{3}$$

where χ is the input feature map, $\sigma(\cdot)$ is the sigmoid activation function, $\mathrm{C1D}_k(\cdot)$ represents a one-dimensional convolution layer with kernel size K, and $\mathrm{GAP}(\cdot)$ is the Global Average Pooling layer. This method achieves adaptive adjustment of different channel importance, not only with high computational efficiency but also effectively capturing channel dependencies, enhancing the network's response capability to key mangrove features.

During the bottom-up downsampling process in the CPFM module, semantic features of small-sized mangrove targets are rapidly lost through layer-by-layer downsampling, resulting in less distinct semantic information in top-layer features. To address this issue, this paper introduces the LSK (Large Selective Kernel) module at the top layer of the feature pyramid. This module works collaboratively with PPM through a spatial selection mechanism, dynamically adjusting feature receptive fields and weighted merging of features processed by deep kernel sequences, with kernel weights determined adaptively based on input. This design enables the model to flexibly utilize convolution kernels of different sizes, specifically adjust the spatial receptive field of deep targets, and further aggregate top-layer context information through cascade fusion with PPM, significantly enhancing the network's ability to extract high-level semantic features of mangroves in deep layers. The specific structure is shown in Figure 9.

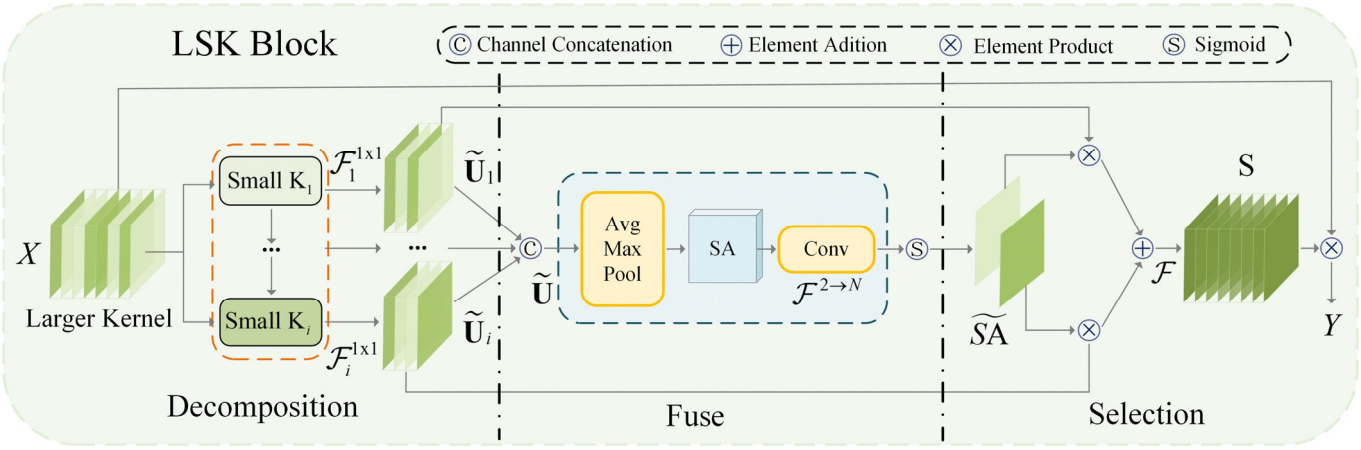


Figure 9. Structure Diagram of LSK Module.

Specifically, the design of the LSK module encompasses three core phases: decomposition, fusion, and selection. During the decomposition phase, large-receptive-field convolution operations are decomposed into multiple small-sized convolution kernels with different dilation rates. This design maintains the original overall receptive field range while providing the network with multiple receptive field options at different scales. The kernel size k, dilation rate d, and receptive field RF expansion of the i-th convolution are defined as shown in Equations (4) and (5):

$$k_{i-1} \leq k_i, d_1 = 1, d_{i-1} < d_i \leq RF_{i-1} \tag{4}$$

$$RF_1 = k_1, RF_i = d_i(k_i - 1) + RF_{i-1} \tag{5}$$

By adjusting the combination of convolution kernel size and dilation rate, different sizes of receptive fields can be flexibly obtained. To avoid sampling gaps in the feature maps caused by dilated convolutions, a reasonable upper limit is set for the dilation rate. As shown in Table 2, a large-sized convolution kernel can be decomposed into 2 to 3 smaller convolution operations, and this decomposition method can achieve theoretical receptive fields of 15 and 23, respectively.

Table 2. Examples of Receptive Field Decomposition.

Receptive Field (RF)	Convolution Kernel and Dilation Rate (K,D) Sequence
15	(15,1) (3,1) → (5,3) (3,1) → (7,2)
23	(23,1) (5,1) → (7,3) (3,1) → (5,2) → (5,3)

Decomposed convolutions can generate multiple feature maps with different receptive field sizes, providing greater flexibility for subsequent kernel selection. Compared to directly using large-sized convolution kernels, the sequential decomposition method can significantly reduce computational complexity. While maintaining the same theoretical receptive field, the decomposition strategy can not only reduce the model's parameter count but also effectively lower the network's computational overhead. The specific calculation process is shown in Equation (6).

$$U_0 = X, U_{i+1} = \mathcal{F}_i(U_i) \tag{6}$$

where $\mathcal{F}_i(\cdot)$ represents the i-th convolution operation with dilation rate, and U_i denotes the i-th input feature map. Additionally, each small-sized convolution kernel is followed by a 1×1 convolution layer FF to reduce the number of channels, thereby decreasing subsequent computational and parameter costs. The calculation process is shown in Equation (7).

$$U_0 = X, U_{i+1} = \mathcal{F}_i(U_i) \tag{7}$$

where $\widetilde{U_i}$ represents the feature map after processing through the 1×1 convolution layer $\mathcal{F}_i^{1\times1}(\cdot)$. In the fusion phase, the contextual features obtained from different receptive field convolution kernels in the decomposition phase are first concatenated along the channel dimension. The calculation process is shown in Equation (8).

$$\widetilde{U} = Concat(\widetilde{U_1}, \ldots, \widetilde{U_i}) \tag{8}$$

where $\widetilde{U}$ represents the feature map after concatenating $\widetilde{U_i}$. Then, channel-based average pooling and max pooling are applied to $\widetilde{U}$ to effectively extract spatial relationships between different positions in the feature map. The calculation process is shown in Equation (9).

$$SA_{avg} = P_{avg}(\widetilde{U}), SA_{\max} = P_{\max}(\widetilde{U}) \tag{9}$$

where $P_{avg}(\cdot)$ and $P_{max}(\cdot)$ represent channel-based average pooling and max pooling, respectively, and SA_{avg} and SA_{max} represent the spatial relationship feature maps after average pooling and max pooling. To achieve information interaction between different spatial relationships, different spatial relationship feature maps are concatenated, and the convolution layer $\mathcal{F}^{2\to N}$ is used to transform the aggregated feature map (2 channels) into two $N \times N$ spatial attention maps (where N is the width of the feature map, assuming input feature dimensions are consistent). The calculation process is shown in Equation (10).

