remote sensing

MDPI

Article

Balancing Accuracy and Efficiency: HWBENet for Water Body Extraction in Complex Rural Landscapes

Pengyu Lei [1], Jiang Zhang [2,*] and Jizheng Yi [1]

[1] Institute of Artificial Intelligence Applications, Central South University of Forestry and Technology, Changsha 410000, China; 255001026@csu.edu.cn (P.L.); t20152279@csuft.edu.cn (J.Y.)
[2] School of Computer and Mathematics, Central South University of Forestry and Technology, Changsha 410000, China
* Correspondence: t19990896@csuft.edu.cn

Highlights

What are the main findings?

- A novel hybrid network called HWBENet is introduced, and it is specifically designed for water extraction in complex rural landscapes, achieving an excellent balance between accuracy and efficiency.
- The CIMM and ERM modules are proposed to improve the network's segmentation accuracy of fragmented water bodies and water body boundaries.

What are the implications of the main findings?

- The introduction of HWBENet resolves the long-standing contradiction between accuracy and efficiency in the field of water body extraction from remote sensing images.
- As an efficient and accurate solution, HWBENet offers significant practical value for large-scale water ripple mapping in complex rural landscapes, meaning it can provide more reliable and timely water information for environmental monitoring.

Abstract

The accurate and timely extraction of water bodies from high-resolution remote sensing imagery is vital for environmental monitoring, yet segmenting small, scattered, and irregularly shaped water bodies in complex rural landscapes remains a persistent challenge. While state-of-the-art deep learning models have advanced segmentation accuracy, they often achieve this at the cost of substantial computational overhead, limiting their practical application for large-scale monitoring. To address this trade-off between precision and efficiency, this paper introduces HWBENet, a novel hybrid network for water body extraction. HWBENet is built upon a lightweight MobileNetV3 encoder to ensure computational efficiency while preserving strong feature extraction capabilities. Its core innovation lies in two specifically designed modules. First, the Contextual Information Mining Module (CIMM) is proposed to enhance the network's ability to learn and fuse both global scene-level context and fine-grained local details, which is crucial for identifying fragmented water bodies. Second, an Edge Refinement Module (ERM) is integrated into the decoder, which uniquely leverages transformer mechanisms to sharpen boundary details by effectively fusing prior feature information with up-sampled features. Extensive experiments on challenging rural water body datasets demonstrate that HWBENet strikes a superior balance between accuracy and computational cost. The experimental results validate the finding that HWBENet is an efficient, accurate, and scalable solution, offering significant practical value for large-scale hydrological mapping in complex rural environments.

check for updates

Academic Editor: John Trinder

Received: 18 September 2025
Revised: 2 November 2025
Accepted: 7 November 2025
Published: 14 November 2025

Citation: Lei, P.; Zhang, J.; Yi, J. Balancing Accuracy and Efficiency: HWBENet for Water Body Extraction in Complex Rural Landscapes. *Remote Sens.* **2025**, *17*, 3711. https://doi.org/10.3390/rs17223711

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/).

Keywords: water extraction; Mamba; transformer; semantic segmentation; remote sensing imagery

1. Introduction

Water bodies play a critical role in ecological and hydrological balance, serving as essential resources for agriculture, drinking water, and habitat maintenance [1]. The timely and accurate segmentation of surface water in high-resolution imagery supports various tasks, such as hydrological monitoring. Segmenting small and scattered water bodies in rural areas remains challenging due to heterogeneous surroundings and complex land cover types [2]. The spectral similarity between water and other land cover types, including vegetation, impervious surfaces, and bare soil, poses a major challenge and highlights the necessity of robust automated extraction methods.

Conventional approaches such as the index-based methods, the Normalized Difference Water Index (NDWI) [3], and its variations, like the Modified NDWI (MNDWI) [4], rely heavily on spectral differences between water and non-water features. While effective in many cases, these approaches often struggle in scenarios with mixed pixels, shadows, or similar spectral features, especially in rural landscapes. Recent developments in deep learning, particularly convolutional neural networks (CNNs), have shown great promise in improving segmentation accuracy by learning complex hierarchical features directly from raw image input. Widely adopted models like FCN [5], UNet [6], and DeepLabV3+ [7] have shown consistent segmentation accuracy across diverse applications. Transformer-based segmentation architectures are increasingly adopted due to their ability to capture long-distance dependencies and global scene semantics in visual data. Vision Transformer (ViT) [8] and its derivatives, such as SegFormer [9] and Swin Transformer [10], have shown remarkable improvements in semantic segmentation. Through self-attention mechanisms, these architectures learn long-range feature dependencies, enabling them to model global interactions across an entire scene. However, this global modeling capability comes at the cost of quadratic computational complexity, posing challenges for high-resolution remote sensing imagery. Seeking to overcome this limitation, recently proposed State Space Models (SSMs) like Mamba [11] offer a promising alternative by capturing long-range dependencies with linear complexity.

Despite this rapid architectural evolution, a critical challenge remains for segmenting water bodies in complex rural regions. First, existing general-purpose backbones, from CNNs to transformers and the latest SSMs, are not inherently optimized for the specific characteristics of rural water bodies, which are often small, scattered, and irregularly shaped. Accurately segmenting them demands a model that can synergistically integrate scene-level global contexts with fine-grained boundary details. Second, achieving this precision often results in computationally intensive models, which may not be suitable for large-scale or real-time applications. While emerging models like Mamba aim to improve efficiency, a pressing need still exists for a lightweight and targeted solution that explicitly balances precision with computational costs for this specific downstream task. This creates a clear demand for a solution designed to be both accurate and efficient in scenarios involving large-area, high-resolution remote sensing images.

To address these two challenges, we propose HWBENet, a hybrid water body extraction network specifically designed for complex rural environments. HWBENet adopts a lightweight MobileNetV3 [12] as the encoder, offering a balance between efficiency and feature extraction capability while incorporating novel context mining and boundary refinement modules to capture both global information and fine-grained details. Unlike prior

models that often trade efficiency for precision, HWBENet addresses rural water segmentation by balancing accuracy and efficiency, delivering scalable results in high-resolution imagery. The key contributions of this work are as follows:

(1) To improve the model's ability to capture both global context and fine-grained spatial details, a Contextual Information Mining Module (CIMM) is proposed. This module enhances the network's ability to learn and fuse both global and local information, which is crucial for accurately segmenting fragmented and diverse rural water bodies.

(2) An Edge Refinement Module (ERM) is designed, and it leverages transformer mechanisms to enhance edge detection. By fusing prior feature information with upsampled features, the network focuses on water boundaries, providing more accurate segmentation of water bodies, especially in complex backgrounds and small regions.

(3) Through extensive experiments and rigorous statistical analysis, we demonstrate that HWBENet achieves a favorable balance between segmentation accuracy and computational efficiency.

The structure of the paper is outlined as follows. Section 2 reviews related work on water body extraction methods. Section 3 describes the architecture of the proposed network in detail. Section 4 presents the datasets, implementation details, and evaluation metrics used in our study. Section 5 reports the experimental results, along with a detailed analysis, and Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1. Conventional Methods for Water Body Extraction

Water body extraction from remote sensing imagery has been an active area of research, driven by the need for accurate and efficient segmentation techniques in environmental monitoring. Traditional methods for water body extraction often involve thresholding techniques or shallow classifiers applied to remote sensing imagery. Thresholding techniques primarily utilize the distinct spectral characteristics of water bodies across various spectral bands to develop models and water indices for extraction. One of the most widely used indices is the Normalized Difference Water Index (NDWI) [3], which enhances water features by exploiting the reflectance differences in the green and near-infrared (NIR) bands. The calculation method of NDWI is described in Equation (1).

$$\text{NDWI} = \frac{GREEN - NIR}{GREEN + NIR} \tag{1}$$

where $GREEN$ represents the reflectance value in the green spectral band, and NIR denotes the reflectance value in the near-infrared spectral band. Inspired by the NDWI, some researchers have proposed other types of index models such as [13,14]. Nevertheless, threshold-based segmentation methods are challenged by feature selection and parameter settings, as they are sensitive to mixed water and background pixels, making it difficult to determine an appropriate threshold for segmentation. particularly in complex rural landscapes.

Shallow classifiers primarily integrate spectral, shape, structural, and textural features from remote sensing images for water body extraction [15,16]. For example, Wu et al. [17] proposed a water body extraction method based on support vector machines, which employs a mathematical model designed to mitigate the impact of background noise by analyzing its characteristics. Approaches based on shallow classifiers mainly stem from traditional machine learning techniques, showing promise in specific scenarios. Yet, they often demand intricate manual classifier designs, making them challenging to generalize across diverse scenarios. Traditional methods predominantly leverage shallow features such as color, shape, spectrum, and texture for water body extraction, thus limiting accuracy

and adaptability to complex backgrounds and varied scales of water body distribution due to the inability to harness deep features.

