Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

MSFANet: multi-scale fusion attention network for mangrove remote sensing Image segmentation using pattern recognition

Lixiang Fu[1], Jinbiao Chen[2], Zhuoying Wang[3], Tao Zang[1], Huandong Chen[1*], Shulei Wu[1*] and Yuchen Zhao[1]

Abstract

Mangroves are ecosystems that grow in the intertidal areas of coastal zones, playing crucial ecological roles and possessing unique economic and social values. They have garnered significant attention and research interest. Semantic segmentation of mangroves is a fundamental step for further investigations. However, mangrove remote sensing images often have large dimensions, with a substantial portion of the image containing mangrove features. Deep learning convolutional kernels may lead to inadequate receptive fields for accurate mangrove recognition. In mangrove remote sensing images, various challenges arise, including the presence of small and intricate details aside from the mangrove regions, which intensify the segmentation complexity. To address these issues, this paper primarily focuses on two key aspects: first, the exploration of methods to achieve a large receptive field, and second, the fusion of multi-scale information. To this end, we propose the Multi-Scale Fusion Attention Network (MSFANet), which incorporates a multi-scale network structure with a large receptive field for feature fusion. We emphasize preserving spatial information by integrating spatial data across different scales, employing separable convolutions to reduce computational complexity. Additionally, we introduce an Attention Fusion Module (AFM). This module helps mitigate the influence of irrelevant information and enhances segmentation quality. To retain more semantic information, this paper introduces a dual channel approach for information extraction through the deep structure of ResNet. We fuse features using the Feature Fusion Module (FFM) to combine both semantic and spatial information for the final output, further enhancing segmentation accuracy. In this study, a total of 230 images with dimensions of 768 pixels in width and height were selected for this experiment, with 184 images used for training and 46 images for validation. Experimental results demonstrate that our proposed method achieves excellent segmentation results on a small sample dataset of remote-sensing images, with significant practical value. This paper primarily focuses on three key aspects: the generation of mangrove datasets, the preprocessing of mangrove data, and the design and training of models. The primary contribution of this paper lies in the development of an effective approach for multi-scale information fusion and advanced feature preservation, providing a novel solution for mangrove remote sensing image segmentation tasks. The best Mean Intersection over Union (MIoU) achieved on the mangrove dataset is 86%, surpassing other existing models by a significant margin.

*Correspondence:
Huandong Chen
chd@hainnu.edu.cn
Shulei Wu
wsl@hainnu.edu.cn
Full list of author information is available at the end of the article

Introduction

The mangrove forest is an ecosystem that thrives in intertidal zones along coastal areas, characterized by evergreen trees and shrubs. China represents the northernmost extent of global mangrove distribution, with primary habitats in Hainan, Guangdong, Guangxi, Zhejiang, Fujian, as well as Macau, Hong Kong, and Taiwan. Mangrove coastal ecosystems offer various essential values. Studies have demonstrated that mangroves play a crucial role in mitigating the impact of tropical storms and hurricanes by effectively reducing storm surges, minimizing inundated areas, and safeguarding inland wetlands [1, 2]. They also possess some resistance to tsunami incursions [3] and promote sediment deposition, thereby preventing coastal erosion and enhancing community safety in coastal regions [4, 5]. Furthermore, mangroves exhibit significant carbon sequestration capabilities, contributing to the mitigation of global warming [6]. The establishment of mangrove coastal ecosystems is economically advantageous compared to conventional coastal engineering practices [7]. In summary, mangroves serve as effective buffers against the adverse effects of storm surges and other disasters, bolster coastal defenses, and enhance the overall climate environment. They hold profound ecological significance, unique economic value, and have attracted considerable attention and research. In recent times, mangroves have faced disturbances due to human activities, rendering them fragile natural ecosystems. The preservation of mangroves has become an urgent issue requiring resolution.

Accurate identification of mangrove areas through efficient methods is of paramount importance for mangrove preservation. Semantic segmentation, a vital technique, allows for the precise delineation of mangrove areas. However, the intricate textures, variations in illumination, indistinct boundaries between tidal flats and water bodies, the dense vegetation within mangrove regions, and significant alterations in the shape and position of mangroves in remote sensing images introduce formidable challenges to semantic segmentation. Moreover, remote sensing images of mangroves are often extensive, with mangrove sections occupying a substantial portion of the entire image. In deep learning, employing small convolution kernels may result in limited receptive fields for mangroves, adversely affecting recognition rates. Furthermore, apart from the mangrove segments, remote sensing images of mangroves often contain numerous intricate features, further complicating the segmentation task.

To address these issues, this study primarily focuses on two aspects: exploring a large receptive field and fusing multi-scale information. Therefore, we propose a multi-scale fusion attention network (MSFANet), a network designed for efficient feature fusion. In the MSFANet network, a multi-scale large receptive field network structure is employed for multi-feature fusion. This network structure combines spatial information between different scales to maximize the retention of spatial features and utilizes separable convolutions to reduce computational complexity. Furthermore, an Attention Fusion Module (AFM) is introduced to reduce the influence of irrelevant information and enhance the quality of segmentation results. To preserve more semantic information, a dual pathway is additionally incorporated, which utilizes the deep-level structure of ResNet for feature extraction and employs a Feature Fusion Module (FFM) for feature fusion, combining semantic and spatial information for output, further improving segmentation performance. On the mangrove dataset, our method achieves a best mean intersection over union (Miou) of 86%, significantly surpassing other models. Experimental results demonstrate that the proposed approach achieves excellent segmentation results on remote sensing image datasets and holds significant practical value. This study's contributions also include the creation of the mangrove dataset, the processing of mangrove image data, and the design and training of the model. Our work offers a novel solution for mangrove forest remote sensing image segmentation tasks.

The application of the Multi-Scale Fusion Attention Network (MSFANet) in the semantic segmentation of remote sensing images in mangrove ecosystems holds profound significance. This method has shown immense potential in addressing the challenges of mangrove image segmentation, potentially impacting environmental monitoring, protection, and other related applications.Firstly, through the multi-scale feature fusion of MSFANet, we can comprehensively capture surface information in mangrove areas. This is crucial for understanding the structure and evolution of ecosystems, providing scientists and environmental researchers with more accurate data to support ecological studies, habitat monitoring, and climate change impact assessments. Secondly, the efficiency and accuracy of MSFANet are crucial for real-time monitoring and protection of the ecological environment. By enhancing the semantic segmentation accuracy of remote sensing images, this method helps promptly detect potential ecological threats, such as illegal logging, land development, or other destructive activities. This provides a powerful tool for protecting mangrove ecosystems, aiding in the formulation of more effective conservation strategies and management plans. Lastly, the application of

MSFANet substantially contributes to the sustainable management and resource planning of mangroves. By providing high-quality remote sensing image segmentation results, decision-makers can better understand land use patterns and ecosystem dynamics in mangrove areas, thereby supporting the achievement of sustainable development goals.

Therefore, the Multi-Scale Fusion Attention Network is not merely an image processing technique but also a powerful tool supporting the management and protection of mangrove ecosystems. It has the potential to impact various fields such as environmental monitoring, sustainable development, and ecological research.

The overall workflow of this specific work is shown in Fig. 1. Firstly, mangrove images used in this study were obtained from Google Earth, with a focus on the mangrove forest in Dongzhaigang and Wenchang River, Hainan Province, China. Image data processing involved the removal of low-quality and duplicate images, resulting in a total of 230 images. All original images were resized to 768×768 pixels, and image categories were defined into six classes based on the most prominent features in the images. Fine labels were added to all images using the Labelme tool. During the model design phase,

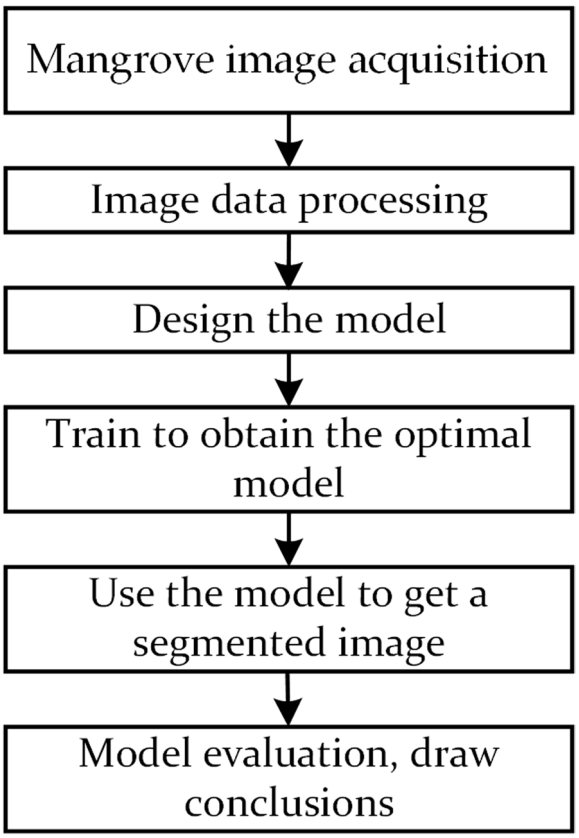

Fig. 1 Overall flowchart of our work

our MSFANet model drew inspiration from several classic models, adapting them to the characteristics of mangrove forest image segmentation. After preparing the data, it was input into the model for training, yielding the optimal model. The best model was then used to predict image categories, ultimately generating segmented mangrove images. These segmented mangrove images were evaluated from both visual and data metric perspectives, leading to conclusions about the performance of the MSFANet model in mangrove forest image segmentation. In this research paper, we present three primary contributions:

- We introduce the Multi-Scale Large Receptive Field Network, designed to effectively preserve spatial information across varying scales. This innovative network architecture capitalizes on a substantial receptive field to optimize the retention of spatial details while concurrently leveraging separable convolution to significantly curtail computational complexity.
- Our work incorporates the Attention Fusion Module, which enhances fusion efficiency and efficacy. Traditional fusion methods such as simple addition or direct channel connections can introduce undesirable interference information, thereby negatively impacting segmentation results. The Attention Fusion Module effectively mitigates this interference, leading to improved segmentation outcomes.
- We introduce a Dual Channel Design that facilitates the comprehensive integration of both semantic and spatial information. This design choice leads to enhanced output results by synergizing these two critical types of information in the segmentation process.

The rest of this paper is organized as follows. "Related work" section introduces the related work of remote sensing image segmentation. "Methodology" section details the methodology of the MSFANet model. "Experimental results and analysis" section shows the extensive experiments and analysis of our proposal. We conclude in "Conclusion" section.

