frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 09 April 2025
DOI 10.3389/fmars.2025.1535917

Check for updates

OPEN ACCESS

EDITED BY
Dongjie Fu,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Bolin Fu,
Guilin University of Technology, China
Tri Soeprobowati,
Diponegoro University, Indonesia

*CORRESPONDENCE
Shulei Wu
wsl@hainnu.edu.cn
Yukai Chen
chenyukai@hainnu.edu.cn

RECEIVED 28 November 2024
ACCEPTED 10 March 2025
PUBLISHED 09 April 2025

CITATION
Fu L, Wang Y, Wu S, Zhuang J, Wu Z, Wu J, Chen H and Chen Y (2025) TCCFNet: a semantic segmentation method for mangrove remote sensing images based on two-channel cross-fusion networks.
Front. Mar. Sci. 12:1535917.
doi: 10.3389/fmars.2025.1535917

COPYRIGHT
© 2025 Fu, Wang, Wu, Zhuang, Wu, Wu, Chen and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

TCCFNet: a semantic segmentation method for mangrove remote sensing images based on two-channel cross-fusion networks

Lixiang Fu[1], Yaoru Wang[1], Shulei Wu[1]*, Jiasen Zhuang[1], Zhongqiang Wu[1], Jian Wu[2], Huandong Chen[1] and Yukai Chen[3]*

[1]School of Information Science and Technology, Hainan Normal University, Haikou, China, [2]China Mobile Group Hainan Company Limited, Haikou, China, [3]School of Life Sciences, Hainan Normal University, Haikou, China

Mangrove ecosystems play a crucial role in coastal environments. However, due to the complexity of mangrove distribution and the similarity among different categories in remote sensing images, traditional image segmentation methods struggle to accurately identify mangrove regions. Deep learning techniques, particularly those based on CNNs and Transformers, have demonstrated significant progress in remote sensing image analysis. This study proposes TCCFNet (Two-Channel Cross-Fusion Network) to enhance the accuracy and robustness of mangrove remote sensing image semantic segmentation. This study introduces a dual-backbone network architecture that combines ResNet for fine-grained local feature extraction and Swin Transformer for global context modeling. ResNet improves the identification of small targets, while Swin Transformer enhances the segmentation of large-scale features. Additionally, a Cross Integration Module (CIM) is incorporated to strengthen multi-scale feature fusion and enhance adaptability to complex scenarios. The dataset consists of 230 high-resolution remote sensing images, with 80% used for training and 20% for validation. The experimental setup employs the Adam optimizer with an initial learning rate of 0.0001 and a total of 450 training iterations, using cross-entropy loss for optimization. Experimental results demonstrate that TCCFNet outperforms existing methods in mangrove remote sensing image segmentation. Compared with state-of-the-art models such as MSFANet and DC-Swin, TCCFNet achieves superior performance with a Mean Intersection over Union (MIoU) of 88.34%, Pixel Accuracy (PA) of 97.35%, and F1-score of 93.55%. Particularly, the segmentation accuracy for mangrove categories reaches 99.04%. Furthermore, TCCFNet excels in distinguishing similar categories, handling complex backgrounds, and improving boundary detection. TCCFNet demonstrates outstanding performance in mangrove remote sensing image segmentation, primarily due to its dual-backbone design and CIM module. However, the model still has limitations in computational efficiency and small-target recognition. Future research could focus on developing lightweight Transformer architectures, optimizing data augmentation strategies, and expanding the dataset to diverse remote sensing scenarios to further enhance generalization capabilities. This study presents a novel mangrove remote sensing image segmentation approach—TCCFNet. By

integrating ResNet and Swin Transformer with the Cross Integration Module (CIM), the model significantly improves segmentation accuracy, particularly in distinguishing complex categories and large-scale targets. TCCFNet serves as a valuable tool for mangrove remote sensing monitoring, providing more precise data support for ecological conservation efforts.

KEYWORDS

mangrove remote sensing image segmentation, TCCFNet, CIM, multi-scale feature fusion, mangrove

1 Introduction

Mangrove is one of the most ecologically valuable ecosystems in nature (Kathiresan, 2021), known as the "natural coastal protection project" and the "cradle of the ocean". These ecosystems play critical roles in maintaining coastline stability, water purification, carbon sequestration and mitigating wave-related natural disasters (Gijón Mancheño et al., 2024; Rahman et al., 2024; Wang et al., 2024; Sun et al., 2025). Mangrove ecosystems harbor exceptional biodiversity through their complex structural characteristics. Their vegetation structure shows remarkable diversity (Anniwaer et al., 2024; Bonet et al., 2024; Liu et al., 2024), creating vital habitats that support numerous fish, bird, and other animal species (Mohamed et al., 2024; Tasneem et al., 2024). This rich biodiversity underscores the irreplaceable ecological value of mangrove ecosystems. The growing recognition of mangroves' ecological importance has led to increased conservation efforts worldwide (Nijamdeen et al., 2024). A key component of these conservation initiatives is the effective monitoring and assessment of mangrove distribution and health.Image segmentation can segment mangrove images into different areas or objects, thus providing information on mangrove distribution, area change, growth and destruction. However, the traditional image segmentation method has some unfavorable trends in the case of mangrove growth (He, 2022; Fu et al., 2024). First of all, traditional methods often rely on hand-designed features, and it is difficult to accurately divide different areas such as trees, sediment and water in the complex and changeable mangrove environment. In addition, the traditional method has poor robustness to noise, which is difficult to adapt to the low resolution and noise interference in remote sensing images, and can not well meet the needs of large-scale and long-term monitoring of mangrove growth. Deep learning image segmentation method (Mei et al., 2025) has significant advantages in image segmentation. It can automatically learn complex features in the image, no need to manually design features, more adaptable. Therefore, deep learning image segmentation provides an important technical means for mangrove research. Deep learning models, such as convolutional neural networks (CNN) and architectures derived from them, such as U-Net (Azad et al., 2024), SegNet (Li et al., 2024), DeepLabv3+ (Wang et al., 2024), etc., can automatically learn multi-level features in images, reduce the dependence on manual feature engineering, and better capture complex spatial patterns and semantic information. So you can perform well in different scenarios and conditions.

However, CNN's inherent limitations still lead to deficiencies in handling some task scenarios:

1. The CNN model does not perform well in handling long-distance dependencies, because its receptive field is limited by the size and number of layers of the convolution kernel, resulting in certain limitations in capturing global context information;
2. When the CNN model deals with scale changes and complex backgrounds, it may lose or over-smooth the detailed information, resulting in decreased segmentation accuracy.

To solve the first problem, Transformer architecture is one of the mainstream research hotspots for feature extraction. Transformer architecture itself is a method based on multi-head self-attention mechanism. Li et al. (Li et al., 2024) introduced an efficient end-to-end architecture called CNSST (Convolutional Network and Spectral Space Transformer). This approach combines CNN's strength in local feature extraction with Transformer's global modeling capabilities to enhance hyperspectral image classification. The CNSST architecture features two main components. First, a 3D-CNN based network performs hierarchical feature fusion to capture local spectral-spatial relationships. Second, a spectral spatial transformer incorporates inductive bias to establish global dependencies while preserving important local information. The researchers validated their method through extensive testing on multiple public hyperspectral datasets. Results consistently demonstrated that CNSST outperforms existing state-of-the-art approaches in classification accuracy. Yang et al. (Yang et al., 2025) introduced the Adaptive Coupling Module (ACM), which combines CNN for local context and Transformer for global dependencies in hyperspectral image processing. The Adaptive Response Fusion Module (ARFM) merges these representations across resolutions, while cosine similarity constraints preserve both aspects during mutual learning. Experiments on three public datasets show ACTN outperforms state-of-the-art CNN and

Transformer methods. The DC-Swin model designed by Wang et al. (Wang et al., 2022) puts forward a new semantic segmentation scheme that uses Swin Transformer(Shifted Window Transformer) as the encoder. A Densely Connected Feature Aggregation Module (DCFAM) is designed as a decoder to restore resolution and generate segmentation maps. The experimental results show that The DC-Swin model performs well on remote sensing semantic segmentation data sets.