2.2. Deep Learning Methods for Water Body Extraction

In recent years, deep learning technology has significantly advanced, bringing new opportunities for addressing the challenges of water body extraction in rural areas. The application of deep learning to semantic segmentation tasks, particularly convolutional neural networks (CNNs), has led to substantial progress in extracting water bodies from remote sensing images. These advancements have shown promising results, although challenges such as over-segmentation and the need for multi-scale contextual understanding remain. Below, we review recent developments and notable methods used in deep learning-based water body extraction.

The introduction of fully convolutional networks (FCNs) by Long et al. [5] marked a significant breakthrough in pixel-level semantic segmentation, including water body extraction. FCNs replaced fully connected layers with convolutional layers to enable end-to-end processing, producing pixel-wise predictions essential for extracting water bodies from remote sensing imagery. Building on this foundation, various CNN-based models, such as Unet [6], have been widely adopted, leveraging hierarchical feature learning to achieve accurate segmentation [18,19]. However, a key challenge that quickly emerged was the need for a multi-scale contextual understanding. This is particularly critical for remote sensing tasks like water body segmentation, where the model must simultaneously perceive the scene at multiple scales: it needs a large receptive field to understand the global context and maintain the continuity of large rivers or lakes while also needing to process fine-grained, high-resolution details to accurately delineate the boundaries of small, fragmented water bodies. To address this, seminal works introduced powerful multi-scale feature extraction modules. For instance, PSPNet [20] proposed the Pyramid Pooling Module to aggregate context from different regions, while the DeepLab series [7,21] introduced Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale information using parallel atrous convolutions. Inspired by these ideas, many subsequent water extraction networks have incorporated similar multi-scale designs. For instance, Kang et al. [22] introduced MSCEnet, which combines Res2Net [23] with dilated convolutions to extract multi-scale global features. Luo et al. [24] employed the ASPP module to enhance multi-scale information capture, while Nie et al. [25] proposed SE-BiSeNet to extract spatial and contextual information in urban water extraction. These methods, while effective in enhancing global feature representation, often struggle with precise boundary localization.

Attention mechanisms have been widely applied to refine boundary segmentation. For example, Miao et al. [26] designed a boundary-weighted loss function to emphasize edge features, and models like FWE-net [27] and MEC-net [28] applied channel attention to focus on water body boundaries. While these methods improve segmentation accuracy, challenges persist in adapting to the complex and fragmented nature of rural water bodies, characterized by varying scales and diverse vegetation. The advent of transformer [29] architectures has provided new opportunities to overcome these challenges. Transformers excel in modeling long-range dependencies, making them highly effective for capturing global contextual information. Dosovitskiy et al. [8] introduced the Vision Transformer (ViT), demonstrating its ability to achieve remarkable results in image classification by processing images as sequences of patches. Building on this, Chen et al. [30] proposed Transunet, which integrates a transformer for feature encoding and a CNN for decoding, achieving notable success in medical image segmentation. In remote sensing, transformer-based approaches have shown promise [31,32]. For instance, Xu et al. [33] developed an adaptive fusion module for a multi-scale feature integration, while Zhong et al. [34]

proposed a multi-level transformer module to capture lake–boundary relationships effectively. Qi et al. [35] ingeniously integrated the geodesic active contour model with the ViT architecture to achieve a more precise extraction of lake water bodies.

2.3. Mamba for Remote Sensing

In image processing tasks, effectively capturing both global and local information while reducing computational complexity has long been a critical research challenge. With the advancement of deep learning technologies, Mamba [11], originally designed based on the state space model (SSM) for natural language processing tasks, has emerged as an effective approach to address this issue. Its introduction to image processing with Vim [36] and VMamba [37] demonstrated remarkable performance, sparking widespread interest in related fields. The Mamba mechanism has recently gained recognition as a promising solution for enhancing efficiency and accuracy in semantic segmentation tasks, particularly for handling high-dimensional and complex input data such as remote sensing imagery. This section reviews the application and evolution of the Mamba mechanism in semantic segmentation, with a special focus on its utility in remote sensing.

Rs-mamba [38] enhances the global connectivity of remote sensing images by introducing a scanning mechanism that captures spatial features with large spatial scales and multiple directions, particularly in ultra-high-resolution imagery. By incorporating a selective state space mechanism, RS-Mamba effectively addresses the intricate details of high-resolution satellite images. CM-Unet [39] integrates the Mamba mechanism into a UNet architecture, utilizing selective state modeling to capture both contextual and local information and resulting in superior performance in semantic segmentation tasks. RS3Mamba [40] establishes a visual spatial state model that maintains linear computational complexity while modeling long-range dependencies, enabling richer global information acquisition in remote sensing image segmentation. ChangeMamba [41] extends the Mamba mechanism by incorporating spatio-temporal state spaces, allowing the model to detect subtle environmental changes over time and making it highly suitable for scenarios involving gradual environmental alterations. PyramidMamba [42] introduces a Mamba-based decoder for the semantic segmentation of remote sensing images, addressing semantic homogeneity issues in multi-scale feature fusion by employing a dense spatial pyramid module and a pyramid fusion Mamba module.

Collectively, these adaptations demonstrate the versatility and scalability of the Mamba mechanism across diverse domains and tasks, underscoring its effectiveness in improving segmentation accuracy and efficiency, particularly in the context of complex and high-resolution remote sensing data.

3. Methods

The architecture of our proposed water body extraction network is designed to balance precision and efficiency in detecting water bodies from high-resolution remote sensing images, particularly in rural regions. This network follows a typical encoder–decoder architecture, in which the encoder, based on MobileNetV3, extracts multi-scale features from the input images, and the decoder refines these features iteratively to output a pixel-wise segmentation map. However, the network's core innovations include two specialized modules: the CIMM module and the ERM module, which together enhance the model's ability to segment water bodies with high precision while maintaining efficiency.

As shown in Figure 1, the structure of the proposed network is detailed. MobileNetV3 is utilized as the encoder to extract semantic features from the input images. Let the input image be presented as $X \in \mathbb{R}^{C \times H \times W}$, where C denotes the number of channels, H the height, and W the width of the image. After processing through dif-

ferent stages of the encoder, we derive multi-level feature representations as follows: $X^{64\times\frac{H}{4}\times\frac{W}{4}}$, $X^{128\times\frac{H}{8}\times\frac{W}{8}}$, $X^{256\times\frac{H}{16}\times\frac{W}{16}}$, $X^{512\times\frac{H}{32}\times\frac{W}{32}}$.

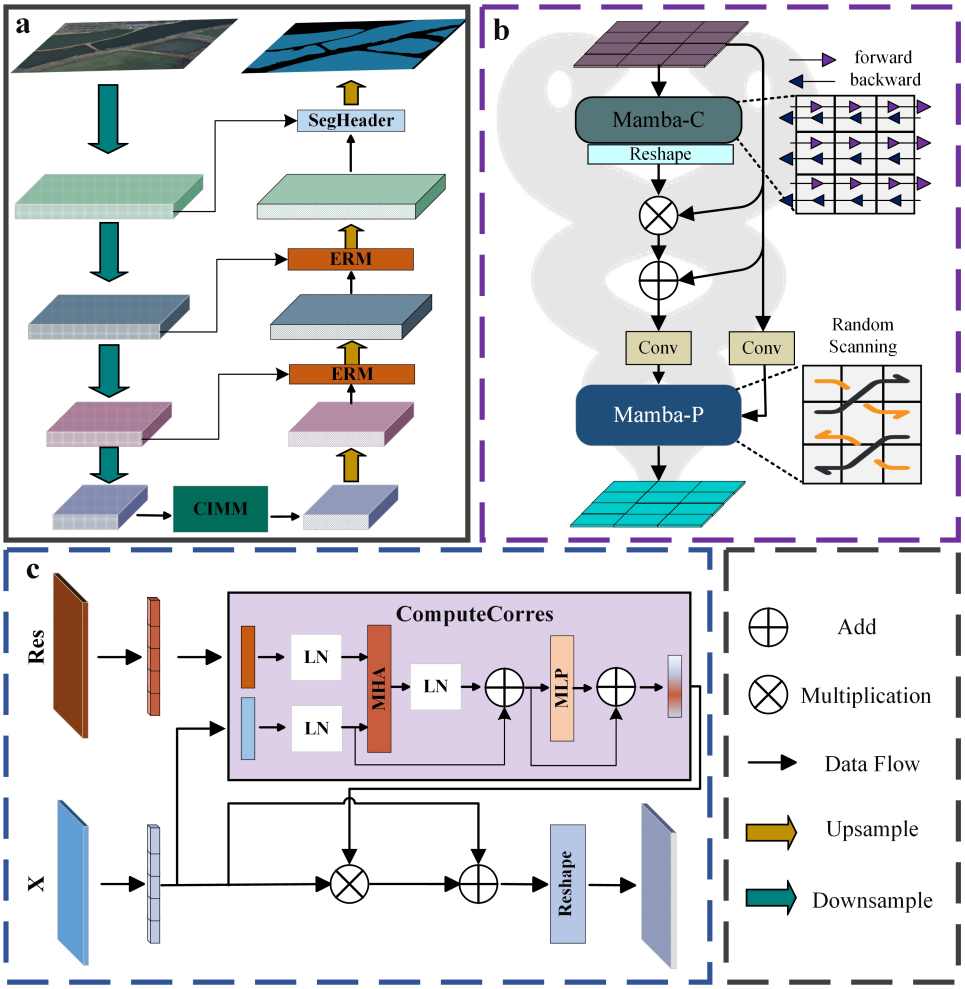


Figure 1. Overview of our proposed architecture. (**a**) Network structure. (**b**) CIMM. (**c**) ERM.