Related work

Traditional machine learning methods have held a significant position in the field of remote sensing image segmentation, encompassing threshold-based segmentation [8–10], edge-based segmentation [11–13], region-based segmentation [14, 15], and clustering-based segmentation methods [16]. Threshold-based segmentation methods, due to their intuition, simplicity, and computational speed, are commonly used in traditional image

segmentation algorithms. The Otsu method [8], known for its simplicity and robustness to variations in image brightness and contrast, has found extensive applications in digital image processing. The significant advantages of Otsu have attracted many researchers. Yang et al. [9] improved the Otsu algorithm, resulting in better segmentation of test images, competitive misclassification errors, Dice Similarity Coefficient (DSC) values, and reduced computational time. Pratiwi et al. [10] applied the Otsu method to segment mangrove ecosystems captured by unmanned aerial vehicles, demonstrating that it effectively separates mangrove areas from others while preserving mangrove details.

Commonly used improved edge detection operators include the Canny operator [11], the Prewitt operator [12], and Sobel operator [13], which not only correctly detect object edges but also suppress image noise effectively, striking a balance between the two. Xue et al. [14] proposed an improved watershed algorithm for accurately extracting land boundaries in high-resolution remote sensing images, resulting in improved segmentation performance and time efficiency. In the realm of region-based segmentation, Dong Yang et al. [15] introduced an enhanced region-growing method that incorporates improved median filtering for smoother image processing, increased internal consistency within targets, texture information retention, and automatic seed selection, followed by fragment merging to obtain diverse object segmentation results. Clustering-based image segmentation methods [16] categorize all pixels in an image into different regions based on grayscale values, spatial positions, and other information, ensuring similar structures within the same region and significant differences between different regions, thereby achieving image segmentation.

Traditional machine learning methods have made notable contributions to image segmentation, but feature extraction often depends heavily on the algorithm designer's expertise. Each method is usually tailored to specific applications, resulting in limited generalization and robustness. In contrast, deep learning relies on data-driven feature extraction, deriving deep, dataset-specific feature representations from extensive sample learning. These abstract features exhibit stronger robustness and better generalization, making them more efficient and accurate. However, traditional deep learning methods are not well-suited to mangrove remote sensing images due to their limited receptive field and the challenges associated with effective high-level feature extraction, along with the presence of numerous difficult samples. Recent advances in computing power, coupled with the widespread use of deep learning specialized hardware such as GPUs and TPUs, have led to remarkable achievements in various tasks, making deep learning-based semantic segmentation a primary focus in remote sensing image analysis.

Deep learning-based semantic segmentation methods typically explore multiple aspects, including multi-scale information and attention mechanisms. Multi-scale exploration involves recognizing that different scales can better identify target information for segmentation, enhancing the accuracy of information for each category. For example, Chen et al. introduced the Atrous Spatial Pyramid Pooling (ASPP) mechanism in the DeepLabv3 [17] and DeepLabv3+[18] series, which utilizes different receptive fields to extract features from feature maps, allowing for the acquisition of information at different scales. Zhao et al. [19] proposed the Pyramid Scene Parsing Network (PSPNet) with a pyramid pooling module that integrates global contextual information. Other models, such as DenseASPP [20], combine ASPP from the DeepLab series and dense connections from DenseNet, resulting in larger receptive fields and denser sampling points. Despite these similarities in pyramid pooling methods, they have drawbacks, as they introduce a large number of channels, which increases parameter count, and global average pooling can lead to a significant loss of positional information [21]. In addition to pyramid structures, many models use multi-scale fusion techniques to improve segmentation accuracy, such as OCNet [22], EMANet [23], which employ multi-scale fusion techniques to enhance accuracy. Attention mechanisms, which weigh the importance of information in the input data, can help networks better understand key information, thereby improving performance, efficiency, interpretability, and adaptability. Attention mechanisms can be categorized into two main types: channel attention, which weights channels in convolution, and pixel attention, which weights pixels. In terms of channel attention, representative models include DANet [24], CCNet [25], DFN [26], DSANet [27], and BiSeNet [28], all of which employ channel attention to improve segmentation accuracy. DANet contributes a dual attention mechanism to handle multiscale information and channel relationships. CCNet introduces a Criss-Cross attention module, modeling cross-channel relationships to improve segmentation accuracy. DFN addresses intraclass inconsistency by introducing attention mechanisms and global average pooling to select more representative features. DSANet and BiSeNet also use channel attention to enhance segmentation accuracy. For pixel-level attention mechanisms, OCNet introduces Object Context Attention (OCA) to capture relationships between objects. This mechanism calculates the similarity between each pixel and all other pixels, adjusting the feature map's weights based on these similarities to capture object relationships

effectively. Fan et al. [29] proposed a mangrove image segmentation model based on domain adaptation, combining self-attention mechanisms and remote sensing spectral indices. Self-attention mechanisms enable the model to focus on more important image channels, while spectral indices address potential edge information loss in domain adaptation. Numerous experiments related to attention mechanisms have demonstrated their effectiveness in improving image segmentation.

In addition to convolutional neural networks (CNNs), transformer-based methods have also gained popularity in deep learning-based image segmentation. Utilizing Transformer structures for pixel-level weighting is a current research focus, starting with the Vision Transformer [30]. Various Transformer variations have been applied to image segmentation, including TransUNet [31], which improves upon the UNet model, and pure Transformer structures like Swin-UNet [32]. SpectralFormer [33] was the first to apply the transformer structure to hyperspectral image classification. SpectralFormer's structure is simple, efficient, and nearly identical to the original transformer structure, yet it has demonstrated outstanding performance and attracted significant attention. Zhong et al.'s model [34] is both simple and efficient, and the study extensively verifies the performance of various transformer combinations, providing valuable insights for researchers. Sun et al. [35] proposed a method that cleverly combines the main CNN and transformer structures. CNN captures low-level spectral-spatial features and transforms them into semantic labels, while the transformer structure models high-level semantic features. In a dual-branch structure, Wang et al. [36] introduced the Hyper-ES2T network, harnessing the power of Transformers. The fusion of Transformer and dual-branch architecture showcased robust generalization capabilities and superior feature representation. Yang et al. [37] proposed the Hyperspectral Image Transformer (HiT) network, incorporating CNN operations within the Transformer framework. This approach effectively captures subtle spectral distinctions and conveys localized spatial context information.

In the domain of mangrove image segmentation [38], the UNet-based neural network model has garnered substantial acclaim. UNet leverages both spatial and spectral information to facilitate semantic segmentation through an encoder-decoder architecture. As an illustration of this paradigm, Dong et al. [39] introduced the GC-UNet model, which incorporates global contextual blocks into the UNet framework, thereby capturing long-range dependencies. Their approach utilizes SPConv to emphasize intrinsic details and employs adaptive spatial feature fusion to address disparate feature levels. A pivotal contribution to the field was made by Moreno [40], who conducted a pioneering deep-learning investigation into mangrove image segmentation using radar time-series data. This study harnessed the UNet model in conjunction with spatial, temporal, and polarization datasets, revealing that when coupled with Efficient-net-B7, UNet outperforms competing architectures such as ResNet-101 and VGG16.

Receptive field research is a crucial direction, with receptive field size significantly impacting neural network performance. Smaller receptive fields capture image details but may sacrifice global contextual information. In contrast, larger receptive fields extract more contextual information at the potential cost of fine-grained details. The size of the receptive field is directly influenced by convolutional kernel size, prompting the exploration of dilated convolutions and large kernel designs. RepLKNet [41] reevaluated large kernel designs in modern Convolutional Neural Networks (CNNs) from a sparsity perspective, proposing an extremely large kernel approach, expanding kernels to 61×61, with improved experimental results. GCN [42] discussed the advantages of large convolutional kernels compared to smaller ones, achieving global convolution effects by using convolutional kernels of the same resolution as feature maps in the deepest network layers. Experiments on standard datasets demonstrated concurrent state-of-the-art segmentation results.

In summary, based on the diverse body of literature encompassing image segmentation methods discussed earlier, it becomes evident that several critical factors significantly influence the efficacy of image segmentation. These key factors include multi-scale considerations, the incorporation of attention mechanisms, the architectural design of the model, and the extent of the receptive field. Notably, when dealing with the challenges posed by the vast dimensions of remote sensing images capturing mangrove ecosystems, an important observation emerges. These images often contain a substantial proportion of mangrove regions, resulting in potential limitations associated with inadequate receptive fields for mangrove areas. This limitation can hinder the extraction of nuanced features and complicate the segmentation process, particularly given the presence of numerous complex samples. To address these aforementioned challenges, our study focuses on the strategic integration of multi-scale methodologies, attention mechanisms, model network structures, and the enhancement of receptive fields. These aspects collectively contribute to a more effective and precise mangrove image segmentation approach.

Methodology

GCN model

The GCN model proposes a global convolutional network to solve classification and localization problems for semantic segmentation. The application of large kernel convolution is considered. The core modules of the GCN model are the GCN and BR modules, which utilize the large receptive field convolution of the GCN module to obtain multi-scale features. The residual-based boundary refinement network BR module is used to further refine the boundaries of the image. Figure 2 shows the network architecture of the entire GCN model.

The detailed process of image segmentation by GCN model is introduced later. Firstly, the backbone network uses ResNet to extract features, and ResNet is adjusted to a five-layer network structure, i.e., Layer0- Layer4. Because the last classification layer of ResNet is not used, so the last layer is discarded. The size of the downsampled feature maps in each layer of the image is 1/2 of the original size, and the number of channels is expanded by two times from 64 in the Layer0 layer, and the number of channels finally reaches 2048 in the Layer4 layer. The dashed part of the figure shows the different scales of feature maps from Layer1 to Layer4 extracted using the GCN module respectively. And it makes the number of output channels equal to the number of categories, which are classified into six categories in this paper. Then BR module is used to receive the output of the GCN module for further feature extraction, it does not change the number of channels and feature map size. The feature map is up-sampled twice in Layer4 using Deconv inverse convolution, so that it is the same size as the feature map of Layer3, 48×48, and the two feature maps are summed up, and then the residuals are connected to the previous layer through the BR module and the inverse convolution to achieve the fusion of different scales. Then the final output is obtaineded after three BR modules and two inverse convolution modules.

Each GCN module uses two 15×15 size convolutional kernels to extract features. Due to the use of separable convolutions [43] in the GCN module instead of commonly used convolutional forms, the advantage is that it reduces computational complexity while ensuring the quality of extracted features. Finally, the features extracted from the two convolutions are added and fused to obtain the final result. The BR module is simpler by using two convolutions with an activation function added to the original feature map to form a residual structure. These two modules have been used multiple times throughout the GCN model, and the experimental data in the GCN of literature [35] shows that good results have been obtained on different datasets.