$$\widehat{SA} = \mathcal{F}^{2\to N}([SA_{avg}, SA_{max}]) \tag{10}$$

where $\widehat{SA}$ represents the set of spatial attention maps. For each spatial attention map $\widehat{SA_i}$, a sigmoid activation function is used to obtain the spatial selection mask $\widetilde{SA_i}$ for each participating decomposed convolution kernel. The calculation process is shown in Equation (11).

$$\widetilde{SA_i} = \sigma(\widehat{SA_i}) \tag{11}$$

where $\sigma(\cdot)$ represents the sigmoid function. In the selection phase, the feature maps in the decomposed convolution kernel sequence are first weighted according to the spatial selection masks. Subsequently, the weighted feature maps are fused through convolution operation $\mathcal{F}_i(\cdot)$ to generate the attention feature map S. The calculation process is shown in Equation (12).

$$S = \mathcal{F}\left(\sum_{i=1}^{N}\left(\widetilde{SA_i} \cdot \widetilde{U_i}\right)\right) \tag{12}$$

Finally, the input feature map X is element-wise-multiplied with the attention feature map S to obtain the output feature map Y. The calculation process is shown in Equation (13).

$$Y = X \cdot S \tag{13}$$

Based on this design, the LSK (Large-Scale Kernel) module can adaptively adjust the receptive field range according to the target's scale, enabling the network to more precisely extract contextual information needed for mangrove targets of different scales. This adaptive mechanism not only effectively addresses the problem of insufficient contextual

information but also reduces interference from redundant background information, thereby improving the accuracy of feature extraction.

Through the above analysis, it can be seen that the semantic decoder proposed in this paper constructs an efficient multi-scale feature fusion and enhancement framework through the collaborative work of two core modules: CPFM and MSKAM. CPFM achieves a seamless fusion of multi-level features based on a cascaded feature pyramid structure and effectively captures global contextual information through the PPM module. MSKAM innovatively integrates the ECA channel attention mechanism and LSK Large Selective Kernel module, achieving precise channel selection and adaptive receptive field adjustment at different levels of the feature pyramid. This multi-level feature processing mechanism fully aligns with the characteristics of mangrove remote sensing images, effectively addressing core challenges in semantic segmentation such as multi-scale feature fusion, key information enhancement, contextual information acquisition, and computational efficiency optimization, significantly improving the performance of mangrove remote sensing image semantic segmentation.

2.5. Loss Function

Due to the high-resolution characteristics of remote sensing images, the number of background pixels far exceeds the number of mangrove pixels, leading to a severe sample distribution imbalance problem. To address this challenge, this chapter designs a hybrid loss function CL_{loss} to optimize network performance. This loss function combines cross-entropy loss and Lovasz-Softmax loss, which not only effectively alleviates the class imbalance problem but also specifically enhances the optimization of segmentation boundary regions, thereby improving the overall segmentation accuracy of the model.

The cross-entropy loss function is used to measure the difference between predicted probabilities and true labels in classification tasks, thereby driving the model to learn correct pixel classification. Its definition is shown in Equation (14):

$$Loss_{CE} = -\frac{1}{N}\sum_{I=1}^{N} y_i * log\hat{y}_i \tag{14}$$

where N represents the total number of pixels, y_i is the true label of pixel i, $\hat{y}_i$ is the model's predicted probability output for pixel i, and log represents the natural logarithm.

The Lovasz-Softmax loss is constructed based on the Jaccard index (IoU) [41]. For category c, the Jaccard index is defined as shown in Equation (15):

$$J_c(y, \widetilde{y}) = \frac{|y = c \cap \widetilde{y} = c|}{|y = c \cup \widetilde{y} = c|} \tag{15}$$

where $y = c$ represents the set of pixels belonging to Category c in the true labels, $\widetilde{y} = c$ represents the set of pixels belonging to Category C in the predicted results, and $|\cdot|$ represents the cardinality (number of elements) of the set. The Jaccard loss is defined as shown in Equation (16):

$$\Delta J_c(y, \widetilde{y}) = 1 - J_c(y, \widetilde{y}) \tag{16}$$

To more precisely characterize misclassified pixels, the set of misclassified pixels for Class c can be defined as shown in Equation (17):

$$M_c(y, \widetilde{y}) = y = c, \widetilde{y} \neq c \cup y \neq c, \widetilde{y} = c \tag{17}$$

Based on the set of misclassified pixels, the Jaccard loss can be rewritten as shown in Equation (18):

$$\Delta J_c(y, \widetilde{y}) = \frac{|M_c|}{|y = c \cup M_c|} \tag{18}$$

Since the Jaccard loss is non-differentiable, it needs to be smoothed for use in deep learning frameworks. Here, the Lovasz extension is introduced, which is a method to extend discrete set functions to continuous domains. For a submodular function F, its Lovasz extension is defined as shown in Equation (19):

$$\overline{F}(m) = \sum_{p=1}^{P} F(\pi_1, \ldots, \pi_p)[\pi_p(m) - \pi_{p-1}(m)] \tag{19}$$

where m is the error vector, P is the vector dimension, π is the permutation that sorts the components of m in descending order, and $\pi_0(m) = 0$. The calculation method for the error vector m is shown in Equation (20):

$$m_i = \begin{cases} 1 - f_c(x_i) & \text{if } y_i = c \\ f_c(x_i) & \text{otherwise} \end{cases} \tag{20}$$

where $f_c(x_i)$ is the model's predicted probability that pixel i belongs to category c. The final Lovasz-Softmax loss is defined as shown in Equation (21):

$$Loss_{Lovasz} = \frac{1}{|C|} \sum_{c \in C} \overline{\Delta J_c}(m(c)) \tag{21}$$

where C is the set of categories, which is 2 in this study. Finally, combining the above two loss functions yields the final hybrid loss function, as shown in Equation (22):

$$Loss_{CL} = \alpha Loss_{CE} + (1 - \alpha) Loss_{Lovasz} \tag{22}$$

where α is a weight parameter used to balance the contribution of the two loss functions. Through experimental validation, when α = 0.6, the model achieves optimal performance. The selection of weights takes into account both the degree of sample imbalance and boundary precision requirements. During the training process, the cross-entropy loss dominates the early optimization process, ensuring basic classification accuracy. Meanwhile, the Lovasz-Softmax loss takes a dominant role in the later stages, further optimizing segmentation boundaries and addressing class imbalance issues. Experimental results demonstrate that this hybrid loss function strategy can effectively improve model performance in mangrove remote sensing image segmentation tasks.