To strengthen the network's ability to capture water body boundary features, the ERM is introduced. As formulated in Equations (2) and (3), ERM integrates preceding features with the current feature map, adaptively adjusting the weight assigned to edge features. This adaptive weighting mechanism enables the network to focus more effectively on water body boundaries, thereby improving segmentation accuracy along the edges.

$$Fusion_X = ERM(X_{cur}, Res_{pre}) \tag{2}$$

$$X_{next} = Upsample(Fusion_X) \tag{3}$$

where $Fusion_X$ represents the feature map generated via ERM, X_{cur} represents the current feature map, Res_{pre} refers to the feature map at the corresponding scale during downsampling in the encoder, and X_{next} indicates the feature map after upsampling.

Before the network outputs the predicted image, a SegHeader module is utilized to enhance the model's accuracy. SegHeader employs a dynamic parameter strategy to assign weights to pixels in the low-level feature maps, thereby facilitating the effective aggregation of shallow semantic information. The algorithmic process of SegHeader is represented in Algorithm 1. Finally, the aggregated low-level features are upsampled to generate the network's predictions.

Algorithm 1: SegHeader process

Input: Ft, Ft_{skip} /* Original and skip feature. */
Output: $AggFeat$ /* Feature map after process. */
Init parameter tensors: A, B
for *epoch in epochs* **do**
  $Ft_{cat} \leftarrow \text{Concat}(Ft, Ft_{skip})$ /* Concatenation for shallow aggregation. */
  $Ft_{inter} \leftarrow A \odot \text{Conv}(Ft_{cat}) + B \odot Ft_{skip}$ /* Deeper aggregation by assigning weights. */
  $AggFeat \leftarrow \text{ResBlock}(Ft_{inter})$ /* Weights refinement. */
  Update A, B
end
return $AggFeat$

3.1. Mamba-Based Contextual Information Mining Module

To comprehensively capture contextual information within deep features, we designed the CIMM module. Unlike models such as VMamba, which utilize Mamba as a general-purpose backbone, the core innovation of CIMM lies in its factorized information enhancement pipeline. This pipeline decomposes the complex task of contextual modeling into two sequential, well-defined stages: first, inter-channel semantic relationship enhancement, followed by inter-pixel spatial structure modeling. This design allows us to address the semantic ambiguity and spatial complexity in water body segmentation in a highly efficient and targeted manner.

The first stage aims to resolve the challenge of semantic ambiguity. In the deeper layers of the network, each channel of the input feature map represents different high-level semantic features. To distinguish between visually similar ground objects, such as water bodies and shadows, the model must understand the complex dependencies between these abstract feature channels. To this end, we first employ the Mamba-C module to enhance inter-channel interactions. The second stage is designed to tackle the challenge of spatial complexity, namely the precise segmentation of irregularly shaped and fragmented water bodies. In this stage, we use the Mamba-P module to model the spatial structural relationships between pixels. Figure 2 clearly illustrates the processing flow of Mamba-C and Mamba-P, and the procedure of CIMM is illustrated in Figure 1b.

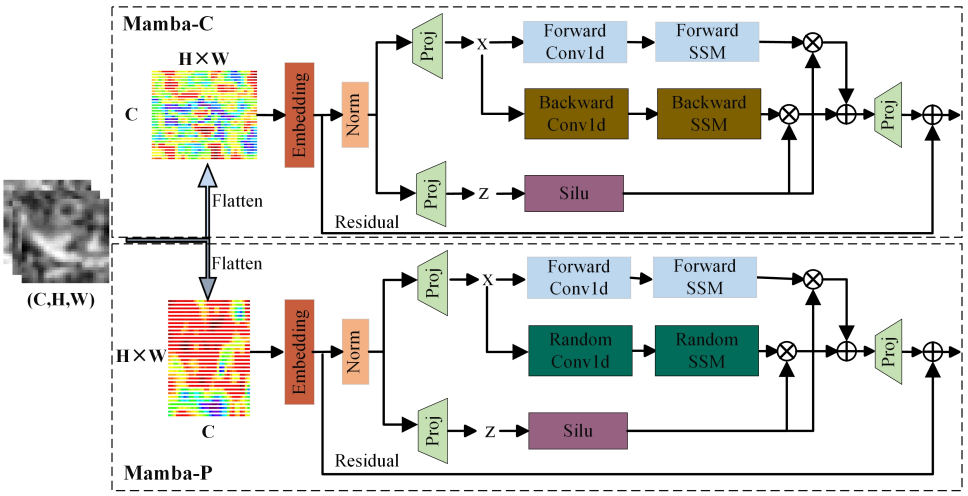


Figure 2. Details about Mamba-C and Mamba-P. The residual of Mamba-P is an external input.

$$X_{seq} = Flatten(X_{final}) \tag{4}$$

$$X^{c}_{mamba} = \sigma(Mamba_{C}(X_{seq})) \odot X + X \tag{5}$$

$$X^{context}_{mamba} = Mamba_{P}\{Conv(X^{c}_{mamba}), Conv(X_{seq})\} \tag{6}$$

In detail, CIMM begins by taking the final feature map from the encoder, denoted as X_{final}, and applying a Flatten operation to convert it into a sequence format, X_{seq}, as shown in Equation (4). After it is transformed into sequence data, a cascaded contextual information extraction mechanism is applied. Initially, the channel dimension is treated as the primary dimension, converting the sequence data into the form (C, H×W). The resulting sequence is subsequently passed into Mamba-C to facilitate the fusion of inter-channel information. As illustrated in Equation (5), to enhance feature integration, element-wise multiplication and addition operations are employed. Element-wise multiplication applies weights to corresponding feature information, thereby amplifying significant features. The addition operation enhances the expression of the fused information while mitigating the risk of gradient explosion.

Given the excessive number of channels in X^{c}_{mamba} and X_{final}, directly performing pixel-level information fusion would significantly increase computational complexity. Therefore, a *Conv* operation is applied to downsample these two feature maps along the channel dimension before pixel fusion. In the Mamba-P process, the channel-fused feature sequence is treated as the primary feature, while the original feature sequence serves as a residual sequence, supplementing the primary sequence. Thus, as shown in Equation (6), the feature $X^{context}_{mamba}$, containing both inter-channel and inter-pixel information, is obtained. Algorithm 2 illustrates the execution process of CIMM.

Algorithm 2: Contextual information mining module

Input: $X \in \mathbb{R}^{C\times H\times W}$

Output: $X_{context} \in \mathbb{R}^{C\times H\times W}$

$T_x \leftarrow \text{Flatten}(X)$ /* Convert to sequence mode */

$M_c \leftarrow Mamba_c(T_x)$ /* Applying Mamba in channel dimension */

$X^{c}_{wt} \leftarrow \sigma(Rsp(M_c)) \odot X + X$ /* Information fusion across channel */

$X_{sq}, X^{c}_{sq} \leftarrow \text{Conv}(X), \text{Conv}(X^{c}_{wt})$ /* Squeeze the number of channel */

$X_{context} \leftarrow Mamba_P(X^{c}_{sq}, X_{sq})$ /* Conduct Mamba-P operation */

return $X_{context}$

It is important to note that the scanning mechanism in Mamba-C differs from that in Mamba-P. In the channel fusion Mamba block, the forward-backward scanning mechanism, originally employed in the VisionMamba network, is adopted to capture comprehensive inter-channel information. However, in the pixel fusion Mamba block, the backward scanning mechanism is replaced with the random scanning mechanism to enhance the generalization capability of the network and accommodate the distribution characteristics of pixels. We have also empirically validated the superiority of this hybrid scanning design in our ablation study in Section 5.2.2.

3.2. Edge Refinement Module

Boundary information plays a crucial role in image processing, particularly in tasks related to water body analysis. It not only includes the noticeable visual differences between water bodies and surrounding land, such as color and texture transitions, but also involves several intricate details, such as the precise shape of the boundary and the spatial relationship with the surrounding environment. To address this, we designed the Edge Refinement Module (ERM) to perform a sophisticated fusion and refinement of features from the encoder and decoder paths. The design of ERM is motivated by the fundamental limitations of existing attention-based fusion mechanisms. Existing approaches can be broadly categorized as performing feature selection or re-weighting. For instance, mechanisms like channel and spatial attention learn to emphasize salient channels or spatial regions based on feature statistics. However, these methods often overlook the complex relationships between features, which are particularly important for accurately delineating water body boundaries.

