remote sensing

MDPI

Article

VMMT-Net: A Dual-Branch Parallel Network Combining Visual State Space Model and Mix Transformer for Land–Sea Segmentation of Remote Sensing Images

Jiawei Wu, Zijian Liu, Zhipeng Zhu, Chunhui Song, Xinghui Wu and Haihua Xing *

School of Information Science and Technology, Hainan Normal University, Haikou 571158, China
* Correspondence: xinghaihua@hainnu.edu.cn

Abstract

Land–sea segmentation is a fundamental task in remote sensing image analysis, and plays a vital role in dynamic coastline monitoring. The complex morphology and blurred boundaries of coastlines in remote sensing imagery make fast and accurate segmentation challenging. Recent deep learning approaches lack the ability to model spatial continuity effectively, thereby limiting a comprehensive understanding of coastline features in remote sensing imagery. To address this issue, we have developed VMMT-Net, a novel dual-branch semantic segmentation framework. By constructing a parallel heterogeneous dual-branch encoder, VMMT-Net integrates the complementary strengths of the Mix Transformer and the Visual State Space Model, enabling comprehensive modeling of local details, global semantics, and spatial continuity. We design a Cross-Branch Fusion Module to facilitate deep feature interaction and collaborative representation across branches, and implement a customized decoder module that enhances the integration of multiscale features and improves boundary refinement of coastlines. Extensive experiments conducted on two benchmark remote sensing datasets, GF-HNCD and BSD, demonstrate that the proposed VMMT-Net outperforms existing state-of-the-art methods in both quantitative metrics and visual quality. Specifically, the model achieves mean F1-scores of 98.48% (GF-HNCD) and 98.53% (BSD) and mean intersection-over-union values of 97.02% (GF-HNCD) and 97.11% (BSD). The model maintains reasonable computational complexity, with only 28.24 M parameters and 25.21 GFLOPs, striking a favorable balance between accuracy and efficiency. These results indicate the strong generalization ability and practical applicability of VMMT-Net in real-world remote sensing segmentation tasks.

Keywords: Vision Mamba; Transformer; dual-branch network; remote sensing; land–sea segmentation

check for updates

Academic Editor: Tania Stathaki

Received: 10 May 2025
Revised: 2 July 2025
Accepted: 14 July 2025
Published: 16 July 2025

Citation: Wu, J.; Liu, Z.; Zhu, Z.; Song, C.; Wu, X.; Xing, H. VMMT-Net: A Dual-Branch Parallel Network Combining Visual State Space Model and Mix Transformer for Land–Sea Segmentation of Remote Sensing Images. *Remote Sens.* **2025**, *17*, 2473. https://doi.org/10.3390/rs17142473

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Coastlines form the boundary between ocean and land, and are recognized by the International Geographic Data Committee as one of the 27 fundamental surface features [1]. The spatiotemporal evolution of coastlines not only alters the physical geography of coastal zones, but also has a significant impact on regional economies and societal development in littoral areas [2]. Various natural factors (such as crustal movements, tidal fluctuations, and climate change) as well as anthropogenic influences (including land reclamation, port construction, and tidal flat encroachment) mean that coastlines are subject to persistent and dynamic changes [3]. Therefore, the ability to rapidly and accurately extract coastline

information and monitor its changes in a timely manner is of great practical significance for coastal resource management, ecological protection, and the sustainable development of marine economies.

In recent years, remote sensing technology has increasingly replaced traditional manual surveying methods owing to its advantages in large-scale coverage, multi-temporal monitoring, and high spatial resolution. Remote sensing has become the primary technique for tracking coastline dynamics [4], with widespread applications in geological surveys [5,6], marine observation [7,8], and disaster assessment [9,10]. However, the land–sea segmentation of remote sensing imagery remains more challenging than general semantic segmentation. The inherent irregularity and ambiguity of coastlines often lead to pixel-level misclassification at the land–sea boundaries, manifesting as edge shifts or segmentation breaks [11]. Additionally, atmospheric conditions, illumination variations, and terrain diversity may result in highly similar spectral characteristics across land and sea regions, making it difficult for models to distinguish between them and increasing the complexity of accurate and timely coastline extraction and monitoring [12]. As illustrated in Figure 1, remote sensing images often contain confusing land–sea transition zones and small-scale objects that are difficult to segment, further emphasizing the challenges faced by this task.

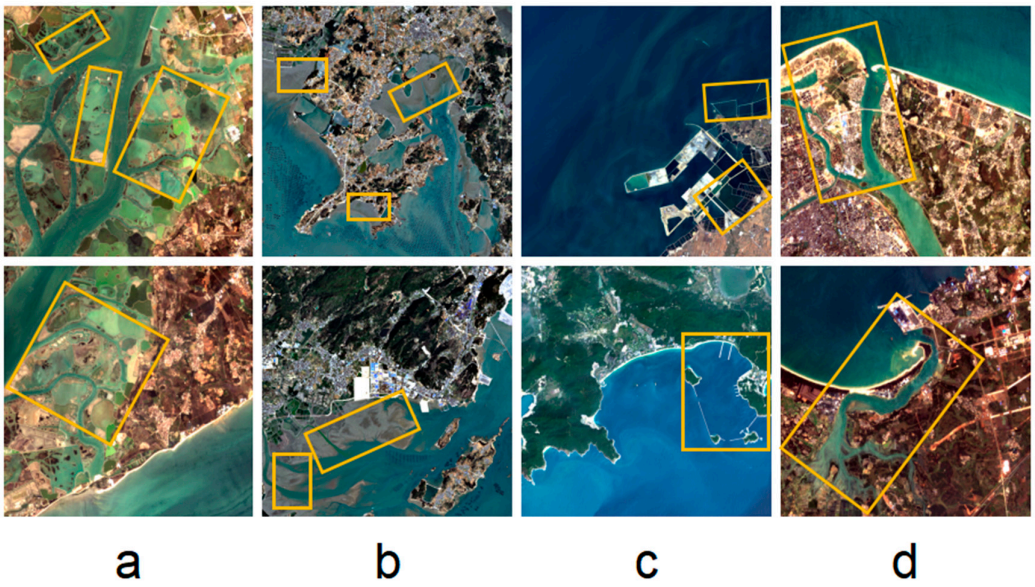


Figure 1. Examples of typical land–sea boundary ambiguities in remote sensing imagery. (**a**) Spectral similarity between land and sea caused by aquaculture activities. (**b**) Blurred land–sea boundary caused by sediment accumulation. (**c**) Coastal scene with dense man-made structures. (**d**) Meandering small-scale river segment.

Traditional coastline-extraction methods applied to remote sensing imagery include threshold-based segmentation [13], edge-detection operators [14], and object-based image analysis [15]. These approaches are structurally simple and easy to implement, and they perform reasonably well when dealing with regular boundaries or homogeneous spectral features. However, in complex coastal scenarios characterized by intricate textures, multi-scale variations, and spectral similarities between land and sea, these methods often fail to effectively capture high-level semantic representations. This limitation hinders their ability to meet the accuracy demands of coastline-extraction tasks. In recent years, deep learning has achieved remarkable progress in semantic segmentation. Convolutional Neural Networks (CNNs) have powerful feature-extraction capabilities and have been widely adopted in land–sea segmentation tasks [16]. Nevertheless, the local receptive fields of

CNNs constrain their capacity to model global contextual information, making it difficult to accurately recognize structures with long-range dependencies. To address this issue, Transformer-based models leveraging self-attention mechanisms have been applied to remote sensing segmentation. These models exhibit superior expressiveness in capturing long-range and multiscale semantic dependencies [17]. However, the computational complexity of traditional self-attention mechanisms scales quadratically with the image resolution, leading to substantial resource demands in practical applications. Against this backdrop, the recently proposed Mamba architecture offers an efficient sequence modeling framework with linear time complexity, while retaining the ability to model long-range dependencies [18]. Inspired by this, VMamba (a Visual State Space Model, VSSM) extends the Mamba architecture to vision tasks, enabling effective modeling of 2D spatial data and offering a promising new direction for high-precision coastline extraction in remote sensing imagery [19].

Given the inherent complexity of coastline remote sensing images—such as irregular shorelines, the coexistence of multiscale targets, and spectral similarity between land and sea regions—land–sea segmentation remains a challenging task. To this end, we propose a novel segmentation network named VMMT-Net, which is tailored for land–sea segmentation in remote sensing imagery. This network integrates the strengths of the Mix Transformer (MiT) and VMamba within a parallel dual-branch encoder framework, aiming to achieve accurate and robust coastline delineation under complex conditions.

The main contributions of this work can be summarized as follows:

(1) We propose VMMT-Net, a dual-branch parallel network combining MiT and VMamba that is designed for the fine-grained segmentation of complex land–sea boundaries in medium- and high-resolution remote sensing images.

(2) We have designed a Cross-Branch Fusion Module (CBFM) that is embedded into each feature-extraction stage of the parallel dual-branch encoder composed of MiT and VMamba. In this way, we achieve information complementarity and guidance between the dual encoders, thereby enhancing the overall discriminative performance for ambiguous pixels.

(3) To facilitate effective feature integration between shallow and deep layers during upsampling, we design a customized decoder block that, together with two residual convolution blocks, constitutes a complete decoder structure for the superior restoration of fine coastline boundary details.

2. Related Work

2.1. Semantic Segmentation

2.1.1. CNN-Based Semantic Segmentation Methods

Semantic segmentation, as one of the important downstream tasks in computer vision, aims to perform pixel-level classification of images to achieve fine-grained recognition and structured representation of objects in scenes. Early research primarily relied on CNNs, driving the continuous development of semantic segmentation technology from the construction of a basic framework to multiscale modeling and the recovery of spatial details. Long et al. [20] proposed Fully Convolutional Network (FCN), which first replaced fully connected layers in classification tasks with convolutional layers, enabling end-to-end pixel-level prediction and pioneering the application of deep learning in semantic segmentation. Subsequently, Ronneberger et al. [16] proposed U-Net, which adopts a symmetric encoder–decoder structure and introduces skip connections to enhance feature reconstruction capability, becoming one of the most representative basic architectures in medical and remote sensing image segmentation.

To enhance the modeling capability for contextual semantics and spatial information, researchers began to introduce multiscale information perception mechanisms. Zhao et al. [21] proposed Pyramid Scene Parsing Network (PSPNet), which aggregates multiscale contextual features through the Pyramid Pooling Module (PPM), significantly improving the parsing capability for complex scenes. Based on DeepLabV3, Chen et al. [22] proposed DeepLabV3+, introducing the Atrous Spatial Pyramid Pooling (ASPP) module to perceive multi-scale semantic information while combining a lightweight decoder to enhance spatial detail restoration. Li et al. [23] proposed ABCNet, which models spatial and semantic features separately in a dual-path architecture, utilizing linear attention mechanisms to achieve efficient cross-layer information fusion, balancing inference efficiency and segmentation accuracy. Alongside attempts to enhance the semantic understanding capability, the effective recovery of spatial details and boundary information has gradually become the focus of research. To improve the recovery quality of spatial structures during upsampling, Badrinarayanan et al. [24] proposed SegNet, which uses pooling indices to guide nonlinear upsampling, improving boundary recognition capability while reducing resource consumption. To address the problem of spatial information loss during downsampling, Wang et al. [25] proposed High-Resolution Network (HRNet), which continuously exchanges information through multi-resolution parallel paths, achieving unified modeling of spatial details and semantic depth. Furthermore, Qin et al. [26] designed a nested U-Net structure (U^2-Net) by introducing Residual U-blocks (RSU) to capture broader context and enhance feature representation capability, demonstrating excellent binary segmentation and saliency-detection performance without requiring pre-training.