2.6. Evaluation Metrics

This paper uses Pixel Accuracy (PA), mean Pixel Accuracy (mPA), mean Intersection over Union (mIoU), number of parameters, and Floating-Point Operations (FLOPs) to evaluate the segmentation performance of different models. Pixel Accuracy (PA) represents the proportion of correctly segmented pixels. It can be calculated by dividing the number of correctly segmented pixels by the total number of pixels, and the result is used to evaluate

the overall accuracy of the model. For $k + 1$ classes, including k foreground classes and 1 background class, PA is defined as shown in Formula (23):

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}}, PA \in (0, 1), \tag{23}$$

where p_{ii} denotes the number of correctly classified pixels and p_{ij} is the number of pixels in Class i that are predicted to be in Class j. PA is often used to evaluate the overall segmentation effect, and the closer PA is to 1, the better the model performance is.

In mangrove segmentation, the PA metric particularly focuses on the model's ability to capture the overall outline of mangrove areas. A higher PA value indicates that the model can accurately identify the main areas of mangroves, which is significant for mangrove resource surveys and area statistics. However, considering the uneven distribution and boundary complexity of mangroves, relying solely on the PA metric might overlook the segmentation quality of small-area mangroves or edge regions. Therefore, mean Pixel Accuracy (mPA) is introduced to represent the average of all PAs, with the mPA calculation formula defined as shown in Formula (24):

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{24}$$

mPA better reflects the model's performance in handling mangroves with different densities and distribution characteristics by averaging the accuracy rates of various categories. This is particularly important for evaluating the model's segmentation performance in complex environments (such as areas where mangroves are mixed with other vegetation, and areas with varying degrees of density).

To more rigorously evaluate segmentation accuracy, especially performance in mangrove boundary regions, this paper also introduces mean Intersection over Union (mIoU) as an evaluation metric for the model, with mIoU defined as shown in Formula (25):

$$mIoU = \frac{\sum_{i=0}^{K} \frac{|A_i \cap B_i|}{|A_i \cup B_i|}}{K+1}, \tag{25}$$

where A and B denote labeling and prediction results, respectively. The mIoU metric is more sensitive to the accuracy of segmentation boundaries and can better evaluate the model's segmentation precision at interfaces between mangroves and other features (such as water bodies and other vegetation). This has significant practical implications for the dynamic monitoring and boundary change analysis of mangrove ecosystems.

Parameters and Floating-Point Operations (FLOPs) are two core metrics for evaluating the complexity of deep learning models. FLOPs quantifies the number of floating-point calculations required for a single forward inference of the model, typically measured in GFLOPs (billion Floating-Point Operations). The FLOP value directly reflects the computational cost of the model, with higher FLOPs indicating greater computational load and longer inference time. The number of parameters characterizes the storage scale of the model. In mangrove species recognition tasks, the comprehensive evaluation of these two metrics—Parameters and FLOPs—provides important guidance for actual model deployment. In mangrove semantic segmentation tasks, balancing these metrics helps achieve an equilibrium between model performance and resource consumption, ensuring both

recognition accuracy and efficient model operation under limited resources. This quantitative evaluation method provides a reliable basis for the engineering implementation of mangrove species recognition systems.

3. Result and Discussion

3.1. Network Training Parameters

The experimental environment configuration for this paper is as follows. For hardware, an Intel Core i7-12700H processor is used, equipped with 32 GB DDR4-3200 MHz memory, and a single NVIDIA RTX 3090 graphics card (24 GB VRAM) (NVIDIA, Santa Clara, CA, USA) for deep learning computation acceleration. The software environment is based on the Ubuntu 18.4 operating system, using Python 3.8 as the development language and the PyTorch 1.9.0 deep learning framework (with CUDA 11.1.1) for related experiments. In deep learning training, the batch size is set to 4, the optimizer is "AdamW", the weight decay is 0.05, the initial learning rate is 1×10^{-4}, the learning rate strategy uses LinearLR and PolyLR adjustment, and the number of training iterations is 160,000.

3.2. Comparison Experiment Results and Analysis

To evaluate the performance of MFA-UperNet in mangrove segmentation, comparative experiments were conducted between the proposed MFA-UperNet and other models on the dataset constructed in this paper. The dataset consists of 1798 training images, 514 validation images, and 256 test images. The values shown in the table represent the average of five independent tests performed on the validation and test sets. Independent sample *t*-test analysis indicates that the performance differences between models are statistically significant ($p < 0.05$). This paper uses mIoU, mPA, Params, and FLOPs as evaluation metrics. The segmentation results of different models for mangroves are shown in Table 3.

Table 3. A performance comparison of different segmentation models on the MSSDBG dataset.

Model	Backbone	Mangrove		Background		mIoU (%)	mPA (%)	Params (M)	FLOPs (G)
		IoU (%)	PA (%)	IoU (%)	PA (%)				
DANet [24]	ResNet50	87.86	93.03	92.80	96.57	90.33	94.80	47.5	216.1
PSANet [25]	ResNet50	89.56	94.43	93.79	96.83	91.67	95.63	56.8	205.1
PSPNet [23]	ResNet50	90.56	95.03	94.39	97.12	92.47	96.07	46.6	**182.6**
DeepLabv3+ [27]	ResNet50	89.61	94.44	93.82	96.86	91.72	95.65	**41.2**	183.3
Swin-UperNet [32]	Swin Transformer tiny	92.05	95.80	95.29	97.62	93.67	96.71	59.4	242.7
K-Net [28]	Swin Transformer tiny	88.31	93.36	93.07	96.67	90.69	95.02	250.1	441.3
PIDNet [29]	PIDNet-S	92.41	95.90	95.52	97.80	93.97	96.85	72.6	258.6
Ours	ConvNeXt-tiny	**93.13**	**96.21**	**95.95**	**98. 07**	**94.54**	**97.14**	59.2	234.2

Note: Bold indicates best data.

As shown in Table 3, compared to methods such as DANet, DeepLab v3+, and Swin-UperNet, the proposed MFA-UperNet method achieves more precise recognition results, particularly excelling in mangrove boundary localization and detail preservation under complex backgrounds. It achieves optimal results across all evaluation metrics, including IoU, PA, mIoU, and mPA, primarily due to the innovative design of three core modules in the network: the feature encoder, Auxiliary Edge Neck (AEN), and the Cascade Pyramid Fusion Module (CPFM) along with Multi-scale Selective Kernel Attention Module (MSKAM) in the semantic decoder. The feature encoder employs the ConvNeXt architecture, whose powerful feature extraction capability enables the network to effectively capture multi-level information from local textures to global structures, providing a solid feature foundation for subsequent processing. The Auxiliary Edge Neck significantly improves segmentation accuracy in edge regions through lightweight fully convolutional network design and the introduction of hybrid loss functions, providing an effective solution to the problem of

blurred boundaries between mangroves and backgrounds. It not only provides additional gradient information for the backbone network but also offers new insights into edge enhancement theory in remote sensing image segmentation. The CPFM and MSKAM in the semantic decoder represent an innovative feature fusion and enhancement solution. CPFM integrates the Feature Pyramid Network (FPN) and Pyramid Pooling Module (PPM), employing a bidirectional feature fusion strategy from top-down and bottom-up, achieving an efficient fusion of multi-scale features. Meanwhile, MSKAM combines ECA channel attention and LSK Large Selective Kernel module to achieve precise channel selection and adaptive receptive field adjustment at different levels of the feature pyramid. This design not only effectively addresses the challenges of complex backgrounds and variable scales in mangrove remote sensing images while maintaining computational efficiency but also provides an innovative solution to the problem of multi-scale feature representation in remote sensing image analysis.