As for the second problem, usually the problem of using attention mechanism to improve information loss, the research on attention mechanism is also very rich. BANet (Tsai et al., 2022) introduced the Blur-aware Attention Mechanism, which is used to improve the accuracy of image segmentation in dynamic scenes. By combining multi-scale features and attention mechanisms, this method can better capture important information in images, enhance the ability to capture target edges and details, and thus improve the accuracy and robustness of segmentation. A^2-FPN (MaL et al., 2021) proposes a more efficient multi-head self-attention mechanism in terms of attention mechanisms to reduce computational complexity. At the same time, the paper discusses the combination of attention mechanism and convolutional neural network (CNN) to improve the effect of image feature extraction and improve the overall performance of the model. Tang et al. (Ma, 2023) studied the application of attention mechanism in image fusion. Specifically, they use attention mechanisms to selectively highlight key features, ensuring that important semantic information is retained during the fusion process. Through the progressive semantic injection strategy, the model can gradually introduce and fuse semantic features from infrared and visible images, thereby improving the scene fidelity and task performance of the fused images.

Based on the above situation, this paper proposes a method based on two-channel cross-fusion network (TCCFNet). TCCFNet addresses these limitations through several key innovations. Its dual backbone architecture represents a significant departure from previous hybrid models that treat CNN and Transformer paths equally. Instead, TCCFNet implements an asymmetric design specifically optimized for mangrove feature extraction, balancing local detail capture through ResNet with global context modeling via Swin Transformer. This design specifically addresses the multi-scale nature of mangrove ecosystems, where both fine-grained texture and broad spatial patterns carry important information. In order to highlight the key points and improve the recognition rate of the model for large target areas such as mangroves, rivers, oceans and large tidal flats, TCCFNet adopts the following strategies without neglecting small targets:

1. The model uses a dual backbone network structure. It incorporates two key components: the ResNet residual network and the Swin Transformer. This design leverages the complementary advantages of both networks. ResNet is good at capturing local details, which is conducive to the identification of small targets. Swin Transformer significantly improves the identification accuracy of large target areas such as mangroves, rivers and oceans through its advantages of hierarchical structure, multi-scale feature extraction, sliding window mechanism and self-attentional global feature modeling. It is able to capture global information while maintaining detail, which is particularly effective for dealing with large and complex natural landscapes. The combination of the two can effectively improve the sensitivity of the model to different target regions, so as to improve the recognition accuracy of the model under complex background.
2. In order to further enhance the feature expression capability of the model, a Cross Integration Module (CIM) is designed in this paper. The module first receives as input feature maps from different levels of ResNet and Swin Transformer. In the process of processing, the deeper feature map is first sampled to the same size as the shallow feature map, and then the features are channelled. Then it is processed by two parallel branches: one branch uses global average pooling (GAP) to extract the channel attention information, and adjusts the channel weight by convolutional layer and activation function; The other branch uses SlimConv (consisting of two depth-separable convolution) for feature extraction.GAP is a technique for reducing the spatial dimension by calculating the average of each feature map. This action helps capture global context information and reduces computational complexity. SlimConv is a lightweight convolution operation consisting of two deep convolution layers. This design significantly reduces computational requirements compared to standard convolution operations, while maintaining efficient feature extraction capabilities. The output of these two branches is then fused with the original feature map. Finally, CFFM (Cross Feature Fusion Module) is used to fuse all the processed feature maps and output a unified size feature map. CFFM is a specialized component that can combine and process feature mappings from different network layers. CFFM systematically integrates spatial and channel information to generate a more comprehensive feature representation. This design enhances the model's understanding of image details and global semantics through cross-attention mechanism and multi-level feature fusion, and effectively improves the segmentation accuracy.

To sum up, this paper proposes TCCFNet, which can effectively identify mangrove areas and improve the model's performance in mangrove remote sensing image semantic segmentation by adopting dual backbone network structure, combining the advantages of ResNet residual network and Swin Transformer, and the feature fusion capability of cross-integration module.

2 Materials

2.1 Research area

The first research area is located in Hainan Dongzhaigang National Nature Reserve (100°32 '-100° 37' E, 19°51' -20° 1' N) shown in Figure 1. With a total area of 3337.6 hectares, the reserve is the largest coastal mangrove forest in China. It includes 35 species in 18 families, including 24 true mangrove species in 10 families and 11 semi-mangrove species in 8 families. It is the first national mangrove nature reserve in Hainan Province. The second study area is Qinglan Port Nature Reserve (110°45 '-110°47' E, 19°34 '-19° 37' N), which is the area with the largest number of mangrove species in China. There are 24 species of true mangrove and 10 species of semi-mangrove, accounting for 88.89% and 100% of the total species of true mangrove and semi-mangrove in China respectively.

2.2 Data

The data set used in this article is the same as that used in the MSFANet (Fu et al., 2024) network. The proportion of each category in the collected data set is shown in Figure 2. The three types of features, mangroves, rivers and oceans, and tidal flats, account for nearly 90% of the total, indicating the dominance of these large targets in the dataset.

This study draws on a dataset of 230 high-resolution remote sensing images. Following standard machine learning practices, we divided the dataset into two parts: 184 images (80%) for model training and 46 images (20%) for validation. The image size is set to 768 pixels, and the mangrove image data source is from Google Maps and consists of a high spatial resolution (0.3 m) remote sensing image taken on June 6, 2021.

3 Methods

3.1 Attention Aggregation based Feature Pyramid Network

Attention Aggregation based Feature Pyramid Network (A^2-FPN) is a feature pyramid network based on attention aggregation. This model mainly reduces the computational complexity by introducing a more efficient multi-head self-attention mechanism. It combines the attention mechanism with convolutional neural network (CNN) to improve the effect of image feature extraction, and aggregates different levels of features through the improved multi-head attention mechanism, thus improving the overall performance of the model.

3.2 Adaptive Bezier Curve Network

Adaptive Bezier Curve Network (ABCNet) is a text detection model based on deep learning, which is especially suitable for detecting curved text in images. It is proposed to solve the problem of curve text detection in natural scenes, and can maintain high precision detection in complex background and irregular shape text. ABCNet can perform end-to-end training, directly from image input to detection result output, requiring manual design of complex post-processing steps. This architecture not only simplifies the training process, but also performs better in precision and speed.

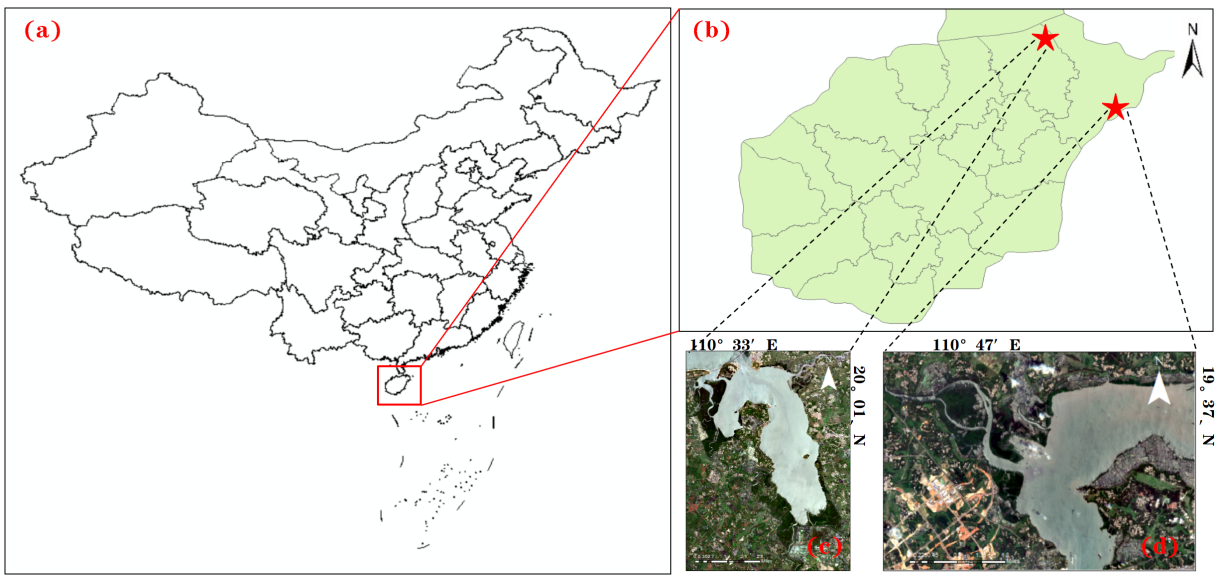

FIGURE 1
Study area diagram. **(a)** China; **(b)** Hainan Province; **(c)** Hainan Dongzhaigang National Nature Reserve; **(d)** Qinglan Port.