2.1.2. Transformer and Hybrid Structure-Based Semantic Segmentation Methods

The Transformer architecture was proposed by Vaswani et al. [17] and achieved tremendous success in the field of natural language processing. Inspired by this, Dosovitskiy et al. [27] proposed Vision Transformer (ViT), which first applied Transformer to image classification tasks. Compared to traditional CNN models, ViT excels at capturing global contextual dependencies, bringing new pathways to semantic segmentation. However, ViT lacks local inductive bias and performs poorly when processing small-scale targets in remote sensing images. To address the shortcomings of ViT, Liu et al. [28] proposed Swin Transformer. As an important variant of ViT, Swin Transformer adopts a hierarchical structure and shifted window attention mechanism, enhancing both local and global modeling capabilities while maintaining low computational complexity, and is widely applied to pixel-level prediction tasks in high-resolution scenarios. In the same year, Zheng et al. [29] proposed the Segmentation Transformer (SETR) model, which adopts a pure Transformer encoder with various decoder structures, re-examining semantic segmentation tasks from a sequence-to-sequence modeling perspective and effectively modeling long-range contextual information. To further improve computational efficiency, Xie et al. [30] proposed a lightweight Transformer model (SegFormer), which integrates the advantages of ViT with lightweight design concepts, constructing a hierarchical Transformer encoder without positional encoding and pairing it with a simple and efficient MLP decoder, achieving leading performance with low computational overhead.

Considering the advantages of CNNs in local feature extraction and the powerful performance of Transformers in global modeling, numerous researchers began exploring the combination of these technologies to achieve the collaborative modeling of local and global information. In existing research, fusion strategies can be broadly categorized into serial fusion and parallel fusion. Serial fusion is represented by TransUNet [31] and FTransUNet [32]. Among them, TransUNet stacks CNNs and Transformers sequentially to form a hybrid encoder, then adopts a UNet-style decoder, constructing a hybrid vi-

sion Transformer structure that significantly improves image segmentation performance. FTransUNet further employs parallel CNN branches in the shallow feature-extraction stage to obtain fine-grained features from different modal data, then feeds the fused shallow features into a specially designed fusion Vit (FVit) module. This approach achieves sufficient fusion and semantic enhancement of deep features from different modal data through attention mechanisms, thus improving the segmentation accuracy of multi-modal remote sensing images in complex scenarios. Unlike serial fusion, parallel fusion methods typically adopt dual-branch or multi-branch structures, introducing CNN and Transformer paths in parallel during the feature-extraction stage, and achieving the joint modeling of multiscale semantics through feature-interaction modules. Representative examples include ST-UNet [33] and CLCFormer [34]. ST-UNet embeds Swin Transformer and CNN in parallel into the U-Net architecture, forming a parallel dual-branch encoder structure, and introduces cross-branch interaction mechanisms to enhance feature expression and fusion capabilities, improving the recognition performance of small-scale ground objects in remote sensing images. CLCFormer combines SwinV2 and EfficientNet-B3 to form a parallel dual-branch network, achieving complementary integration of dual-branch features through Bilateral Feature Fusion Module (BiFFM) at multiple scales. This method introduces auxiliary supervision strategies to accelerate model convergence, and achieves improved overall segmentation accuracy and robustness in complex remote sensing scenarios.

2.1.3. Mamba and Hybrid Structure-Based Semantic Segmentation Methods

Although Transformers and their derivative structures have achieved significant results in the field of semantic segmentation, their computational complexity and resource consumption in processing high-resolution remote sensing images remain urgent problems to be solved. To address this, Gu et al. [18] proposed an efficient sequence modeling method (Mamba) based on Selective State Space Models (SSMs). Unlike traditional Transformers, Mamba completely abandons attention mechanisms and Multilayer Perceptron (MLP), significantly reducing computational complexity. Subsequently, Liu et al. [19] proposed VMamba, which migrates the Mamba architecture to the computer vision field by introducing a 2D Selective Scan (SS2D) mechanism to adapt to the spatial structure of visual data, achieving efficient modeling of long-range dependency information in images and demonstrating performance potential that surpasses traditional CNN and Transformer models. Building on this foundation, Ruan et al. [35] proposed VM-UNet, the first pure VSSM medical image segmentation network based on the VMamba architecture, constructing a complete model with VSS Block and establishing the research foundation for Mamba in visual segmentation tasks. Furthermore, addressing the insufficient global dependency modeling and inadequate frequency domain feature utilization in hyperspectral image classification (HSI), Zhuang et al. [36] proposed a novel Frequency-Aware Hierarchical Mamba network (FAHM), which achieves collaborative modeling of local textures and global semantics by fusing frequency domain attention mechanisms with hierarchical Mamba structures, further promoting in-depth research and application of Mamba in remote sensing image interpretation tasks.

The Mamba architecture demonstrates good modeling capabilities in semantic segmentation tasks. Hence, researchers have explored its combination with mainstream structures such as CNNs or Transformers to form complementary hybrid architectures. For the remote sensing domain, Ma et al. proposed RS3Mamba [37], which represents the first introduction of Visual State Space Models (VSSM) into remote sensing image semantic segmentation research. It adopts VMamba as an auxiliary branch combined with ResNet18 to form a dual-branch backbone network, and achieves global and local semantic fusion through the Collaborative Completion Module (CCM), demonstrating the application potential of the

Mamba architecture in remote sensing image processing tasks. Subsequently, Hatamizadeh et al. [38] first combined the advantages of Mamba and Transformer, proposing a hybrid visual backbone network MambaVision, expanding the depth and breadth of Mamba applications in the computer vision field.

2.2. Land–Sea Segmentation

Land–sea segmentation, as a specific task in semantic segmentation, aims to segment ocean and land regions from remote sensing images, providing reliable data support for subsequent coastline detection, marine environment monitoring and analysis, and various other ocean-related tasks. Initially, researchers primarily relied on CNN architectures for land–sea segmentation. Cheng et al. [39] proposed SeNet, which uses DeconvNet as the basic architecture and introduces local smoothness regularization terms and structured edge-detection branches to construct a multi-task joint optimization framework, demonstrating strong land–sea boundary perception capabilities in port remote sensing images. Li et al. [40] proposed DeepUNet based on U-Net, achieving feature compression and restoration through customized DownBlock and UpBlock modules while introducing Plus connection mechanisms to enhance feature transmission and information flow capabilities, improving the expression accuracy of land–sea detail regions. To further integrate multi-scale semantic information, Shamsolmoali et al. [41] proposed RDU-Net, which integrates dense residual modules based on U-Net and achieved satisfactory segmentation results.

With ongoing research, attention mechanisms and multiscale modeling have been extensively used to enhance land–sea boundary discrimination and contextual awareness capabilities. Cui et al. [42] proposed SANet, combining adaptive multi-scale feature learning module (AML) with channel attention mechanisms (SE modules), enhancing the model's boundary discrimination capability in complex land–sea scenarios. Gao et al. [43] designed MSRNet based on Res2Net to construct multi-scale feature-extraction paths, integrating squeeze-and-excitation attention modules with deep supervision mechanisms, significantly improving the expression effectiveness of weak land–sea boundary regions. Ji et al. [44] proposed a Dual-Branch Ensemble Network (DBENet), utilizing parallel dense branches and residual branches for feature extraction, and enhancing inter-branch feature interaction and information transmission through Ensemble Attention learning strategies (EAM), effectively improving discrimination capability for irregular land–sea boundaries. Gao et al. [45] proposed A2RDNet, which alleviates the problems of large intra-class differences and small inter-class differences in weak boundary regions by enhancing spatial localization information and fusing multi-scale contextual semantic features, significantly improving the accuracy and stability of land–sea segmentation. Ji et al. [11] improved the traditional U-shaped structure and proposed a lightweight and efficient E-shaped Convolutional Neural Network (E-Net), introducing Contextual Aggregation Attention Mechanism (CA2M) to effectively enhance cross-layer feature interaction and fuzzy coastline boundary recognition capabilities.

In recent years, some researchers have introduced Transformer structures into land–sea segmentation tasks. For example, Xiong et al. [46] combined CNNs and Transformers to design a lightweight dual-branch parallel network (TCUNet), achieving collaborative modeling of local and global information, and constructed a land–sea segmentation dataset covering the Yellow Sea region. Tong et al. [47] combined Swin Transformer with improved inverted residual networks and proposed a new method with global supervised feature learning and layer-by-layer feature-extraction capabilities (STIRUNet), applied to coastline-extraction tasks in Hainan Island waters, providing new insights for mixed pixel identification. Furthermore, the application of multi-source remote sensing data in land–sea segmentation tasks has become increasingly widespread. Chen et al. [48] proposed a

semantic segmentation network MPG-Net for fine extraction of coastal aquaculture ponds, integrating multi-scale feature-extraction structure (MS) with polarization-global context attention mechanism (PGC) to achieve high-precision extraction of nearshore water targets in multi-source remote sensing images. Ai et al. [49] used dual-polarization SAR images and AIS data as data sources and proposed a novel Pyramid Vision Transformer (PVT) assisted by long-term AIS data (AIS-PVT) for land–sea segmentation, which can capture global multi-scale information and enhance the model's data fusion capabilities.

In summary, semantic segmentation techniques have been widely applied in remote sensing image analysis, and the methodological development of land–sea segmentation has undergone a progressive evolution from traditional CNN-based models to hybrid architectures that integrate Transformers and VMamba. CNNs are adept at modeling local textures, Transformers excel in capturing global contextual features, while the Mamba architecture achieves a balance between long-sequence dependency modeling and computational efficiency. The continuous fusion of these technologies has significantly advanced the accuracy of coastline extraction in remote sensing imagery. However, existing approaches still encounter considerable limitations when addressing challenges such as irregular coastline geometries, blurred land–sea boundaries, and spectral ambiguities in large-scale remote sensing images. These challenges highlight the necessity for deeper architectural innovation and modeling mechanisms. To this end, we propose VMMT-Net, a novel framework that incorporates a parallel dual-branch encoder and a cross-branch feature fusion mechanism to further enhance the modeling and recognition of complex land–sea boundaries.

3. Method

VMMT-Net has a classical encoder–decoder architecture, which enables the effective extraction of multiscale feature information while enhancing the modeling of spatial structural continuity. In this section, we provide a comprehensive overview of the overall network design of VMMT-Net, including its key components such as MiT, VMamba, CBFM, and a custom decoder.

3.1. Overall Network Architecture

The overall architecture of the proposed VMMT-Net is illustrated in Figure 2. In the encoder part, we construct a parallel dual-branch structure based on MiT [30] and VMamba [19]. Specifically, the encoder is composed of four hierarchical stages, each operating in parallel through the MiT and VSS branches. This design allows the model to simultaneously capture high-resolution shallow features and low-resolution deep semantic representations. To effectively integrate the multiscale features extracted by the dual-branch encoders, we introduce a CBFM at each stage. The CBFM serves to aggregate and enhance the complementary features from both the MiT and VSS branches, improving the representational capacity of the encoder. The fused multiscale features from each stage are then forwarded via skip connections to the corresponding layers of the decoder, ensuring that the decoder can fully leverage the contextual information acquired by the encoder.

In the decoder part, we design a custom decoder block to efficiently merge the shallow features from the encoder with the deeper features from the previous decoding layer. These decoder blocks, together with residual convolutional blocks, constitute a complete decoder architecture. Through a series of upsampling operations, the decoder progressively reconstructs spatial details while simultaneously integrating semantic information from various encoder levels. This ensures that the model retains both fine-grained boundary details and global semantic coherence. Moreover, skip connections help preserve low-level spatial details that may otherwise be lost during deep convolutional operations, providing

the decoder with richer information for reconstruction. Finally, a segmentation head is applied to generate the final land–sea classification map.