To address this challenge, ERM is designed to refine edge features by fusing downsampled features, which carry semantic information, with upsampled features, which contain spatial details. The upsampled features generated via the decoder layer are enriched with spatial details, playing a crucial role in accurately delineating water boundaries. As the upsampling process progresses, these features are progressively restored to the resolution of the input image, clearly presenting the approximate contours and local details of the water boundaries. In contrast, the downsampled features extracted via the encoder layer are focused on capturing the semantic information of the image, allowing the relationships between water bodies and their surrounding environments to be understood. Subsequently, connections between features at different positions are established using the self-attention mechanism of the transformer, enabling the capture of long-range dependencies among features.

The details of ERM are shown in Figure 1c. Initially, the upsampled feature map (denoted as X_{up}) and the corresponding downsampled feature map (denoted as X_{dn}) are converted into vector forms through the Flatten operation. The Flatten operation is defined in Equations (7) and (8). Subsequently, the resulting Vec_{up} and Vec_{dn} are input into the ComputeCorres (as shown in Algorithm 3) processor to compute the correlation coefficients between features.

Within ComputeCorres, LayerNorm is applied to normalize the input distributions for standardization. The features are then processed via the multi-head attention mechanism (MHA), as defined in Equation (9).

$$Vec_{up} = \mathcal{F}(X_{up}) \tag{7}$$

$$Vec_{dn} = \mathcal{F}(X_{dn}) \tag{8}$$

$$X_{MHA} = \mathrm{MHA}(Vec_{up}, Vec_{dn}) \tag{9}$$

where $\mathcal{F}$ represents the Flatten operation, and MHA denotes the multi-head attention mechanism. In MHA, the upsampled feature Vec_{up}, which contains abundant spatial details, is utilized as the Query, as these details are critical for capturing subtle variations in water body boundaries. Meanwhile, Vec_{dn}, which contains high-level semantic information, serves as the Key and Value to support the determination of boundary relationships with the surrounding environment. To prevent gradient vanishing and accelerate model convergence, residual connections and normalization operations are applied after the MHA. Finally, the correlation coefficient *Corres* is computed as shown in Equation (10).

$$Corres = (Vec_{up} + X_{MHA}) + MLP(Vec_{up} + \mathrm{X_{MHA}}) \tag{10}$$

After the computation of correlation coefficients between features is completed, the primary feature Vec_{up} is multiplied element-wise by *Corres* to obtain a feature map weighted by the correlation coefficients. Subsequently, an element-wise addition operation is applied to enhance the spatial representation of edge information, as shown in Equation (11).

$$X_{Edg} = Corres \odot Vec_{up} + Vec_{up} \tag{11}$$

where $\odot$ indicates element-wise multiplication. Finally, a reshape operation is performed to restore the refined feature map to its original shape.

Algorithm 3: ComputeCorres processor

Input: Vec_{up}, Vec_{dn}
Output: X_{corres}
$Vec_{up} = LayerNorm(Vec_{up})$
$Vec_{dn} = LayerNorm(Vec_{dn})$
/* Normalization using LayerNorm */
let Vec_{up} as Q; let Vec_{dn} as K and V
/* Setting Query, Key, Value */
$X_{MHA} = MHA(Q, K, V)$
/* Applying Mamba in channel dimension */
$X_{MHA} = LayerNorm(X_{MHA})$
/* Normalization using LayerNorm */
$X_{skip} = Vec_{up} + X_{MHA}$
/* Residual connection */
$X_{mlp} = MLP(X_{skip})$
/* Processing through MLP */
$X_{corres} = X_{skip} + X_{mlp}$
return X_{corres}

4. Materials and Experiments

In this section, we present the materials and experimental setup of our proposed water body segmentation network. We first introduce the datasets used in our experiments, followed by the details of the experimental setup, including training configuration, network implementation, and hardware specifications. Lastly, we describe the metrics used for performance evaluation.

4.1. Dataset Introduction

For the evaluation of our water body segmentation network, we use publicly available remote sensing datasets that feature high-resolution remote sensing imagery with detailed annotations for water body segmentation. The datasets particularly focus on rural regions, where water bodies are often irregular, and their boundaries are difficult to detect due to complex land–water interfaces. Both datasets are split into training, validation, and testing sets. The training set consists of 70% of the total data, the validation set comprises 15%, and the testing set contains the remaining 15%. Detailed information about the two datasets is illustrated in Table 1.

Table 1. Detailed information about LoveDA and GLH-water datasets.

Dataset	DataSource	GSD	Resolution	Amount	Train	Eval	Test
LoveDA	Google Earth	0.3 m	1024 × 1024	5987	1262	158	158
GLH-water	Google Earth	0.3 m	12,800 × 12,800	250	1584	198	198

4.1.1. LoveDA Dataset

The LoveDA dataset [43] is a high-resolution land-cover dataset, originally designed for land-cover classification tasks. It contains images acquired from a variety of rural and urban areas, and it includes precise annotations for water bodies in the context of segmentation. Although primarily used for land-cover classification, the dataset's rich spectral and spatial features make it highly suitable for water-body segmentation tasks. We utilize the water body annotations from this dataset for the training and evaluation of our network, which provides a challenging environment for the detection of water bodies in complex rural regions.

4.1.2. GLH-Water Dataset

The GLH-Water dataset [44] is another important dataset in remote sensing, and it is specifically designed for water body segmentation in high-resolution satellite imagery. The dataset includes both rural and urban water bodies, with corresponding ground-truth annotations marking the water areas. This dataset is particularly useful for testing the robustness of segmentation networks in rural and semi-urban environments, where water bodies may involve intricate shapes and challenging boundaries. We employ this dataset to further evaluate the performance of our proposed model in detecting water bodies with high spatial and spectral variance.

4.2. Experimental Setup

The network is trained end to end with the objective of minimizing pixel-wise segmentation loss, using a combination of cross-entropy loss and Dice loss. The training process is conducted on all three datasets mentioned above, ensuring generalizability across different environmental conditions.

The Python 3.10.13 environment and the PyTorch 2.1.1 framework are used to implement our network. The training is conducted on an Intel i9-11900K 3.50-GHz CPU (64 GB of RAM) with an NVIDIA GeForce RTX 3090 workstation featuring 24 GB of memory. During training, the SGD optimizer with a weight decay of 0.0005 and a momentum of 0.9 was employed to expedite model convergence. The BCE_Dice loss was adopted as the loss function to assess model performance and enhance its stability. The formulation of the BCE_Dice loss is defined below:

$$BCE_Diceloss = \lambda_1 \cdot BCEloss + \lambda_2 \cdot Diceloss \tag{12}$$

$$BCEloss = -(\mathrm{y}\log(\hat{\mathrm{y}}) + (1-\mathrm{y})\log(1-\hat{\mathrm{y}})) \tag{13}$$

$$Diceloss = 1 - \frac{2y\hat{y}+1}{y+\hat{y}+1} \tag{14}$$

where λ_1 and λ_2 are the weights for each loss component. Following a common and empirically robust practice in the remote sensing segmentation literature [2,18], we set both λ_1 and λ_2 to 1. Here, image labels are represented as y, while the model's predictions are represented as $\hat{y}$.

To dynamically control the speed and extent of learning rate decay during model training, the 'Poly' learning rate updating strategy was employed with an exponent of 0.9 and an initial learning rate of 0.0001. Our network underwent training for 300 epochs on the LoveDA and GLH-water datasets. Additionally, the batch size was set to 4 due to GPU limitations.

Evaluation metrics of segmentation are used to assess the performance of our proposed model in water body segmentation tasks. The metrics include the intersection over union (IoU), the F1-score (F1), precision, recall, and pixel accuracy (PA). These metrics provide

comprehensive insights into the model's ability to accurately segment and classify water bodies. The calculation formulas for these metrics are as follows:

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

$$\text{F1} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{16}$$

$$\text{PA} = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

where TP (true positive) is the number of correctly predicted pixels belonging to the water class. TN (true negative) represents the number of correctly predicted as non-water pixels. FP (false positive) is the number of pixels incorrectly predicted as water when they are not. FN (false negative) is the number of pixels incorrectly predicted as non-water when they actually belong to the water class.

These metrics collectively offer a robust evaluation framework, covering different aspects of model performance. The IoU and the F1-score provide a comprehensive evaluation by accounting for both over-segmentation (false positives) and under-segmentation (false negatives). Pixel accuracy provides an overall accuracy assessment, making it a reliable measure for general performance comparisons.

By utilizing this set of metrics, we can effectively compare the performance of our proposed model against baseline methods, offering insights into its strengths and areas for improvement in accurately segmenting water bodies from remote sensing images.

