SwinDefNet: A Novel Surface Water Mapping Model in Mountain and Cloudy Regions Based on Sentinel-2 Imagery

Xinyue Chen , Haiyan Pan * , Jun Liu

Posted Date: 4 June 2024

doi: 10.20944/preprints202406.0084.v1

Keywords: remote sensing images; water body extraction; deep learning; semantic segmentation; transformer

Article

SwinDefNet: A Novel Surface Water Mapping Model in Mountain and Cloudy Regions Based on Sentinel-2 Imagery

Xinyue Chen [1], Haiyan Pan [1,*] and Jun Liu

[1] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2152309@st.shou.edu.cn
[2] National Earthquake Response Support Service; liujun_nerss@sina.com
* Correspondence: hy-pan@shou.edu.cn

Abstract: Surface water plays a crucial role in climate change, human production, and life, making accurate monitoring and observation of surface water particularly important. However, due to the significant diversity and complexity of water distribution in surface space, accurate mapping of surface water faces considerable challenges. When extracting water bodies from medium-resolution satellite remote sensing images, CNN methods may suffer from limitations in receptive fields and insufficient context modeling capabilities, leading to the loss of water body boundary details and poor fusion of multiscale features. Currently, there is relatively little research on this issue; therefore, it is necessary to explore new combinations of deep learning networks to address these challenges. The purpose of this study is to address the above issues. We propose a new combination of deep learning networks that fully utilize multiscale information to enhance water features. Specifically, we first combine deformable convolutions with the Swin Transformer to increase effective receptive fields while better integrating global semantic information. This combination can capture features of water bodies at different scales, improve the accuracy and integrity of water extraction, and provide reliable technical support for detailed water body extraction. We tested the newly constructed model using Sentinel-2 satellite images. Our model achieved results of over 90%, with an average accuracy of 97.89%, average precision of 94.98%, average recall of 90.05%, and an average F1 score of 92.33%. In addition, our model achieved an accuracy of 98.03% in mountainous areas. Our experiments and results validate the potential of combining the Swin Transformer and deformable convolutions in detailed water body extraction.

Keywords: remote sensing images; water body extraction; deep learning; semantic segmentation; transformer

1. Introduction

Surface water is an indispensable part of the Earth's ecosystem and a significant influencing factor in climate change, ecological environmental protection, and human production and life [1]. It plays crucial roles in various aspects such as environmental monitoring and management [2], ecological protection and restoration, disaster prevention and control, emergency response, agricultural production, land use, and people's water safety [3]. By obtaining and analyzing the spatial distribution and area information of water bodies, we can guide and support humans to adopt better ways of living [4]. Consequently, precise mapping of surface water holds significant importance for both environmental surveillance and societal advancement.

Due to its expansive coverage, relatively high spatial and temporal resolutions, as well as the distinct advantage of uninterrupted Earth surface monitoring, Remote Sensing (RS) technology has become a ubiquitous tool in the extraction of surface water [5]. The methods utilized for water extraction from satellite remote sensing images can be categorized into three principal groups: (1) threshold-based methods, (2) machine learning methods, and (3) hybrid methods. During the process of water extraction, threshold-based methods primarily depend on the spectral reflectance properties of water bodies within specific bands, encompassing both single-band threshold techniques and

 doi:10.20944/preprints202406.0084.v1

multi-band threshold methodologies. The former utilizes information from a single band [6] for water body identification, while the latter uses a set of multi-band data to detect water bodies through mathematical and logical operations, such as Normalized Difference Water Index (NDWI) [7], Modified Normalized Difference Water Index (MNDWI) [8], Multi-band Water Index (MBWI) [9], Background Difference Water Index (BDWI) [10], Normalized Difference Water Fraction Index (NDWFI) [11], Composite Normalized Difference Water Index (CNDWI) [12], etc. Fuzzy C-means, K-means clustering, support vector machine, decision tree, random forest and other machine learning methods are used to identify and extract water bodies. Hybrid methods integrate water features and machine learning classifiers or multi-classifier ensembles to achieve high-precision water body extraction. The image processing involved in hybrid methods is complex, with multiple influencing factors and high uncertainty. Typically, traditional methods heavily depend on the expertise of domain experts and may have limited abilities to express features, which makes it challenging to fully grasp intricate semantic details and spatial connections between pixels.

Compared to traditional water extraction methods, deep learning methods possess the capability to learn and explore deep features, enabling the acquisition of more complex and nonlinear water characteristics [13]. They can avoid the need for manual adjustment of optimal thresholds, adapt to large-scale learning, and exhibit higher flexibility and generality, thus finding widespread applications in water extraction research. Isikdogan [14] introduced a unique CNN design named DeepWaterMap, utilizing a fully convolutional network structure that minimizes the number of parameters requiring training and enables comprehensive, large-scale analysis. The network embeds the shape, texture, and spectral features of water bodies to eliminate interfering features such as snow, ice, clouds, and terrain shadows. Chen [15] presented a novel approach for detecting open surface water in urbanized areas, employing unequal and physical size constraints to recognize water bodies in urban environments. This method addresses the serious confusion errors of traditional water resource indices in high spatial resolution images. The potential application of the method in large-scale water detection tasks is evidenced through experimental verification on spectral libraries and genuine high spatial resolution RS imagery. Kang et al. [16] introduced a multi-scale context extraction network, MSCENet, aimed at precise and efficient extraction of water bodies from high-resolution optical RS images. This network incorporates multi-scale feature encoders, feature decoders, and context feature extraction modules. Specifically, the feature encoder employs Res2Net to capture rich multi-scale details of water bodies, effectively handling variations in their shape and size. The context extraction module, comprising an expanded convolutional unit and a sophisticated multi-kernel pooling unit, further distills multi-scale contextual information to produce refined high-level feature maps. Luo et al. [17] proposed an automated method for surface water mapping and constructed a novel surface water mapping model called WatNet. This model addresses the issue of diminished mapping precision caused by the resemblance of non-water features to water features, employing a tailored design for mapping surface water to achieve precise identification of smaller water bodies. The study also constructed the Earth Surface Water Knowledge Base (ESWKB), a freely available dataset based on Sentinel-2 images. Li et al. [18] proposed a water index-driven deep fully convolutional neural network (WIDFCN) method that achieves precise water delineation without relying on manually collected samples. WIDFCN effectively handles scale and spectral variations of surface water and demonstrates robustness in experiments involving different types of shadows, such as those from buildings, mountains, and clouds. The most important aspect of this method is the extraction of high-precision but incomplete water membranes from water spectral indices, which are then expanded to enhance completeness. This approach realizes an efficient strategy for automatically generating training samples without the need for manual labeling, significantly reducing economic costs. Zhang et al. [19] proposed an end-to-end CNN water segmentation network, MRSE-Net, based on multi-scale residual and squeeze-excitation (SE) attention. The network enhances prediction results using the SE-attention module to alleviate water boundary ambiguity and reduces the number of model parameters using multi-scale residual modules to accurately extract water pixels, addressing the problem of fuzzy boundaries of small river water bodies. Yu et al. [20] proposed a network called WaterHRNet, which is composed of multi-branch