MSFANet model

Inspired by the GCN model and combined with the characteristics of mangrove remote sensing images, this paper proposes a multi-scale fusion attention network MSFANet (Multi-Scale Fusion Attention Network) suitable for the segmentation of mangrove remote sensing images. Usually, the networks we design tend to use small

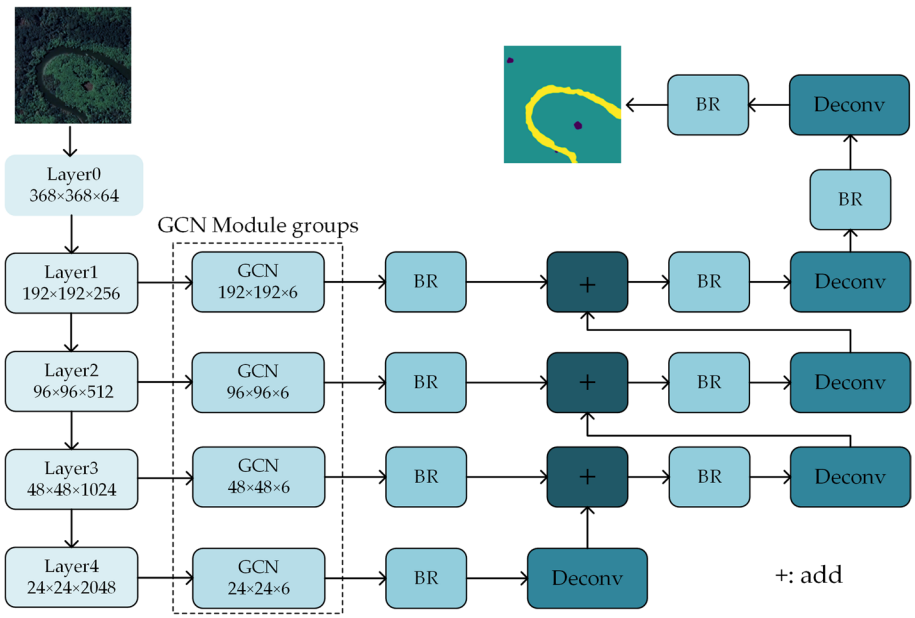

Fig. 2 Diagram of GCN network architecture

convolutional kernels (such as 1×1 or 3×3). However, the biggest difference between segmentation and classification is that image segmentation not only classifies pixel points but also locates image pixels. Convolutional neural networks have been proven to have no translation invariance, which inevitably results in the loss of more positional information for small convolutional kernels compared to large convolutional kernels. Therefore, this paper will also consider how to use large convolutional kernels. After extracting sufficient information, how to preserve and fully utilize it is also a problem. This paper will use attention mechanisms to solve this problem. The MSFANet model was ultimately obtained by combining the idea of large convolutional kernels with the attention fusion module.

The detailed process of image segmentation using the MSFANet model is illustrated in Fig. 3. Firstly, ResNet-50 was used as the backbone network, and the ResNet layer was adjusted to reduce the output feature map size of Layer0 to half of the original size, with a channel count of 64. The combination of the ResNet maximum pooling layer and ResNet Layer1 forms a new Layer1. The entire ResNet is divided into five layers: Layer0, Layer1, Layer2, Layer3, and Layer4. This paper will use the first four layers, with a resolution reduction by half, for each layer, while featuring channel counts of 64, 256, 512, and 1024, respectively. As depicted in group (a) of Fig. 3, three Atrous Large Receptive Field Convolutional Network (ALRFCN) modules are utilized, forming a multi-scale large receptive field network. For the outputs of Layer0, Layer1, and Layer2, feature extraction is performed using convolutions with dilated rates of 12, 6, and 3, respectively, and the corresponding convolution kernel sizes with dilated convolutions are 49, 25, and 13, whose original convolution kernel size is 5. This can preserve the position information as much as possible, and the number of output channels is unified to 64. Then, using the same three ALRFCN network modules as shown in group (b) in Fig. 4, adjust the dilated rates to 8, 4, and 2, and unify the output channels to 6, which exactly corresponds to the six categories to be divided.

Layer 3, as a deep-level network, has rich advanced features and semantic information. It performs double upsampling through transposed convolution, and the number of channels is adjusted from 1024 to 512. Then, through the ARM (Attention Refinement Module) attention refinement module, weights are assigned to the input channels, and the output channel is set to 256. The output results of the ARM module are upsampled and transmitted to the AFM module and ARM module respectively. The AFM module fuses the output of ALRFCN and the output of ARM and assigns weights to the fused channels. Simultaneously, another branch passes through two ARM modules, achieving four-fold upsampling, attaining half of the original image resolution, and maintaining a channel count of 6. Both branches converge at the Feature Fusion Module (FFM), where two-fold upsampling is applied to the final output result. The output of the two branches is fused at the Feature Fusion Module (FFM), where two-fold upsampling is applied to achieve the final output result. This model incorporates two upsampling methods: bilinear interpolation and transpose convolution. Feature maps processed by the ARM module exclusively use transpose convolution-based

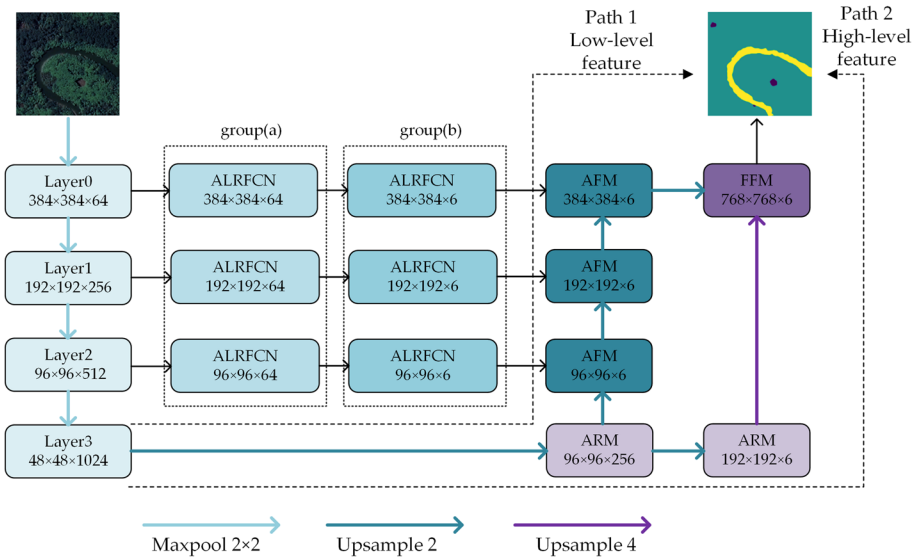

Fig. 3 MSFANet network architecture

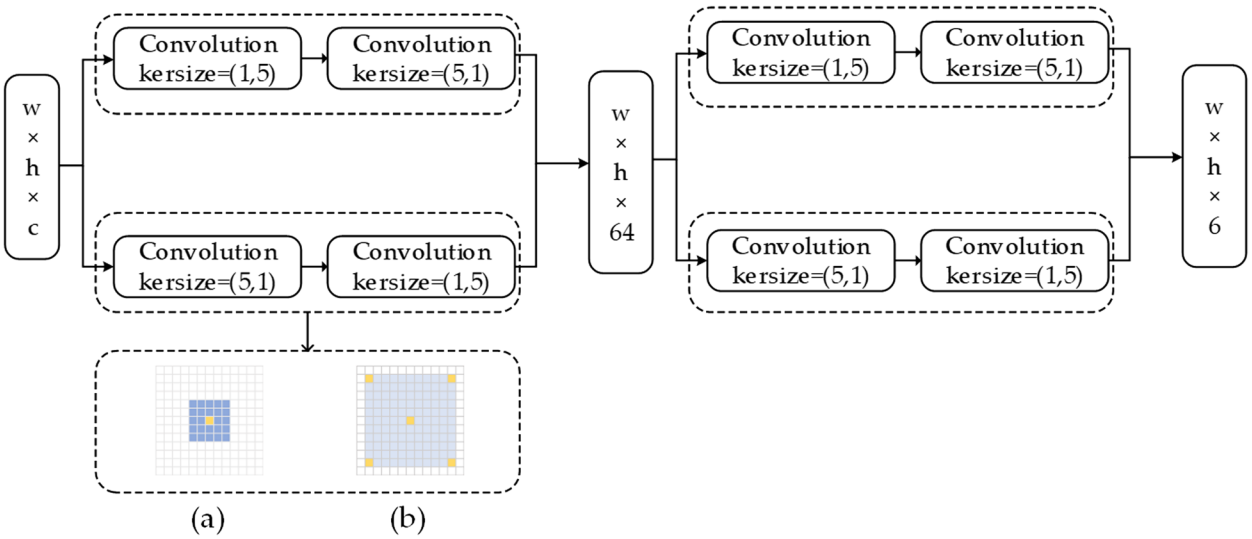

Fig. 4 ALRFCN modul

upsampling, while other parts of the network employ bilinear interpolation.

The dual-path design in the MSFANet network aims to maximize the utilization of high-level and low-level feature information, ensuring both accurate preservation of positional information and effective extraction of low-level features such as image contours. Moreover, high-level features are fused with low-level features using channel attention mechanisms after deep-level network extraction, as illustrated by Path1 and Path2 in Fig. 3, enhancing the quality of image segmentation.

ALRFCN model

The ALRFCN (Atrous Large Receptive Field Convolutional Network) module, compared to the GCN module in the GCN paper and the ASPP (Atrous Spatial Pyramid Pooling) module in Deeplabv3+, offers a unique combination of advantages from both the GCN module and the ASPP module, effectively mitigating their respective limitations. In contrast to the GCN module, which significantly reduces the number of channels to a small classification count for different-scale feature maps, thus accelerating computation but also causing severe information loss, and the ASPP module, which requires the computation of five feature maps followed by concatenation, thus greatly preserving the multi-scale information of the maps, but also leading to substantial computational load. The MSFANet method employs the ALRFCN module in a two-step, two-component approach.

In this approach, three ALRFCN modules form a multi-scale large receptive field network, as depicted in Fig. 4, group (a). These modules process feature maps from Layer0, Layer1, and Layer2, extracting features and maintaining an equal number of channels. The first component yields an output channel count of 64, preventing severe channel reduction and significant information loss. Subsequently, another set of three ALRFCN modules processes the output from group (a), equalizing the output channel count with the classification count, which is set to 6 in this paper.

Moreover, while using large convolutional kernels directly for achieving a large receptive field is effective, it inevitably incurs a substantial increase in computational load. To address this concern, the MSFANet method improves upon this concept by employing dilated convolutions instead of large convolutional kernels. This approach leverages the advantages of atrous convolutions, which expand the receptive field without increasing kernel size, while also enhancing computational efficiency through the use of separable convolutions.