To more intuitively understand the differences between the proposed method and other methods, visualization results are presented in Figure 10, where red regions represent identified mangroves, blue regions represent non-mangrove areas, and in the second column, black and gray regions correspond to background and Ground Truth, respectively. Through comparative analysis, the method in this chapter demonstrates clear advantages in several key aspects. The generated segmentation boundaries more closely align with the actual mangrove distribution range. (Figure 10a,b) In handling complex background areas, this chapter's method shows stronger robustness, effectively avoiding the common category confusion problems seen in other models, which is particularly evident in the results shown in the third row. (Figure 10c) Notably, in the completeness test shown in the fourth row, only this chapter's method successfully achieved complete coverage of the target area, while other models showed varying degrees of omission. (Figure 10d) Considering all metrics comprehensively, the method in this chapter demonstrates significant advantages in both segmentation accuracy and boundary precision.

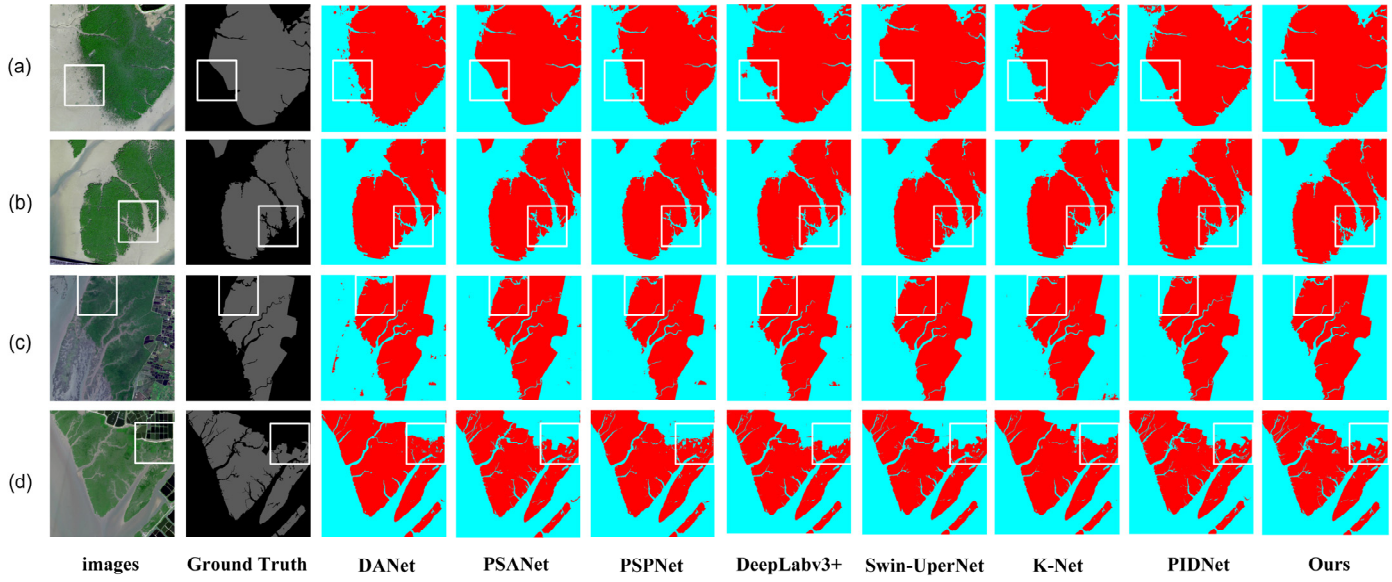


Figure 10. A plot of the segmentation results of different models on the MSSDBG dataset. (White boxes are used to highlight the differences in the results under different methods.)

However, through an in-depth analysis of the segmentation results, we also found that MFA-UperNet still has room for improvement in certain challenging scenarios. In areas with dense mangrove distribution, due to spectral confusion caused by vegetation canopy overlap in remote sensing images, the model's delineation of mangrove boundaries is not

precise enough. In areas with complex lighting conditions, especially in remote sensing images with strong shadows or uneven illumination, the segmentation performance shows slight degradation. In transition zones between mangroves and other coastal vegetation, due to the similarity of the spectral features in remote sensing images, there exists some uncertainty in the model's discrimination between mangrove and non-mangrove areas. Furthermore, in coastal areas affected by tides, spectral variations caused by water coverage also pose challenges to accurate mangrove identification. From a computational efficiency perspective, the Parameters and FLOPs of our proposed model have increased compared to lightweight networks such as DANet and PSANet. This suggests that one important direction for future research is to further reduce the model's parameter count and computational complexity while maintaining or improving current segmentation accuracy through optimizing network structure design, exploring more efficient feature extraction and fusion strategies, and adopting model compression techniques, thereby achieving a better balance between efficiency and performance. This has significant practical implications for model deployment in real-world applications.

As shown in Figure 11, the curves illustrating changes in Loss values and accuracy are presented. These graphs comprehensively reflect the training process and performance of the model.

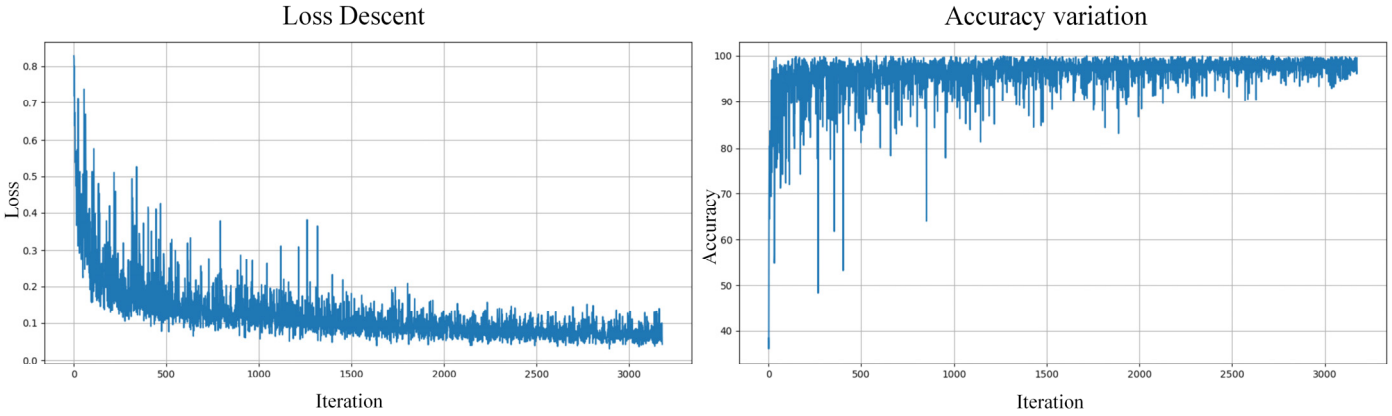


Figure 11. Plot of Loss and Accuracy.