 doi:10.20944/preprints202406.0084.v1

high-resolution feature extractor (HRNet), feature attention module and segmentation header module. It is a hierarchical focus high-resolution network that can provide high-quality, strong semantic feature representation for precise segmentation of water bodies in various scenarios. Xin Lyu et al. [21] proposed a multi-scale normalized attention network, MSNANet, for accurate water body extraction in complex scenes. The network incorporates the Multi-Scale Normalized Attention (MSNA) module to fuse multi-scale water body features, highlighting feature representations. It utilizes an optimized spatial pyramid pooling (OASPP) module to refine feature representations using contextual information, improving segmentation performance. Kang et al. [22] proposed WaterFormer, a combination of transformer and convolutional neural network, for accurate water detection tasks. The network includes dual-stream CNNs, Cross-Level Visual Transformers (CL-ViT), lightweight attention modules (LWA), and sub-pixel upsampling modules (SUS).The network includes dual-stream CNNs, Cross-Level Visual Transformers (CL-ViT), lightweight attention modules (LWA), and sub-pixel upsampling modules (SUS). The dual-stream network abstracts water features from multiple perspectives and levels, embeds cross-level visual transformers in the dual stream to capture long-range dependencies between foundational spatial information and high-order semantic features, and enhances feature abstraction and generates high-resolution, high-quality class-specific representations using lightweight attention modules and sub-pixel upsampling modules.

From the above analysis of the current mainstream water extraction techniques, we can clearly understand that most of the methods are based on Convolutional Neural Networks (CNN). Although CNNs possess powerful feature extraction capabilities, the diversity of water body spatial distributions and the complexity of environmental backgrounds can lead to the loss of boundary details and affect the accuracy of water body extraction when using CNNs on medium-resolution satellite remote sensing images. Thus, they exhibit certain limitations in water body extraction. In recent years, Transformers have attracted attention due to their outstanding semantic representation capabilities and advantages in modeling global information relationships. Particularly, the Swin Transformer [23] has demonstrated strong feature extraction capabilities, contextual modeling capabilities, and multiscale feature fusion capabilities, providing a strategy for precise water body extraction from remote sensing images. However, research in this area is currently limited. The aim of this research, therefore, is to investigate a novel integration of deep learning networks that leverages multiscale information to the fullest extent, thereby enhancing the identification of water body features. For the first time, we combine deformable convolutions [24] with the Swin Transformer to increase effective receptive fields and better integrate global semantic information.

This network skillfully combines the powerful local feature extraction capabilities of CNNs with the extensive global feature extraction capabilities of Swin Transformers, enabling high-precision extraction of water bodies. Our main contributions in this research are outlined as follows:

(1) We designed a new combination of deep learning networks, which combines CNNs and Swin Transformers for the first time. The refined model emphasizes the extraction of water body features, particularly the accurate delineation of water body boundaries. To achieve this goal, we capture the image's details and edge information through CNNs and model global contextual information using Swin Transformers to better capture image semantic information. This hybrid model can consider both detailed information in images and global contextual information, thereby improving the accuracy and performance of semantic segmentation.

(2) Considering the complex morphology and size variations of water body boundaries, we use deformable convolutions for the precise extraction of water body boundary features. Deformable convolutions, by introducing offsets, can adaptively adjust the receptive fields, allowing convolutional kernels to adaptively deform on input feature maps according to target shapes, thereby capturing water body features more accurately.

(3) To ascertain the efficacy of our approach, we apply it to water bodies of different sizes and in different environments. We utilized the image dataset of Sentinel-2 to conduct tests in both high mountainous and cloud-covered areas. The results indicate that the model exhibits a high level of accuracy.

 doi:10.20944/preprints202406.0084.v1

The content distribution of the remaining sections of this paper is as follows: Section 2 introduces the details of the data and methods used in this study. Section 3 analyzes the experimental results and provides the experimental configuration. Section 4 discusses the ablation experiments. Finally, our conclusions are presented in Section 5.

2. Materials and Methods

2.1. Materials

We utilized the Earth Surface Water Knowledge Base (ESWKB). This dataset is publicly available and can be downloaded from Zenodo (https://zenodo.org/records/5205674). The ESWKB dataset fully utilizes Sentinel-2 satellite imagery resources, meticulously selecting 95 different scenes. Surface water labeling was conducted using six bands: blue, green, red, near-infrared (NIR), mid-infrared 1, and mid-infrared 2, covering various types of water bodies under different environmental conditions.

The Sentinel-2 mission has been designed, constructed, and managed by the ESA's Copernicus Programme since 2015 [25]. This mission 's primary objective is to observe the Earth's surface, offering crucial services like forest surveillance, detection of alterations in land cover, and effective management of natural disasters. The Sentinel-2 satellite is equipped with a Multi-Spectral Instrument (MSI) capable of capturing images in 13 bands, with resolutions ranging from 10 meters to 60 meters.

To assess the effectiveness of SwinDefNet, we selected 38 images from the ESWKB dataset as the test dataset, using the remaining images for training. We chose six images from the test set to illustrate the model's performance (see Figure 1), covering mountainous areas and cloud-covered areas (see Table 1).

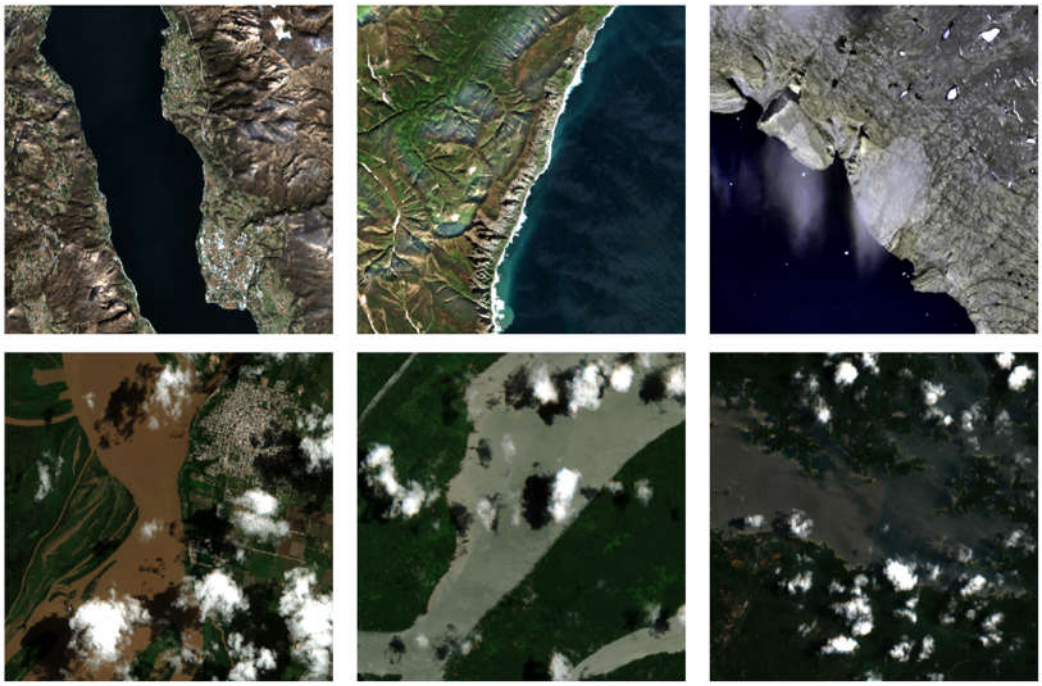


Figure 1. Representation of Regions in Images (The first line represents mountains, and the second line represents cloud regions.).