The ALRFCN network adopts a serial fusion approach, as illustrated in Fig. 4. In each ALRFCN module, the size of a feature map is w×h×c as initial input. This feature map undergoes feature extraction via two separable convolutions, utilizing convolution kernel sizes of (1,5) and (5,1) to replace a traditional (5,5) convolution kernel. The adoption of depth-wise separable convolutions significantly enhances computational efficiency. The parameter count can be analyzed as follows. suppose an input feature map with the size w×h×c and convolution kernel sizes of k×k×ck, the parameter count per convolution layer can be calculated using Eq. 1:

$$\textbf{Params} = C_0 \times (k_w \times k_h \times C_i + 1); \quad (1)$$

Where c0 represents the input channel count, kw represents the convolution kernel width, and kh represents

the convolution kernel height. The "+1" term accounts for bias parameters. For depth-wise separable convolutions, the parameter count is significantly reduced. Given that the values of k are all greater than 5, it becomes evident that depth-wise separable convolutions result in significantly fewer parameters compared to regular convolutions.

The first stage of output is w×h×64, serving as a transitional phase to ensure effective information retention while avoiding excessive information loss. The second stage yields an output of w×h×6, representing the final classification count, or the final number of output channels.

Figure 4 illustrates the ALRFCN module alongside standard 5×5 convolution (rate=1) and atrous convolutions of the same size but with different dilation rates (rate=3). Figure 4a illustrates the receptive field of the standard 5×5 convolution, while Fig. 4b shows the receptive field of the convolution with a corresponding dilation rate of 3. It is evident from the figure that the receptive field with dilated convolution significantly increases. The ALRFCN module is composed of two groups, denoted as group(a) and group(b). Each group consists of three modules. In group(a), the modules receive the outputs from Layer0, Layer1, and Layer2, and employ dilations of 12, 6, and 3, respectively, with an original kernel size of 5 for feature extraction. According to Eq. (2), this results in dilated convolution kernel sizes of 49, 25, and 13, and receptive field sizes of 48, 24, and 12, as determined by Eq. (3). In group(b), the dilation rates are 8, 4, and 2, leading to convolution kernel sizes and receptive field sizes, as calculated by the formulas, to be 33, 17, and 9, and 80, 56, and 20, respectively.

The formula for calculating the receptive field with atrous convolutions in the module is shown in Eq. 2.

$$k_{new} = k_{ori} + (k_{ori}\text{-}1)(\text{rate-}1) \tag{2}$$

The convolution kernel size corresponding to the dilation rate represents the actual calculated kernel size, which is the original kernel size., and "rate" denotes the dilation rate. The calculation of the receptive field is shown in Eq. 3.

$$RF_{i+1} = RF_i + (\textbf{kernel\_size} - 1) \times \textbf{stride} \tag{3}$$

Where RF_{i+1} is the actual receptive field, RF_i is the receptive field of the previous layer, kernel_size is the size of the convolutional kernel, and stride is the stride of the convolution, with a default value of one.

In summary, the two groups, group (a) and group (b), composing the ALRFCN module can effectively address the challenge of insufficient receptive fields for mangrove image recognition associated with a small convolutional kernel.

ARM module

During the process of feature extraction, a significant number of channels are generated. Taking ResNet as an example, it can even utilize up to 2048 channels. However, it's essential to recognize that not all channels contain the same amount of information. This necessitates the identification of important channels while downplaying less critical ones, focusing attention on the significant parts. This concept is the core of the channel attention mechanism [17]. The aforementioned process can be expressed as shown in Eq. 4.

$$\textbf{Attention} = \mathbf{f}(\mathbf{g}(x), \mathbf{x}) \tag{4}$$

Wherein g(x) represents the process of generating attention over the input feature x, while f(g(x),x) represents the operation of enhancing the input feature x based on the generated attention g(x), thus strengthening the features in the salient regions.

The ARM module leverages the fundamental principles of the attention mechanism, where the weights across different channels are unequal, enhancing important features while attenuating less important ones, thereby improving feature quality. As illustrated in Fig. 5, the computation process of the ARM module involves a h×w×c feature map passing through convolution and average pooling to yield a 1×1×c vector. Subsequently, further feature extraction occurs through convolution,

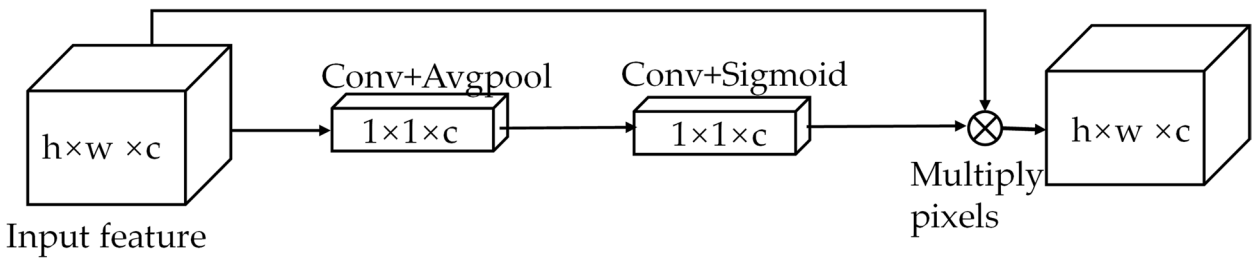

Fig. 5 ARM module

employing the Sigmoid function for weight allocation and normalization. The resulting weight vector is then multiplied element-wise with the original feature map, resulting in an output with the size of h×w×c. The core of this module lies in the weight allocation across channels using the Sigmoid function, which embodies the central idea of channel attention. The specific procedure of the ARM structure is shown in Table 1. In summary, the Attention Refinement Module (ARM) adjusts the importance of different regions in the feature map by introducing an attention mechanism, directing the module's focus toward relevant information in mangrove areas. The utilization of adaptive average pooling in ARM simultaneously reduces spatial dimensions while preserving overall feature information, facilitating adaptability to varying input image sizes. Through element-wise multiplication after Sigmoid activation, ARM diminishes attention to background or irrelevant details, ensuring a heightened focus on mangrove regions. This contributes to enhancing the differentiation from other categories in mangrove remote sensing image segmentation under challenging conditions, such as complex textures and variations in illumination. Additionally, it aids in addressing issues related to the ambiguous boundaries of mudflats and water bodies.

Table 1 The specific procedure of the ARM

Algorithm: Attention Refinement Module
Input: input
Output: x
1. Apply Conv2d operation to input, resulting in input1
2. Apply AdaptiveAvgPool2d operation to input1, obtaining x
3. Apply Conv2d operation to x. x = Conv2d(x, kernel_size = 1)
4. Apply Sigmoid activation to x
5. Element-wise multiply input1 by x
6. Return x

The Sigmoid function used in this module is represented as Eq. 5, wherein x corresponds to the vector after average pooling. The output of the Sigmoid function falls within the range of 0 to 1. Because the output values are confined to this range, it normalizes the output of each neuron.

$$\mathbf{s}(\mathrm{x}) = \frac{1}{1 + \mathrm{e}^{\mathrm{x}}} \tag{5}$$

AFM module

The AFM is responsible for fusing high-level and low-level features by redistributing channel weights through an attention mechanism. As depicted in Fig. 6a, it initially takes two feature maps, high-level feature f1, and low-level feature f2, and concatenates them to obtain the f3 feature map. This direct concatenation implies equal weights for different channels. However, in reality, features at different stages possess varying degrees of discriminative power, leading to differing prediction consistency. To achieve intra-class prediction consistency, it's crucial to extract discriminative features and suppress non-discriminative ones. To achieve this, an average pooling is performed, generating a weight vector WV. This weight vector undergoes convolution, ReLU activation, and another convolution, followed by sigmoid activation for weighting. The convolution is used to match the channel count of the weight vector to that of the high-level feature f1. The unweighted weight vector is then multiplied element-wise with the low-level feature f2 to produce a new weighted feature map WF. The WF is added to f1 to obtain the final fused feature FF. The entire process can be represented by Eqs. 6, 7, 8, 9 and 10:

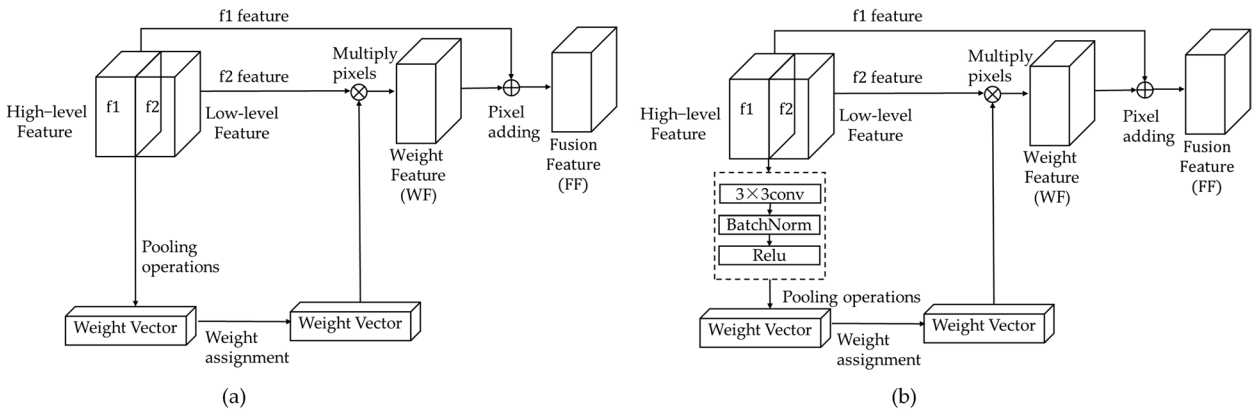

Fig. 6 AFM and FFM **a** AFM **b** FFM

$$\mathbf{f3} = \mathbf{concat}[\mathbf{f1}, \mathbf{f2}] \tag{6}$$

$$\mathbf{WV} = \mathbf{AvgPool}(\mathbf{f3}) \tag{7}$$

$$\mathbf{WV} = \boldsymbol{\beta}(\boldsymbol{\delta}(\alpha(\delta(\mathrm{Wv})))) \tag{8}$$

$$\mathbf{WF} = \mathbf{Wv} \otimes \mathbf{f1} \tag{9}$$

$$\mathbf{FF} = \mathbf{Wf} \oplus \mathbf{f2} \tag{10}$$

Wherein AvgPool represents global average pooling, WV is the weight vector, WF is the weighted feature map, FF is the final output, δ signifies a 1×1 convolution, α denotes the ReLU activation function, β represents the sigmoid activation function, ⊗ represents element-wise multiplication, and ⊕ signifies element-wise addition.

In summary, through global average pooling, AFM can capture global information, facilitating the fusion of multi-scale features and enhancing the model's adaptability to different scale information. The weights generated by the Sigmoid activation function are used to adjust the importance of input features, directing the model's attention to regions contributing to the segmentation of mangroves, thereby improving the segmentation performance.