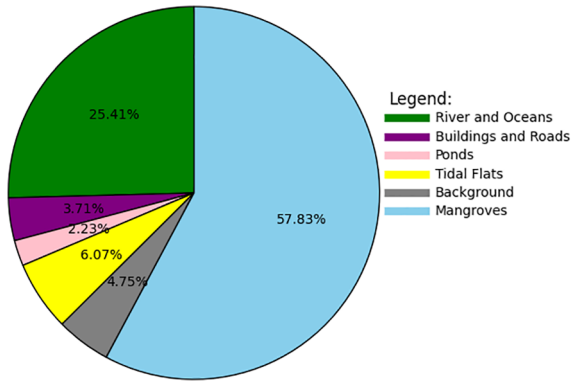

FIGURE 2
Category distribution of mangrove remote sensing image segmentation dataset.

3.3 Boundary-Aware Network

Boundary-aware Network (BANet) is a deep learning model for object segmentation and Boundary detection, which is mainly used to improve the boundary accuracy in image segmentation tasks. By introducing the boundary sensing module, BANet makes the model pay more attention to the boundary region of the target, so as to improve the detail processing ability of segmentation results.

3.4 Dual-channel Shifted Window Transformer

Dual-channel Shifted Window Transformer (DC-Swin) is an improved model based on Swin Transformer designed to enhance the performance of Transformer in computer tasks such as image visual recognition and segmentation. DC-Swin introduced a two-channel structure and innovated Transformer's window mechanism in Swin to be more efficient and accurate in capturing global and local features.

3.5 UNetFormer

UNetFormer is a deep learning model combining U-Net and Transformer, which is mainly used for image segmentation tasks, especially for high-precision segmentation of medical images and remote sensing images. UNetFormer combines U-Net's multi-scale feature extraction capabilities with Transformer's global attention mechanism, making the model excellent at capturing detail and long-distance dependencies.

3.6 Multi-Scale Feature Aggregation Network

Multi-Scale Feature Aggregation Network (MSFANet) is a deep learning model designed for image segmentation, especially for fine segmentation in complex scenes. Through multi-scale feature aggregation and attention mechanism, MSFANet enhances the model's adaptability to objects at different scales, and can better process details and boundaries in images.

3.7 Proposed method

3.7.1 Two Channel Cross Fusion Network

TCCFNet consists of two parts: the ResNet residual network on the left and the Swin Transformer architecture on the right, as shown in Figure 3.

The TCCFNet architecture consists of two parallel processing paths: a ResNet-based path on the left and a Swin Transformer path on the right. Both paths begin with the same input image of size $H \times W \times 3$.

The left path employs a ResNet structure similar to MSFANet, processing the input through four sequential layers (Layer 0-3). At each layer, the feature map undergoes downsampling, reducing its spatial dimensions by half. These features then pass through convolutional layers (conv1-conv4), which adjust the channel dimensions to match the number of target categories (N).

The right path implements a lightweight version of Swin Transformer ("SWin-T") for efficient feature processing. This path divides its processing into four stages (Stage 1-4). Initially, the input image is partitioned into 16 non-overlapping patches per channel, creating 48 patches total for the three-channel input. The first stage performs Linear Embedding, converting these 48 patches into 96 feature vectors, which then proceed through the Swin Transformer Block.

In Stages 2, 3, and 4, Patch Merging operations progressively combine patches while doubling the channel count. After each stage, a convolutional layer (conv5-conv8) adjusts the output features to match the dimensions of the corresponding left path features.

The two paths merge through the Cross Integration Module (CIM), which enables effective information exchange between different scales of features from both paths. The final processing stage generates feature maps at 1/8, 1/4, and 1/2 scales of the original input. These undergo successive upsampling and channel fusion operations, denoted by 'C' in the architecture diagram, shown in Figure 4. A final upsampling step produces the segmented image at the original resolution.

3.7.2 Swin Transformer Block

Figure 5 shows the network structure of the Swin Transformer Block. By using the W-MSA and MLP (Multilayer Perceptron) modules alternately, and combining with the Layer Normalization (LN) layer, the module can effectively capture the local and global information of images and achieve efficient extraction of image features. Among them, the W-MSA module is responsible for feature interaction in the local window, while the SW-MSA module is responsible for feature interaction across the window. In this way, the model can effectively capture the global information of the image and enhance the understanding of image features. On the basis of the self-attention mechanism, MLP module further

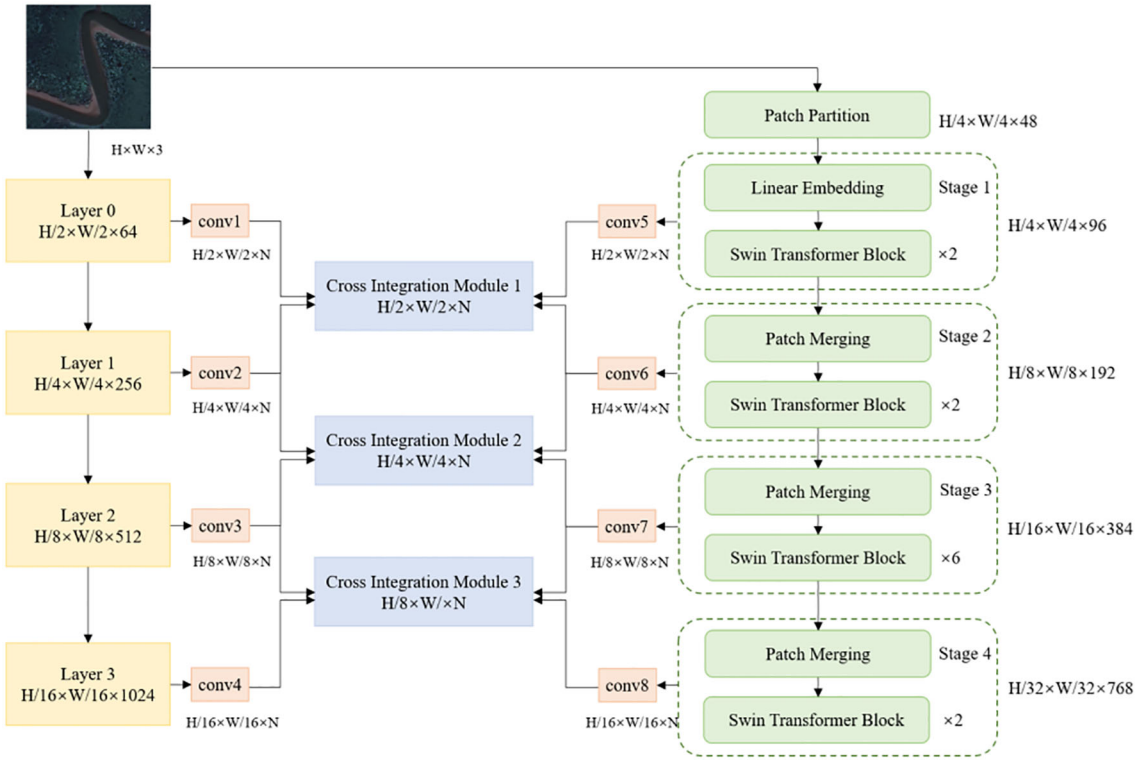

FIGURE 3
TCCFNet architecture.

enhances the expressive ability of the model through nonlinear transformation. LN layer can improve the training stability and help the model converge faster by normalizing the input features. Swin Transformer Block calculation process is shown in Equation 1.

$$\begin{aligned} Z &= W - MSA(LN(z^{l-1})) + z^{l-1} \\ \hat{z}^{l} &= MLP(LN(\hat{z}^{l})) + \hat{z}^{l} \\ \hat{Z}^{l+1} &= SW - MSA(LN(z^{l})) + z^{l} \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{1}$$

Where, $\hat{z}^{l}$ and z^{l} represent the output characteristics of W-MSA and SW-MSA modules and MLP modules respectively, and l represent the number of layers of Swin Transformer Block.

In summary, by combining the advantages of residual neural networks ResNet and Swin Transformer, the TCCFNet model uses multi-scale feature fusion and attention mechanism to improve the semantic segmentation performance of mangrove remote sensing images.

3.7.3 Cross Integration Module

CIM is a specialized neural network component that combines and enhances feature information from multiple network layers. The module acts as a bridge between different parts of the network, enabling more efficient information flow and feature refinement.The CIM module can be used to integrate feature information from different levels to improve the model's ability of target recognition. The CIM module consists of two parts: the attention mechanism part and the feature fusion part. Figure 6 shows the architecture of the cross-integration module CIM.