Table 1. Characteristics and Challenges of Different Regions.

Region	Characteristics	Challenges
Mountainous Area	Large undulating terrain, complex landforms, scattered water bodies, and high-altitude	Water bodies are often obstructed by mountains, resulting in incomplete extraction information; their

 doi:10.20944/preprints202406.0084.v1

	areas that may be covered by ice and snow.	scattered distribution leads to a small extraction scale range; there is also potential interference from ice and snow.
Cloudy Area	Under conditions of frequent clouds, rain, and fog, cloud cover areas are large and last for long durations. Clouds exhibit spectral characteristics similar to water bodies in some bands.	Cloud cover affects the transmission and reflection characteristics of remote sensing images, increasing the difficulty of water body extraction. Due to their spectral similarities, confusion between clouds and water bodies is prone to occur.

2.2. Methods

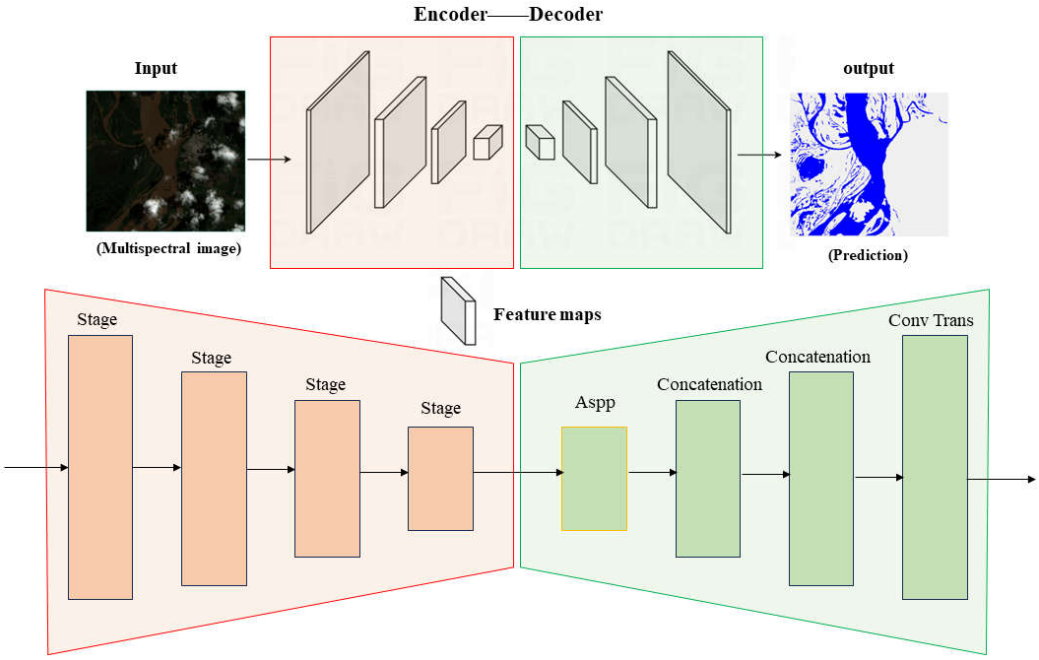
2.2.1.

The encoder-decoder architecture plays an important role in feature extraction, data compression, sequence mapping, and multi-level feature fusion. It enables models to handle various complex data transformation and generation tasks and is widely used in natural language processing (NLP), digital image processing, time series prediction, speech recognition, and more. Modern deep learning models such as CNNs, RNNs, and Transformers all originate from this architecture. Many semantic segmentation models such as U-net [26], DeeplLabV3 [27], and Pspnet [28] are built based on the encoder-decoder architecture. The encoder-decoder architecture consists of two parts:

(1) Encoder: Maps high-dimensional input images to low-dimensional representations. It extracts features from input images and condenses the spatial dimensions of the feature maps.

(2) Decoder: Reconstructs the low-dimensional representations mapped by the encoder to the original input. It restores the spatial scale and target details of the original image.

The goal of semantic segmentation is to label each pixel in the image with the corresponding semantic category. In this study, our aim is to achieve accurate mapping of surface water bodies and refine their extraction from remote sensing images. Both tasks essentially involve pixel-level understanding of the image and assigning a semantic label to each pixel. Therefore, we adopt the encoder-decoder architecture, commonly used in semantic segmentation, to construct the network for mapping surface water bodies. The overall framework is illustrated in Figure 2.



 doi:10.20944/preprints202406.0084.v1

Figure 2. Use of encoder-decoder architecture in semantic segmentation. (Encoder: the spatial dimensions of features gradually decrease while the depth (number of channels) increases. Decoder: the spatial dimensions of feature maps gradually increase while the depth decreases).

Considering the scale of feature extraction and model computational efficiency, we adopt Swin Transformer and DeepLabV3+ as the encoder and decoder of the model, respectively. Swin Transformer employs a hierarchical structure, using different numbers of transformer layers at different stages to capture relationships between various regions through the self-attention mechanism, effectively handling image information at different scales. Moreover, by using the windowed self-attention mechanism, it reduces the computational cost of the model. DeepLabV3+ is a model designed for semantic segmentation, employing dilated convolutions as its core component, which enhances segmentation accuracy through multi-scale feature fusion while efficiently recovering detailed object boundaries.

2.2.2. Water Extraction Network Based on Swin Transformer

The Transformer, originally proposed for natural language processing (NLP), differs from CNNs in feature extraction by using self-attention. The Transformer also possesses powerful global information relationship modeling capabilities. However, it encounters high computational complexity issues with long sequence inputs. In 2021, the Swin Transformer was introduced to address the challenges of large-scale visual entities and computational complexity. This technology has shown great potential in tackling various visual tasks, such as image classification, object detection, and semantic segmentation [29]. In this study, the Swin Transformer is employed to better capture water features at different scales and establish contextual global information for the fine extraction of water boundaries. The network structure for surface water mapping using Swin Transformer for feature extraction is illustrated in Figure 3, employing an encoder-decoder architecture.