The FFM structure, illustrated in Fig. 6b, serves to receive the ultimate high-level and low-level features and performs the final feature fusion. It operates based on the same principles as the AFM module. The key difference lies in the part where high-level feature f1 and low-level feature f2 are concatenated, as it involves feature extraction. This primarily involves a convolution block, depicted by the dashed line in Fig. 6a, which includes a 3×3 convolution, BatchNorm layer, and ReLU layer. This enhances the feature information of the two pathways to ensure a more comprehensive final output, ultimately improving the overall output quality. The specific procedure of the FFM structure is shown in Table 2.

FFM integrates low-level features from Path1 and high-level features from Path2, ensuring a balance between semantic information from high-level features and spatial information from low-level features. The combination of both through the FFM module enhances the recognition of small targets in mangrove images while preserving the contour information of large targets.n general, the Feature Fusion Module provides an effective mechanism by integrating features at different levels or channels, enhancing the modeling capability of deep learning models for input data. This contributes to improving the performance and generalization ability of the model.

Experimental results and analysis

Study area

Mangrove vegetation is widespread in Southeast Asian coastal areas and subtropical regions of China, including those found in Hainan Province, which serves as the primary study area for this research. Remote sensing images of mangroves from within Hainan Province, such as Dongzhaigang and Wenchang River located at the red star in Fig. 7, were selected for analysis. To diversify the dataset, some foreign mangrove images were also included as supplementary data. Figure 8 shows a schematic diagram of the labels used for different classifications in the images. Image labels were meticulously annotated using the labelme tool. The dataset was divided into six classes based on the content in the images, including mangroves, rivers and oceans, buildings and roads, ponds, tidal flats, and background. Mangroves are depicted in light green, rivers and oceans in green, buildings and roads in blue, ponds in magenta, tidal flats in yellow, and the background in black. This dataset employs six criteria for classification. Firstly, the identification of mangrove areas is based on a scale of approximately 100 m on Google Earth. Images are selected to ensure the presence of multiple categories, with a focus on delineating the boundaries of entire mangrove forest regions

Table 2 The specific procedure of the FFM

Algorithm: Feature Fusion Module
Input: input_1, input_2
Output: x
1. Concatenate input_1 and input_2 along the channel dimension to create a new tensor x
2. Apply ConvBlock operation to x, yielding a feature tensor
3. Utilize the AdaptiveAvgPool2d operation on the feature tensor, resulting in x
4. Apply Conv2d with ReLU and Sigmoid activations to x x = Sigmoid(Conv2d(ReLU(Conv2d(AdaptiveAvgPool2d(ConvBlock([input_1,input_2])),kernel_size = 1)), kernel_size = 1))
5. Perform element-wise multiplication of the result of ConvBlock([input_1, input_2]) by x
6. Conduct element-wise addition of the above result to ConvBlock([input_1, input_2])
7. Return x

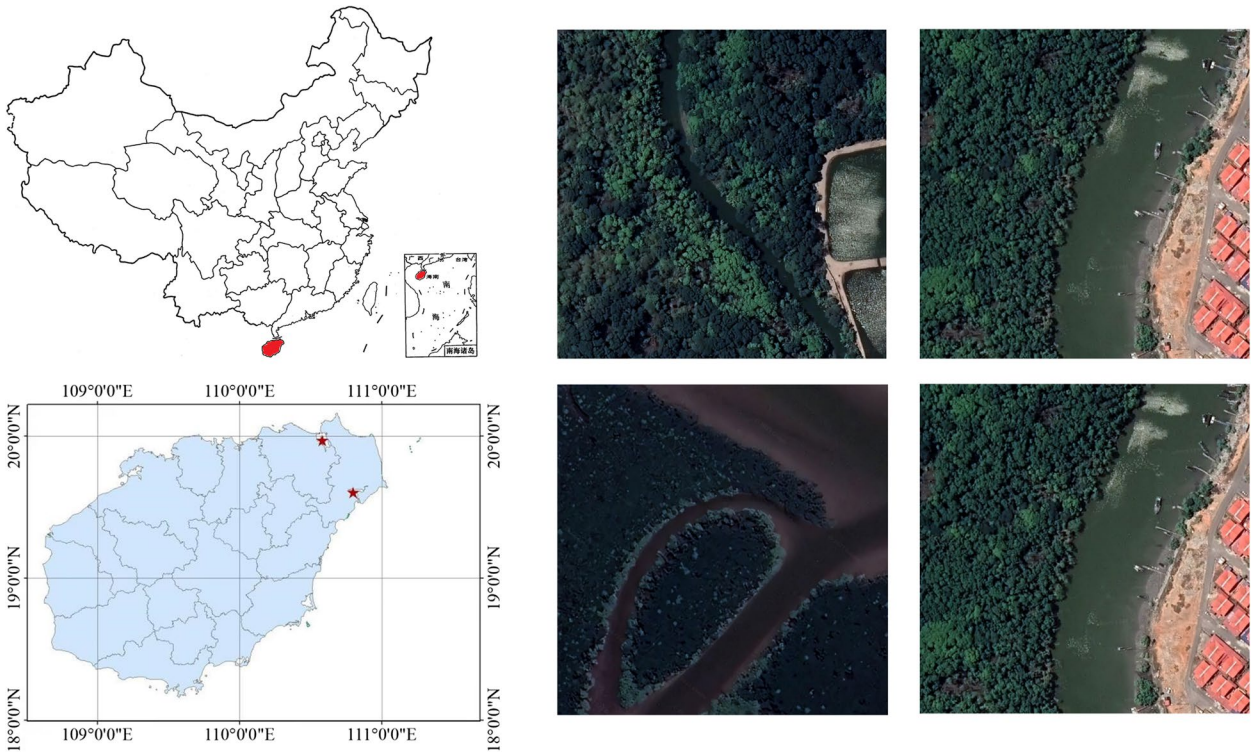

Fig. 7 Location of the study and Mangrove images **a** Location of the study **b** Mangrove images

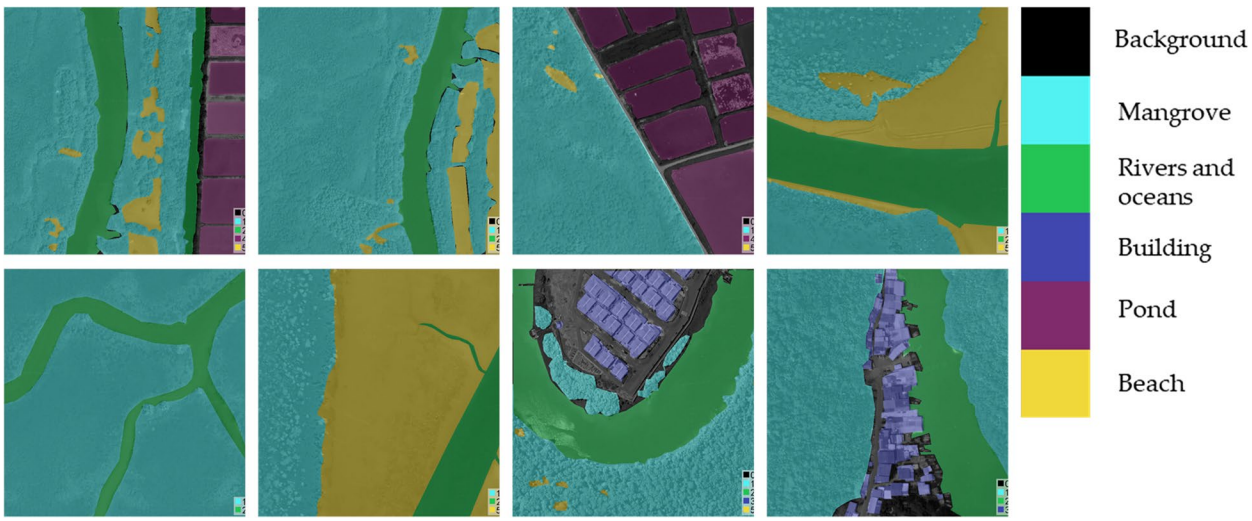

Fig. 8 Schematic diagram of fine labels

rather than isolated trees. Open areas within mangrove regions are classified as mudflats, including the intertidal zones between mangroves and rivers or the ocean. Ponds are delineated by the boundaries of their water bodies. Buildings are primarily identified by the structures of houses and are delimited by the boundaries around the buildings. The dataset is designed to represent the most common scenarios in mangrove ecosystems, capturing typical features to ensure it is representative of real-world scenes.

A total of 230 images were selected for this experiment, with 184 images used for training and 46 images for validation, divided in an 80:20 ratio. The image dimensions were set to 768 pixels, and the mangrove image data sources were obtained from Google Earth [44], consisting of high spatial resolution (0.3 m) remote sensing images captured on June 6, 2021.

The research area of Dongzhaigang is situated at the intersection of Wenchang City and Haikou City in the northeastern part of Hainan Province, China, whose longitude ranges from E110°32' to E110°37' and latitude ranges from N19°51' to N20°1'. Covering a total area of 3337.6 hectares, the core zone occupies 1635 hectares and the mangrove coverage extends over a substantial 1771 hectares, coexisting harmoniously with the coastline and spanning a length of 28 km.

Experiment and evaluation indicators

The experiment of this study was implemented on the operating system of Windows 11 Professional Edition, version 21H2. The employed CPU is an Intel(R) Xeon(R) E5-2680 v3, with a memory size of 32.0 GB. The storage disk was a Samsung SSD 870 EVO with a capacity of 500 GB. The used GPU is an NVIDIA GeForce RTX 4090 with 24 GB of dedicated graphics memory.

The overall procedural flowchart for the deep learning-based experiments conducted in this paper is depicted in Fig. 9. The dataset for mangroves is divided into two parts: the training set and the validation set. The procedure begins by initializing model parameters, followed by loading the training data into the model for iterative computations of the loss values and parameter updates to refine the model. Subsequently, the validation set is input into the model with updated parameters, and the Mean Intersection over Union (Miou) evaluation metric is employed to select the model with the highest value after each epoch as the optimal model. In addition to the Miou indicator, this experiment also selected several important indicators such as pixel accuracy (PA), F1 score (F1_Score), and class pixel accuracy (CPA). Through these different evaluation indicators, the model's advantages and disadvantages can be more comprehensively evaluated. The model uses ResNet as the backbone network and initializes the model parameters using ResNet's pre-trained model. The optimizer selected a more stable Adam algorithm. The cross-entropy loss function was chosen as the loss function for this experiment, and the expression is shown in Eq. 11. Cross entropy loss has the characteristics of stability and fast learning speed, and is also the mainstream loss function in current deep learning. The selected batch size was 8, and the training duration was approximately four hours.The learning rate is reduced by 40% using a fixed iteration of 80 times, with an initial value of 0.001 and a total of 450 iterations. The selection of the optimal model is based on Miou as the only indicator. Select the Miou with the largest validation set as the optimal model.