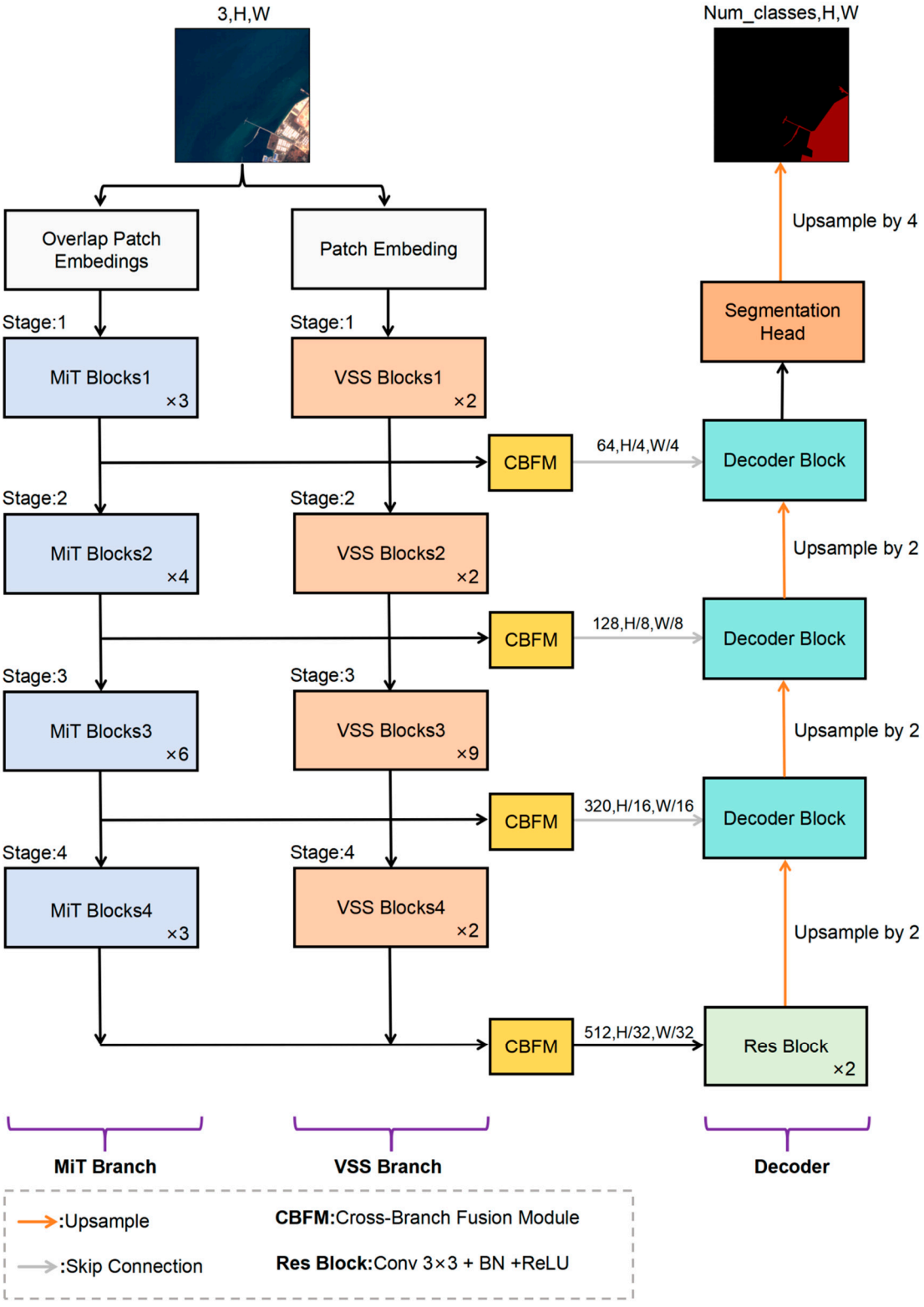


Figure 2. Overall architecture of the proposed VMMT-Net.

3.2. Dual-Branch Encoder Based on Mamba and Transformer

In recent years, hierarchical visual Transformers such as the MiT [30] and Swin Transformer [28] have gained traction in the computer vision community, and have emerged as effective backbones for a wide range of visual tasks. As a lightweight hierarchical

Transformer encoder, MiT adopts an efficient self-attention mechanism that captures dependencies between distant pixels. This reduces the computational complexity from $O(N^2)$ (in standard Transformers) to $O(N^2/R)$. Furthermore, MiT employs a progressive downsampling structure and a positional-encoding-free design, which enables robust multiscale feature extraction while eliminating the interpolation errors typically introduced by resolution changes during testing. This enhances the model's adaptability to inputs of varying resolutions and improves generalization in semantic segmentation tasks. Additionally, the use of overlapping patch merging preserves local continuity across patch boundaries, thereby enhancing the local feature representation.

VMamba is a vision backbone network based on VSSM [19], which extends the SSM architecture proposed in Mamba [18]. The VMamba framework is designed to address the quadratic computational complexity problem of Transformers in visual tasks. Instead of modeling long-range dependencies via self-attention, VMamba adopts a linear-time Selective Scan mechanism that significantly reduces the complexity from $O(N^2)$ to $O(N)$, while effectively capturing long-range interactions in images. Specifically, VMamba first converts the input image into embedded representations using a 2D patch embedding layer. These embeddings are then processed through a series of VSS blocks. The core computational unit in each VSS block is the SS2D module, which captures long-range dependencies across spatial dimensions. The SS2D module comprises three stages: Scan Expanding, S6 Block, and Scan Merging. In the Scan Expanding stage, the input features are unfolded into four independent feature sequences, allowing for the comprehensive extraction of information from different directions. These sequences are then processed by S6 blocks based on the SSM, enabling each element to interact with previously scanned elements via a compressed hidden state. Finally, the Scan Merging operation fuses the four directional outputs to restore the original spatial resolution, achieving efficient global context modeling. Moreover, each VSS block includes a Patch Merging layer for the progressive downsampling of feature maps, facilitating effective multiscale feature extraction. The computation within the SS2D module can be formalized as follows:

$$X_v = \text{ScanExpand}(X, v), \tag{1}$$

$$\overline{X}_v = \text{S6}(X_v), \tag{2}$$

$$\overline{X} = \text{ScanMerge}\left(\overline{X}_1, \overline{X}_2, \overline{X}_3, \overline{X}_4\right), \tag{3}$$

where $v \in V = \{1, 2, 3, 4\}$ denotes the scanning direction and S6 refers to the Selective Scan State Space Model of Mamba [18]. $\text{ScanExpand}(\cdot)$ and $\text{ScanMerge}(\cdot)$ correspond to the Scan Expanding and Scan Merging operations, respectively [19].

Considering the lightweight nature of MiT and its ability to model both local and global information, as well as the advantages of VMamba in capturing spatial structural continuity and directional awareness, we construct a parallel heterogeneous dual-branch encoder by integrating MiT and VMamba. In terms of implementation details, both branches adopt a hierarchical structure comprising four stages for multiscale feature extraction. In the MiT branch, each stage consists of multiple MiT blocks—specifically, 3, 4, 6, and 3 blocks in Stages 1–4, respectively. The number of feature channels for each stage is set to 64, 128, 320, and 512, and the corresponding spatial resolutions are (H/4,W/4), (H/8,W/8), (H/16,W/16), and (H/32,W/32), respectively. The overall architecture of the MiT block is illustrated in Figure 3a. The input feature, denoted as X_{in}, first undergoes Layer Normalization (LN) to stabilize the feature distribution and improve the training efficiency. The normalized feature is then fed into the efficient self-attention module, wherein spatial reduction is applied to the Key and Value representations to reduce computational complexity while capturing global dependencies. Subsequently, attention scores between the

Query, Key, and Value are computed to generate enhanced feature representations, which are then added to the input via a residual connection to yield the intermediate feature X_{att}. The computation can be formulated as:

$$Q = X_{in}W_Q, K = SR(X_{in})W_K, V = SR(X_{in})W_V, \tag{4}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5}$$

$$X_{att} = \text{LN}(\text{Attention}(Q, K, V) + X_{in}), \tag{6}$$

where W_Q, W_K, and W_V denote the linear projection weight matrices for the Query, Key, and Value, respectively. $SR(\cdot)$ represents the Spatial Reduction operation and d_k is the reduction factor for the dimensionality of the Key.

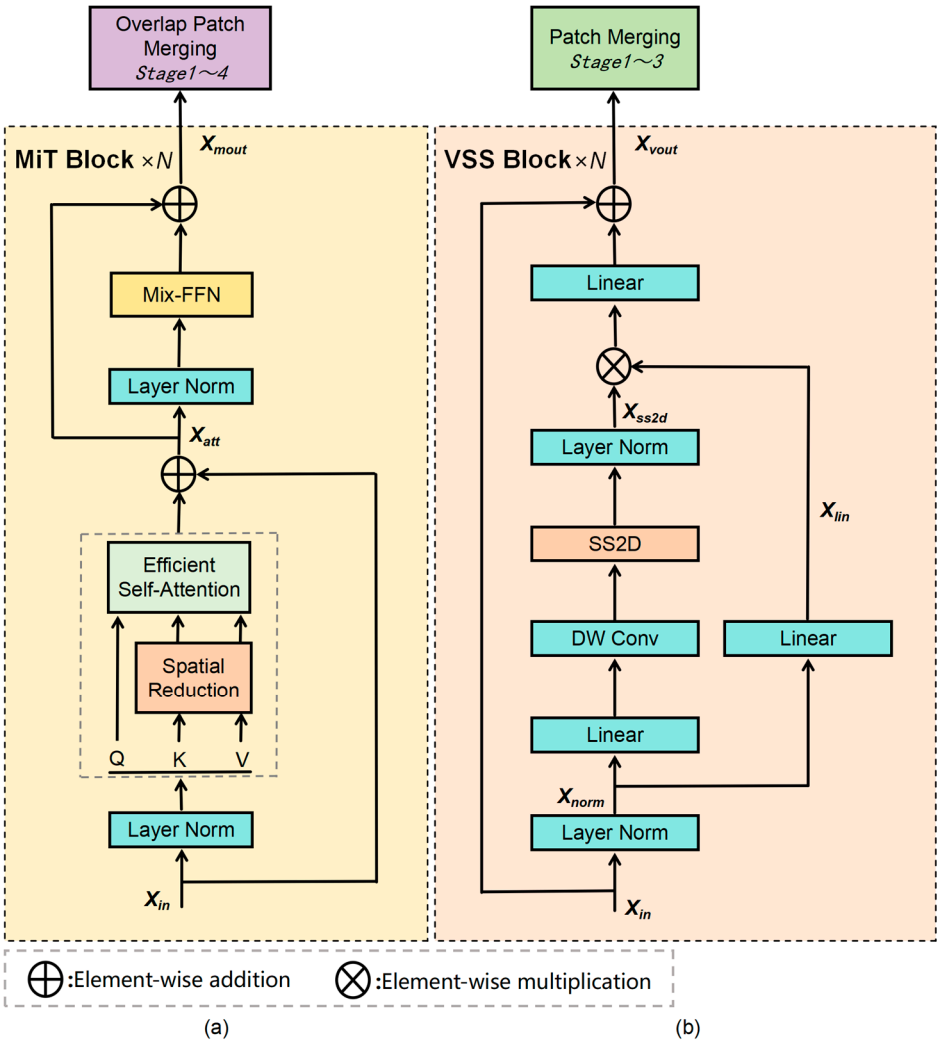


Figure 3. (**a**) Detailed architecture of the MiT Block. (**b**) Detailed architecture of the VSS Block.

Subsequently, X_{att} is normalized again via LN and passed through a Mix Feed-Forward Network (Mix-FFN) for further feature enhancement. The Mix-FFN consists of a 1×1 convolution for channel expansion, a 3×3 depth-wise separable convolution (DWConv) for local feature extraction, and another 1×1 convolution for channel compres-

sion. The output is then added to X_{att} via a residual connection, yielding the final feature map X_{mout}. The computation process is formulated as:

$$F' = \mathrm{ReLU}(\mathrm{Conv}_{1\times1}(X_{att})), \tag{7}$$

$$F'' = \mathrm{Conv}_{1\times1}\big(\mathrm{ReLU}\big(\mathrm{DWConv}_{3\times3}(F')\big)\big), \tag{8}$$

$$X_{mout} = \mathrm{LN}(F'' + X_{att}), \tag{9}$$

where $\mathrm{Conv}_{1\times1}$ denotes 1×1 convolution, $\mathrm{DWConv}_{3\times3}$ denotes 3×3 depth-wise separable convolution, ReLU is the activation function, and LN denotes Layer Normalization.