The encoder is established based on the Swin Transformer, which directly takes the original image size as input. We segment the images into patches and reshape them into feature vectors through patch embedding. The Swin Transformer is a hierarchical feature extraction network, constructing hierarchical feature mappings with linear computational complexity related to image size. It consists of four stages, each composed of patch merging and Swin Transformer blocks. Patch merging performs downsampling operations at the beginning of each stage to reduce resolution and adjust channel numbers, which helps save computational costs. Each Swin Transformer block comprises windowed multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) layers for local feature extraction. Notably, the Swin Transformer performs both intra-window and inter-window attention computations to extract global information. Through the shifted window operation, feature maps are shifted with a set mask, solving the issue of the number of windows after shifting and obtaining equivalent computational results. It achieves multi-scale feature extraction while enhancing model computational efficiency by implementing cross-window connections.

In the encoder section of our model, each stage outputs a feature layer, resulting in four feature maps of different sizes, corresponding to 1/16, 1/32, 1/64, and 1/128 of the input size. We designate the feature layers output from the first, second, and fourth stages as basic, intermediate, and advanced feature layers, forming a group of feature maps representing different scales of water features, which are then input into the decoder section.

In this study, we use Deeplabv3+ to decode the features extracted by the encoder. Deeplabv3+ is a semantic segmentation model based on dilated convolutions that, through the setting of dilation factors, achieves different-scale receptive fields to freely extract multi-scale information [30]. It possesses excellent feature extraction capabilities, aiding the network in distinguishing differences between water bodies and backgrounds. Importantly, this model can efficiently recover detailed boundary information of targets. Additionally, we introduce deformable convolutions into the ASPP

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2024 doi:10.20944/preprints202406.0084.v1

module for water feature extraction in the advanced feature layer. The internal structure of ASPP is illustrated in Figure 4.

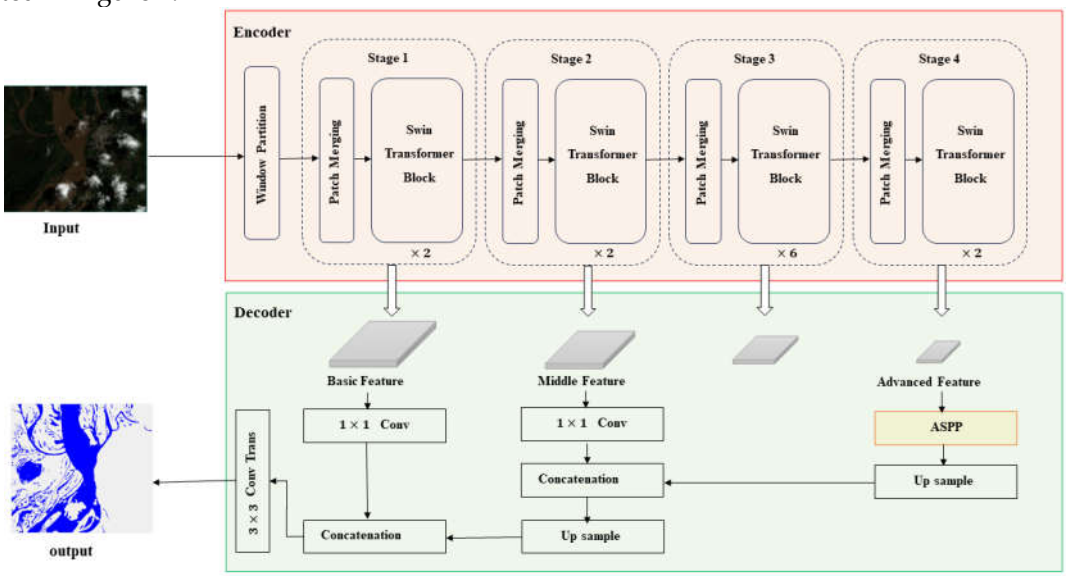


Figure 3. SwinDefNet's network structure diagram. Normalization and ReLU activation layers follow each convolution operation in the network.

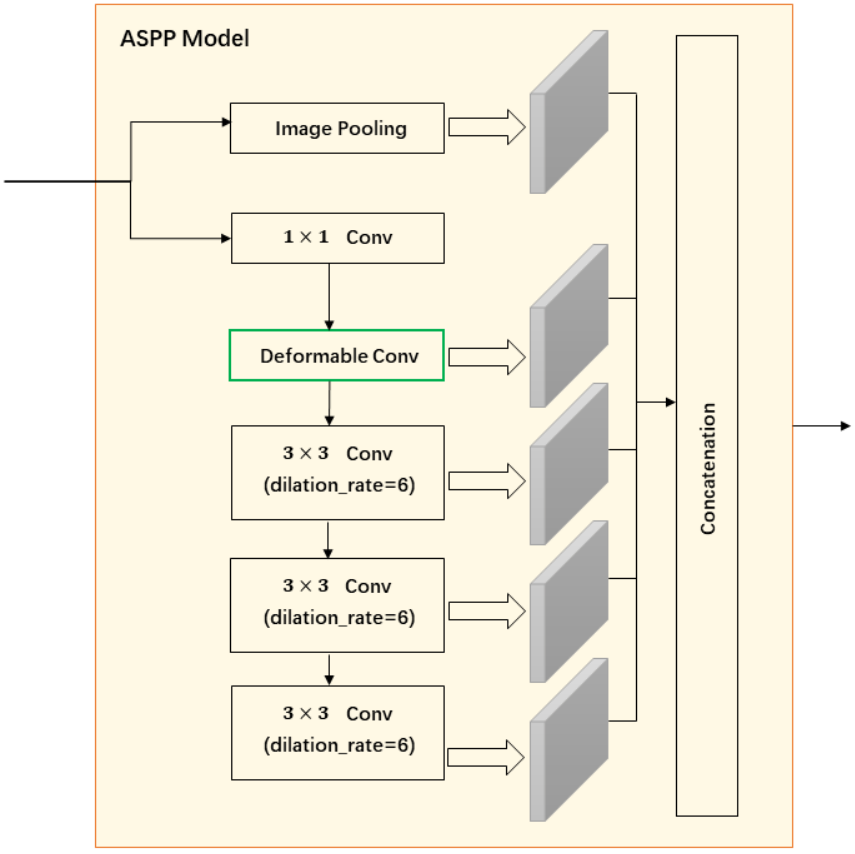


Figure 4. Internal details of the ASPP module.

In this study, we decided on a patch size of 4 and an embedding dimension of 96 for the images. We used Swin Transformer blocks in each stage, with a total of 2, 2, 6, and 2 for the different stages. Each of the blocks had 3, 6, 12, and 24 multi-head self-attention heads, adding to the complexity of the model. The window size was set to 7, creating a sense of unpredictability. These settings were

 doi:10.20944/preprints202406.0084.v1

carefully chosen to work smoothly with our processor and deliver excellent accuracy in a relatively short timeframe.