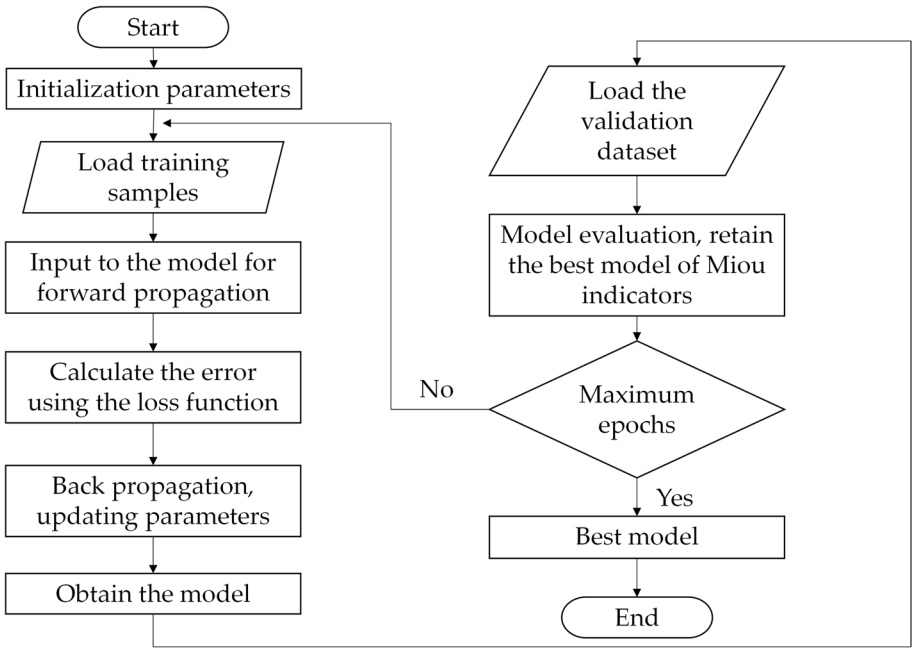

Fig. 9 Flow chart of the experiment in this study

The formula for the cross-entropy loss function is as follows:

$$L=-\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{11}$$

Where M is the number of classes, which is 6 for classification in this paper; N represents the total number of samples, signifying the number of images; $\mathbf{y_{ic}}$ takes a value of 0 or 1, if the true class of sample i is equal to c, take 1, otherwise take 0; $\mathbf{p_{ic}}$ is the prediction probability of the observed sample i belonging to class c.

The evaluation metrics for this model include the Mean Intersection over Union (Miou) for average class IoU, Pixel Accuracy (PA), Class Pixel Accuracy (CPA) for individual classes, and the F1 score. These metrics are computed using a confusion matrix. The terms used in these metrics are defined as follows: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

The Intersection over Union (IoU) represents the ratio of the intersection to the union of the model's prediction and the ground truth for a specific class. Miou, which stands for the mean IoU, represents the average IoU among the model's predictions and the ground truth for all classes. It is calculated as the sum of IoU values for all classes divided by the number of classes, as shown in Eq. 12, where k represents the number of categories.

$$\text{Miou} = \frac{1}{k}\sum_{i=1}^{k}\frac{TP}{(TP+FP+FN)} \tag{12}$$

PA measures the proportion of correctly predicted pixels for all classes over the total number of pixels. It is defined as shown in Eq. 13. CPA is a metric that specifically assesses the accuracy of predicting pixels belonging to a particular class. In this study, the primary focus was on the mangrove class.

$$PA = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{13}$$

The F1 Score (F1_Score) is a statistical metric used to evaluate the accuracy of a model, taking into account both the precision and recall of a classification model. It can be seen as a weighted average of a model's precision and recall, with a maximum value of 1 and a minimum value of 0. A higher F1_Score indicates better overall performance of the model. The definition of F1_Score is shown in Eq. 16, where Eq. 14 represents precision, and Eq. 15 represents recall.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{15}$$

$$\text{F1\_score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{16}$$

Comparison of experimental results for different models

Firstly, eight models were selected for different comparative experiments, and the experiments of different models on mangroves were evaluated from multiple indicators. The experimental data of these eight models are shown in Table 3, displaying the data for five metrics: Miou, PA, F1_score, Mangrove CPA, and River & Ocean CPA. The backbone networks for MSFANet, GCN, and ExFuse are either ResNet-50 or ResNet-101. UNET and DeepLabv3+are models that establish their backbone networks. From the experimental results, it can be seen that the Miou of the MSFANet model is significantly higher than that of the other models, reaching over 80%, and the performance of PA is also one percentage point ahead of other best models. From the experimental results, it can

Table 3 Experimental results of different models on the mangrove dataset

Model	Miou(%)	PA(%)	F1_score(%)	Mangrove(CPA) (%)	River & Ocean(CPA) (%)
MSFANet ResNet-50	0.86015	0.96419	0.92029	0.98981	0.96425
MSFANet ResNet-101	0.84928	0.96261	0.91238	0.98951	0.94620
UNET	0.68159	0.91982	0.77046	0.98508	0.91875
GCN ResNet-50	0.78231	0.95376	0.86321	0.98847	0.94395
GCN ResNet-101	0.78443	0.95078	0.86935	0.98013	0.93630
DeepLabv3 +	0.73638	0.92917	0.83042	0.98569	0.93651
ExFuse ResNet-50	0.72884	0.94070	0.80712	0.98487	0.93891
ExFuseResNet-101	0.78559	0.94418	0.86893	0.98364	0.94242

be seen that MSFANet has a significant improvement in experimental data when using ResNet-50 as the backbone network compared to ResNet-101 as the backbone network. The experimental data of different backbone networks in GCN models do not differ significantly, and the main reason for this result is that MSFANet has too many parameters when using ResNet-101 as the backbone network, resulting in poor overfitting results.

In the testing results of these eight models, the Miou metric is the most critical indicator for image segmentation experiments. The bar chart of the data, as shown in Fig. 10, illustrates the Miou values. From these data, it can be observed that the MSFANet ResNet-50 model performs the best in terms of the Miou metric, achieving a score of 0.86015. The MSFANet ResNet-101 model also yields good results, with a Miou of 0.84928. The performance of the GCN ResNet-50 and GCN ResNet-101 models is relatively good as well, with scores of 0.78231 and 0.78443, respectively. However, the UNET model exhibits relatively poorer performance, with a Miou of only 0.68159. The MIoU scores for the DeepLabv3+and ExFuse ResNet-50 models are 0.73638 and 0.72884, respectively, slightly lower than the other models. The ExFuse ResNet-101 model achieves a Miou score of 0.78559, which is comparable to the GCN ResNet-101 model.

In summary, the MSFANet ResNet-50 and MSFANet ResNet-101 models perform the best in this dataset, while the UNET model exhibits relatively poorer performance. However, to comprehensively assess the performance of these models, other metrics such as PA and F1_Score need to be considered. The bar chart in Fig. 10 illustrates the PA and F1_Score metrics. From these data, it can be observed that the MSFANet ResNet-50 and MSFANet ResNet-101 models perform the best in terms of PA and F1_Score metrics, exhibiting high accuracy and comprehensive performance. The performance of the GCN ResNet-50 and GCN ResNet-101 models is also relatively good, approaching the performance of the MSFANet models. The UNET model performs relatively poorly in terms of PA and F1_Score, indicating lower accuracy and overall performance. The DeepLabv3+and ExFuse ResNet-50 models score slightly lower on PA and F1_Score, while the performance of the ExFuse ResNet-101 model is relatively good, approaching that of the GCN ResNet-101 model.

In conclusion, the MSFANet ResNet-50 and MSFANet ResNet-101 models perform the best across all metrics. They achieve high scores in Miou, PA, and F1_Score, as well as high scores in the mangrove (CPA) metric. This indicates their strong accuracy and overall performance in the mangrove segmentation task. The performance of other models is relatively lower.

Figure 11a displays the training and validation loss variation for the MSFANet ResNet-50 model. The loss values are sampled every 25 epochs. From the experimental data, it is evident that the validation set initially exhibits significant deviation during the early stages of training. After approximately 100 epochs, the loss curve gradually stabilizes and slowly reduce. Beyond 400 epochs, the validation loss approaches that of the training set, and both losses converge to near-zero values. This indicates that the MSFANet model fits the data well, demonstrating its excellent performance.