In the VSS branch, each stage comprises multiple VSS Blocks (with 2, 2, 9, and 2 blocks in stages 1–4, respectively). The corresponding feature channel dimensions are 96, 192, 384, and 768, while the spatial resolutions are (H/4,W/4), (H/8,W/8), (H/16,W/16), and (H/32,W/32), respectively. The overall architecture of the VSS Block is illustrated in Figure 3b. First, the input feature X_{in} is normalized via LN to stabilize the feature distribution and improve training convergence. The normalized features are then processed along two parallel computational paths. In the first path, a linear layer adjusts the channel dimension, followed by SiLU activation to apply a nonlinear transformation, resulting in the intermediate feature X_{lin}. In the second path, the features are first passed through a linear layer for channel adjustment, followed by DWConv for local feature extraction, and then SiLU activation. The resulting features are then fed into the SS2D module, which leverages a Selective Scan mechanism to model long-range dependencies with linear computational complexity. The output of SS2D is again normalized by LN. The outputs from the two paths are fused via element-wise multiplication to integrate complementary information. A subsequent linear layer adjusts the feature channels, and the result is added back to the original input X_{in} through a residual connection to produce the final output X_{vout}, which is then used for Patch Merging at the end of the first three stages. These operations increase the channel dimension while reducing the spatial resolution of the features. The entire computational process of the VSS Block is formulated as:

$$X_{norm} = \mathrm{LN}(X_{in}), \tag{10}$$

$$X_{lin} = \mathrm{SiLU}(\mathrm{Linear}(X_{norm})), \tag{11}$$

$$X_{ss2d} = \mathrm{LN}(\mathrm{SS2D}(\mathrm{SiLU}(\mathrm{DWConv}_{3\times3}(\mathrm{Linear}(X_{norm}))))), \tag{12}$$

$$X_{vout} = \mathrm{Linear}(X_{lin} \otimes X_{ss2d}) + \mathrm{Xin} \tag{13}$$

where LN refers to Layer Normalization, SiLU is the activation function, Linear denotes the linear transformation layer, $\mathrm{DWConv}_{3\times3}$ represents 3×3 depthwise separable convolution, SS2D denotes the 2D Selective Scan computation process, and $\otimes$ indicates element-wise multiplication.

3.3. Cross-Branch Fusion Module

To enhance the accuracy of the model in land–sea segmentation tasks, we address the feature discrepancy between the dual-branch encoders by applying different attention mechanisms tailored to the characteristics of each branch. This results in a parallel dual-branch attention structure. The overall architecture of the proposed CBFM is illustrated in Figure 4. Given that the VSS branch excels at long-range spatial perception, we design a multiscale spatial attention mechanism to capture key spatial regions and emphasize fine-grained details. Simultaneously, for the MiT branch, we introduce a channel attention mechanism [50] to strengthen the explicit modeling of inter-channel relationships and suppress redundant or irrelevant information. Building upon this, we further promote

deep feature interaction between the two branches by cross-multiplying the attention-enhanced features from both. Specifically, the channel-level attention from the MiT branch is used to improve the semantic representation of the VSS branch, while the spatial-level attention from the VSS branch is leveraged to enrich the spatial detail expression of the MiT branch. This strategy ultimately leads to a more comprehensive feature fusion outcome. Moreover, to preserve the feature independence of the two encoders, prevent mutual interference, and reduce noise accumulation during the long-range modeling process [51], the features fused by the CBFM are not fed into subsequent layers of the encoders. Instead, they are directly transmitted to the corresponding decoder stages via skip connections.

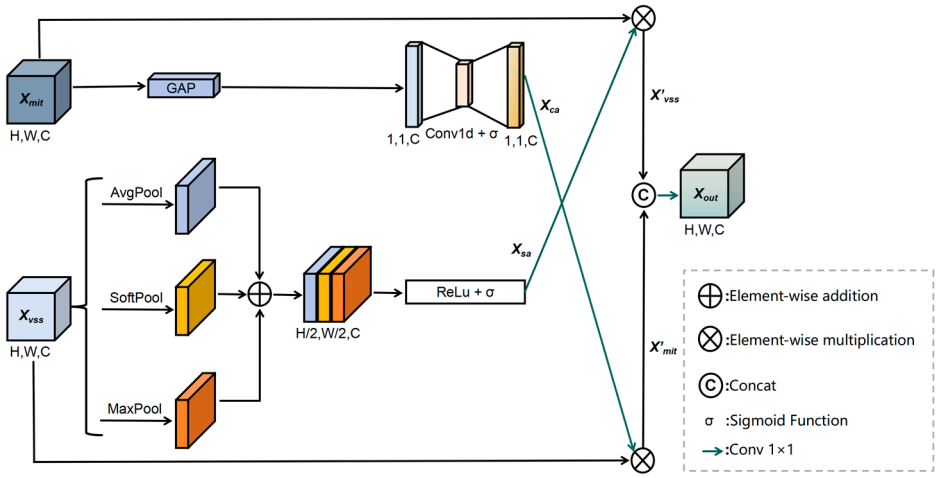


Figure 4. Principle of the proposed Cross-Branch Fusion Module.

We denote the feature maps from the MiT and VSS encoder branches as X_{mit} and X_{vss}, respectively. First, global average pooling is performed along the channel dimension of X_{mit} to extract global channel-wise information. This is followed by a one-dimensional convolution operation with a learnable kernel to model the inter-channel dependencies. A Sigmoid activation function is then applied to generate the channel attention weights. Next, we perform channel alignment using a 1×1 convolution, and conduct element-wise multiplication with X_{vss} to enhance the sensitivity of the VMamba branch to global semantic features. For the VSS branch, we apply a multiscale pooling operation consisting of average pooling, soft pooling [52], and max pooling, which extracts spatial information at different scales and generates a spatial attention distribution. The three pooling methods each have their own focus: average pooling emphasizes global statistical features, max pooling highlights locally salient regions, while soft pooling balances smoothness and detail preservation, enabling better retention of complex information. The three methods are complementary and can comprehensively model feature maps at different scales, thereby enhancing the feature representation capability. This spatial attention is then passed through a 1×1 convolution layer to align its channel dimensions with those of the MiT branch. The resulting spatial attention is applied to X_{mit} through element-wise multiplication, thereby enhancing the ability of the MiT branch to model fine-grained spatial features. Finally, we concatenate the outputs along the channel dimension, followed by 1×1 convolution for dimensionality reduction, yielding the final fused feature map X_{out}. The entire computation process can be described as follows:

$$X_{ca} = \mathrm{Conv}_{1\times1}(\sigma(\mathrm{Conv1D}(\mathrm{GAP}(X_{mit})))), \tag{14}$$

$$X_{sa} = \mathrm{Conv}_{1\times1}(\sigma(\mathrm{ReLU}(\mathrm{AvgPool}(X_{vss}) + \mathrm{MaxPool}(X_{vss}) + \mathrm{SoftPool}(X_{vss})))), \tag{15}$$

$$X'_{mit} = X_{ca} \otimes X_{vss}, \tag{16}$$

$$X'_{vss} = X_{sa} \otimes X_{mit}, \tag{17}$$

$$X_{out} = Conv_{1\times1}(Concat(X'_{mit}, X'_{vss})), \tag{18}$$

where $Conv_{1\times1}$ denotes 1×1 convolution, σ represents the Sigmoid activation function, ReLU denotes the ReLU activation function, $Conv1d(\cdot)$ indicates 1D convolution with a kernel size of K, $GAP(\cdot)$ denotes global average pooling, AvgPool, MaxPool, and SoftPool represent average pooling, max pooling, and soft pooling, respectively, and Concat refers to concatenation along the channel dimension.

3.4. Decoder Design

Following the four-stage feature-extraction process of the dual-branch encoder, the network obtains contextual information at multiple scales. Deeper layers capture rich global semantic features, which are beneficial for higher-level cognitive and understanding tasks. In contrast, shallower layers contain crucial edge and shape details, which are essential for precise coastline boundary localization. Therefore, the effectiveness of feature fusion across different levels plays a pivotal role in determining the overall performance of the network. Conventional decoder designs typically rely on simple upsampling and layer-by-layer fusion. However, such approaches are prone to information loss and excessive smoothing of edges, thereby degrading the segmentation performance in boundary-sensitive tasks such as land–sea segmentation [53]. To address these limitations, we propose a customized decoder structure aimed at enhancing the feature fusion ability. The decoder primarily comprises skip connections, dual residual convolutional blocks, and three customized decoder blocks. The structure of the decoder block is illustrated in Figure 5.

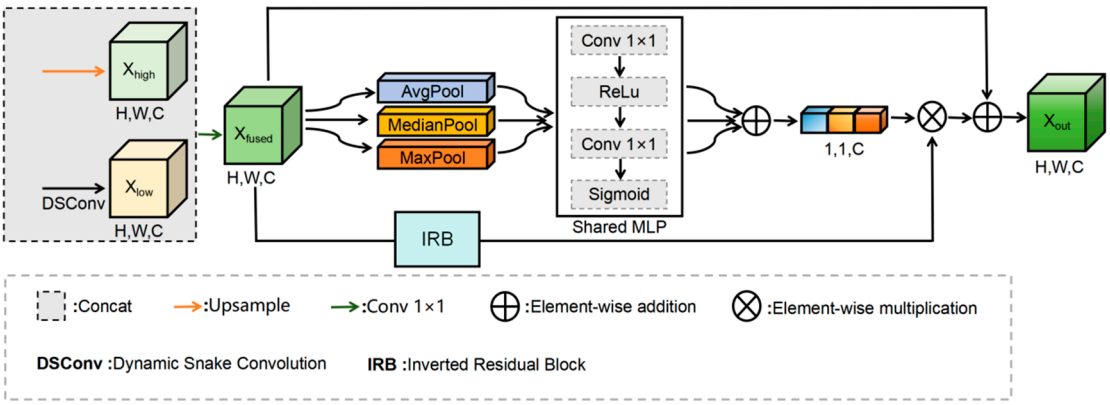


Figure 5. Schematic of the proposed decoder block.

In the decoding process, we first enhance the deepest-level features using a Dual Residual Convolution Block to fully leverage global semantic information and alleviate the vanishing gradient problem. Subsequently, three customized decoder blocks are employed to progressively integrate multiscale features across layers, ensuring the effective combination of semantic information and spatial details. In each decoder block, X_{high} and X_{low} denote deep features and shallow features, respectively. The deep features of X_{high} are first upsampled, while the shallow feature of X_{low} are transmitted via skip connection and undergo spatial refinement through a Dynamic Snake Convolution (DSConv) [54]. Next, the two feature maps are concatenated along the channel dimension and passed through a 1×1 convolution layer to reduce the dimensionality, resulting in the fused feature map X_{fused}. To further enhance the representation capability, we introduce a channel attention mechanism that determines the channel importance by applying average pooling (AvgPool), median pooling (MedianPool), and max pooling (MaxPool). The inclusion of median pooling is beneficial for denoising while preserving critical feature information [55].

The resulting attention weights are then used to reweight the channels of X_{fused}, and the weighted features are multiplied element-wise with the features refined by an Inverted Residual Block [56]. Finally, the result is added back element-wise to the original X_{fused}, reinforcing the feature representation and yielding the final output X_{out}. A Segmentation Head is then used to generate the class prediction map, which is subsequently upsampled to the input resolution to obtain the final land–sea segmentation result. The computation process of the decoder block can be formulated as follows:

$$X_{fused} = \mathrm{Conv}_{1\times1}\Big(\mathrm{Concat}\Big(X_{high}, \mathrm{DSC}(X_{low})\Big)\Big), \tag{19}$$

$$X_{ca} = \mathrm{MLP}(\mathrm{AvgPool}(X_{fused})) + \mathrm{MLP}(\mathrm{MedianPool}(X_{fused})) + \mathrm{MLP}(\mathrm{MaxPool}(X_{fused})), \tag{20}$$

$$X_{irb} = \mathrm{IRB}(X_{fused}), \tag{21}$$

$$X_{out} = X_{ca} \otimes X_{irb} + X_{fused}, \tag{22}$$

where Concat denotes concatenation along the channel dimension, DSC refers to dynamic snake convolution, $\mathrm{Conv}_{1\times1}$ represents 1×1 convolution, and AvgPool, MedianPool, and MaxPool denote average pooling, median pooling, and max pooling, respectively. MLP consists of two 1×1 convolution layers, ReLU and Sigmoid are activation functions, IRB refers to the Inverted Residual Block, and $\otimes$ denotes element-wise multiplication.