The attention mechanism component processes feature maps from multiple network layers. It takes inputs F1 and F3 from convolutional neural networks, and F2 and F4 from Swin Transformer. Before processing, F3 and F4 are upsampled to match the dimensions of F1 and F2. The processing follows these steps:

1. F1 and F2 undergo concatenation to create a 2C×H×W feature map;
2. Global Average Pooling (GAP) extracts channel-wise global information, producing a C-dimensional vector;
3. Sequential processing through convolution layers and ReLU activation adjusts the channel count to C;
4. A Sigmoid activation function computes channel weights;
5. These weights are applied to the original feature map to emphasize important channel information;
6. The weighted feature map is combined with F2 to produce feature map F6.

In a parallel path, the 2C×H×W feature map passes through SlimConv convolution, outputting a C×H×W feature map. This path follows similar processing steps to generate feature map F7. F3 and F4 undergo identical processing after upsampling. This dual-path approach enables effective fusion of features from different network layers, enhancing model performance.

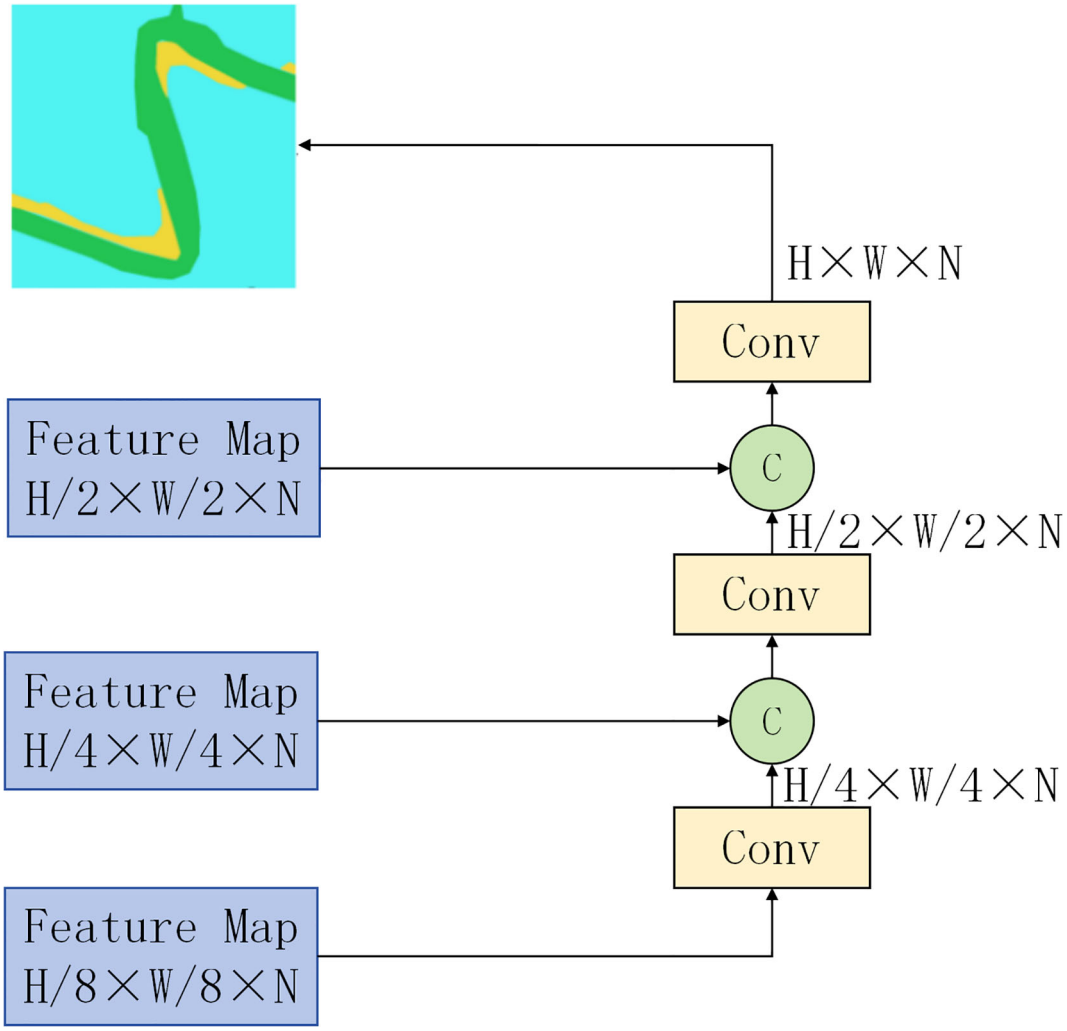

FIGURE 4
TCCFNet output structure diagram.

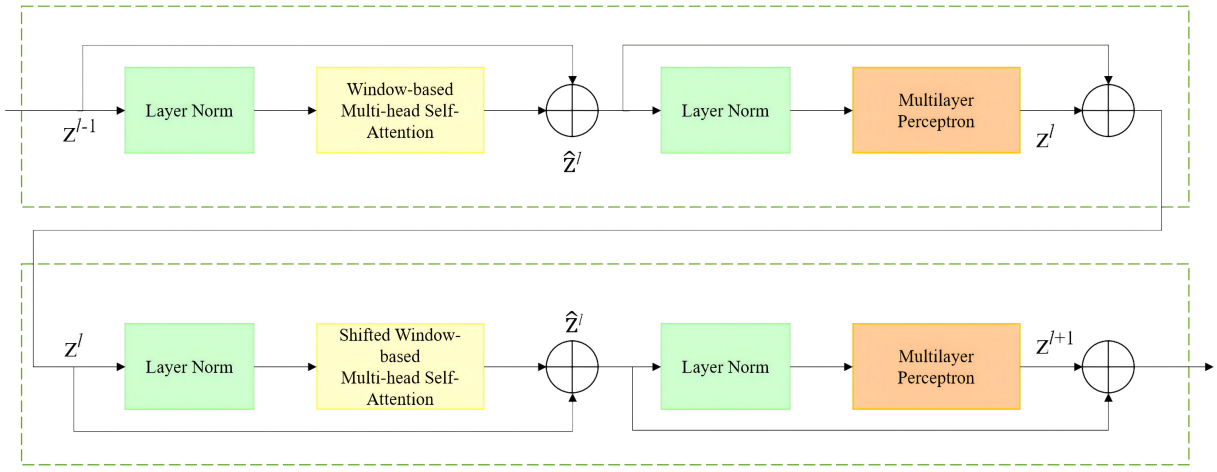

FIGURE 5
Swin transformer block.