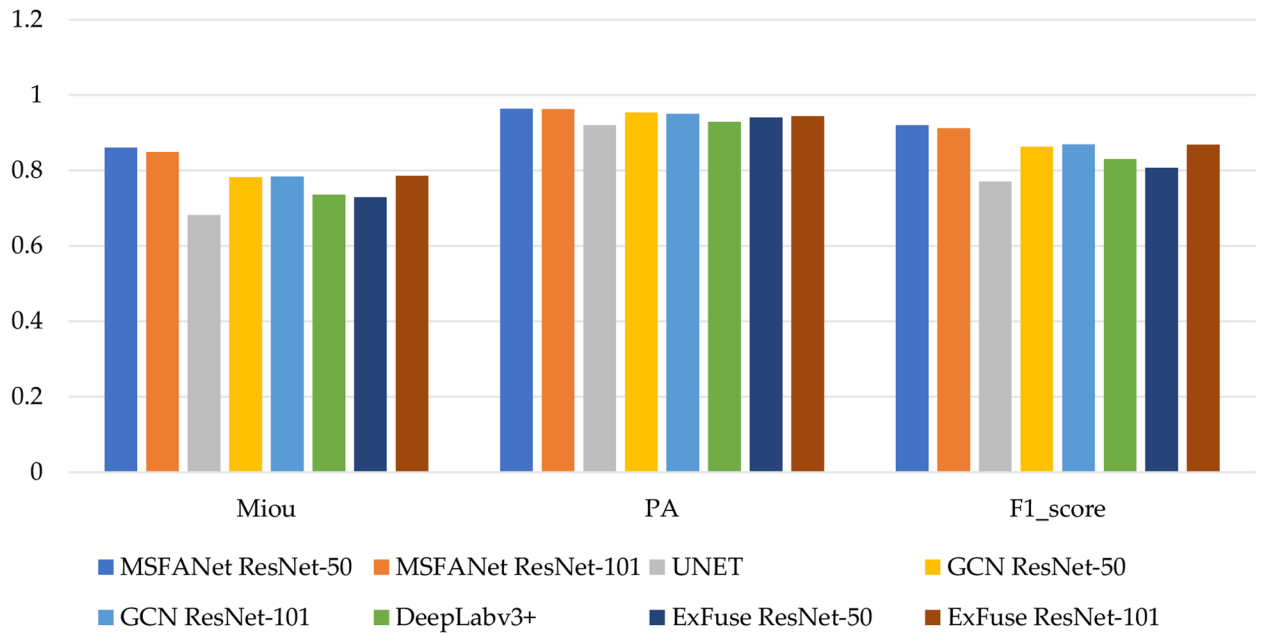

Fig. 10 Comparison of Miou, PA, and F1_score for different models

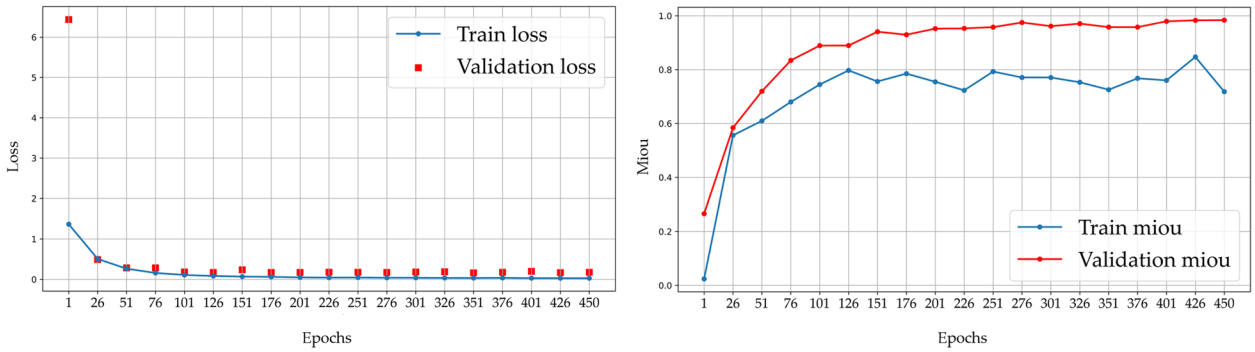

Fig. 11 Loss value and Miou value change graphs for MSFANet ResNet-50 training and verification

Figure 11b depicts the Miou values, also sampled every 25 epochs during training. The Miou values exhibit an overall upward trend during the training phase, stabilizing at over 90% towards the end. While there is some fluctuation in the Miou values during the validation phase, the overall trend remains positive. These fluctuations may be attributed to the relatively small dataset.

Figure 12 shows the visual segmentation results of the MSFANet and GCN models on four images using ResNet-50 and ResNet-101. In Fig. 12a, it is evident that MSFANet performs best on ResNet-50. It accurately classifies ponds, red mangroves, and buildings, with clear category contours. It also identifies the tidal flats accurately, although some less distinct tidal flats remain unclassified. However, on ResNet-101, there are noticeable classification errors, such as classifying dried portions of ponds as tidal flats. While it generally maintains correct category contours, its performance is slightly inferior to that of ResNet-50.

The GCN model on ResNet-50 exhibits numerous classification errors. While it performs well in the mangrove category, its pond classification is relatively poor, with unclear boundaries between the background and ponds. On ResNet-101, it similarly encounters many classification errors, misclassifying some ponds as rivers and recognizing only a small portion of the buildings. Nevertheless, its mangrove classification remains relatively accurate, most of them can be correctly identified.

In Fig. 12b, the two models with ResNet-50 as the backbone network have made significant errors in pond classification. However, these four models exhibit excellent recognition of mangroves and rivers in this image. In Fig. 12c, these four models perform well, demonstrating accurate recognition of tidal flats and rivers, indicating their strong performance in simpler scenes. It can be seen in Fig. 12d that the MSFANet ResNet-50 model has a better recognition effect on smaller tidal flats, while the performance of the other three models has little difference.

In summary, the comparative data from above different models indicate that the MSFANet ResNet-50 model exhibits the best overall performance. It accurately identifies mangroves and rivers, achieves good results in recognizing buildings and ponds, and performs well in identifying smaller targets. The MSFANet ResNet-101 model is slightly inferior to the ResNet-50 model in identifying ponds but excels in accurately identifying crucial targets like red mangroves and rivers, delivering an overall good performance. The two GCN models struggle with pond recognition, displaying several classification errors, but they do fairly well in identifying mangroves and rivers, albeit with an overall lower performance compared to the MSFANet model.

Comparison experiment of ALRFCN module ablation for different void rates

To validate the effectiveness of the ALRFCN module while ensuring experimental fairness and preserving the network structure, three alternative schemes were employed to replace the ALRFCN module in this study. Scheme one involved using three 5×5 convolutions to receive the outputs of Layer0, Layer1, and Layer2, replacing both groups (a) and (b) of modules. This was done primarily to maintain the integrity of the network structure and verify the effectiveness of the two sets of ALRFCN modules. The second option was to employ two sets of three 5×5 convolutions to replace groups (a) and (b). The main purpose of this was to assess the impact of separable convolutions on parameter quantity and the advantages of large receptive fields provided by dilated convolutions. The ResNet-50 architecture was chosen as the backbone network, while other parameters remained constant. The experimental results are presented in Table 4.

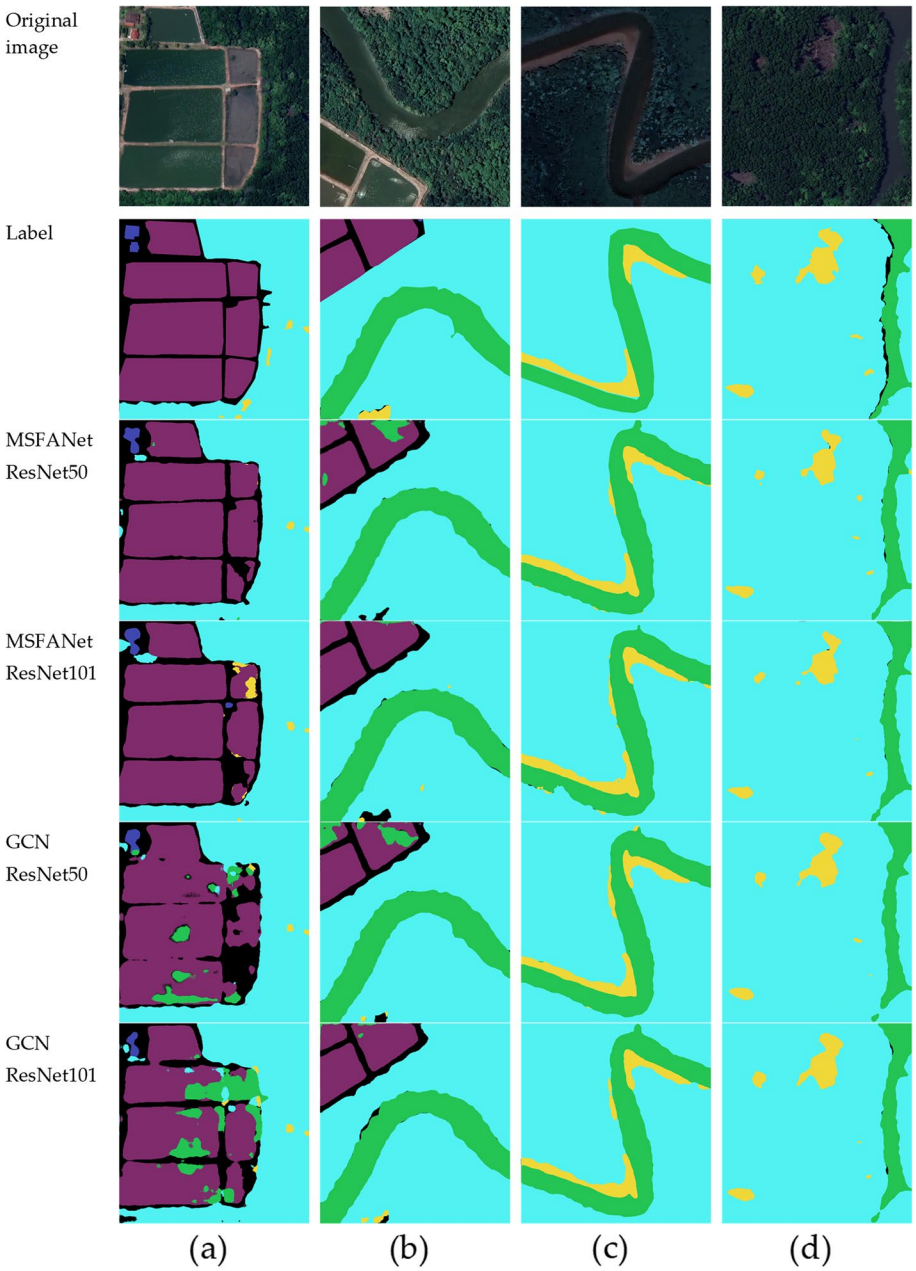

Fig. 12 Segmentation results of MSFANet and GCN models in different backbone networks

The Miou score of Scheme One was 85%, which was one percentage point lower than the ALRFCN module's 86%. The parameter count of the ALRFCN module was lower than that of scheme one. This result validates that the ALRFCN module effectively improves performance. In scheme two, with the addition of two sets of modules, Miou did not show a significant improvement, and compared to ALRFCN, the parameter quantity of the two sets of modules increased by 6 MB. Both schemes exhibited similar trends in the PA (%) indicator as Miou. The

Table 4 Experimental comparison results of ALRFCN module ablation

Scheme	Backbone network	Miou(%)	PA(%)	Parameter quantity
ALRFCN	ResNet-50	0.86015	0.96419	129 M
5 × 5 monogroup	ResNet-50	0.85064	0.95294	130 M
5 × 5 bilgroup	ResNet-50	0.85259	0.95680	135 M
GCN monogroup	ResNet-50	0.84909	0.95869	130 M

experimental results confirmed that separable convolutions can reduce parameter count and that dilated convolutions are effective for performance.

The third scheme used the GCN module, but due to the large convolutional kernel of the global receptive field, this experiment sets the convolution size of the GCN to 15×15. The experimental results show that Miou is close to 85%, and the parameter quantity has not significantly increased due to the separable convolution. Although the receptive field has increased, it is still not as effective as the ALRFCN module.

The selection of dilation rates also has an undeniable impact on the results. In this study, three different groups of dilation rates were selected to verify the optimal combination, and the results are shown in Table 5. Group 1 selected a smaller void fraction, Group 3 selected a larger void fraction, and Group 2 chose a compromise between the two extremes. From the experimental data, it can be observed that the Miou and PA scores for Group 2 were significantly better, with a Miou of 86% and a PA of 96%, outperforming the other two sets in terms of these indicators. Therefore, the optimal model in this study also happens to be the one that chose the Group 2 of dilation rates.