Finally, the overall information flow of VMMT-Net is illustrated in Figure 6. In the encoder stage, the input image first undergoes feature extraction through the parallel dual-branch encoder, with cross-branch feature fusion achieved through CBFM. In the decoder stage, the deepest features are processed by residual convolution blocks, and are then progressively fused with shallow features transmitted through skip connections in the decoder block. This step continuously enhances the spatial resolution, ultimately generating the land–sea segmentation result map.

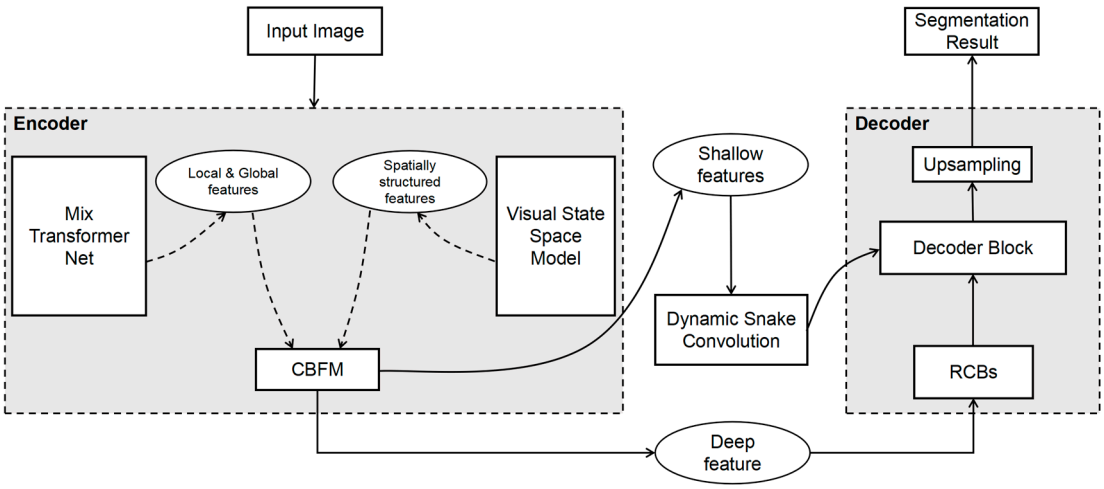


Figure 6. Overview of VMMT-Net data flow.

4. Experiments

All experiments were conducted on a Linux system running Ubuntu 22.04. The hardware configuration included a 16-core AMD EPYC 7542 CPU, an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM, and 56 GB of RAM. The software environment comprised Python 3.10.13, PyTorch 2.0.1, and the mamba-ssm (version 1.2.0) for state space modeling. The AdamW optimizer was employed during training, with an initial learning rate set to 3×10^{-4}, a batch size of 6, and a total of 150 training epochs. The Cross-Entropy Loss function was used.

4.1. Datasets

4.1.1. Benchmark Sea–Land Dataset

The Benchmark Sea–Land Dataset (BSD) [57] is a coastline dataset constructed from Landsat-8 OLI imagery, covering coastal regions of China. The dataset contains a total of 1950 training images and 1411 images for validation and testing. Each image has a spatial resolution of 512 × 512 pixels. Rivers and lakes located in land areas are categorized as land, resulting in two target classes: sea and land. For the experiments, the red–green–blue bands (bands 4, 3, and 2) were selected. Sample images from the BSD dataset are shown in Figure 7a.

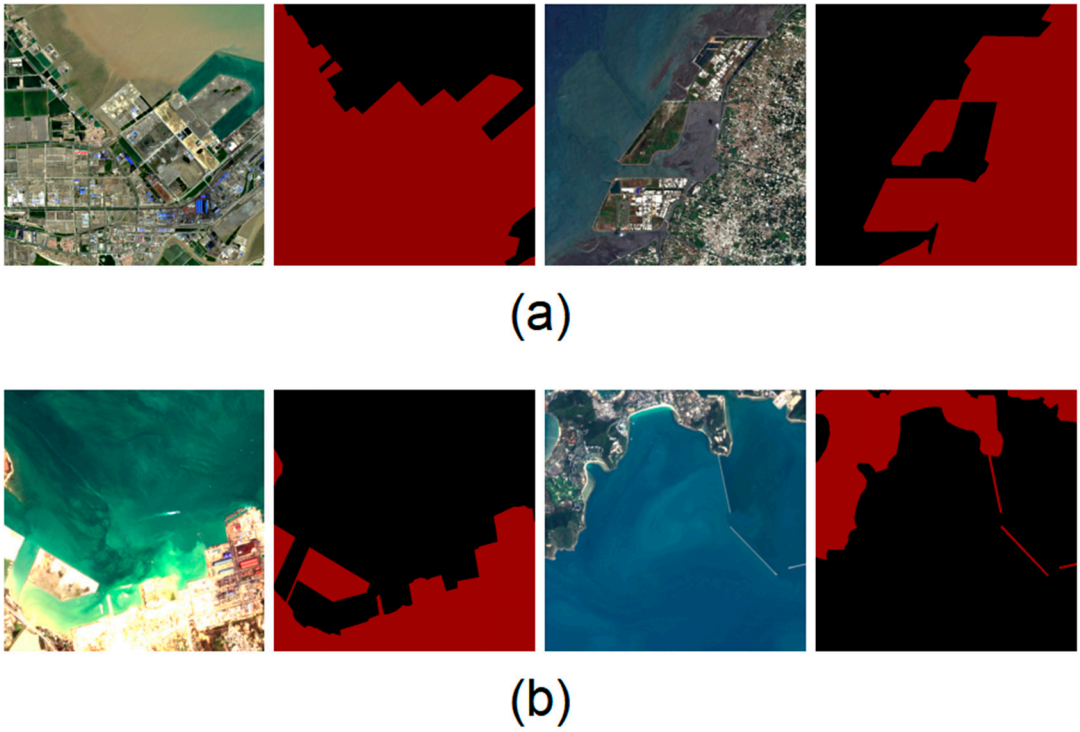


Figure 7. Sample images from BSD and GF-HNCD datasets. (**a**) Sample images from the BSD dataset, where red and black pixels represent land and ocean, respectively. (**b**) Sample images from the GF-HNCD dataset, where red and black pixels represent land and ocean, respectively.

4.1.2. GF-HNCD

The GF-HNCD dataset [47] was specially constructed for the fine-grained land–sea segmentation of Hainan Island and associated tasks. The dataset covers the entirety of Hainan Island, as well as partial coastal areas of Guangdong and Guangxi provinces. The dataset contains eight original images with dimensions of 12,000 × 13,400 pixels, acquired by the GF-1 WFV satellite at a spatial resolution of 16 m. The original images are first cropped into 512 × 512-pixel image patches using a sliding window approach, and then samples affected by cloud occlusion are removed. The 5010 sample images retained after this procedure are divided into training, validation, and test sets at a ratio of 8:1:1. Data augmentation operations (random cropping, scaling, flipping, and color space transformation) are applied during the training phase. The images adopt a red–green–blue (bands 4, 3, 2) combination and contain two categories: ocean and land. Sample images from the GF-HNCD dataset are shown in Figure 7b.

4.2. Evaluation Metrics

To comprehensively evaluate the performance of the proposed VMMT-Net in land–sea segmentation tasks, we adopt the mean Intersection over Union (MIoU) and mean F1-score

(mF1) as primary accuracy metrics. These metrics are derived from the confusion matrix and are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{23}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{24}$$

$$\text{F1score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{25}$$

$$\text{MIoU} = \frac{1}{\text{k} + 1} \sum_{\text{i}=0}^{\text{k}} \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \tag{26}$$

where TP, FP, FN, and TN represent the number of true positives, false positives, false negatives, and true negatives, respectively.

The numbers of parameters and floating-point operations per second (FLOPs) are used to evaluate the computational complexity of the model. The corresponding formulations are as follows:

$$\text{Params} = \text{K}_\text{h} \times \text{K}_\text{w} \times \text{C}_\text{in} \times \text{C}_\text{out}, \tag{27}$$

$$\text{FLOPs} = \text{Params} \times \text{H}_\text{out} \times \text{W}_\text{out}, \tag{28}$$

where $\text{K}_\text{h} \times \text{K}_\text{w}$ denotes the kernel size, C_in and C_out represent the number of input and output channels, respectively, and $\text{H}_\text{out} \times \text{W}_\text{out}$ corresponds to the spatial dimensions of the output feature map.

4.3. Comparative Experiments and Results Analysis

To comprehensively validate the performance advantages of the proposed VMMT-Net model in land–sea segmentation tasks, this paper conducts comparative experiments with eight representative semantic segmentation models on the two remote sensing coastline datasets described earlier. These models cover different structural types, including CNN-based UNet [16] and PSPNet [21], pure Transformer-based SegFormer [30], CNN-Transformer hybrid architectures including TCUNet [46], CLCFormer [34], and FTransUNet [32], Visual State Space Model (VSSM)-based VM-UNet [35], and RS3Mamba [37] which combines CNN with VSSM. Specifically, UNet adopts the original fully convolutional network as its backbone, PSPNet uses ResNet50, SegFormer is configured as MiT-B0, TCUNet combines PVT V2 with ResNet to construct a dual-branch encoder, CLCFormer combines SwinV2 and EfficientNet-B3 to form a parallel dual-branch structure, FTransUNet uses ResNet50 in both of its dual branches as well as FVit to form the backbone network, VM-UNet employs VMamba as the backbone, and RS3Mamba fuses ResNet18 with VMamba. To ensure fairness, all models are trained under the same experimental environment and uniformly without using any pre-trained weights.

Table 1 compares the performance of the different semantic segmentation models and the proposed VMMT-Net on the GF-HNCD and BSD remote sensing coastline datasets. The results demonstrate that VMMT-Net achieves optimal performance on all datasets, exhibiting strong generalization capability and robustness. Specifically, the mF1 scores reach 98.48% and 98.53% on the GF-HNCD and BSD datasets, respectively, with corresponding MIoU scores of 97.02% and 97.11%. From the perspective of model architecture analysis, hybrid structures generally outperform single structures on both datasets. For example, TCUNet, CLCFormer, and FTransUNet, which fuse CNN and Transformer modules, not only surpass the pure CNN-based U-Net and PSPNet on both datasets, but also outperform the pure Transformer-based SegFormer. RS3Mamba, which combines CNN with VMamba, also demonstrates superior performance to VM-UNet, which uses only VMamba as its backbone, on both datasets, further illustrating that hybrid structures possess stronger

modeling capabilities in terms of capturing local details and global contextual information. Furthermore, the computational complexity indicators (Parameters and FLOPs) indicate that, although FTransUNet achieves outstanding performance on the GF-HNCD dataset, it requires 184.14 M parameters and 57.28 G FLOPs, significantly more than other models. While CLCFormer and RS3Mamba are relatively lightweight, their parameter counts still exceed that of VMMT-Net. Although TCUNet has the smallest parameter count at only 1.72 M, its performance is inferior to that of our model. Therefore, VMMT-Net not only achieves state-of-the-art performance in terms of segmentation accuracy, but also realizes a better balance between accuracy and complexity.

Table 1. Performance and segmentation results comparison of different methods. The best results are highlighted in bold.