2.2.3. Deformable Convolution

Remote sensing images contain rich surface information. However, due to their varying resolutions and the diversity of water body shapes, significant differences can exist in the information captured by remote sensing images. For instance, a lake covering 100 km² may be represented by just one pixel in the image, making it difficult for traditional convolution operations to accurately extract its features. Traditional convolutional kernels have fixed receptive fields and sizes, making them unable to adapt to geometric deformations. They struggle to extract effective features when dealing with targets that are too large or too small, although they perform better with regularly shaped water bodies subject to human interference. Some studies have used dilated convolutions to overcome this issue, aiming to expand the effective receptive field and capture multi-scale information. However, this approach may result in the loss of some detailed image features. To better extract features from the input remote sensing images, we introduce deformable convolutions.

Compared to traditional convolution and dilated convolution, deformable convolution predicts offsets for feature sampling points, adaptively changes sampling positions, allows target points to fall on targets as much as possible, adapts to irregular situations, and better handles geometric deformations for feature extraction. Deformable convolution adds a direction vector to each convolution kernel, enabling adaptive shape changes, automatically adjusting scale and receptive field, and aligning more closely with the shape and size of objects. It introduces an offset parameter based on traditional convolution. Firstly, based on the input feature map, it generates X and Y direction offset layers (2N), then combines the offset layers to obtain the output feature map through deformable convolution. The implementation process of deformable convolution is illustrated in Figure 5.

First, we define a standard convolution kernel R, denoted w as the weighted sum of sampling values, obtaining Equation (1), where R represents the regular grid of traditional convolution with a 3×3 convolution kernel of stride 1. Then, a standard convolution feature y(p_0) matrix can be obtained as shown in Equation (2). Finally, an offset Δp_n is introduced into R to realize the offset of feature points. The value of Δp_n is shown in Equation (3), and the eigenmatrix of the deformable convolution kernel is obtained as shown in Equation (4).

$$\mathrm{R} = \{(-1,-1),\quad (-1,0),\quad \cdots\quad ,(0,1),\quad (1,1)\} \tag{1}$$

$$y(p_0) = \sum_{p_n \in R} w(x_n) \cdot x(p_0 + p_n) \tag{2}$$

$$\{\Delta p_n | n = 1, \cdots, N\}, N = |R| \tag{3}$$

$$y(p_0) = \sum_{p_n \in R} w(x_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{4}$$

When the feature map from the previous layer undergoes 3×3 convolution, we first define another 3×3 convolutional layer (offset field) with the same size as the input feature map and a channel number of 2N, representing the offsets in the X and Y directions. Using this offset field for interpolation, we then perform standard convolution operations. To address the challenge of potentially non-integer offsets, bilinear interpolation is employed, computing the weighted average of the four neighboring pixel values around each sample point to estimate the pixel value at the new position. For each point in the feature map, we need to consider the four neighboring pixels it may correspond to after offsetting. Thus, within the 3×3 convolution kernel range, each sample point may be associated with up to 36 different pixel values. Due to the effect of the offset field, these positions vary with changes in the offset, thereby achieving feature extraction capabilities for multiscale and irregular shapes. This mechanism effectively enlarges the receptive field of the convolution operation,

 doi:10.20944/preprints202406.0084.v1

allowing the network to capture complex structures and subtle changes in remote sensing images more finely and accurately.

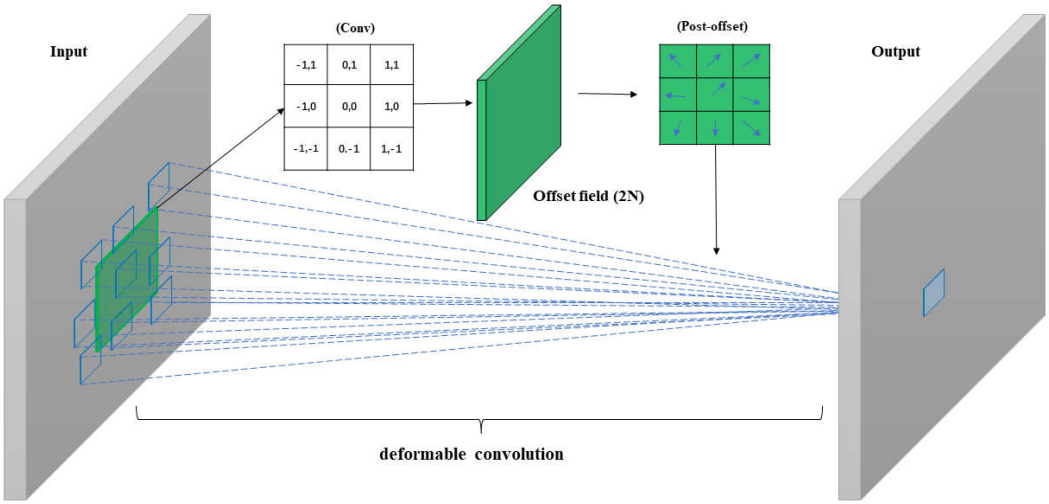


Figure 5. Using 3×3 convolution as an example, this demonstrates the offset process of deformable convolution and shows the corresponding effective receptive field.

3. Results

3.1. Experimental Environment and Parameter Settings

The experiments were conducted on an Intel(R) Core(TM) i7-12700H 2.30 GHz processor with 16.0 GB of RAM, an NVIDIA GeForce RTX 3060 Laptop GPU, and CUDA 11.2. The input image size was set to 256×256, batch size to 2, epochs to 200, and the learning rate to 0.002.

We trained and validated our model on the ESWKB dataset. We divided the dataset into a training set, a test set, and a validation set in a ratio of 6:2:2, with images randomly cropped to the specified pixel size of 256×256. Additionally, simple data augmentation techniques were applied to the training set, including image flipping and random rotations by multiples of 90°.

3.2. Evaluation Metrics

To assess the SwinDefNet performance, we utilized four widely recognized metrics in the field of semantic segmentation: accuracy, precision, recall, and the F1 score.

The fraction of accurately predicted pixels relative to all pixels is known as accuracy. The fraction of accurately predicted water body pixels among all pixels projected to be water bodies is expressed as precision. Out of all actual water body pixels, recall is the proportion of water body pixels that were accurately anticipated. The F1 score, which is a measure of the model's overall performance, is the harmonic mean of accuracy and recall.

In these evaluation metrics, accuracy represents the proportion of pixels that are correctly predicted relative to the total number of pixels. The precision quantifies the proportion of pixels that are labeled as bodies of water but are also bodies of water. Recall, on the other hand, measures the proportion of pixels correctly identified in the actual water body. F1 score is the harmonic average of accuracy and recall and can be used as a comprehensive measure of the overall performance of the model.