Discussing the computational efficiency of the proposed method

In the presented experiment, we evaluated the spending time of our method under various image sizes, as shown in Table 6. The results demonstrate a clear relationship between the image size and the corresponding time. Specifically, as the image size increases, the spending time also exhibits an upward trend. For instance, at the smallest size of 128×128, the method achieves a millisecond time of 58.84, while at the largest size of 1024×1024, the time extends to 2639.95 ms.

Discussing the computational efficiency of our proposed method, we observe that the inference time grows with larger image sizes. This indicates a trade-off between computational speed and image resolution. While our method performs efficiently for smaller image sizes, there is an increase in computational cost for larger images. It's essential to note that the method remains effective in handling a variety of image sizes, offering flexibility for diverse applications. However, for larger datasets, computational resources need to be considered, and optimization strategies may be explored to enhance efficiency.

Table 5 Experimental comparison results of different dilation rates

Group	Dilation rates for group(a) and group(b)	Backbone network	MIoU(%)	PA(%)
Group 1	(6, 3, 1) (4, 2, 1)	ResNet-50	0.84718	0.95696
Group 2	(12, 6, 3) (8, 4, 2)	ResNet-50	0.86015	0.96419
Group 3	(18, 9, 5) (12, 6, 3)	ResNet-50	0.84286	0.95985

Table 6 Spending time under different image sizes

Image size	128×128	256×256	512×512	768×768	1024×1024
Millisecond	58.84	178.52	634.27	1402.79	2639.95

Conclusion

This paper focuses on the mangrove forest in the Dongzhaigang area of Hainan Province, China, and utilizes satellite data from Google Earth to construct a dataset for mangrove samples. A multi-Scale fusion attention network MSFANet is proposed by the improved GCN model. Through training on the dataset, an optimal network model with ResNet-50 as the backbone is obtained. When compared to various classic models, MSFANet consistently outperforms them in all evaluated metrics. Moreover, visual comparisons of the segmentation results for the study area clearly demonstrate the significant superiority of the MSFANet network model over other models.

The main contribution of this paper: one is to propose the multi-scale large receptive field network, which can preserve spatial information between different scales. The large receptive field maximizes the retention of spatial information while using separable convolution to significantly reduce computational complexity. Secondly, the introduction of the attention fusion module improves fusion efficiency and improves fusion effectiveness. Simple additions or connections with a low channel number can introduce more interference information, leading to worse segmentation results due to this unnecessary data. And the attention fusion module reduces interference information. Thirdly, the dual channel design allows for the full integration of semantic and spatial information, thereby achieving better output results. The experiments on the mangrove dataset show that the model designed in this paper has a significant improvement in segmentation data compared to other models, whether in Miou, PA, F1_ Score or mangrove (CPA) are superior to other models.

In summary, MSFANet proves to be an excellent neural network model suitable for mangrove segmentation. However,there is still room for further improvement in this research. MSFANet's performance in identifying

boundary regions with complex architectural features is relatively poor, possibly due to the insufficient representation of such samples in the dataset, leading to incomplete learning. ResNet-101 is prone to overfitting too early, resulting in poor performance. Future work could involve augmenting the dataset and balancing class samples to enhance model accuracy.

Acknowledgements
Not applicable.

Institutional review board statement
Not applicable.

Authors' contributions
Conceptualization, S.W. and L.F.; methodology, S.W.; software, L.F.; validation, J.C., Z.W. and T.Z.; formal analysis, J.C.; investigation, H.C.; resources, H.C.; data curation, Y.Z.; writing—original draft preparation, L.F.; writing—review and editing, S.W., J.C. and Z.W.; visualization, L.F. and T.Z.; supervision, H.C.; project administration, S.W.; All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by National Natural Science Foundation of China (No.61966013), Hainan Natural Science Foundation of China (No.620RC602) and Hainan Provincial Key Laboratory of Ecological Civilization and Integrated Land-sea Development.

Availability of data and materials
The data presented in this study is available on request from the corresponding author.

Declarations

Ethics approval and consent to participate
All data presented in this paper are derived from publicly available data on Google Earth, containing no sensitive information and posing no adverse impact on the privacy of specific individuals or communities. The data acquisition adheres to relevant regulations and ethical standards.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Information Science and Technology, Hainan Normal University, Haikou 571158, China. [2]Smart Police College, People's Police University of China, Langfang 065000, China. [3]Fine Arts Academy, Hainan Normal University, Haikou 571158, China.

Received: 18 October 2023 Accepted: 5 December 2023
Published online: 26 January 2024

References

1. Krauss KW, Doyle TW, Doyle TJ, Swarzenski CM, From AS, Day RH, Conner WH (2009) Water level observations in mangrove swamps during two hurricanes in Florida. Wetlands 29:142–149
2. Zhang K, Liu H, Li Y, Xu H, Shen J, Rhome J, Smith TJ III (2012) The role of mangroves in attenuating storm surges. Estuar Coast Shelf Sci 102:11–23
3. Zhang X, Lin P, Gong Z, Li B, Chen X (2020) Wave attenuation by Spartina alterniflora under macro-tidal and storm surge conditions. Wetlands 40:2151–2162
4. Thampanya U, Vermaat J, Sinsakul S, Panapitukkul N (2006) Coastal erosion and mangrove progradation of Southern Thailand. Estuar Coast Shelf Sci 68:75–85
5. Guannel G, Arkema K, Ruggiero P, Verutes G (2016) The power of three: coral reefs, seagrasses and mangroves protect coastal regions and increase their resilience. PLoS One 11:e0158094
6. Li CH, Cai R, Yan X (2020) Analysis on the changes of carbon budget of mangrove wetland in Hainan Dongzhaigang during 2010–2018. Bull Mar Sci 39:488–497
7. Temmerman S, Meire P, Bouma TJ, Herman PM, Ysebaert T, De Vriend HJ (2013) Ecosystem-based coastal defence in the face of global change. Nature 504:79
8. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9:62–66
9. Yang P, Song W, Zhao X, Zheng R, Qingge L (2020) An improved Otsu threshold segmentation algorithm. Int J Comput Sci Eng 22:146–153
10. Pratiwia NMD, Widiarthaa IM (2021) Mangrove ecosystem segmentation from drone images using otsu method. Jurnal Elektronik Ilmu Komputer Udayana p-ISSN 2301:5373
11. Rong W, Li Z, Zhang W, Sun L (2014) An improved CANNY edge detection algorithm. In: Proceedings of the 2014 IEEE international conference on mechatronics and automation. pp 577–582
12. Yang L, Wu X, Zhao D, Li H, Zhai J (2011) An improved Prewitt algorithm for edge detection based on noised image. In: Proceedings of the 2011 4th International congress on image and signal processing. pp 1197–1200
13. Gao W, Zhang X, Yang L, Liu H (2010) An improved Sobel edge detection. In: Proceedings of the 2010 3rd International conference on computer science and information technology. pp 67–71
14. Xue Y, Zhao J, Zhang M (2021) A watershed-segmentation-based improved algorithm for extracting cultivated land boundaries. Remote Sensing 13:939
15. Dong-yang Y, Dong-ping M (2017) Object-oriented remote sensing image segmentation based on automatic multiseed region growing algorithm. Chin J Eng 39:1735–1742
16. Wang T (2021) Segmentation of cervical cell cluster by multiscale graph cut algorithm. In; Proceedings of the Business Intelligence and Information Technology: Proceedings of the International Conference on Business Intelligence and Information Technology BIIT 202. pp 131–140
17. Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587
18. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the Proceedings of the European conference on computer vision (ECCV). pp 801–818
19. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2881–2890
20. Yang M, Yu K, Zhang C, Li Z, Yang K (2018) Denseaspp for semantic segmentation in street scenes. In: Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3684–3692
21. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA (2020) Shortcut learning in deep neural networks. Nat Mach Intell 2:665–673
22. Yuan Y, Wang J (2018) Object context network for scene parsing
23. Li X, Zhong Z, Wu J, Yang Y, Lin Z, Liu H (2019) Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision. pp 9167–9176
24. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 3146–3154
25. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision. pp 603–612
26. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Learning a discriminative feature network for semantic segmentation. In: Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1857–1866

27. Elhassan MA, Huang C, Yang C, Munea TL (2021) DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes. Expert Syst Appl 183:115090
28. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the Proceedings of the European conference on computer vision (ECCV). pp 325–341
29. Fan Y, Zeng Q, Mei Z, Hu W (2022) Semantic segmentation for mangrove using spectral indices and self-attention mechanism. In: Proceedings of the 2022 7th International Conference on Signal and Image Processing (ICSIP). pp 436–441
30. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y (2022) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45:87–110
31. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y et al (2021) Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306
32. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European conference on computer vision. pp 205–218
33. Hong D, Han Z, Yao J, Gao L, Zhang B, Plaza A, Chanussot J (2021) SpectralFormer: rethinking hyperspectral image classification with transformers. IEEE Trans Geosci Remote Sens 60:1–15
34. Zhong Z, Li Y, Ma L, Li J, Zheng W-S (2021) Spectral–spatial transformer network for hyperspectral image classification: a factorized architecture search framework. IEEE Trans Geosci Remote Sens 60:1–15
35. Sun L, Zhao G, Zheng Y, Wu Z (2022) Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–14
36. Wang W, Liu L, Zhang T, Shen J, Wang J, Li J (2022) Hyper-ES2T: efficient spatial–spectral transformer for the classification of hyperspectral remote sensing images. Int J Appl Earth Obs Geoinf 113:103005
37. Yang X, Cao W, Lu Y, Zhou Y (2022) Hyperspectral image transformer classification networks. IEEE Trans Geosci Remote Sens 60:1–15
38. Ronneberger O, Fischer P, Brox T (2022) Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Conference Proceedings
39. Dong Y, Yu K, Hu W (2021) GC-UNet: an improved UNet model for mangrove segmentation using Landsat8. In: Proceedings of the The 2021 3rd International Conference on Big Data Engineering. pp 58–63
40. de Souza Moreno GM, de Carvalho Júnior OA, de Carvalho OLF, Andrade TC (2023) Deep semantic segmentation of mangroves in Brazil combining spatial, temporal, and polarization data from Sentinel-1 time series. Ocean Coastal Management 231:106381
41. Ding X, Zhang X, Han J, Ding G (2022) Scaling up your kernels to 31x31: revisiting large kernel design in cnns. In: Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 11963–11975
42. Bhatti UA, Tang H, Wu G, Marjan S, Hussain A (2023) Deep learning with graph convolutional networks: an overview and latest applications in computational intelligence. Int J Intell Syst 2023:1–28
43. Fran C (2017) Deep learning with depth wise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
44. Wang J, Zheng Z, Ma A, Lu X, Zhong Y (2021) LoveDA: a remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv preprint arXiv:2110.08733

Publisher's Note