3.3. Ablation Experiment Analysis

In this section, we conduct ablation experiments and performance analysis on the key modules in MFANet to validate their effectiveness. First, we discuss the performance differences between the ConvNeXt network and the Swin Transformer as encoders for the MFANet. Then, we evaluate the effectiveness of the CPF module, MECA module, and AEN module in MFANet while also assessing the impact of different loss functions on the mangrove segmentation results.

The encoder used in the MFANet is the ConvNeXt network. Figure 12 presents a comparison of the improvement effects between the ConvNeXt network and the Swin Transformer.

Figure 12 illustrates the improvement effects of the five components in the ConvNeXt network and provides a direct comparison with the Swin Transformer. Compared to Swin-T, ConvNeXt-T achieves a 0.7% improvement in accuracy on the ImageNet dataset under the same GFlOPs. Moreover, it outperforms the Swin Transformer on the COCO detection dataset and the ADE20K segmentation dataset. Therefore, we employ ConvNeXt-T for the semantic segmentation task of mangrove remote sensing images.

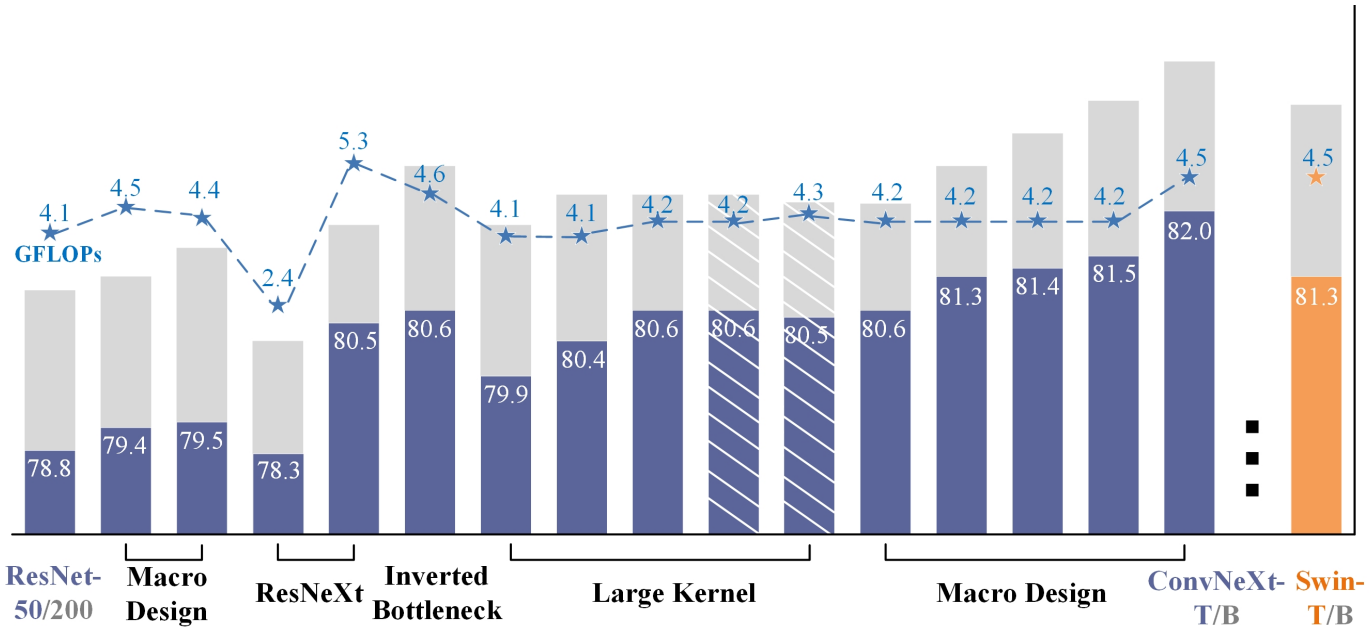


Figure 12. ConvNeXt Improvement and Effectiveness Comparison Chart.

To validate the effectiveness of the backbone network in the MFA-UperNet model, experiments were performed using ConvNeXt-small, ConvNeXt-tiny, and ConvNeXt-base as the backbone networks. This was performed to compare the impact of different backbone network levels on the segmentation results, which aimed to achieve a better balance between segmentation accuracy and inference speed. Table 4 presents the experimental results and the number of parameters for each level of the backbone networks.

Table 4. The experimental results of different backbone networks on the MRSDBG dataset.

Backbone	Mangrove		Background		mIoU (%)	mPA (%)	Params (M)	Flops (G)
	IoU (%)	PA (%)	IoU (%)	PA (%)				
ConvNeXt-T	**93.03**	**96.14**	95.52	98.02	94.27	**97.08**	**59.24**	**234**
ConvNeXt-S	92.90	96.05	**95.82**	**98.03**	**94.36**	97.04	80.88	256
ConvNeXt-B	92.59	95.95	95.63	97.88	94.11	96.92	121.02	292

Note: Bold indicates best data.

The experimental results show that ConvNeXt-T achieved relatively optimal segmentation results on the MSSDBG dataset, and it was worth noting that ConvNeXt-T achieves the lowest parameter count of 59.24 M, which was 21.64 M and 61.82 M less than ConvNeXt-S and ConvNeXt-B, respectively. Although ConvNeXt-T's FLOPs was 234 G, which was lower than ConvNeXt-S and ConvNeXt-B, the combination of its lowest number of parameters and highest accuracy proved that ConvNeXt-T better processed the semantic features of the mangrove forests in the MSSDBG, so ConvNeXt-tiny was finally chosen as the encoder backbone network for the model.

To comprehensively evaluate the network architecture proposed in this paper, we conducted systematic ablation experiments on key modules using our constructed dataset and performed an in-depth analysis of the results. This section focuses on evaluating the effectiveness of several core modules proposed in this paper, including CPFM, CLloss, AEN, and MSKAM modules. Using FPN as the baseline, we verified their effectiveness through a progressive addition of each module. The results are shown in Table 5.