The calculation methods for accuracy, precision, recall, and F1 score are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

 doi:10.20944/preprints202406.0084.v1

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$f1_{score} = \frac{2 * precision - recall}{prscision - recall} \tag{8}$$

In the formulas, TP represents true positives, indicating the number of samples where the model correctly predicts water body when the actual class is water body; TN represents true negatives, indicating the number of samples where the model correctly predicts non-water body when the actual class is non-water body; FP represents false positives, indicating the number of samples where the model incorrectly predicts water body when the actual class is non-water body; FN represents false negatives, indicating the number of samples where the model incorrectly predicts non-water body when the actual class is water body.

3.3. Method Comparison

To evaluate the performance of the proposed model, we compared it with four commonly used methods: U-Net, ResNet, DeepLabv3+, and DeepWaterMapv2. Here are detailed descriptions of these methods:

U-Net: U-Net is a fully convolutional neural network architecture with an encoder-decoder structure that is widely used for image segmentation tasks due to its ability to capture contextual information and precise localization. The encoder extracts image features through convolutional layers, while the decoder progressively restores the spatial resolution of the image. By introducing skip connections, it merges the features from the encoder and decoder to improve segmentation accuracy. U-Net has been widely applied in remote sensing image segmentation due to its efficient and reliable performance. The model explored in this study also adopts an encoder-decoder structure, hence we chose this model for comparison.

ResNet: ResNet (Residual Network) [33] is a CNN designed to address the issues of gradient vanishing and representation bottlenecks in deep networks. By introducing Residual Blocks, the model allows the network to learn residual representations between input and output, optimizing deep networks. This structure enables ResNet to construct deeper network models while maintaining lower error rates. ResNet comes in multiple versions, and in this paper, we utilize ResNet-50.

DeepLabv3+: DeepLabv3+ is an advanced semantic segmentation model that integrates multi-scale features through an encoder-decoder structure, combined with dilated convolutions and ASPP modules to expand the receptive field and capture contextual information. The decoder module effectively merges high and low-resolution features to refine segmentation results, particularly focusing on object boundaries.

DeepWaterMapv2: DeepWaterMapv2 focuses on surface water mapping tasks and adopts the U-Net network structure. By iteratively applying convolutional and pooling operations, it effectively extracts key features from images for the precise identification and extraction of water body regions.

 doi:10.20944/preprints202406.0084.v1

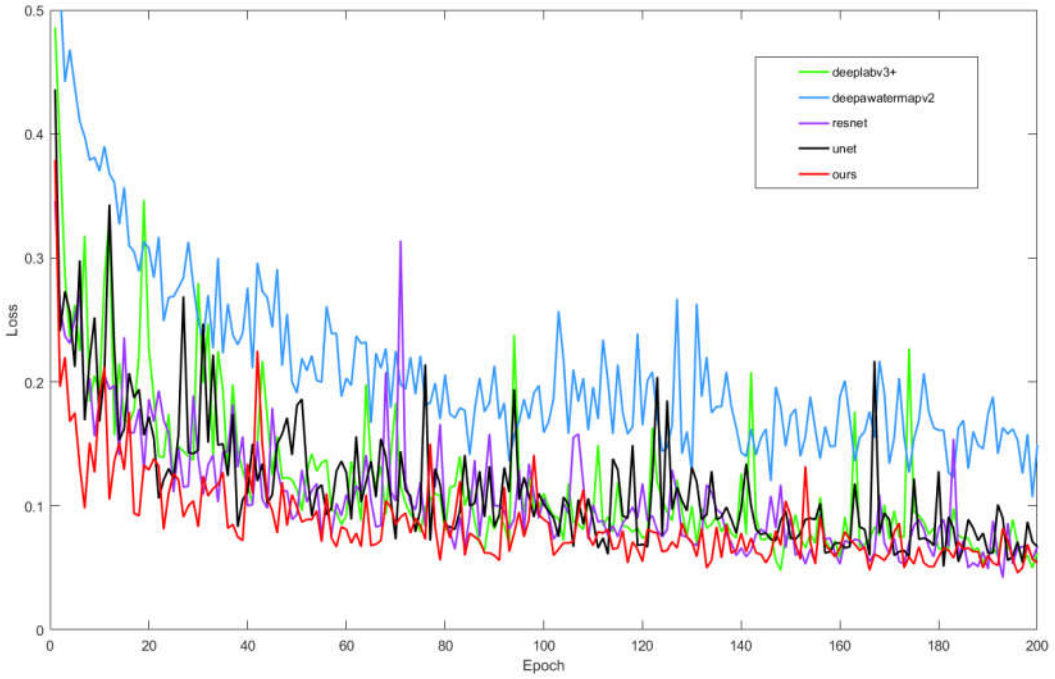


Figure 6. Visualization of the loss for U-Net, ResNet, DeepLabv3+, DeepWaterMapv2, and our proposed model after 200 epochs of training.

3.4. Analysis of Experimental Results

After completing model training, we used 20% of the dataset for validation. In the validation set, we selected six images to demonstrate the prediction results (as shown in Figure 7), including mountainous regions with large variations in terrain and areas with high cloud cover. To accurately evaluate the performance of SwinDefNet ,we compare it with other advanced water body detection methods, Table 2 presents the average values of four evaluation metrics for all images in the test set. Additionally, compared to other methods, our model achieved results of over 90%, with values of 97.89% for accuracy, 94.98% for precision, 90.05% for recall, and 92.33% for F1 score, respectively.

Through comparison and analysis of the four metrics, we found that our model achieved the highest accuracy among all methods. This indicates that our model has the best overall predictive ability for mapping surface water and can better extract water bodies from remote sensing images. Additionally, when compared to other methods, our model also achieved the highest recall, indicating that it can capture more water pixels during surface water mapping, with the lowest degree of omission in extracting water pixels.

Analyzing Table 2, we found that in the test, ResNet achieved an F1 score of 92.68%, which is 0.35% higher than our proposed model, indicating that ResNet has better overall performance. However, its recall was 88.96%, indicating that it would miss many water pixels during extraction, showing a certain gap in the fine extraction of water bodies. DeepWaterMapv2 achieved a precision as high as 99.07%, but with a recall of only 91.89%. This suggests that the DeepWaterMapv2 method sacrifices recall for precision in model prediction, indicating room for improvement in the fine extraction of water bodies.

Figure 7 compares the predicted images generated by different methods, while Figure 8 shows the true labels of the predicted images. Figures 9 and 10 present partial results of surface water mapping in mountainous and cloudy regions, with other result comparisons provided in the Appendix at the end of the document. We observe that our proposed model exhibits excellent noise suppression in the background (see Figure 9), outperforming ResNet and DeepLabv3+ methods. Moreover, satisfactory extraction results are achieved for winding small rivers and complex urban areas (see Figure 10). Compared to other methods, the boundaries are clearer, and in regions with

 doi:10.20944/preprints202406.0084.v1

higher cloud cover, our model has the lowest probability of misclassifying the background as water, compared to U-Net and ResNet methods.