Model	Backbone	GF-HNCD		BSD		Parameters (/M)	FLOPs (/G)
		mF1 (%)	MIoU (%)	mF1 (%)	MIoU (%)		
U-Net	-	94.79	90.10	97.02	94.22	28.98	203
PSPNet	ResNet50	96.47	93.18	97.82	95.75	46.69	179
SegFormer	MiT-B0	97.01	94.20	97.70	95.51	3.75	8.50
TCUNet	PVT V2-ResNet	97.94	95.96	98.08	96.23	1.72	3.24
CLCFormer	SwinV2-EfficientNet-B3	97.17	94.50	97.88	95.86	38.01	31.06
FTransUNet	ResNet50-FVit	98.29	96.65	97.95	95.99	184.14	57.28
VM-UNet	VSS	97.44	95.02	97.95	96.00	22.03	16.45
RS3Mamba	VSS-ResNet18	98.24	96.54	98.42	96.89	31.65	43.32
VMMT-Net (Ours)	**-**	**98.48**	**97.02**	**98.53**	**97.11**	**28.24**	**25.21**

To provide a more intuitive validation of the effectiveness of the proposed VMMT-Net, Figures 8 and 9 present the segmentation results of all compared methods on the GF-HNCD and BSD remote sensing coastline datasets. As can be observed, our model achieves the best visual segmentation performance, particularly in the areas highlighted by the orange rectangles. Based on the distinct characteristics of land–sea features in remote sensing imagery and their segmentation outcomes, we focus on the following two aspects:

(1) Perception of complex morphology and fine-scale structures

Figure 8 (images 1–5 and image 8) and Figure 9 (images 1–3 and image 8) display coastal scenes characterized by complex morphologies and fine-grained structures. These include meandering small-scale water bodies and elongated man-made structures such as piers and breakwaters. Common to these features are their small scale, blurred boundaries, and spatial structural complexity. These properties make them particularly prone to misclassification or omission in land–sea segmentation tasks, often resulting in contour discontinuities, blurred boundaries, or structural loss. This places high demands on the model's ability to capture spatial structures and preserve fine details. A comparison of the segmentation results across methods reveals that VMMT-Net demonstrates superior performance in such scenarios. The proposed method accurately reconstructs the geometric continuity of small-scale water bodies and clearly delineates the edges and contours of man-made structures, preserving their spatial integrity. In contrast, other models suffer from issues such as indistinct boundaries, structural fragmentation, or incomplete recognition. Therefore, in land–sea segmentation tasks with high spatial complexity and fine-grained boundary requirements, our model exhibits stronger adaptability and representational power.

(2) Modeling of blurred land–sea transition zones

Figure 8 (images 6–7) and Figure 9 (images 4–7) also depict coastline scenes in which the transition between land and sea is visually ambiguous. These regions are characterized by broad spectral mixing zones, gradual grayscale transitions along the boundaries, and interference from natural factors such as cloud cover and water disturbances in remote sensing imagery. Such areas lack clearly defined geometric edges and often exhibit visually indistinct or partially occluded information, posing significant challenges for precise land–sea segmentation. In practical results, VMMT-Net restores the land–sea boundaries more clearly, avoiding edge discontinuities and false contours. It also accurately localizes coastlines in low-contrast and visually ambiguous transition zones, demonstrating strong robustness under the uncertain environmental conditions commonly encountered in remote sensing. In contrast, other models frequently exhibit boundary drift and misclassification in these regions, struggling to cope with the uncertainties introduced by boundary ambiguity. This performance advantage can be attributed to our model's exceptional global modeling and multiscale feature representation capabilities, which reduce misjudgment in low-contrast coastal regions. Consequently, our model exhibits greater practical potential for application in coastal scenes with weak boundary characteristics.

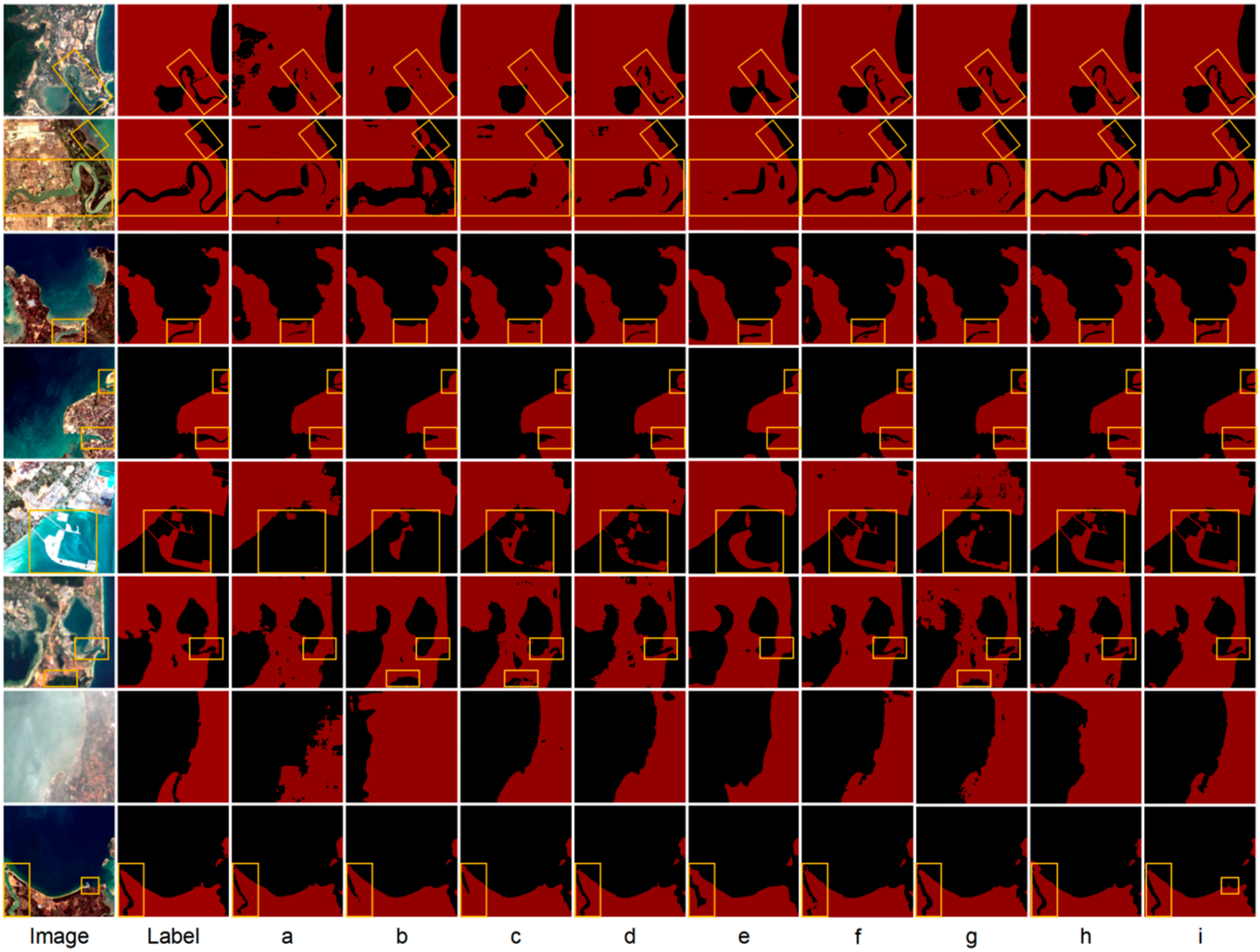


Figure 8. Segmentation results based on GF-HNCD: (**a**) UNet, (**b**) PSPNet, (**c**) SegFormer, (**d**) TCUNet, (**e**) CLCFormer, (**f**) FTransUNet, (**g**) VM-UNet, (**h**) RS3Mamba, and (**i**) VMMT-Net.

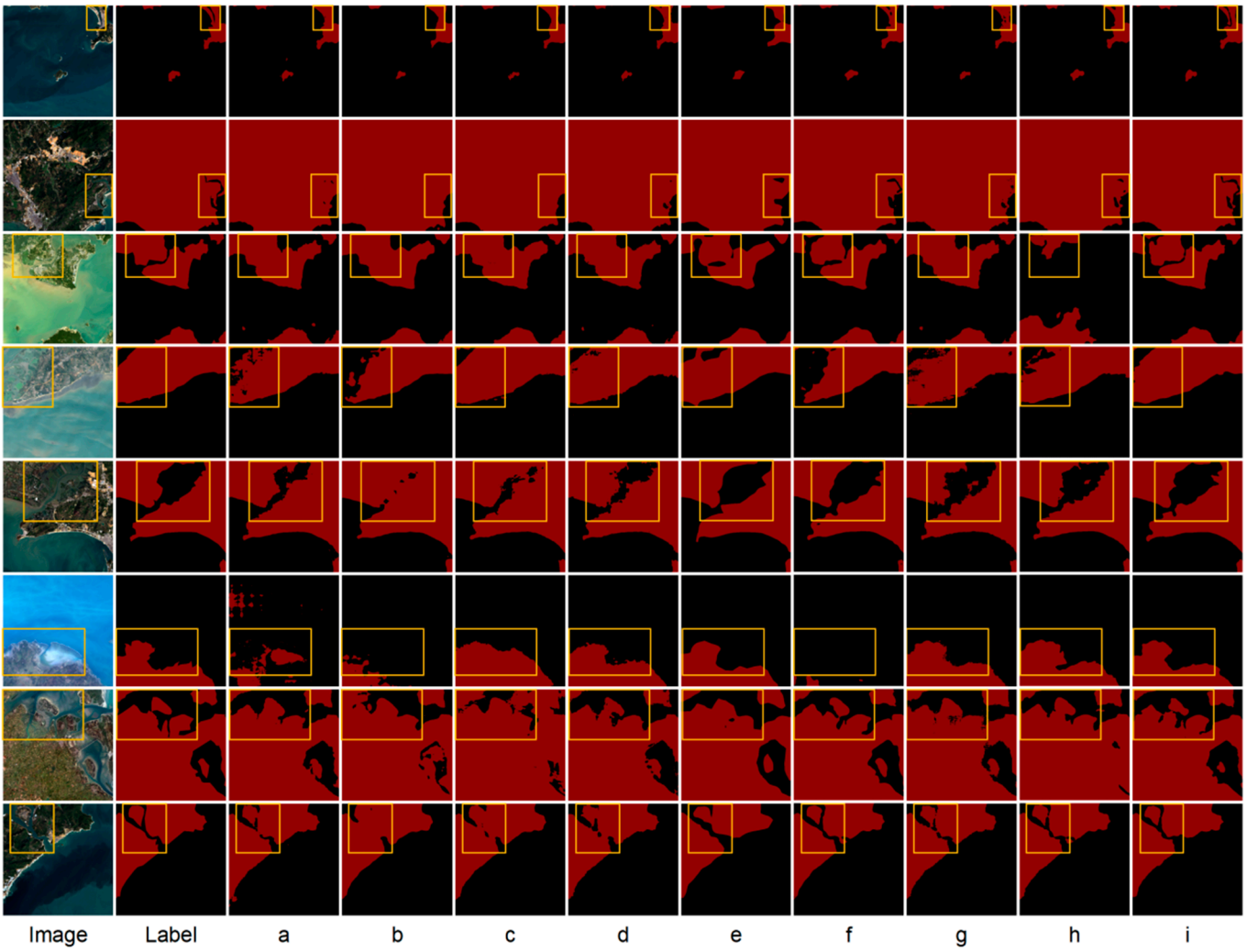


Figure 9. Segmentation results based on BSD: (**a**) UNet, (**b**) PSPNet, (**c**) SegFormer, (**d**) TCUNet, (**e**) CLCFormer, (**f**) FTransUNet, (**g**) VM-UNet, (**h**) RS3Mamba, and (**i**) VMMT-Net.

4.4. Ablation Study

To systematically evaluate the contribution of each component within the proposed VMMT-Net framework, we conducted a series of ablation experiments using the GF-HNCD dataset. The results are summarized in Table 2. Using the complete VMMT-Net model (denoted as "Full") as the baseline, we sequentially removed key modules or combinations of structures while keeping the remaining architecture unchanged. This setup allowed us to isolate and analyze the individual impact of each component on the overall model performance. It is important to note that, in all ablation configurations, we maintained the same number of MiT Blocks and VSS Blocks per stage as in the full model, ensuring fair and consistent comparisons. The ablation settings are as follows:

- VMMT-Net (Full): The complete model developed in this study, which includes MiT, VSS, CBFM, and the customized decoder.
- w/o VSS and CBFM: The VSS branch is removed from the encoder, with only the MiT branch retained as the backbone. The CBFM is also removed accordingly.
- w/o MiT and CBFM: The MiT branch is removed from the encoder, with only the VSS branch retained as the backbone. The CBFM is likewise removed in this configuration.