Furthermore, to ensure a rigorous comparison, we assess the statistical significance of the performance differences between our proposed HWBENet and other methods. We employ the Wilcoxon signed-rank test for this purpose. As a non-parametric and paired test, it is ideally suited for comparing the distributions of per-image scores from two models on the same test set without assuming normality. For all tests, we adopt a significance level of $p < 0.05$ to determine whether the observed improvements are statistically significant and not due to random chance.

5. Results and Discussion

5.1. Methods Comparison

To verify the effectiveness of the model that we proposed, we selected a series of advanced network models for extensive comparisons. These networks include the current mainstream semantic segmentation network models and water extraction methods. The mainstream semantic segmentation models include the following: (1) CNN-based semantic segmentation networks—UNet [6], U2Net [45], PSPNet [20], HRNet [46], BANet [47], DeeplabV3+ [7], ABCNet [48]; (2) transformer-based semantic segmentation networks—SegFormer [9], SwinUperNet [10], Unetformer [49], and Segmenter [50]; and (3) Mamba-based semantic segmentation networks—PyrimidMamba [42], RS3Mamba [40], Swin-UMamba [51], and Mamba-Unet [52]. The network architectures designed for water extraction are as follows: WatNet [24], SR-Segnet [53], MEC-net [28], and DeepWaterMapV2 [54].

Firstly, to compare the training phase of different models concisely and without monotony, we selected the best-performing models based on CNN, transformer, and Mamba architectures with the LoveDA dataset for comparison. Figure 3 presents the training loss curves for some comparison methods and the combined training and evaluation loss curves for the proposed network in the LoveDA dataset, respectively. The analysis of these figures provides a comprehensive understanding of the model's learning behavior and optimization performance. As Figure 3 shows, the training loss of our

model exhibits a steady and smooth decline, indicating effective optimization throughout the training process. In comparison with the training loss curves of other network models, our network achieved the lowest loss value. Although PyrimidMamba also reached a similarly low loss value, our network converged faster. As an observation of Figure 3 reveals, the gap between training loss and evaluation loss remains minimal throughout the training process. This indicates that our model is not overfitting and maintains strong generalization capabilities.

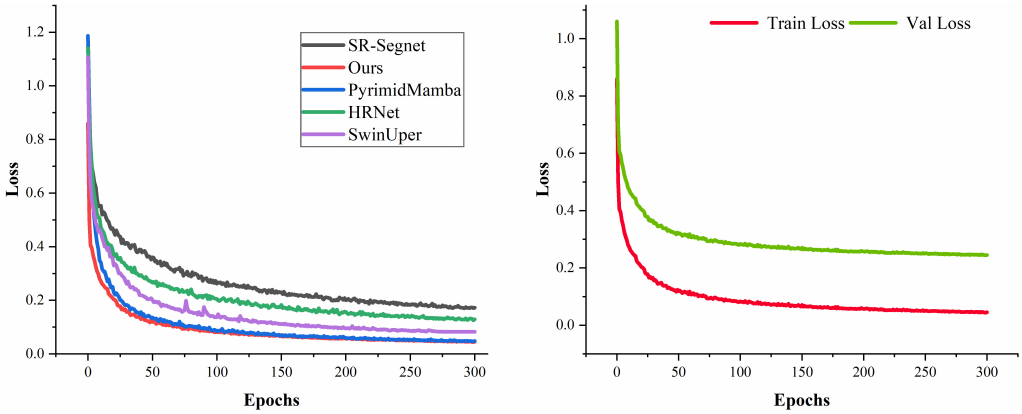


Figure 3. Loss curves of representative networks (**left**); the train and evaluation loss curves of our network (**right**).

Secondly, in order to more clearly compare the effects of different models, we respectively compare from two aspects of quantification and visualization. From Table 2, we show the results of all comparison networks in various evaluation indicators. As can be seen from the table, whether it is CNN-based, transformer-based, or Mamba-based network architecture, the network model that we proposed has achieved the best results on the LoveDA dataset. On the GLH-water dataset, although our network did not achieve the best performance, we achieved an excellent balance in computational complexity and important evaluation metrics. In order to show the contents of the table more intuitively, as compared with the above training process, we selected the best model of different architectures for visualization. As can be seen from Figure 4, the networks we proposed on the LoveDA dataset have obviously achieved the best indicators.

Moreover, to provide a qualitative assessment of the segmentation performance, Figure 5 presents a visual comparison of our HWBENet against other leading methods on several challenging samples. While our model performs competitively across all examples, its advantages are most pronounced in scenes with extremely fine-grained features or complex boundaries. To this end, we provide zoomed-in insets for two particularly representative examples. It can be clearly observed from Col 1, Col 3, Col 7, and Col 10 in Figure 5 that, when faced with a complex rural geographical environment, our network can achieve better results in mining global information. Therefore, our network can well preserve the continuity and integrity of the water in space during the water extraction process. As can be seen from Col 2, Col 4, Col 8, and Col 9 in Figure 5, our network can segment the water boundary more accurately when the water boundary is close to or mixed with other ground objects. From the visualized results of all comparison methods, it is evident that our network maintains continuity and integrity, particularly in large or elongated water bodies. This is crucial for the accurate mapping and monitoring of extensive water areas, for which maintaining the shape and structure of the water body is essential for a reliable analysis.

Table 2. The comparison results of model performance between different network architectures on the LoveDA and GLH-water datasets.

Methods		IoU (%)		F1 (%)		PA (%)	
		D1	D2	D1	D2	D1	D2
CNN-based	Unet [6]	62.90	76.73	72.63	84.41	92.77	94.74
	U2Net [45]	58.03	76.75	66.29	84.49	92.24	94.65
	ABCNet [48]	64.05	55.31	73.78	65.98	94.01	88.65
	BANet [47]	62.46	75.58	71.98	83.63	92.67	94.50
	DeeplabV3+ [7]	50.75	67.43	60.72	76.91	88.60	91.78
	PSPNet [20]	67.12	78.81	77.00	86.51	95.15	95.18
	HRNet [46]	67.46	78.42	76.89	85.60	93.93	94.51
Transformer-based	SegFormer [9]	54.68	52.45	65.41	55.31	89.65	88.12
	SwinUperNet [10]	63.49	45.22	73.66	55.02	93.58	83.75
	Unetformer [49]	62.86	77.15	73.43	85.10	92.64	94.75
	Segmenter [50]	57.71	53.34	67.85	66.32	91.26	89.74
Mamba-based	PyrimidMamba [42]	70.86 *	**80.53**	79.48 *	**87.49**	95.46	95.43
	Swin-UMamba [51]	55.39	72.56	66.52	80.14	89.96	94.10
	Mamba-unet [52]	42.50	73.65	53.43	81.97	82.18	94.05
	RS3Mamba [40]	65.03	58.30	75.32	69.57	94.33	83.39
Water Extraction methods	WatNet [24]	60.75	72.83	71.42	81.87	91.08	93.93
	SR-Segnet [53]	66.32	71.65	76.26	80.47	94.75	93.58
	MEC-net [28]	58.68	78.99	67.81	86.44	91.99	95.29
	DeepWaterMapV2 [54]	52.85	69.12	63.42	78.27	89.27	92.84
Ours		**72.21**	79.67 *	**80.71**	87.00	**95.73**	**95.45**

Bold indicates the best performance, and underline indicates the second-best. An asterisk [*] indicates that the difference between the best and second-best performance is statistically significant (Wilcoxon signed-rank test, $p < 0.05$). This analysis was applied to the IoU and F1 metrics only. The absence of a symbol indicates no statistically significant difference.

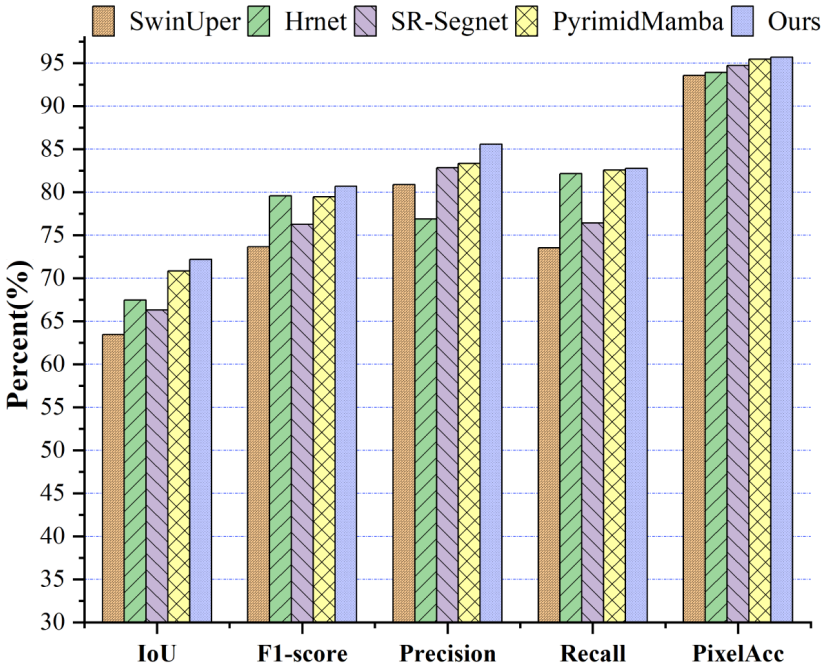


Figure 4. Histogram of results compared with representative networks.