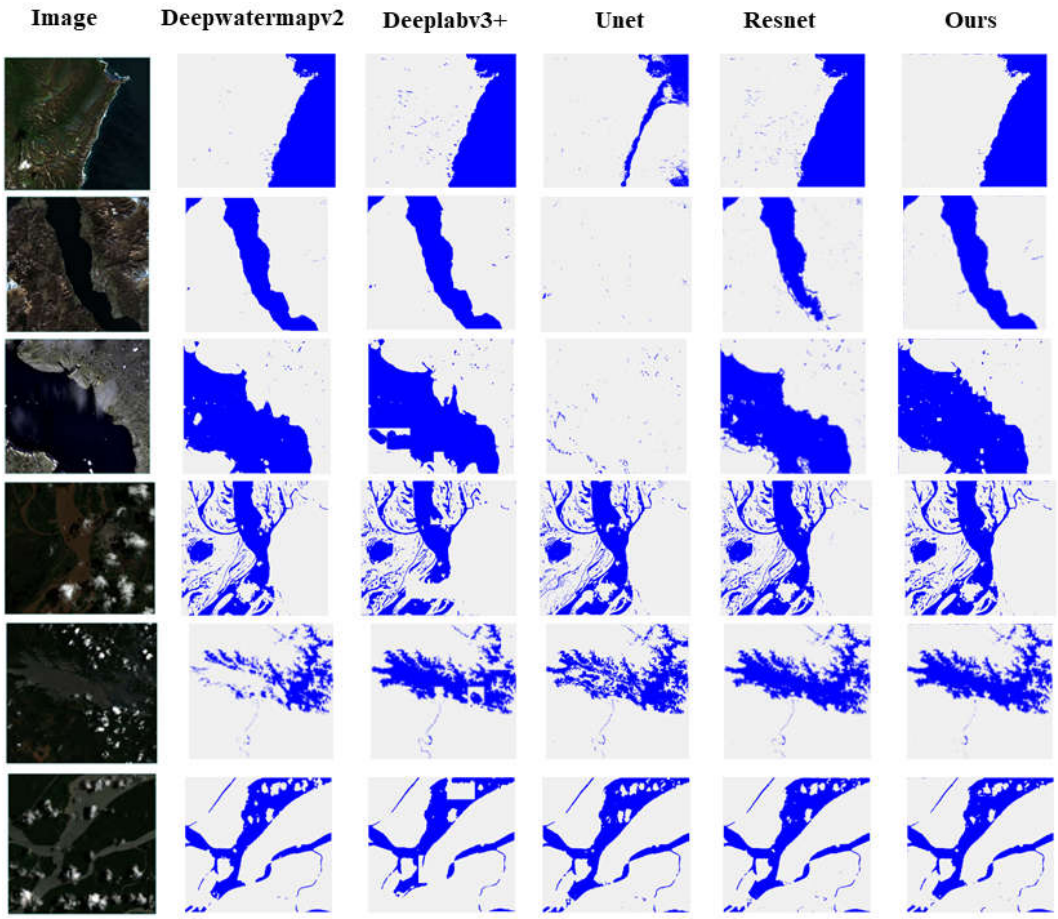


Figure 7. Comparison of predicted results using different methods on test images, with red circles indicating areas of comparison between the predicted images generated by each method.

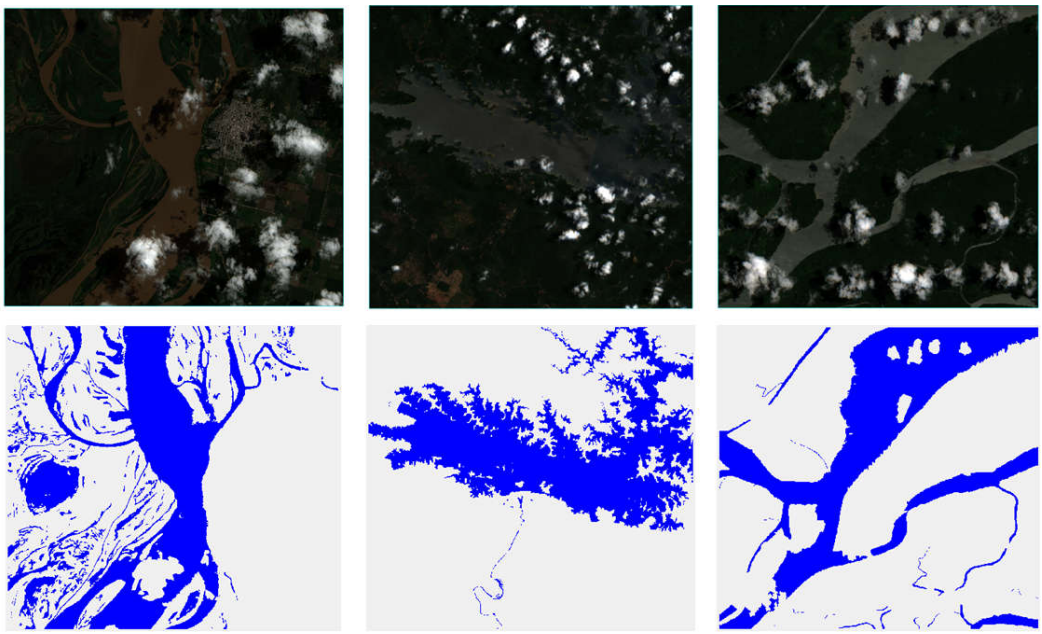


Figure 8. Labels data for illustrating the prediction results.

 doi:10.20944/preprints202406.0084.v1

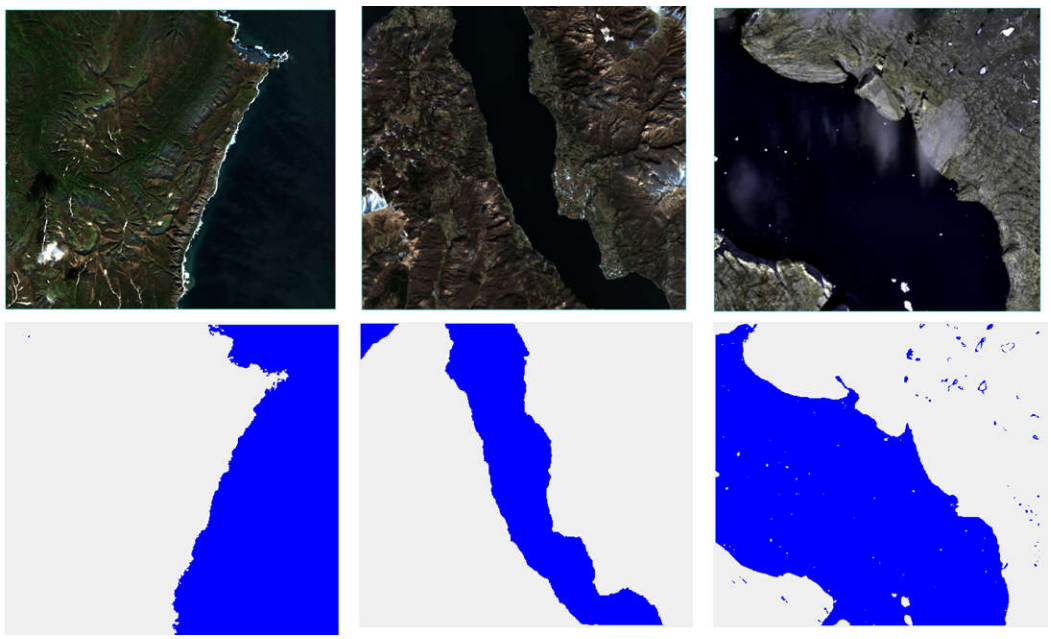


Figure 8. (continued).