Table 2. Overview of ablation experiments on the GF-HNCD dataset. The best results are highlighted in bold.

Method	mF1 (%)	MIoU (%)
VMMT-Net (Full)	**98.48**	**97.02**
w/o VSS and CBFM	97.55	95.22
w/o MiT and CBFM	97.99	96.06
w/o CBFM	98.22	96.51
w/o Decoder	98.17	96.40

The CBFM is specifically designed to facilitate interaction between the dual-branch encoder. Thus, it becomes functionally irrelevant when either the MiT or VSS branch is excluded. Therefore, to preserve the architectural clarity and prevent fusion-related mechanisms from influencing the results, the CBFM is removed in such scenarios, and skip connections are used to directly forward features from each stage to the decoder.

- w/o CBFM: The CBFM is removed from all stages and replaced with a simple concatenation followed by 1×1 convolution.
- w/o Decoder: The customized decoder is replaced with a UNet-style decoder to assess the contribution of the decoder design.

As illustrated by the ablation results in Table 2 and Figure 10, the complete VMMT-Net model achieves the best performance on the GF-HNCD dataset, with an mF1 score of 98.48% and an MIoU of 97.02%, demonstrating its strong ability in land–sea segmentation tasks. First, when the VSS branch and CBFM are removed, the model performance drops significantly, with mF1 and MIoU reduced to 97.55% and 95.22%, respectively. As shown in Figure 10a, the segmentation results exhibit broken water boundaries and structural discontinuities, highlighting the critical role of the VSS branch in modeling spatial structural continuity. This advantage is attributed to the SS2D mechanism within the VSS module, which excels at capturing long-range dependencies and enforcing spatial consistency. Conversely, when the MiT branch and CBFM are removed, the mF1 and MIoU drop to 97.99% and 96.06%, respectively. As illustrated in Figure 10b, the model's ability to recognize small-scale structures is diminished, indicating that although MiT enhances the segmentation accuracy, it is less effective than VSS in modeling global structural coherence. This finding confirms the complementary nature of the dual-branch architecture for joint modeling. Subsequently, when CBFM is replaced with a simple concatenation and convolution operation, the mF1 and MIoU decrease by 0.26% and 0.51%, respectively. Figure 10c shows that the simplified feature fusion mechanism results in inadequate integration between branches, leading to suboptimal perception of fine structures. This underscores the importance of CBFM in facilitating cross-branch feature interaction and enhancing structural detail. Finally, replacing the customized decoder with a standard U-Net-style decoder reduces the mF1 and MIoU to 98.17% and 96.40%, respectively. As depicted in Figure 10d, the segmentation quality deteriorates, further validating the effectiveness of the proposed decoder in multiscale feature integration and spatial detail restoration. Overall, the prediction results of the full model, as shown in Figure 10e, demonstrate the highest accuracy in land–sea boundary delineation and small-scale object recognition, confirming the effectiveness of each component in collaborative modeling.

To further validate the advantages of the proposed CBFM module in cross-branch feature fusion, we replaced CBFM with two current mainstream feature fusion modules, CCM [37] and BiFFM [34], and conducted comparative experiments on the GF-HNCD dataset. The specific experimental settings are as follows:

- VMMT-Net (Full): The complete model structure proposed in this paper, using the CBFM module for cross-branch feature fusion;
- r CCM: Replacing CBFM in the model with CCM;
- r BiFFM: Replacing CBFM in the model with BiFFM.

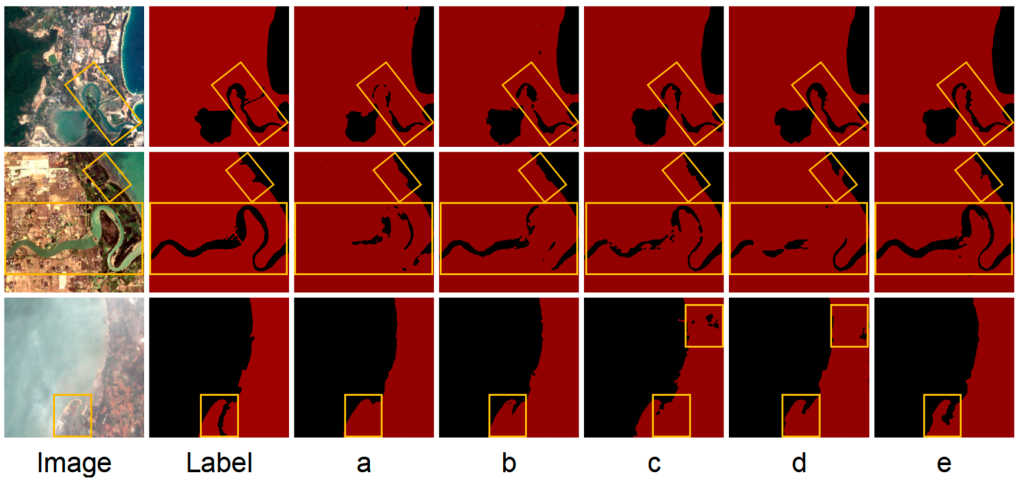


Figure 10. Visual comparison of ablation results on the GF-HNCD dataset: (**a**) w/o VSS and CBFM, (**b**) w/o MiT and CBFM, (**c**) w/o CBFM, (**d**) w/o Decoder, and (**e**) VMMT-Net (Full).

The results in Table 3 demonstrate that when CBFM is replaced with CCM or BiFFM, the VMMT-Net model experiences varying degrees of decline in both segmentation accuracy and computational efficiency on the GF-HNCD dataset. The complete model (VMMT-Net (Full)) achieves mF1 and MIoU scores of 98.48% and 97.02%, respectively, with 28.24 M parameters and 25.21 G FLOPs. This represents the optimal segmentation performance and relatively low computational complexity. When CBFM is replaced with CCM, the mF1 and MIoU scores decrease to 98.24% and 96.55%, respectively, while the parameter count and FLOPs increase to 36.39 M and 31.75 G, respectively. Although the segmentation accuracy exhibits only a slight decline, the model complexity increases significantly, indicating that CCM's information-interaction efficiency in multi-branch structures is inferior to that of CBFM. Furthermore, when CBFM is replaced with BiFFM, the mF1 and MIoU scores are 98.40% and 96.86%, respectively, with a parameter count as high as 51.21 M and FLOPs reaching 43.03 G. The model complexity increases substantially, but the accuracy improvement is limited and still fails to surpass that when using CBFM. Figure 11 presents the visualization results given by the different feature fusion modules on GF-HNCD, further validating the advantages of CBFM. Overall, although different feature fusion modules show similar performance in terms of segmentation accuracy, CBFM significantly reduces the number of model parameters and the computational load while achieving the same or even better performance, fully demonstrating the efficiency and lightweight advantages of its structural design.

Table 3. Performance comparison of different feature fusion modules on GF-HNCD. Best results are shown in bold.

Method	mF1 (%)	MIoU (%)	Parameters (/M)	FLOPs (/G)
VMMT-Net (Full)	**98.48**	**97.02**	**28.24**	**25.21**
r CCM	98.24	96.55	36.39	31.75
r BiFFM	98.40	96.86	51.21	43.03

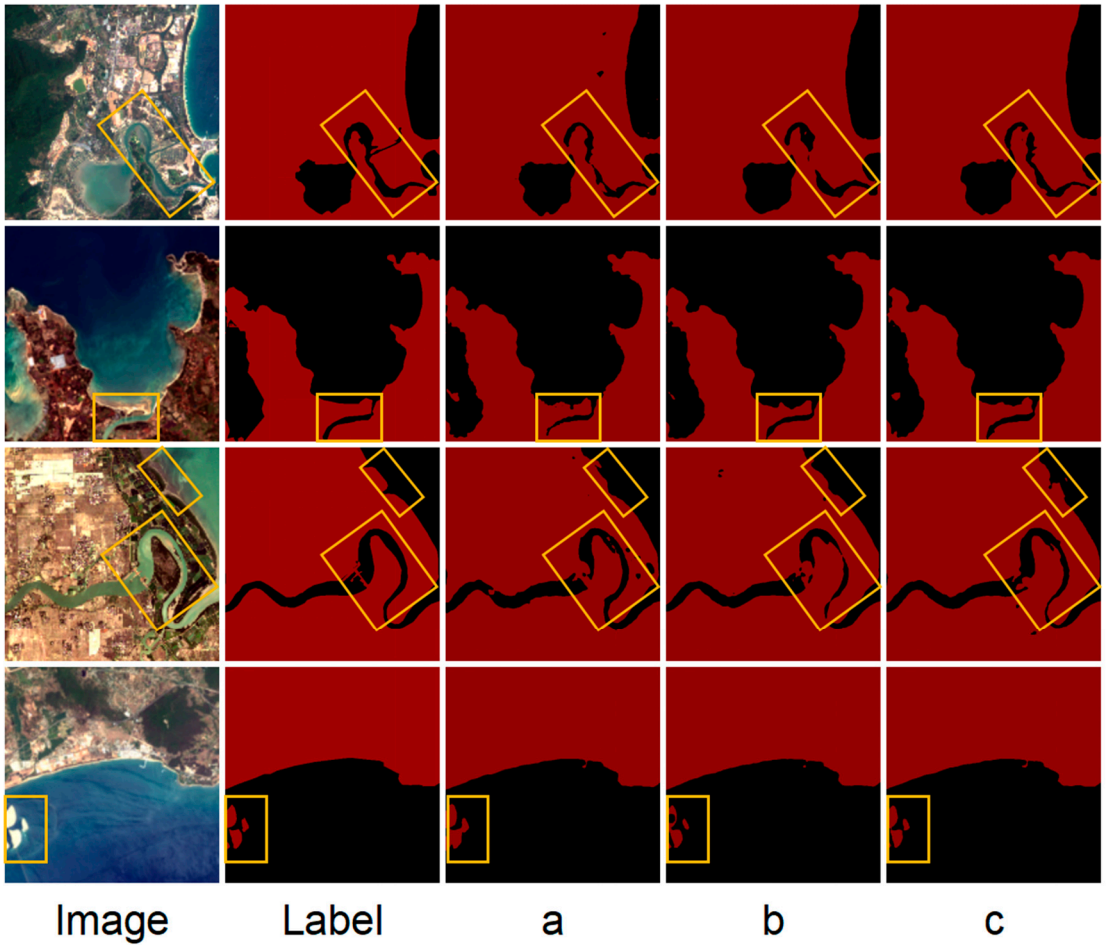


Figure 11. Visualization results of different feature fusion modules on GF-HNCD: (**a**) r CCM, (**b**) r BiFFM, (**c**) VMMT-Net (Full).

5. Discussion

Semantic segmentation technology has been extensively used for image processing in the field of remote sensing, but there has been no research on the efficient collaborative modeling between Transformer and Mamba modules. Moreover, given the requirements of land–sea segmentation tasks for complex coastal remote sensing images, existing models have many limitations. Therefore, we have developed VMMT-Net to explore the feasibility of parallel collaborative modeling using the complementary advantages of the MiT and VMamba, and have applied the proposed approach to the complex task of land–sea segmentation of remote sensing images. Based on comprehensive analysis of the experimental results, we believe that the three core innovations of VMMT-Net contribute to the realization of efficient and accurate land–sea segmentation. The main contributions of this study are reflected in the following aspects:

(1) The dual-branch encoder balances the modeling of multiscale features and spatial structures. Unlike traditional feature-extraction strategies, VMMT-Net adopts MiT and VMamba to construct a parallel dual-branch encoder structure. This strategy fully leverages the advantages of MiT in multiscale semantics and detail capture, and uses VMamba to enhance the modeling of spatial structural continuity. The two components complement each other synergistically, addressing the limitations of traditional CNN models in global dependency modeling and the constraints of Transformer structures in directional perception modeling. Ablation experiments

(Table 2) show that removing either branch results in a decline in the mF1 and MIoU metrics. This demonstrates the complementary performance of the two architectures.