Table 5. Ablation study results of the model under different settings.

Method	mIoU (%)	mPA (%)
Baseline	88.31	93.36
Baseline + CPFM	92.55	95.88
Baseline + CPFM + CLloss	92.89	96.07
Baseline + CPFM + CLloss + AEN	93.07	96.11
Baseline + CPFM + CLloss + AEN + MSKAM	**94.54**	**97.14**

Note: Bold indicates best data.

As shown in the results from the above table, after incorporating CPFM, mIoU increased by 4.24% and mPA by 2.52%, indicating that CPFM effectively integrated multi-scale feature information of mangrove wetlands through the cascaded fusion of Feature Pyramid Network and Pyramid Pooling Module. Detail textures preserved in lower-level features improved localization accuracy, while semantic information from higher-level features enhanced category recognition capability, thereby significantly improving the overall segmentation performance of mangrove wetlands. After adding CLloss, mIoU increased by 0.34% and mPA by 0.19%, demonstrating that the introduction of CLloss hybrid loss function effectively addressed the sample distribution imbalance issue. By combining cross-entropy loss and Lovasz-Softmax loss, it maintained overall classification accuracy while optimizing boundary region segmentation precision, showing particular advantages in handling cases with significant quantity disparities between mangrove and background categories. Following the addition of AEN, mIoU increased by 0.18% and mPA by 0.04%, showing that the introduction of the AEN auxiliary edge segmentation branch enhanced the model's ability to recognize mangrove wetland boundary regions. Through specialized edge learning and auxiliary supervision of the third-layer features, it effectively improved the blurring issues at boundaries between mangroves and surrounding vegetation, enhancing segmentation refinement. After incorporating MSKAM, mIoU increased by 1.47% and mPA by 1.03%, indicating that MSKAM significantly enhanced the model's ability to recognize mangrove wetlands in complex backgrounds. Through the adaptive enhancement of shallow feature channels by the ECA module and the dynamic adjustment of deep feature receptive fields by the LSK module, it achieved more precise feature extraction and spatial information capture, effectively addressing issues such as scale variations and shape differences of mangrove wetland targets in remote sensing images. Therefore, through these ablation experiments, this paper validated the effectiveness of the proposed modules.

Meanwhile, to more intuitively demonstrate the effectiveness of the modules proposed in this paper, we visualized the results of both individual modules and the combination of multiple modules. Through this approach, we can clearly showcase each module's contribution to the model's performance improvement, as shown in Figure 13.

In the baseline method, the model showed notable limitations in mangrove identification, exhibiting misclassification phenomena and obvious blurring effects in boundary regions. After introducing CPFM, due to its effective fusion of multi-scale feature information, the model's overall segmentation performance improved significantly, with notably reduced misclassification areas, although some inaccuracies in boundary processing remained. Following the addition of CLloss, through the optimization of mixed loss functions, the model demonstrated better performance in handling sample imbalance issues and improved boundary region segmentation accuracy, though the capture of detailed features in complex scenarios still needed improvement. With the further introduction of the AEN module, the model's ability to recognize edge regions was enhanced with clearer boundary contours, but some difficulties remained in processing small-target mangroves. Finally, after incorporating MSKAM, the model's performance reached its optimal state. Through the synergistic effect of ECA and LSK modules, feature extraction became more precise and

spatial information acquisition more comprehensive, enabling the model to better handle targets of different scales, particularly achieving significant improvements in small target recognition and boundary detail processing in complex backgrounds. The segmentation results became more refined, boundaries clearer, and misclassification phenomena substantially reduced, fully validating the effectiveness and synergy of each module proposed in this paper.

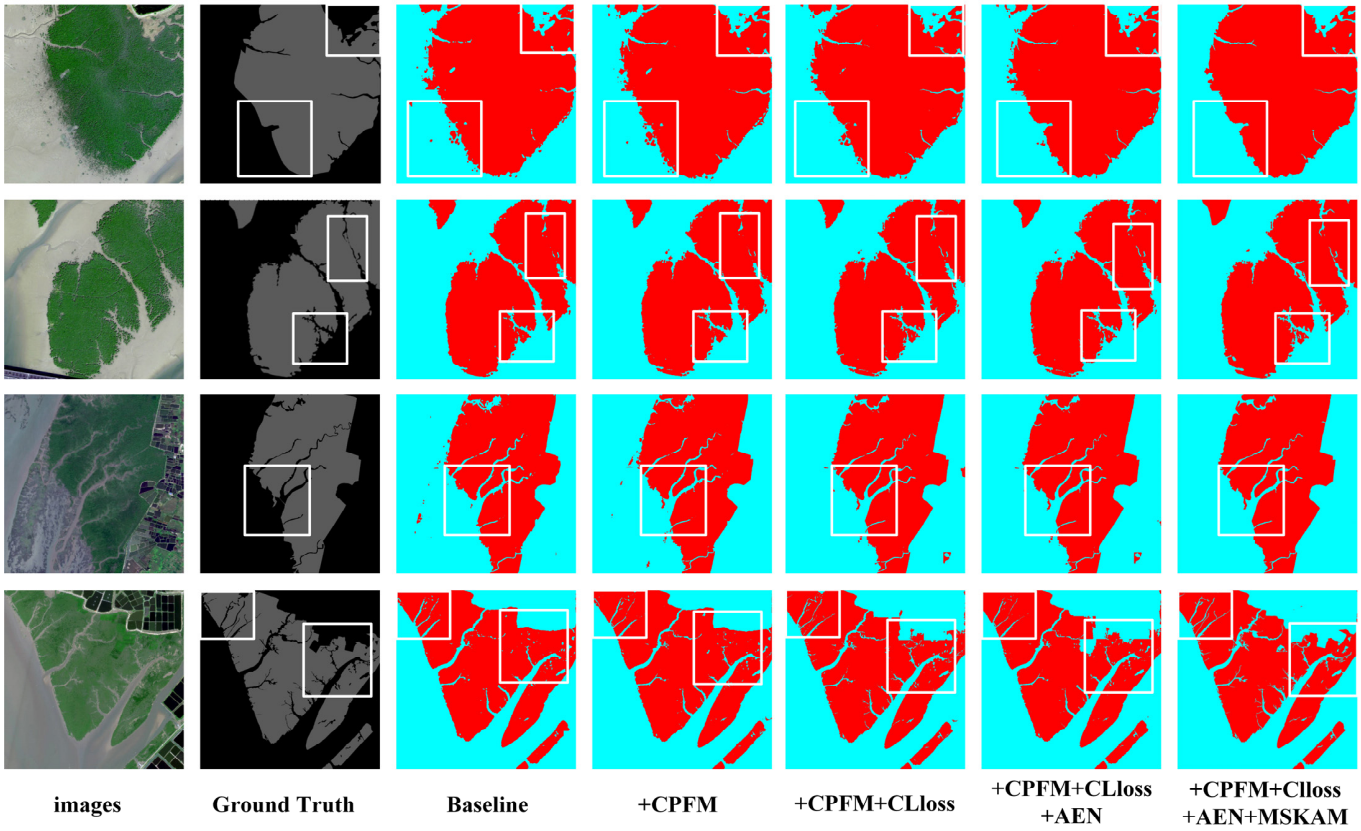


Figure 13. The segmentation result graph of the different modules superimposed on the MRSDBG dataset. (White boxes are used to highlight the differences in the results under different methods.)