In complex environments or when dealing with small water bodies, our network demonstrates superior boundary delineation compared to other methods. This capability

is particularly important in rural regions where water bodies are often fragmented and surrounded by diverse land cover types. The enhanced edge detection ensures that even small and irregularly shaped water bodies are accurately segmented, providing a more detailed and precise representation.

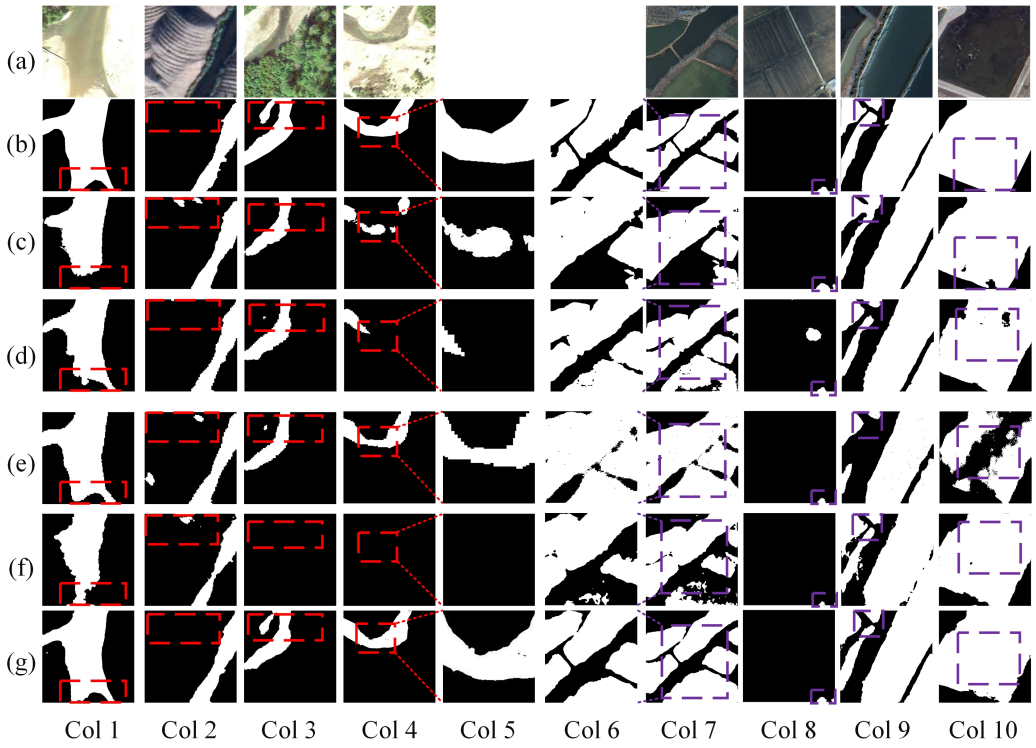


Figure 5. Visualization of segmentation results from different models. Col (1–4) concern the GLH-water dataset, and Col (7–10) concern the LoveDA dataset. Col (5, 6) represents a zoomed-in view. (**a**) Images; (**b**) masks; (**c**) PyrimidMamba; (**d**) SwinUperNet; (**e**) HRNet; (**f**) SR-Segnet; (**g**) Ours. The rectangular colored boxes are used to highlight critical regions where the performance differences between models are most apparent.

The novelty of our approach lies in the integration of global information extraction and edge awareness. By leveraging the CIMM module, our network effectively captures both global context and local details, ensuring comprehensive feature representation. Additionally, the ERM module, which utilizes transformer mechanisms, enhances the network's ability to focus on water boundaries during the up-sampling process. This dual focus on global information and edge awareness significantly improves the accuracy and efficiency of water body extraction, making our method particularly suitable for high-resolution remote sensing applications in rural areas.

Finally, to provide a detailed analysis of the trade-off between accuracy and efficiency, we present a comprehensive comparison of model complexity and practical inference speed in Table 3. All inference speeds were benchmarked on a single NVIDIA GeForce RTX 3090 GPU with a 512 $\times$ 512 input to ensure a fair and direct comparison. The results clearly demonstrate the exceptional efficiency of our proposed HWBENet. In terms of model size, HWBENet is the most lightweight, with only 3.31 M parameters, which is even smaller than the highly efficient SegFormer (3.71 M). Additionally, HWBENet achieves a low computational complexity of 13.32 G FLOPs, making it suitable for deployment in resource-constrained environments. In terms of inference speed, HWBENet processes images at 123.7 FPS, which is significantly faster than many other models, including HRNet (52.3 FPS) and SwinUperNet (13.7 FPS). This high speed is particularly advantageous for real-time applications in remote sensing, for which the rapid processing of large datasets is

often required. While some specialized real-time networks like ABCNet and Unetformer achieve higher FPS, Table 2 shows that they do so at the cost of significantly lower segmentation accuracy. To provide a clear visual summary of the trade-off between accuracy and efficiency, we selected some networks with relatively low computational complexity from Table 3 for a visual display. Figure 6 illustrates a comprehensive comparison of model efficiency and segmentation performance across all evaluated networks. Each dot in the scatter plot represents a network, with its position determined by the number of parameters (horizontal axis) and the computational cost measured in FLOPs (vertical axis), while the IoU is labeled on each dot. From Figure 6, it is evident that our HWBENet achieves a superior balance between model efficiency and segmentation accuracy. It occupies a favorable position in the lower left quadrant of the plot, indicating both a low parameter count and low computational complexity while also achieving one of the highest IoU scores among all models.

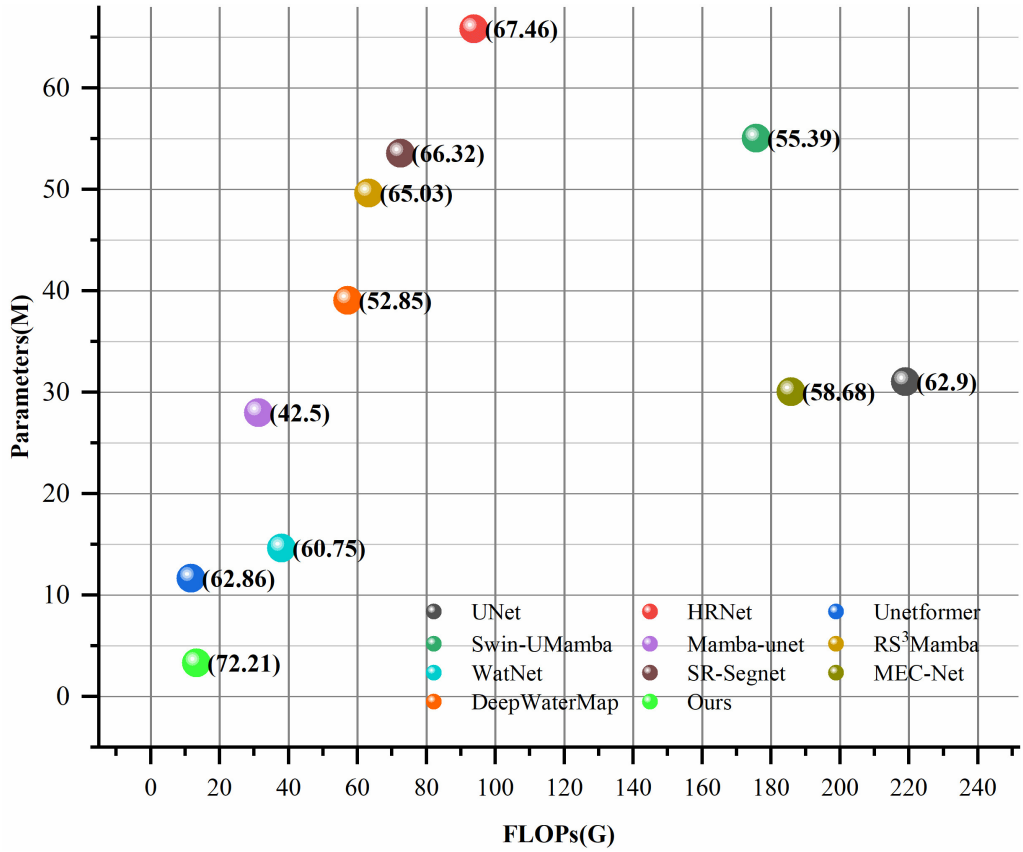


Figure 6. The parameters and computational complexity of different network models. The contents in parentheses are IoU indicators.