(2) CBFM enhances the feature interaction and semantic compensation between different branches. This module employs a dual-branch structure with attention cross-fusion mechanisms to achieve the simultaneous perception of complex land–sea backgrounds and key regions, which is significant for land–sea segmentation in ambiguous areas and discrimination of small-scale structures. Moreover, the experimental data in Table 3 demonstrate that CBFM is more lightweight than existing feature fusion modules (such as CCM and BiFFM), effectively conserving computational resources. This further demonstrates that feature fusion is not necessarily better when more complex—the key lies in precise feature selection and fusion strategies.

(3) The multi-level customized decoder promotes detail restoration and spatial consistency. The decoder block employs a strategy combining multiscale feature fusion, dynamic snake convolution, and channel attention mechanisms, which further enhances the model's ability to restore land–sea boundary details, ensures the integrity of land–sea structures, and overcomes the issue of overly smooth transitions in land–sea edge segmentation results observed in traditional models. Table 2 and Figure 10 indicate that the custom decoder plays a crucial role in restoring boundary details and ensuring the integrity of the land–sea structure.

Overall, under the collaborative modeling of the three key components described above, VMMT-Net ensures the extraction and fusion of features from different dimensions while maintaining appropriate computational complexity. However, compared with some advanced methods, VMMT-Net has not yet achieved optimality in metrics such as parameter count and FLOPs. Furthermore, although our model demonstrates high segmentation accuracy in ambiguous land–sea regions and fine-grained structure recognition, its performance still needs further improvement when processing densely interwoven land–sea scenarios. To address these limitations, future research will attempt to optimize the model structure, explore more efficient and lightweight designs, and extend the approach to knowledge graph-guided multi-modal remote sensing data-driven models to meet the application requirements for larger-scale, higher-precision remote sensing image land–sea segmentation.

6. Conclusions

This study has addressed the complexity and challenges associated with automatic coastline extraction from remote sensing imagery by proposing a novel semantic segmentation architecture, VMMT-Net. The proposed framework employs a heterogeneous dual-branch encoder that integrates MiT and VMamba, effectively combining the strengths of MiT in modeling local details and global semantics with the advantages of VMamba in capturing spatial structural continuity and directional awareness through visual state space modeling. To enable efficient feature interaction between the two branches, we designed the CBFM, which facilitates complementary and collaborative feature representation at various stages of the network. In addition, a customized decoder was introduced to integrate multiscale semantic information with fine-grained boundary features, further enhancing the segmentation accuracy and boundary restoration capabilities. Extensive experiments were conducted on two coastal remote sensing datasets, GF-HNCD and BSD, and comparative analyses were performed against multiple mainstream methods. The results demonstrate that VMMT-Net outperforms existing state-of-the-art approaches across a range of quantitative metrics and qualitative evaluations. Specifically, the proposed model shows superior performance in handling complex morphological structures, fine-

scale features, and ambiguous land–sea transition zones, thereby highlighting its strong generalization capability and practical applicability in real-world scenarios.

Author Contributions: Conceptualization, H.X. and J.W.; Methodology, H.X. and J.W.; software, J.W. and Z.L.; validation, Z.L., Z.Z. and C.S.; formal analysis, H.X. and J.W.; investigation, X.W. and C.S.; resources, H.X.; data curation, J.W. and Z.L.; writing—original draft preparation, J.W. and H.X.; writing—review and editing, J.W., H.X. and X.W.; visualization, Z.Z. and C.S.; supervision, H.X.; project administration, H.X.; funding acquisition, H.X. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number: 62066013) and the Hainan Provincial Natural Science Foundation of China (grant numbers: 622RC674 and 623RC480).

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

Acknowledgments: The author would like to thank the anonymous reviewers for their comments and constructive suggestions for improving the paper.

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

References

1. Yang, G.; Huang, K.; Zhu, L.; Sun, W.; Chen, C.; Meng, X.; Wang, L.; Ge, Y. Spatio-temporal changes in China's mainland shorelines over 30 years using Landsat time series data (1990–2019). *Earth Syst. Sci. Data Discuss.* **2024**, *2024*, 1–26. [CrossRef]
2. Zhang, L.; Li, G.; Liu, S.; Wang, N.; Yu, D.; Pan, Y.; Yang, X. Spatiotemporal variations and driving factors of coastline in the Bohai Sea. *J. Ocean Univ. China* **2022**, *21*, 1517–1528. [CrossRef]
3. Hou, X.Y.; Wu, T.; Hou, W.; Chen, Q.; Wang, Y.; Yu, L. Characteristics of coastline changes in mainland China since the early 1940s. *Sci. China Earth Sci.* **2016**, *59*, 1791–1802. [CrossRef]
4. Zhou, X.; Wang, J.; Zheng, F.; Wang, H.; Yang, H. An overview of coastline extraction from remote sensing data. *Remote Sens. Sens.* **2023**, *15*, 4865. [CrossRef]
5. Shirmard, H.; Farahbakhsh, E.; Müller, R.D.; Chandra, R. A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sens. Environ.* **2022**, *268*, 112750. [CrossRef]
6. Han, W.; Zhang, X.; Wang, Y.; Wang, L.; Huang, X.; Li, J.; Wang, S.; Chen, W.; Li, X.; Feng, R.; et al. A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 87–113. [CrossRef]
7. McCarthy, M.J.; Colna, K.E.; El-Mezayen, M.M.; Laureano-Rosario, A.E.; Méndez-Lázaro, P.; Otis, D.B.; Toro-Farmer, G.; Vega-Rodriguez, M.; Muller-Karger, F.E. Satellite remote sensing for coastal management: A review of successful applications. *Environ. Manag.* **2017**, *60*, 323–339. [CrossRef] [PubMed]
8. Yang, Z.; Yu, X.; Dedman, S.; Rosso, M.; Zhu, J.; Yang, J.; Xia, Y.; Tian, Y.; Zhang, G.; Wang, J. UAV remote sensing applications in marine monitoring: Knowledge visualization and review. *Sci. Total Environ.* **2022**, *838*, 155939. [CrossRef] [PubMed]
9. Xu, Q.; Zhao, B.; Dai, K.; Dong, X.; Li, W.; Zhu, X.; Yang, Y.; Xiao, X.; Wang, X.; Huang, J.; et al. Remote sensing for landslide investigations: A progress report from China. *Eng. Geol.* **2023**, *321*, 107156. [CrossRef]
10. Kucharczyk, M.; Hugenholtz, C.H. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sens. Environ.* **2021**, *264*, 112577. [CrossRef]
11. Ji, X.; Tang, L.; Chen, L.; Hao, L.-Y.; Guo, H. Toward efficient and lightweight sea–land segmentation for remote sensing images. *Eng. Appl. Artif. Intell.* **2024**, *135*, 108782. [CrossRef]
12. Lu, C.; Wen, Y.; Li, Y.; Mao, Q.; Zhai, Y. Sea-land segmentation method based on an improved MA-Net for Gaofen-2 images. *Earth Sci. Inform.* **2024**, *17*, 4115–4129. [CrossRef]
13. Lu, S.; Wu, B.; Yan, N.; Wang, H. Water body mapping method with HJ-1A/B satellite imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 428–434. [CrossRef]
14. Yang, C.S.; Park, J.H.; Harun-Al Rashid, A. An improved method of land masking for synthetic aperture radar-based ship detection. *J. Navig.* **2018**, *71*, 788–804. [CrossRef]
15. Ge, X.; Sun, X.; Liu, Z. Object-oriented coastline classification and extraction from remote sensing imagery. In *Remote Sensing of the Environment: 18th National Symposium on Remote Sensing of China*; SPIE: Bellingham, WA, USA, 2014; Volume 9158, pp. 131–137.

16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
18. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**. [CrossRef]
19. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
22. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
23. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
24. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
25. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
26. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
29. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
30. Xie, E.; Wang, W.; Yu, Z. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
31. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**. [CrossRef]
32. Ma, X.; Zhang, X.; Pun, M.-O.; Liu, M. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403215. [CrossRef]
33. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
34. Long, J.; Li, M.; Wang, X. Integrating spatial details with long-range contexts for semantic segmentation of very high-resolution remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
35. Ruan, J.; Li, J.; Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv* **2024**. [CrossRef]
36. Zhuang, P.; Zhang, X.; Wang, H.; Zhang, T.; Liu, L.; Li, J. FAHM: Frequency-Aware Hierarchical Mamba for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 6299–6313. [CrossRef]
37. Ma, X.; Zhang, X.; Pun, M.O. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5.
38. Hatamizadeh, A.; Kautz, J. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv* **2024**, arXiv:2407.08083.
39. Cheng, D.; Meng, G.; Cheng, G.; Pan, C. SeNet: Structured edge network for sea–land segmentation. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 247–251. [CrossRef]
40. Li, R.; Liu, W.; Yang, L.; Sun, S.; Hu, W.; Zhang, F.; Li, W. DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3954–3962. [CrossRef]
41. Shamsolmoali, P.; Zareapoor, M.; Wang, R.; Zhou, H.; Yang, J. A novel deep structure U-Net for sea-land segmentation in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3219–3232. [CrossRef]

42. Cui, B.; Jing, W.; Huang, L.; Li, Z.; Lu, Y. SANet: A sea–land segmentation network via adaptive multiscale feature learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 116–126. [CrossRef]
43. Gao, H.; Yan, X.D.; Zhang, H. Multi-Scale Sea-Land Segmentation Method for Remote Sensing Images Based on Res2Net. *Acta Opt. Sin.* **2022**, *42*, 1828004.
44. Ji, X.; Tang, L.; Lu, T.; Cai, C. Dbenet: Dual-branch ensemble network for sea–land segmentation of remote-sensing images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–11. [CrossRef]
45. Gao, J.; Zhou, C.; Xu, G.; Sun, W. Multiscale sea-land segmentation networks for weak boundaries. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4205511. [CrossRef]
46. Xiong, X.; Wang, X.; Zhang, J.; Huang, B.; Du, R. Tcunet: A lightweight dual-branch parallel network for sea–land segmentation in remote sensing images. *Remote Sens.* **2023**, *15*, 4413. [CrossRef]
47. Tong, Q.; Wu, J.; Zhu, Z.; Zhang, M.; Xing, H. STIRUnet: SwinTransformer and inverted residual convolution embedding in unet for Sea–Land segmentation. *J. Environ. Manag.* **2024**, *357*, 120773. [CrossRef] [PubMed]
48. Chen, Y.; Zhang, L.; Chen, B.; Zuo, J.; Hu, Y. MPG-Net: A Semantic Segmentation Model for Extracting Aquaculture Ponds in Coastal Areas from Sentinel-2 MSI and Planet SuperDove Images. *Remote Sens.* **2024**, *16*, 3760. [CrossRef]
49. Ai, J.; Xue, W.; Zhu, Y.; Zhuang, S.; Xu, C.; Yan, H.; Chen, L.; Wang, Z. AIS-PVT: Long-time AIS Data assisted Pyramid Vision Transformer for Sea-land Segmentation in Dual-polarization SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3449894. [CrossRef]
50. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
51. Wei, K.; Dai, J.; Hong, D.; Ye, Y. MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *135*, 104241. [CrossRef]
52. Stergiou, A.; Poppe, R.; Kalliatakis, G. Refining activation downsampling with SoftPool. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10357–10366.
53. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
54. Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; Yang, G. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 6070–6079.
55. Liang, L.; Deng, S.; Gueguen, L.; Wei, M.; Wu, X.; Qin, J. Convolutional neural network with median layers for denoising salt-and-pepper contaminations. *Neurocomputing* **2021**, *442*, 26–35. [CrossRef]
56. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
57. Yang, T.; Jiangde, S.; Hong, Z.; Zhang, Y.; Han, Y.; Zhou, R.; Wang, J.; Yang, S.; Tong, X.; Kuc, T.-Y. Sea-land segmentation using deep learning techniques for landsat-8 OLI imagery. *Mar. Geod.* **2020**, *43*, 105–133. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.