3.4. Overall Results and Discussion

Based on China's new generation of high spatial resolution Earth observation satellites, this paper constructs the MRSDBG dataset and quantitatively analyzes the impact of different improvement strategies of MFA-UperNet on mangrove segmentation accuracy. For feature extraction, MFA-UperNet uses ConvNeXt, a pure convolutional network, as the encoder. Compared to Swin Transformer, ConvNeXt achieves a 0.7% accuracy improvement on the ImageNet dataset and demonstrates stronger feature extraction capabilities for high-resolution remote sensing images, performing better than different backbone networks like ResNet and PIDNet. In terms of multi-scale and feature pyramid improvements, the CPFM and MSKAM modules designed through the Feature Pyramid Network can effectively address the complex and variable background of mangroves and large-scale variations, with the addition of CPFM increasing mIoU and mPA by 4.64% and 0.59%, respectively, and the MSKAM improving them by 1.47% and 1.03%. For edge segmentation optimization, the AEN module enhances the model's segmentation performance through fully convolutional neural networks and hybrid CLloss function, with the addition of AEN module and CLloss improving mIoU and mPA by 0.52% and 0.23%, respectively. In comprehensive comparison experiments with existing methods, our MFA-UperNet model achieved improvements of 0.87% and 0.43% in mIoU and mPA, respectively, compared to Swin-UperNet; compared

to DeepLabv3+, it achieved improvements of 2.82% and 1.49%. Furthermore, compared to representative methods such as PIDNet, K-Net, and PSPNet, our model also achieved significant performance improvements. These comprehensive experimental results fully demonstrate the excellent performance and effectiveness of MFA-UperNet in mangrove semantic segmentation tasks. Meanwhile, this research has also improved the mangrove monitoring and assessment system in Beihai City, promoting the enhancement of dynamic monitoring capabilities for mangrove ecosystems.

4. Limitations and Future Work

Although MFA-UperNet demonstrates excellent performance in mangrove remote sensing image semantic segmentation tasks, there are still some limitations and areas for improvement. First, the current experimental validation is primarily based on the MSSDBG dataset from specific geographical regions, and the model's generalization performance across different geographical locations, seasons, and growth states of mangroves needs further validation. Second, mangrove remote sensing images are easily affected by environmental factors such as weather conditions, tidal changes, and atmospheric conditions, and the potential impact of these factors on model performance has not been fully explored. Additionally, there is room for improvement in the model's recognition effectiveness for sparsely distributed or small-area mangrove regions. In practical applications, the model's computational complexity and hardware requirements may also become limiting factors. Future work will focus on expanding the geographical and temporal coverage of the dataset, enhancing the model's adaptability to environmental changes, improving small target recognition performance, and exploring unsupervised methods based on multispectral remote sensing images and drone imagery to further enhance the model's performance in practical applications such as mangrove change detection.

5. Conclusions

This paper proposes a Multi-scale Fusion Attention Unified Perceptual Network (MFA-UperNet) for the semantic segmentation of mangrove remote sensing images. The main conclusions are as follows:

(1) The construction of the MSSDBG dataset based on multi-source high-resolution remote sensing images represents a significant contribution to mangrove monitoring research. By incorporating multi-temporal and multi-scale characteristics from multiple satellite sources, this dataset provides essential support for algorithm development and validation in mangrove semantic segmentation tasks.

(2) The proposed MFA-UperNet demonstrates innovative architectural design through its key components: a ConvNeXt-based feature encoder for superior feature extraction, a Cascade Pyramid Fusion Module (CPFM) for effective multi-scale feature handling, a Multi-scale Selective Kernel Attention Module (MSKAM) for adaptive feature enhancement, and an Auxiliary Edge Neck (AEN) for improved boundary accuracy. This comprehensive design effectively addresses the challenges of complex backgrounds and scale variations in mangrove remote sensing images.

(3) The experimental results convincingly validate the effectiveness of MFA-UperNet, achieving outstanding performance with 94.54% mIoU and 97.14% mPA on the MSSDBG dataset. The model significantly outperforms existing methods, including Swin-UperNet, DeepLabv3+, and PIDNet, while ablation studies confirm the substantial contribution of each proposed module to the overall performance improvement.

The practical applications and implications of this research extend beyond technical achievements, demonstrating the feasibility of high-precision mangrove mapping using deep learning approaches. This work not only contributes to improving regional mangrove

monitoring and assessment systems but also provides valuable technical support for mangrove protection and management efforts, advancing the field of remote sensing-based coastal ecosystem monitoring.

Author Contributions: X.W.: Conceptualization, Supervision, Funding acquisition. Y.Z.: Data curation, Methodology, Software, Writing. W.X.: Methodology, Software, Writing—original draft. X.W.: Formal analysis. H.W.: Supervision, Writing—review and editing. J.C.: Methodology. Q.Q.: Formal analysis, Investigation. Q.W.: Investigation. J.Z.: Formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Guangxi Science and Technology Major Project [grant number AA19254016]; the Guangxi Graduate Student Innovation Project [grant number YCSW2021174]; the Beihai City Science and Technology Planning Project [grant number 202082033]; and the Beihai City Science and Technology Planning Project [grant number 202082023].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable suggestions. We also thank Yixuan Wang for her contribution to this work.