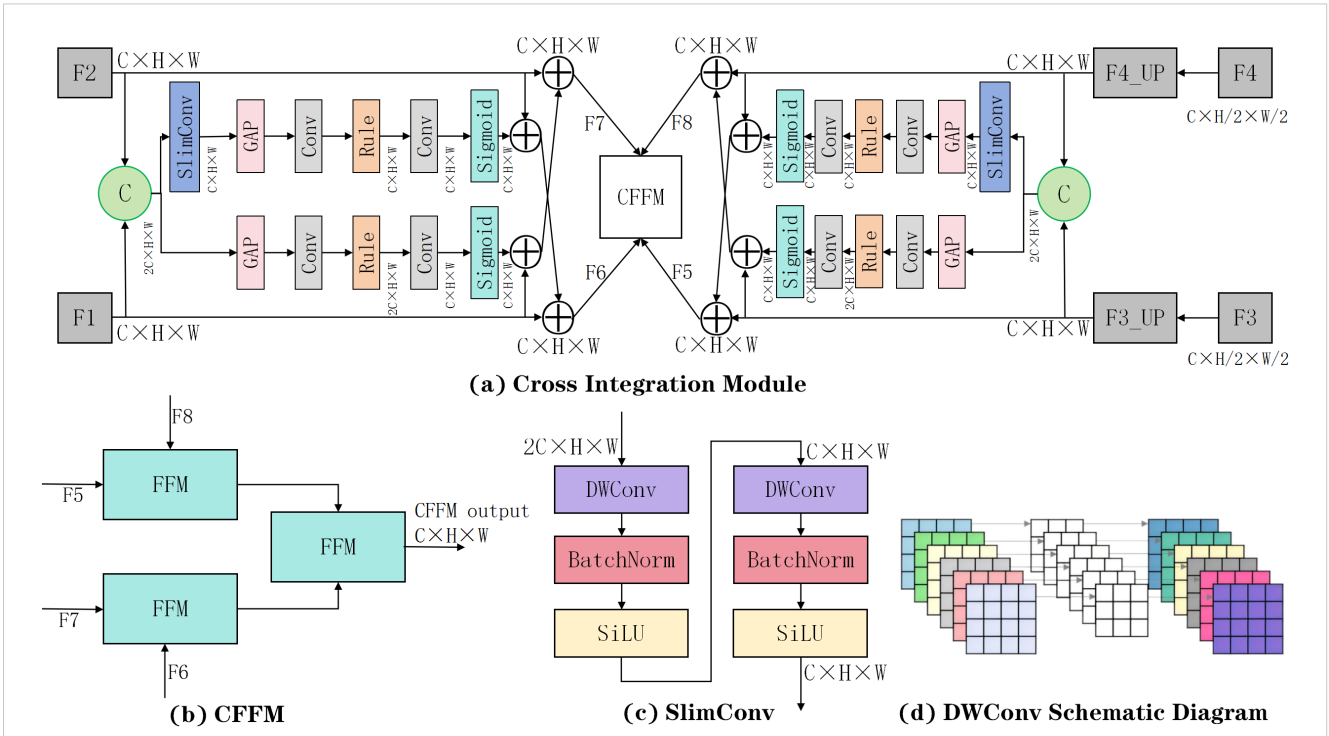

FIGURE 6
CIM architecture diagram. **(a)** Cross Integration Module; **(b)** CFFM; **(c)** SlimConv; **(d)** DWConv Schematic Diagram.

The feature fusion component uses the Cross Feature Fusion Module (CFFM) to combine feature maps F5, F6, F7, and F8. The CFFM consists of three FFM modules: one processing F5 and F8, another processing F6 and F7, and a final module combining their outputs into a C×H×W feature map.

The SlimConv module employs two Depthwise Convolution layers for efficient feature processing. It transforms a 2C×H×W input feature map into a C×H×W output map. This design offers significant computational advantages over standard convolution operations. The first convolution layer extracts local features and reduces channel dimensions, while the second layer refines these features to produce the final output map.

3.8 Experimental setup and evaluation method

In this paper, mangrove remote sensing images are used to segment the dataset, and the dataset is divided into a training set and a validation set with a ratio of 8:2. During the training process, a learning rate strategy with an initial learning rate of 0.0001 and 40% reduction every 80 iterations was adopted, and a total of 450 iterations were carried out. In the experiment, a stable Adam optimizer was used, and the cross-entropy loss function was used as the training objective to minimize the difference between the model prediction results and the real labels.

Model performance evaluation indicators include mean crossover ratio (MIoU), pixel accuracy (PA), F1-Score, and category pixel accuracy to fully evaluate the segmentation capability of the model. The calculation methods of the evaluation metrics used in this study, including MIoU, PA, F1-score, and class-specific accuracies, are detailed in Equations 1–6. These equations form the foundation for the quantitative assessment of model performance. The specific calculation formula is shown in Equations 2-6 respectively.

$$\mathrm{MIoU} = \frac{1}{k}\sum_{i=1}^{k}\frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN})} \tag{2}$$

$$\mathrm{PA} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{3}$$

$$\mathrm{F1} - \mathrm{score} = \frac{2\tilde{n}\mathrm{Precision}\tilde{n}\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{4}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

The four basic elements involved are: true cases (TP), which are positive samples that are correctly predicted by the model to be positive; True negative example (TN), that is, a negative sample that is correctly predicted by the model to be a negative class; False positive examples (FP), that is, negative samples incorrectly predicted by the model to be positive; False negative cases (FN), i.e. positive samples that are incorrectly predicted by the model as negative classes.

4 Results

In order to illustrate the advantages of TCCFNet model, we selected five models with excellent performance in recent years, including A^2-FPN, ABCNet Wang, 2020, BANet, DC-Swin and

UNetFormer, and compared them with MSFANet model. As shown in Table 1. The corresponding optimum values of the parameters have been shown in bold.

Performance evaluation of TCCFNet revealed superior results across all key metrics when compared to the six benchmark models. Through comprehensive testing, we assessed the model's capabilities using three critical performance indicators: mean intersection ratio (MIoU), pixel accuracy (PA), and F1-score. These metrics provide a thorough evaluation of the model's effectiveness in mangrove remote sensing image segmentation.

Quantitative analysis shows that TCCFNet achieved an MIoU of 88.34%, a significant improvement over existing approaches. This result notably surpasses both MSFANet (86.02%) and DC-Swin (86.26%), which previously represented the state-of-the-art performance in this domain. The enhanced MIoU demonstrates TCCFNet's superior ability to balance class intersection and union ratios, a crucial factor in accurate image segmentation.

The model exhibits exceptional capabilities in handling complex scenarios and boundary detection. TCCFNet successfully processes challenging scenes and accurately distinguishes boundaries between different feature types. This ability is particularly valuable in mangrove remote sensing applications, where precise feature delineation is essential for accurate analysis.

Additional performance metrics further validate TCCFNet's effectiveness. The model achieved a pixel accuracy of 97.35% and an F1-score of 93.55%, surpassing all comparative models. These results demonstrate TCCFNet's comprehensive excellence in both pixel-level classification accuracy and overall segmentation performance, establishing it as a robust solution for mangrove remote sensing image segmentation tasks.

In order to analyze the advantages of TCCFNet in more depth, we further compared the performance of the various models in different categories, and the following points can be drawn:

1. Mangrove class accuracy: The pixel accuracy (PA) of TCCFNet in mangrove class reached 99.04%, which was significantly better than other models. This result shows that TCCFNet can accurately identify and segment mangrove areas, greatly reducing the phenomenon of missing and mismarking.
2. Robustness of the beach class: the pixel accuracy (PA) of TCCFNet in the beach class reaches 90.25%. The tidal flat area is difficult to identify because of its complex texture, fuzzy boundary and easy to be confused with the surrounding ocean and mangrove areas. TCCFNet's outstanding performance in this category demonstrates its ability to accurately segment targets under complex background conditions.
3. Advantages of model structure: TCCFNet's dual backbone network structure and cross-integration module design enable it to have a stronger performance in capturing image details and global semantic information. This design effectively suppresses the interference of background noise, thus improving the performance of the model on the whole index.

Figure 7 shows the comparison between TCCFNet model and DC-Swin, A2FPN and UNetFormer model in six types of ground object segmentation results. Four representative pictures were selected for analysis in the experiment. The six types of features are: background, mangroves, rivers and oceans, buildings and roads, ponds and beaches. The results show that the TCCFNet model performs well in the recognition of these six types of objects, and has advantages in distinguishing similar types, dealing with the size difference of objects, and coping with the class imbalance.

First, the TCCFNet model performed well in identifying mangrove, river, and tidal flat areas. In the first row of images, the TCCFNet model accurately identified all three areas, and was particularly accurate in the division of the river. In contrast, the other three models have varying degrees of spacing when dividing the river, which indicates that the TCCFNet model has a better understanding of the target as a whole. In addition, the color of the red box area in the original image is similar to that of the mangrove area, which is prone to category errors. Both A^2-FPN and UNetFormer models showed such misclassification in the

TABLE 1 Experimental results of different models on mangrove dataset.

Model	MIoU(%)	PA(%)	F1-score(%)	PA(%)		
				Mangrove	River and ocean	Mudflat
A^2FPN	84.48	96.11	91.20	98.48	95.81	87.29
ABCNet	81.57	95.95	88.94	98.13	**96.48**	82.20
BANet	79.49	93.67	87.23	98.72	96.07	82.09
DC-Swin	86.26	96.83	92.42	98.88	96.01	89.69
UNetFormer	80.77	95.13	88.27	98.83	94.78	83.41
MSFANet	86.02	96.42	92.03	**99.80**	96.44	83.61
TCCFNet	**88.34**	**97.35**	**93.55**	99.04	95.29	**90.25**

Bold values represent the optimal performance for each metric.