Table 3. Comparison of model efficiency of different networks, including theoretical complexity (parameters and FLOPs) and practical inference speed (frames per second, FPS).

Methods		Params (M)	FLOPs (G)	Speed (FPS)
CNN-based	Unet [6]	31.04	218.95	50.9
	U2Net [45]	44.01	150.61	65.2
	ABCNet [48]	13.31	15.29	**379.4**
	BANet [47]	12.64	12.73	<u>358.4</u>
	DeeplabV3+ [7]	58.82	73.77	51.8
	PSPNet [20]	73.09	279.51	43.1
	HRNet [46]	65.84	93.82	52.3

Table 3. *Cont.*

Methods		Params (M)	FLOPs (G)	Speed (FPS)
Transformer	SegFormer [9]	3.71	**6.76**	125.2
	SwinUperNet [10]	125.16	1092	13.7
	Unetformer [49]	11.68	11.73	322.4
	Segmenter [50]	36.13	63.70	58.8
Mamba-based	PyrimidMamba [42]	109.99	100.90	97.3
	Swin-UMamba [51]	55.05	175.73	47.5
	Mamba-unet [52]	28.00	31.24	88.9
	RS3Mamba [40]	49.65	63.29	75.6
Water Extraction methods	WatNet [24]	14.65	38.01	87.1
	SR-Segnet [53]	53.55	72.55	54.7
	MEC-net [28]	30.06	185.84	55.8
	DeepWaterMapV2 [54]	39.08	57.20	58.1
Ours	**HWBENet**	**3.31**	13.32	123.7

Bold indicates the best performance, and underline indicates the second-best. The complexity and speed are measured via a 512 × 512 input on a single NVIDIA GTX 3090 GPU.

5.2. Ablation Study

5.2.1. Effectiveness of CIMM and ERM

The ablation study was conducted to quantify the individual contributions of the proposed CIMM and ERM modules. To this end, we systematically evaluated the network with and without these modules, observing their impact on key performance metrics and feature representation.

Table 4 presents the evaluation metrics for the full model, the model without CIMM, and the model without ERM. Metrics such as the IoU and the F1-score are reported. The full model consistently achieves the highest scores across all metrics, indicating that the combination of both modules is essential for optimal performance. Notably, removing CIMM or ERM results in a significant drop in the IoU and the F1-score, highlighting the effectiveness of rural water extraction tasks.

Table 4. The ablation experimental results of each component on both datasets.

Methods	CIMM	ERM	IoU		F1	
			LoveDA	GLH-Water	LoveDA	GLH-Water
I	✔	✔	72.21	79.67	80.71	87.00
II	✗	✔	66.49	77.26	75.64	85.00
III	✔	✗	68.42	76.75	77.24	84.74

Figure 7 visualizes feature maps generated at different stages of the ablation study. The heatmaps illustrate how CIMM enhances global feature awareness, capturing large-scale patterns crucial for segmenting extensive water bodies. ERM, on the other hand, excels at refining edge features, ensuring precise delineation even in small or irregularly shaped water areas. The absence of either module results in a notable reduction in feature clarity and segmentation accuracy, particularly in challenging scenarios.

The quantitative results in Table 4, combined with the qualitative insights from Figure 7, validate the necessity of both CIMM and ERM for superior performance. The full model achieves a harmonious balance of global context awareness and local edge sensitivity, resulting in significant improvements in segmentation accuracy and robustness. This

demonstrates that each module addresses complementary aspects of the segmentation task, and their integration is key to the network's overall success.

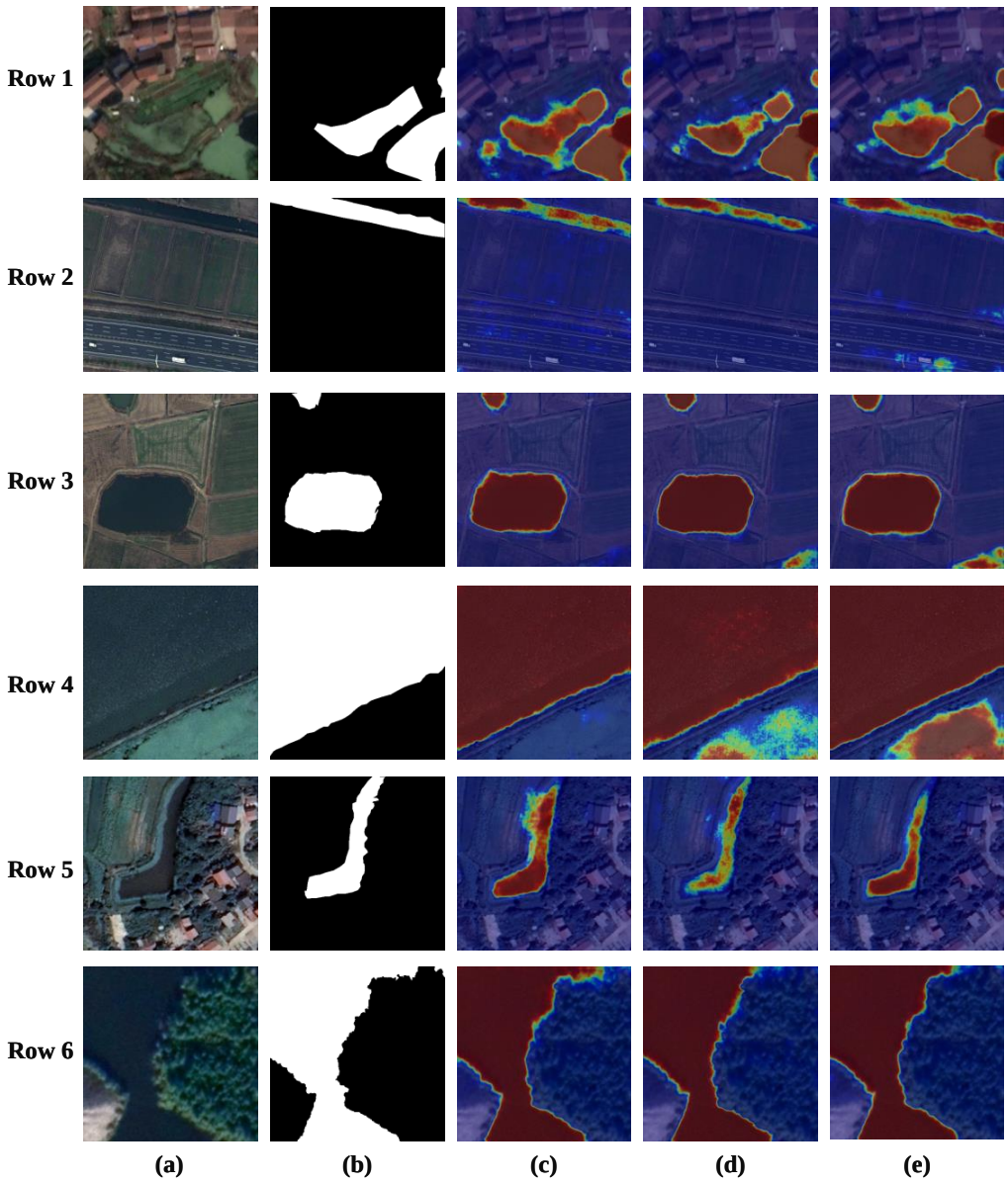


Figure 7. Feature representation visualization embedded in different modules. (**a**) Input image; (**b**) Ground truth; (**c**) Method I; (**d**) Method II; (**e**) Method III.

5.2.2. Analysis of Scanning Strategies in CIMM

To empirically validate the rationale for our choice of scanning strategies within the CIMM module, we conducted a dedicated ablation study. We compared our proposed hybrid strategy against two homogeneous variants: (1) All Forward-Backward, where both Mamba-C and Mamba-P utilize forward-backward scanning, and (2) All Forward-Random, where both components use forward and random scanning. All variants were trained and evaluated on the LoveDA dataset under identical experimental settings.

As shown in Table 5, our proposed hybrid strategy consistently outperformed both homogeneous variants across all metrics. Our analysis revealed that the All Forward-Backward variant, due to its fixed scanning patterns, lacks flexibility in modeling irregular spatial shapes. Conversely, the All Forward-Random variant likely disrupts the struc-

tured nature of the channel dimension, thereby failing to effectively learn stable global semantic patterns.

Table 5. Ablation study on different scanning strategy combinations within the CIMM module on the LoveDA dataset. The best results are in bold.

Method Variant	Mamba-C Scan	Mamba-P Scan	IoU (%)	F1 (%)
All Forward-Backward	Forward-Backward	Forward-Backward	70.54	78.12
All Forward-Random	Forward-Random	Forward-Random	69.12	76.34
Ours	Forward-Backward	Forward-Random	**72.21**	**80.71**

These results experimentally confirm that our task-driven hybrid design, which matches different scanning mechanisms to distinct sub-problems, is indeed the optimal choice for addressing the dual challenges of rural water body segmentation.