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

References

1. Agaton, C.B.; Collera, A. Now or later? optimal timing of mangrove rehabilitation under climate change uncertainty. *For. Ecol. Manag.* **2022**, *503*, 119739. [CrossRef]
2. Aslan, A.; Rahman, A.F.; Robeson, S.M.; Ilman, M. Land-use dynamics associated with mangrove deforestation for aquaculture and the subsequent abandonment of ponds. *Sci. Total Environ.* **2021**, *791*, 148320. [CrossRef] [PubMed]
3. Feng, Z.; Tan, G.; Xia, J.; Shu, C.; Chen, P.; Wu, M.; Wu, X. Dynamics of mangrove forests in Shenzhen Bay in response to natural and anthropogenic factors from 1988 to 2017. *J. Hydrol.* **2020**, *591*, 125271. [CrossRef]
4. Davidson, I.C.; Cott, G.M.; Devaney, J.L.; Simkanin, C. Differential effects of biological invasions on coastal blue carbon: A global review and meta-analysis. *Glob. Chang. Biol.* **2018**, *24*, 5218–5230. [CrossRef] [PubMed]
5. Marx, S.K.; Knight, J.M.; Dwyer, P.G.; Child, D.P.; Hotchkis, M.A.; Zawadzki, A. Examining the response of an eastern Australian mangrove forest to changes in hydro-period over the last century. *Estuar. Coast. Shelf Sci.* **2020**, *241*, 106813. [CrossRef]
6. Xiao, H.; Su, F.; Fu, D.; Yu, H.; Ju, C.; Pan, T.; Kang, L. *10-m Global Mangrove Classification Products of 2018-2020 Based on Big Data*; Science Data Bank: Beijing, China, 2021.
7. Friess, D.A.; Rogers, K.; Lovelock, C.E.; Krauss, K.W.; Shi, S. The state of the world's mangrove forests: Past, present, and future. *Annu. Rev. Environ. Resour.* **2019**, *44*, 1–27. [CrossRef]
8. Wang, L.; Jia, M.; Yin, D.; Tian, J. A review of remote sensing for mangrove forests: 1956–2018. *Remote Sens. Environ.* **2019**, *231*, 111223. [CrossRef]
9. Chen, B.; Xiao, X.; Li, X.; Pan, L.; Doughty, R.; Ma, J.; Giri, C. A mangrove forest map of China in 2015: Analysis of time series Landsat 7/8 and Sentinel-1A imagery in Google Earth Engine cloud computing platform. *ISPRS J. Photogramm. Remote Sens.* **2017**, *131*, 104–120. [CrossRef]
10. Cao, J.; Xu, X.; Zhuo, L.; Liu, K. Investigating mangrove canopy phenology in coastal areas of China using time series Sentinel-1/2 images. *Ecol. Indic.* **2023**, *154*, 110815. [CrossRef]
11. Jiang, Y.; Zhang, L.; Yan, M.; Qi, J.; Fu, T.; Fan, S.; Chen, B. High-resolution mangrove forests classification with machine learning using worldview and uav hyperspectral data. *Remote Sens.* **2021**, *13*, 1529. [CrossRef]
12. Deng, L.; Chen, B.; Yan, M.; Fu, B.; Yang, Z.; Zhang, B.; Zhang, L. Estimation of Species-Scale Canopy Chlorophyll Content in Mangroves from UAV and GF-6 Data. *Forests* **2023**, *14*, 1417. [CrossRef]
13. Lin, Y.C.; Cheng, Y.T.; Zhou, T.; Ravi, R.; Hasheminasab, S.M.; Flatt, J.E.; Troy, C.; Habib, A. Evaluation of UAV LiDAR for mapping coastal environments. *Remote Sens.* **2019**, *11*, 2893. [CrossRef]
14. Freckleton, R.P.; Watkinson, A.R. Large-scale spatial dynamics of plants: Metapopulations, regional ensembles and patchy populations. *J. Ecol.* **2002**, *90*, 419–434. [CrossRef]

15. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4287–4306. [CrossRef]
16. Cheng, Y.; Yan, J.; Zhang, F.; Li, M.; Zhou, N.; Shi, C.; Jin, B.; Zhang, W. Surrogate modeling of pantograph-catenary system interactions. *Mech. Syst. Signal Process.* **2025**, *224*, 112134. [CrossRef]
17. Jin, B.; Gonçalves, N.; Cruz, L.; Medvedev, I.; Yu, Y.; Wang, J. Simulated multimodal deep facial diagnosis. *Expert Syst. Appl.* **2024**, *252*, 123881. [CrossRef]
18. Hu, L.; Li, W.; Xu, B. Monitoring mangrove forest change in China from 1990 to 2015 using Landsat-derived spectral-temporal variability metrics. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *73*, 88–98. [CrossRef]
19. Fu, C.; Song, X.; Xie, Y.; Wang, C.; Luo, J.; Fang, Y.; Qiu, Z. Research on the spatiotemporal evolution of mangrove forests in the Hainan Island from 1991 to 2021 based on SVM and Res-UNet Algorithms. *Remote Sens.* **2022**, *14*, 5554. [CrossRef]
20. Yan, J.; Cheng, Y.; Wang, Q.; Liu, L.; Zhang, W.; Jin, B. Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2827–2842. [CrossRef]
21. Wang, H.; Liu, Z.; Hu, G.; Wang, X.; Han, Z. Offline Meta-Reinforcement Learning for Active Pantograph Control in High-Speed Railways. *IEEE Trans. Ind. Inform.* **2024**, *20*, 10669–10679. [CrossRef]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, 3431–3440. [CrossRef]
23. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. *IEEE Comput. Soc.* **2016**, 2881–2890. [CrossRef]
24. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154. [CrossRef]
25. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283. [CrossRef]
26. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmen-tation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
27. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818. [CrossRef]
28. Zhang, W.; Pang, J.; Chen, K.; Loy, C.C. K-net: Towards unified image segmentation. *Adv. Neural-Form. Process. Syst.* **2021**, *34*, 10326–10338. [CrossRef]
29. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19529–19539. [CrossRef]
30. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434. [CrossRef]
31. Liu, M.; Fu, B.; Xie, S.; He, H.; Lan, F.; Li, Y.; Fan, D. Comparison of multi-source satellite images for classifying marsh vege-tation using DeepLabV3 Plus deep learning algorithm. *Ecol. Indic.* **2021**, *125*, 107562. [CrossRef]
32. Wang, Z.; Li, J.; Tan, Z.; Liu, X.; Li, M. Swin-UperNet: A Semantic Segmentation Model for Mangroves and Spartina alterniflora Loisel Based on UperNet. *Electronics* **2023**, *12*, 1111. [CrossRef]
33. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. *arXiv* **2023**, arXiv:2303.09030. [CrossRef]
34. Wang, X.; Kang, M.; Chen, Y.; Jiang, W.; Wang, M.; Weise, T.; Zhang, C. Adaptive Local Cross-Channel Vector Pooling Attention Module for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1980. [CrossRef]
35. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986. [CrossRef]
36. Zhang, T.; Hu, S.; He, Y.; You, S.; Yang, X.; Gan, Y.; Liu, A. A fine-scale mangrove map of China derived from 2-meter resolution satellite observations and field data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 92. [CrossRef]
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
38. Zhang, M.; Li, Z.; Tian, B.; Zhou, J.; Tang, P. The backscattering characteristics of wetland vegetation and water-level changes detection using multi-mode SAR: A case study. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *45*, 1–13. [CrossRef]
39. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]

40. Ma, A.; Wang, J.; Zhong, Y.; Zheng, Z. FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
41. Berman, M.; Triki, A.R.; Blaschko, M.B. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4413–4421. [CrossRef]