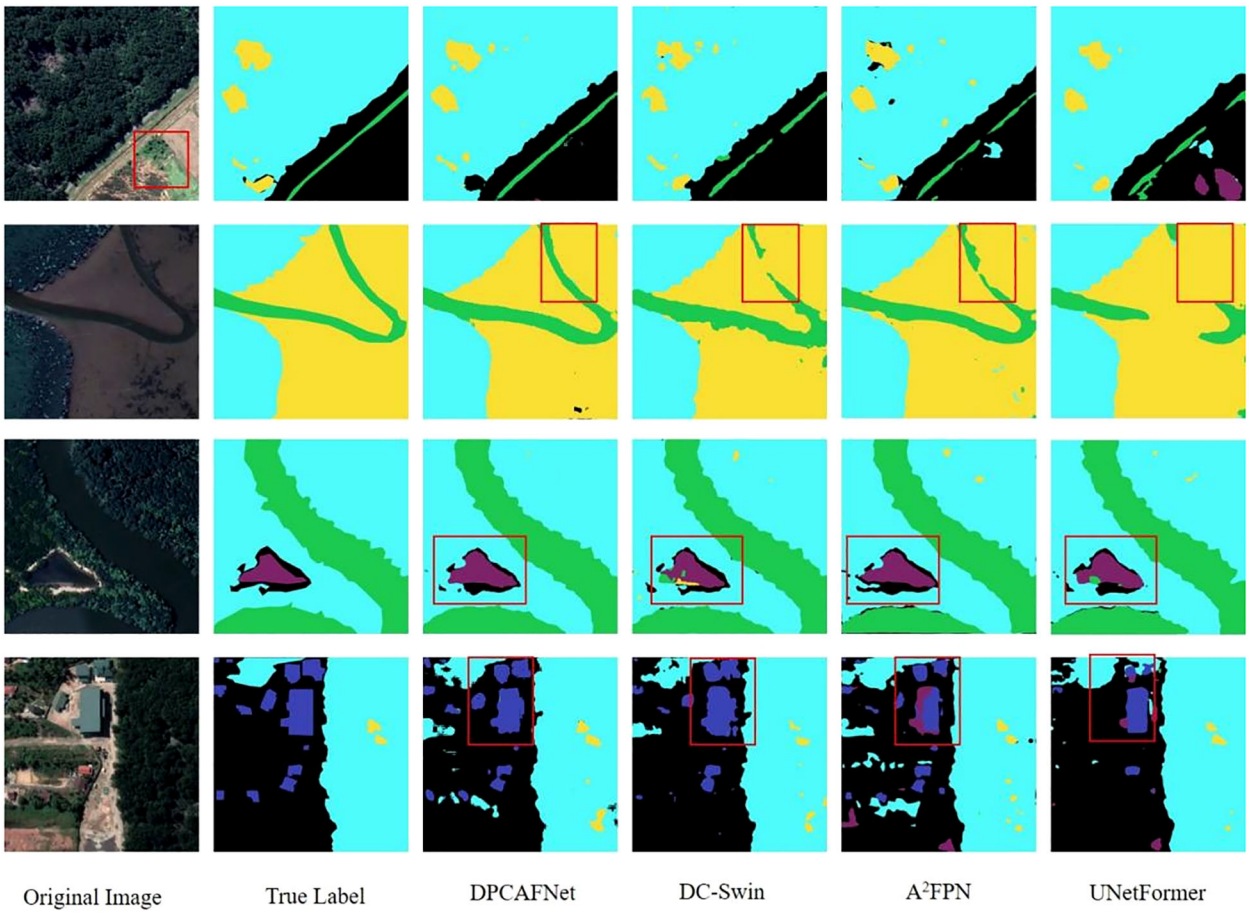

FIGURE 7
Comparison of segmentation results of TCCFNet model.

background region, while TCCFNet model performed well without misjudgment, which indicates that TCCFNet model has stronger performance in distinguishing similar categories. Secondly, the TCCFNet model also performs well in handling the recognition of different target sizes. In the picture in the second row, TCCFNet model performs well in river segmentation, while the other three models all have the problem of inaccurate river segmentation. In particular, UNetFormer model performs worst, identifying large rivers as tidal flat. In the third row of images, the TCCFNet model was able to accurately identify mangroves, rivers, ponds and backgrounds. In contrast, DC-Swin and UNetFormer models identified ponds as rivers, and A^2-FPN models performed well. In addition, except for the TCCFNet model, the other three models have small beach identification errors, indicating that the TCCFNet model has a high overall recognition of large targets and a strong understanding of small targets. Finally, the TCCFNet model still performs well against categorically unbalanced datasets. In the fourth row picture, the TCCFNet model can also accurately identify buildings with less data, which shows that the model can still maintain good performance in the case of unbalanced categories.

5 Discussion

5.1 CIM availability

In order to verify the effectiveness of cross-integration module (CIM) and make a comparison under fair experimental conditions, this paper designs the following experiment on the premise of keeping the overall structure of TCCFNet model unchanged: After upsampling the F3 and F4 feature maps, concatenate them with the F1 and F2 feature maps in channel dimension. In order to ensure the smooth operation of subsequent feature fusion, a 1×1 convolution layer is used to process the spliced feature map and unify the number of channels. This design preserves the core structure of the TCCFNet and enables independent evaluation of the effectiveness of the CIM module to ensure the reliability and persuasibility of the experimental results.

The ablation experimental data of CIM module are shown in Table 2. After the CIM module is removed, the model shows a significant decline in various evaluation indicators, especially in category pixel accuracy and MIoU indicators. Specifically, the F1-

TABLE 2 Comparison of ablation experimental data of CIM modules.

Index		Ablation test data of the CIM module	Complete experimental data
F1-score (%)		90.68	93.55
PA (%)		96.02	97.35
MIoU (%)		83.41	88.34
PA(%)	Background	90.22	92.44
	Mangrove	98.69	99.04
	River and Ocean	94.89	95.29
	Buildings and Roads	74.26	83.51
	Pond	85.73	96.19
	Mudflat	89.25	90.25

score dropped from 93.55% to 90.68%, pixel accuracy dropped from 97.35% to 96.02%, and the average IoU dropped from 88.39% to 83.41%. In terms of pixel accuracy across categories, there was a drop in accuracy across all categories, especially in the "buildings and roads" and "ponds" categories. The absence of CIM module weakens the model's ability in fine-grained feature extraction and target recognition, indicating that CIM module plays a key role in improving the model's ability to distinguish complex scenes and similar categories.

5.2 Generalization performance of TCCFNet

In order to further verify the generalization performance of TCCFNet model in the field of segmentation and the effectiveness of solving image segmentation problems. This paper selected the Urban Drone Dataset (UDD) dataset (Oh and Yoon, 2024) collected and labeled by the Graphics and Interaction Laboratory of Peking University for testing. The UDD dataset is a collection of drone image datasets collected at Peking University, Huludao City, Henan University and Cangzhou City, including vegetation, buildings, roads, vehicles and five other types of ground objects. In this paper, UDD5 version is used, which contains 160 images and is divided into 120 training sets and 40 verification sets according to the ratio of 3:1 to evaluate the segmentation effect of the model on the UDD5 data set.

According to the comparative analysis of experimental data in Table 3, it can be seen that TCCFNet has the best performance in MIoU index, reaching 74.56%. Compared with the new model A^2-FPN, DC-Swin and UNetFormer, TCCFNet has a significant advantage in the MIoU index, which is about 3% higher than that of A^2-FPN (71.46%) and DC-Swin (71.14%) with better data. It shows its excellent performance in semantic segmentation task. This shows that TCCFNet has stronger segmentation ability in complex image environment, and can capture and recognize different types of detail features more accurately.

In terms of PA index, TCCFNet performs well, reaching 89.17%. Compared with other new models, such as A^2-FPN (88.07%) and DC-Swin (88.35%), TCCFNet still has a significant advantage in PA. This shows that TCCFNet has a higher classification accuracy at the pixel level, can segment images more accurately, and improves the overall segmentation effect. Especially in the processing of complex mangrove remote sensing images, TCCFNet can effectively reduce misclassification and improve the reliability of segmentation results.

With a model size of 196MB, TCCFNet is not a lightweight model (Liu et al., 2024; Shi et al., 2024), but it strikes a good balance between performance and complexity. Compared with large models such as DC-Swin (256MB) (Hüseyin Ü et al., 2024), the TCCFNet model size is relatively moderate while maintaining high performance. This balance makes TCCFNet highly practical in practical applications, which not only ensures efficient segmentation effect, but also does not affect deployment and use because the model is too large.

As shown in Figure 8, in the comparison of segmentation results on the UDD5 dataset, TCCFNet model performs particularly well in various scenes, showing its advantages in complex image segmentation tasks. By comparing the segmentation effects of different models, we can more clearly see the advantages of TCCFNet.