6. Conclusions

This study has introduced a novel water body extraction network that leverages the CIMM module for global feature modeling and the ERM module for enhanced edge awareness. Extensive evaluations on benchmark datasets demonstrate the proposed model's ability to achieve a favorable balance between segmentation accuracy and computational efficiency. While heavyweight models like PyrimidMamba may achieve marginally higher accuracy on certain datasets, our HWBENet provides statistically comparable or strong second-best performance while being an order of magnitude more efficient in terms of parameters, FLOPs, and practical inference speed.

Through the ablation study, the significance of the CIMM and ERM modules was clearly demonstrated. The CIMM module effectively captures global contextual information, crucial for segmenting large and continuous water areas, while the ERM module excels in edge preservation, particularly for small or complex water regions. Visualizations further highlight the network's ability to balance global feature extraction with precise edge delineation, ensuring continuity and integrity in water body boundaries.

Despite its promising performance, we acknowledge certain limitations that open avenues for future research. Our model's validation was conducted on high-resolution optical imagery from Google Earth. Future work should focus on evaluating HWBENet's generalizability across a wider range of sensor types, seasonal conditions, and imaging geometries. Furthermore, exploring network quantization or pruning could further enhance its suitability for deployment on resource-constrained edge devices.

In summary, HWBENet provides a robust, accurate, and highly practical solution for large-scale hydrological monitoring. It establishes a strong foundation for further research and development in remote sensing applications requiring precise and efficient segmentation.

Author Contributions: Conceptualization, P.L. and J.Z.; methodology, P.L. and J.Z.; validation, J.Y.; formal analysis, J.Z. and J.Y.; resources, J.Z.; writing—original draft, P.L.; writing—review and editing, P.L. and J.Z.; visualization, J.Y.; supervision, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Program of the Hunan Provincial Department of Education (No. 22B0258) and the Natural Science Foundation of Hunan Province (No. 2024JJ5649).

Data Availability Statement: The data presented in this study will be available upon request.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-4o) for the purpose of improving grammar and readability. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

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

References

1. Wang, M.; Li, C.; Yang, X.; Ban, Y.; Chu, D.; Zhou, Z.; Lau, R.Y. QTU-Net: Quaternion transformer-based U-Net for water body extraction of RGB satellite image. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5634816. [CrossRef]
2. Lei, P.; Yi, J.; Li, S.; Li, Y.; Lin, H. Agricultural surface water extraction in environmental remote sensing: A novel semantic segmentation model emphasizing contextual information enhancement and foreground detail attention. *Neurocomputing* **2025**, *617*, 129110. [CrossRef]
3. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
4. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
6. 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: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
7. 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) 2018, Munich, Germany, 8–14 September 2018; pp. 801–818.
8. 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.
9. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
10. 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 2021, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
11. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
12. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
13. Guo, Q.; Pu, R.; Li, J.; Cheng, J. A weighted normalized difference water index for water extraction using Landsat imagery. *Int. J. Remote Sens.* **2017**, *38*, 5430–5445. [CrossRef]
14. Zhou, Y.; Zhao, H.; Hao, H.; Wang, C. A new multi-spectral threshold normalized difference water index (mst-ndwi) water extraction method—A case study in yanhe watershed. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 2557–2564. [CrossRef]
15. Jia, K.; Jiang, W.; Li, J.; Tang, Z. Spectral matching based on discrete particle swarm optimization: A new method for terrestrial water body extraction using multi-temporal Landsat 8 images. *Remote Sens. Environ.* **2018**, *209*, 1–18. [CrossRef]
16. Nguyen, D. Water body extraction from multi spectral image by spectral pattern analysis. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *39*, 181–186. [CrossRef]
17. Wu, L.; Wang, L.; Min, L.; Hou, W.; Guo, Z.; Zhao, J.; Li, N. Discrimination of algal-bloom using spaceborne SAR observations of Great Lakes in China. *Remote Sens.* **2018**, *10*, 767. [CrossRef]
18. Li, L.; Yi, J.; Fan, H.; Lin, H. A Lightweight Semantic Segmentation Network Based on Self-attention Mechanism and State Space Model for Efficient Urban Scene Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4703215. [CrossRef]
19. Wang, Z.; Yi, J.; Chen, A.; Chen, L.; Lin, H.; Xu, K. Accurate semantic segmentation of very high-resolution remote sensing images considering feature state sequences: From benchmark datasets to urban applications. *ISPRS-J. Photogramm. Remote Sens.* **2025**, *220*, 824–840. [CrossRef]
20. 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 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]

22. Kang, J.; Guan, H.; Peng, D.; Chen, Z. Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102499. [CrossRef]
23. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef]
24. Luo, X.; Tong, X.; Hu, Z. An applicable and automatic method for earth surface water mapping based on multispectral images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102472. [CrossRef]
25. Nie, P.; Cheng, X.; Song, Z.; Mao, M.; Wang, T.; Meng, L. Rethinking BiSeNet: A lightweight network for urban water extraction. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–10. [CrossRef]
26. Miao, Z.; Fu, K.; Sun, H.; Sun, X.; Yan, M. Automatic water-body segmentation from high-resolution satellite images via deep networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 602–606. [CrossRef]
27. Wang, J.; Wang, S.; Wang, F.; Zhou, Y.; Wang, Z.; Ji, J.; Xiong, Y.; Zhao, Q. FWENet: A deep convolutional neural network for flood water body extraction based on SAR images. *Int. J. Digit. Earth* **2022**, *15*, 345–361. [CrossRef]
28. Zhang, Z.; Lu, M.; Ji, S.; Yu, H.; Nie, C. Rich CNN features for water-body segmentation from very high resolution aerial and satellite imagery. *Remote Sens.* **2021**, *13*, 1912. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; MIT Press: Cambridge, MA, USA, 2017; Volume 30.
30. 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**, arXiv:2102.04306. [CrossRef]
31. Zhang, T.; Qin, C.; Li, W.; Mao, X.; Zhao, L.; Hou, B.; Jiao, L. Water body extraction of the Weihe River basin based on MF-SegFormer applied to Landsat8 OLI data. *Remote Sens.* **2023**, *15*, 4697. [CrossRef]
32. Wang, S.; Guo, K.; Wang, N.; Tang, L. FDEN: Frequency-Band Decoupling Detail Enhancement Network for High-Fidelity Water Boundary Segmentation. *Remote Sens.* **2025**, *17*, 3062. [CrossRef]
33. Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; Zhang, X. RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 1052–1064. [CrossRef]
34. Zhong, H.F.; Sun, Q.; Sun, H.M.; Jia, R.S. NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
35. Qi, H.; Kong, X.; Cheng, L.; Hu, J.; Gu, J. Addressing fine-grained lake water body extraction: A hybrid approach combining vision transformer and geodesic active contour. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
36. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv* **2024**, arXiv:2401.09417. [CrossRef]
37. 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.
38. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. Rs-mamba for large remote sensing image dense prediction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5633314. [CrossRef]
39. Tang, F.; Wang, L.; Ning, C.; Xian, M.; Ding, J. Cmu-net: A strong convmixer-based medical ultrasound image segmentation network. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 18–21 April 2023; pp. 1–5.
40. 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*, 6011405. [CrossRef]
41. Chen, H.; Song, J.; Han, C.; Xia, J.; Yokoya, N. Changemamba: Remote sensing change detection with spatio-temporal state space model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4409720.
42. Wang, L.; Li, D.; Dong, S.; Meng, X.; Zhang, X.; Hong, D. PyramidMamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery. *arXiv* **2024**, arXiv:2406.10828. [CrossRef]
43. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
44. Li, Y.; Dang, B.; Li, W.; Zhang, Y. Glh-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery. In Proceedings of the AAAI Conference on Artificial Intelligence 2024, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 22213–22221.
45. 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]
46. 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]
47. Zhou, Q.; Qiang, Y.; Mo, Y.; Wu, X.; Latecki, L.J. Banet: Boundary-assistant encoder-decoder network for semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25259–25270. [CrossRef]

48. 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]
49. 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]
50. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
51. Liu, J.; Yang, H.; Zhou, H.Y.; Xi, Y.; Yu, L.; Li, C.; Liang, Y.; Shi, G.; Yu, Y.; Zhang, S.; et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Marrakesh, Morocco, 6–10 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 615–625.
52. Wang, Z.; Zheng, J.Q.; Zhang, Y.; Cui, G.; Li, L. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv* **2024**, arXiv:2402.05079.
53. Weng, L.; Xu, Y.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. Water areas segmentation from remote sensing images using a separable residual segnet network. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 256. [CrossRef]
54. Isikdogan, L.F.; Bovik, A.; Passalacqua, P. Seeing through the clouds with deepwatermap. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1662–1666. [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.