As shown in the red box area in the first and second lines of images, TCCFNet performs well in edge and shape recognition on buildings, backgrounds and large areas of vegetation, with clear and accurate segmentation boundaries. In contrast, the segmentation boundaries of other models such as A2FPN and UNetFormer are fuzzy, and it is difficult to distinguish the target accurately. This is mainly due to TCCFNet's use of Swin Transformer's long-distance dependency establishment capability and ResNet's local detail capture capability, which effectively improves the recognition rate of large target areas, especially when dealing with complex urban structures. The third line of images shows the performance of TCCFNet in complex urban architecture scenes, and its segmentation effect of buildings and roads is very accurate, with clear boundaries, and it can accurately distinguish different categories. Other models are weaker in this regard, with blurred edges and difficulty distinguishing details. In the red box area in the figure, TCCFNet is also very accurate in distinguishing vegetation and roads, and the other three models all have misclassification. TCCFNet improves feature representation capabilities through cross-integration modules (CIM), enabling it to more accurately identify and segment different categories in complex contexts. The fourth line of images further demonstrates TCCFNet's detail processing capabilities in urban scenes. The boundary of the segmentation result is clear, and the details of the road and the vehicle can be accurately segmented. In contrast, UNetFormer and DC-Swin are poor, as shown in the red box area in the figure, and slightly lacking in detail. TCCFNet enhances the understanding of image details and global semantics through cross-integration

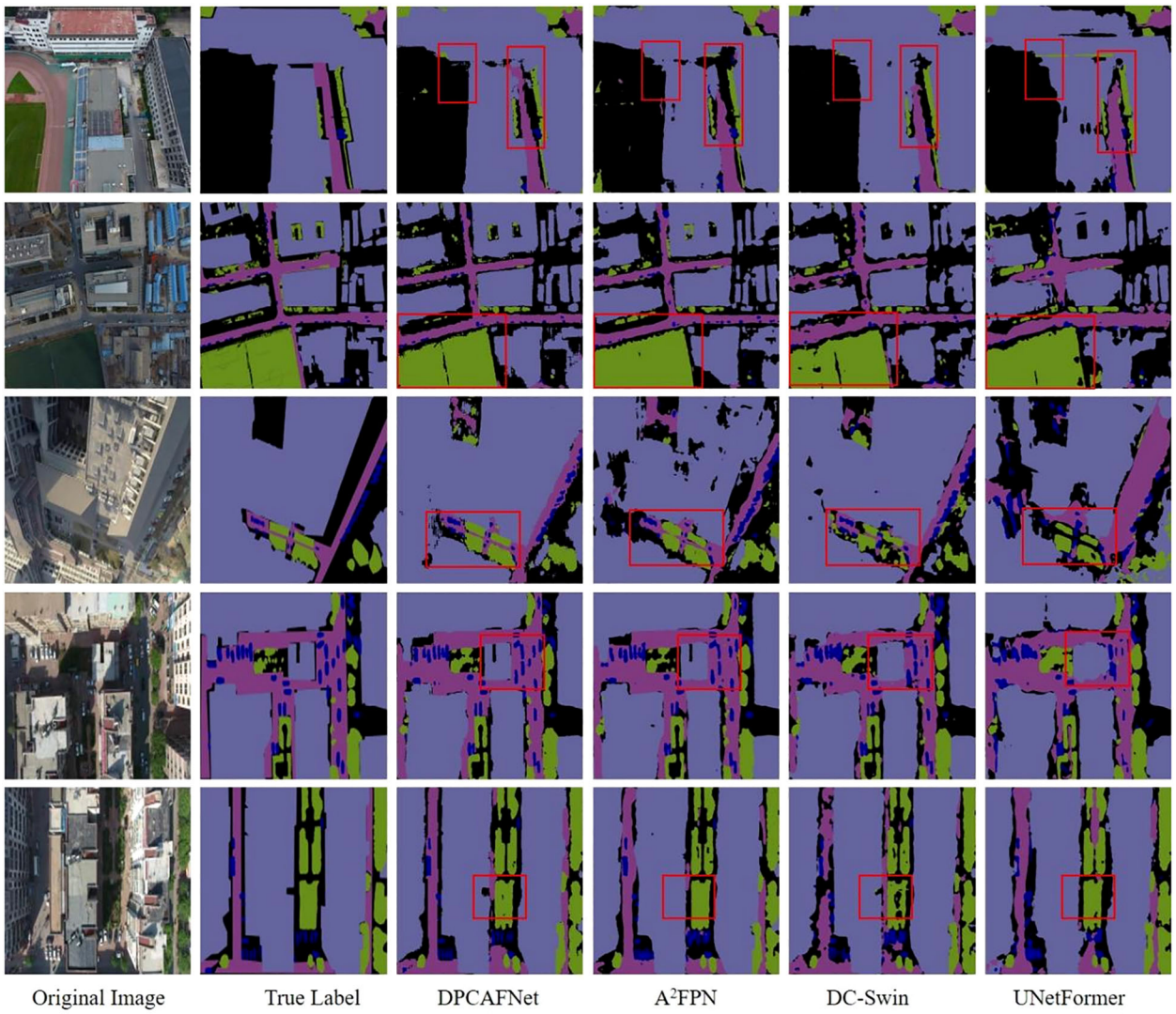

FIGURE 8
Comparison of segmentation results of TCCFNet model.

module (CIM), inhibits the influence of background noise, and improves the segmentation accuracy. In the fifth row image, TCCFNet still performs well in the boundary recognition of roads and buildings, and the segmentation results are clear and accurate. The small background category of the red box area in the figure can also be accurately identified, while other models are weak in this respect and cannot be accurately distinguished. TCCFNet's dual backbone network structure and cross integration module (CIM) enable it to achieve high precision segmentation results when dealing with complex backgrounds and different size target regions.

TABLE 3 Comparison of MIoU, PA and model sizes of different models on the UDD5 dataset.

Model	MIoU (%)	F1-score (%)	PA (%)	Model size (MB)
A^2-FPN	71.46	82.45	88.07	46.5
DC-Swin	71.14	70.99	88.35	256
UNetFormer	60.27	70.73	83.59	44.9
TCCFNet	74.56	84.67	89.17	196

5.3 Limitations and challenges

TCCFNet model performs well in mangrove remote sensing image segmentation, but there is still room for improvement in the following aspects:

1. Model complexity and computational efficiency: TCCFNet adopts the dual backbone structure of ResNet and Swin Transformer, which improves segmentation performance but requires a large amount of computation. You can try to introduce lightweight Transformer or optimize the convolutional layer design to reduce the model complexity and improve the reasoning speed, so as to better adapt to the actual application needs.
2. Generalization ability of diverse datasets: This paper shows the model's performance on mangrove datasets and UDD5 datasets, but does not include the broader remote sensing datasets. Follow-up studies can validate the generalization and application potential by testing the model on more different types of data sets.
3. Segmentation performance of small targets: Although TCCFNet performs well in the recognition of large target areas, there is still room for improvement in the detail segmentation effect of some small targets. The multi-scale feature enhancement module can be considered to further improve the accuracy of the model on small targets.
4. Optimization of data enhancement methods: conventional data enhancement methods such as rotation and noise are used in the model training process, but for the particularity of ecosystems such as mangroves, more targeted enhancement methods may be introduced, such as simulation of light changes or more random occlusion, to improve the robustness of the model in different environments.

6 Conclusion

In this paper, a two-channel cross-fusion network (TCCFNet) is proposed to improve the accuracy and robustness of mangrove remote sensing image segmentation. In order to solve the problems of similar categories, complex details and blurred boundaries in mangrove images, TCCFNet adopts a dual trunk network structure and combines ResNet and Swin Transformer to capture local details and global semantic information of images respectively, so as to better adapt to different target scales and improve segmentation accuracy. In order to enhance the feature fusion effect, TCCFNet designed a cross-integration module (CIM), which uses the cross-attention mechanism and feature fusion strategy to further strengthen the model's recognition ability of similar categories under complex background, and effectively suppress background noise.

Experiments were conducted on mangrove remote sensing image datasets and UDD5 urban datasets, and validated the significant advantages of TCCFNet on several evaluation indicators such as mean intersection ratio MIoU, pixel accuracy PA, and F1-score. Compared with mainstream algorithms such as A2FPN, BANet, DC-Swin and UNetFormer, TCCFNet has shown higher accuracy and robustness in complex scenes such as mangrove areas, rivers and beaches, especially in large-area targets and similar category differentiation. Ablation experiments further prove the effectiveness of CIM module. After the removal of CIM module, the segmentation performance of the model is significantly reduced, especially in complex background and similar category segmentation tasks.

However, the dual backbone structure of TCCFNet makes the model more complex and computates a lot, which limits its applicability in real-time applications. Therefore, the optimization and simplification of the model is the focus of future improvement. In addition, in order to further improve the generalization ability of TCCFNet, it can be tested on more types of remote sensing data sets in the future, and enhance its segmentation accuracy for small target regions, such as the introduction of multi-scale feature fusion module to improve detail resolution.

Data availability statement

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

Author contributions

LF: Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. YW: Visualization, Writing – review & editing. SW: Funding acquisition, Resources, Validation, Writing – original draft, Writing – review & editing, Methodology. JZ: Conceptualization, Writing – review & editing. ZW: Methodology, Resources, Writing – review & editing. JW: Investigation, Writing – review & editing. HC: Investigation, Methodology, Resources, Writing – review & editing. YC: Data curation, Funding acquisition, Methodology, Resources, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. 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.

Conflict of interest

Author JW was employed by the company China Mobile Group Hainan Company Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

References

Anniwaer, N., Li, X., Wang, K., Xu, H., Hong, S. J. A., and Meteorology, F. (2024). Shifts in the trends of vegetation greenness and photosynthesis in different parts of Tibetan Plateau over the past two decades. *Agric. For. Meteorol.* 345, 109851. doi: 10.1016/j.agrformet.2023.109851

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., et al. (2024). Medical image segmentation review: The success of u-net. doi: 10.1109/TPAMI.2024.3435571

Bonet, A., Peña, J., Bellot, J., Cremades, M., and Sánchez, J. (2024). Environment t. Changing vegetation structure and landscape patterns in semi-arid Spain 46.

Fu, L., Chen, J., Wang, Z., Zang, T., Chen, H., Wu, S., et al. (2024). MSFANet: multi-scale fusion attention network for mangrove remote sensing lmage segmentation using pattern recognition. *Int. J. Appl. Earth Obs. Geoinf.* 13, 27. doi: 10.1186/s13677-023-00565-w

Fu, B., Zhang, S., Li, H., Yao, H., Sun, W., Jia, M., et al. (2024). Exploring the effects of different combination ratios of multi-source remote sensing images on mangrove communities classification. *J. Cloud Comput.* 134, 104197. doi: 10.1016/j.jag.2024.104197

Gijón Mancheño, A., Vuik, V., van Wesenbeeck, B., Jonkman, S., van Hespen, R., Moll, J., et al. (2024). Integrating mangrove growth and failure in coastal flood protection designs. *Sci. Rep.* 14, 7951. doi: 10.1038/s41598-024-58705-4

He, B. (2022). Comparison of RFE-DL and stacking ensemble learning algorithms for classifying mangrove species on UAV multispectral images. *Int. J. Appl. Earth Obs. Geoinf.* 112, 102890.

Hüseyin Ü, FIRAT H, Atila, O., and ŞENGÜR, A. (2024). Swin transformer-based fork architecture for automated breast tumor classification. *Expert Syst. Appl.* 256, 125009. doi: 10.1016/j.eswa.2024.125009

Kathiresan, K. (2021). Mangroves: types and importance. *JMe biodiversity Manage.*, 1–31.

Li, W., Lambert-Garcia, R., Getley, A. C., Kim, K., Bhagavath, S., Majkut, M., et al. (2024). AM-SegNet for additive manufacturing in *situ* X-ray image segmentation and feature quantification. *Remote Sens.* 19, e2325572. doi: 10.1080/17452759.2024.2325572

Li, S., Liang, L., Zhang, S., Zhang, Y., Plaza, A., and Wang, X. J. R. S. (2024). End-to-end convolutional network and spectral-spatial Transformer architecture for hyperspectral image classification. *Virtual Phys Prototyp.* 16, 325. doi: 10.3390/rs16020325

Liu, H.-I., Galindo, M., Xie, H., Wong, L.-K., Shuai, H.-H., Li, Y.-H., et al. (2024). Lightweight deep learning for resource-constrained environments: A survey. *ACM Computing Surveys* 56, 1–42. doi: 10.1145/3657282

Liu, Z., He, D., Shi, Q., and Cheng, X. (2024). NDVI time-series data reconstruction for spatial-temporal dynamic monitoring of Arctic vegetation structure. *JG-sIS*, 1–19. doi: 10.1080/10095020.2024.2336602

Ma, L. (2023). Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Inf Fusion* 99, 101870.

MaL, Hu, Fang, Y., Wang, Lu, and Shengjin, (2021). "A2-FPN: attention aggregation based feature pyramid network for instance segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (Los Alamitos, CA, USA: IEEE), 15338–15347.

Mei, J., Zhou, T., Huang, K., Zhang, Y., Zhou, Y., Wu, Y., et al. (2025). A survey on deep learning for polyp segmentation: Techniques, challenges and future trends. 3, 1.

Mohamed, M. K., Adam, E., and Jackson, C. M. J. R. (2024). Assessing the perception and contribution of mangrove ecosystem services to the well-being of coastal communities of Chwaka and Menai Bays. *Zanzibar* 13, 7. doi: 10.3390/resources13010007

Nijamdeen, TWGFM, Ratsimbazafy, H. A., Kodikara, K. A. S., Nijamdeen, T. A., Thajudeen, T., Peruzzo, S., et al. (2024). Delineating expert mangrove stakeholder perceptions and attitudes towards mangrove management in Sri Lanka using Q methodology. *Environ. Sci. Policy* 151, 103632. doi: 10.1016/j.envsci.2023.103632

Oh, S., and Yoon, Y. (2024). Urban drone operations: A data-centric and comprehensive assessment of urban airspace with a Pareto-based approach. *JTRPAP Pract.* 182, 104034. doi: 10.1016/j.tra.2024.104034

Rahman, Lokollo, F. F., Manuputty, G. D., Hukubun, R. D., and Krisye, Maryono (2024). A review on the biodiversity and conservation of mangrove ecosystems in Indonesia. *Biodivers. Conserv.* 33, 875–903. doi: 10.1007/s10531-023-02767-9

Shi, J., Zhong, J., Zhang, Y., Xiao, B., Xiao, L., Zheng, Y. J. R. E., et al. (2024). A dual attention LSTM lightweight model based on exponential smoothing for remaining useful life prediction. *Reliab. Eng. Syst. Saf.* 243, 109821. doi: 10.1016/j.ress.2023.109821

Sun, Y., Ye, M., Ai, B., Lai, Z., Zhao, J., Jian, Z., et al. (2025). Annual change in the distribution and landscape health of mangrove ecosystems in China from 2016 to 2023 with Sentinel imagery. *Glob. Ecol. Conserv.* 57, e03355. doi: 10.1016/j.gecco.2024.e03355

Tasneem, S., Ahsan, M. N. J. O., and Management, C. (2024). A bibliometric analysis on mangrove ecosystem services: Past trends and emerging interests. *Ocean Coast. Manag.* 256, 107276. doi: 10.1016/j.ocecoaman.2024.107276

Tsai, F.-J., Yan-Tsung,, Tsai, C.-C., Lin, Y.-Y., and Lin, C.-W. (2022). BANet: A blur-aware attention network for dynamic scene deblurring. *IEEE Trans. Image Processing.* 31, 6789–6799. doi: 10.1109/TIP.2022.3216216

Wang, Y. (2020). "ABCNet: real-time scene text spotting with adaptive bezier-curve network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (Los Alamitos, CA, USA: IEEE), 9806–9815.

Wang, L. A. L., Duan, R., Zhang, C., Meng, Ce, Xiaoliang,, and Shenghui, F. (2022). A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3143368

Wang, Y., Yang, L., Liu, X., and Yan, P. J. S. R. (2024). An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. *Sci. Rep.* 14, 9716. doi: 10.1038/s41598-024-60375-1

Wang, W., Zhai, D., Li, X., Fang, H., and Yang, Y. (2024). Conflicts in mangrove protected areas through the actor-centred power framework-Insights from China. *JFP Economics* 158, 103122. doi: 10.1016/j.forpol.2023.103122

Yang, X., Cao, W., Tang, D., Zhou, Y., YJIToG, Lu, and Sensing, R. (2025). ACTN: adaptive coupling transformer network for hyperspectral image classification. doi: 10.1109/TGRS.2025.3528411