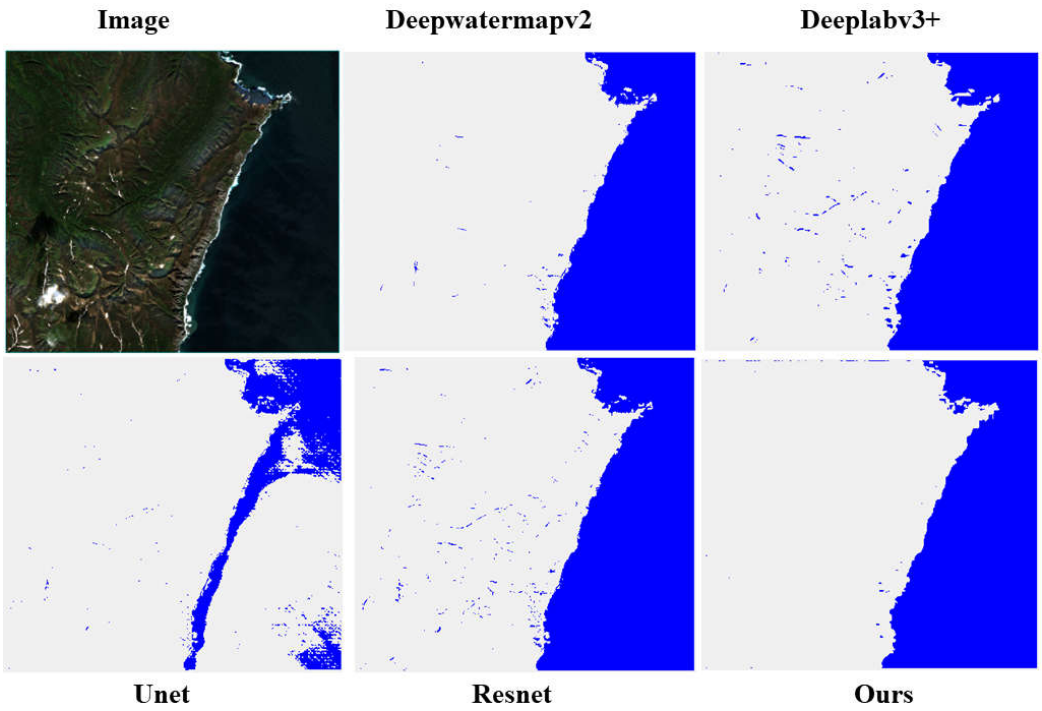


Figure 9. Results of water extraction in hilly regions are compared with existing approaches and our suggested model in mountainous regions.

 doi:10.20944/preprints202406.0084.v1

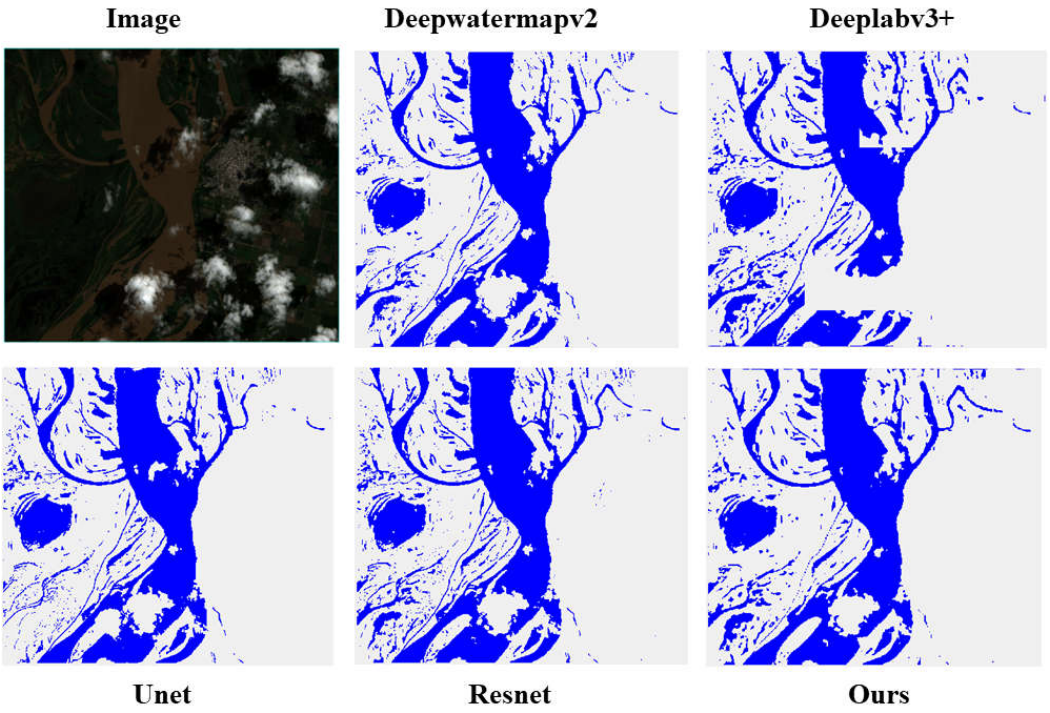


Figure 10. Comparison of our suggested model's and other approaches' water extraction outcomes in cloudy regions.

Table 2. Evaluation metrics of different methods. Average values for validation set images.

Method	Accuracy(%)	Precision(%)	Recall(%)	f1_score(%)
Ours	**97.89**	94.98	**90.05**	92.33
Unet	90.79	95.24	72.17	77.03
Resnet	97.65	97.26	88.96	**92.68**
Deeplabv3_plus	97.27	93.80	86.53	89.22
Deepwatermapv2	97.41	**99.07**	81.89	88.69

In order to further confirm the reliability and accuracy of our method, we conducted tests in various regions, comparing the performance of U-Net, ResNet, DeepLabv3+, and DeepWaterMapv2 in cloudy and mountainous areas. The results for accuracy and F1 score are shown in Figures 11 and 12, respectively. Table 3 provides detailed numerical values of the four evaluation metrics for the five methods in the three regions and the entire validation dataset.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2024 doi:10.20944/preprints202406.0084.v1

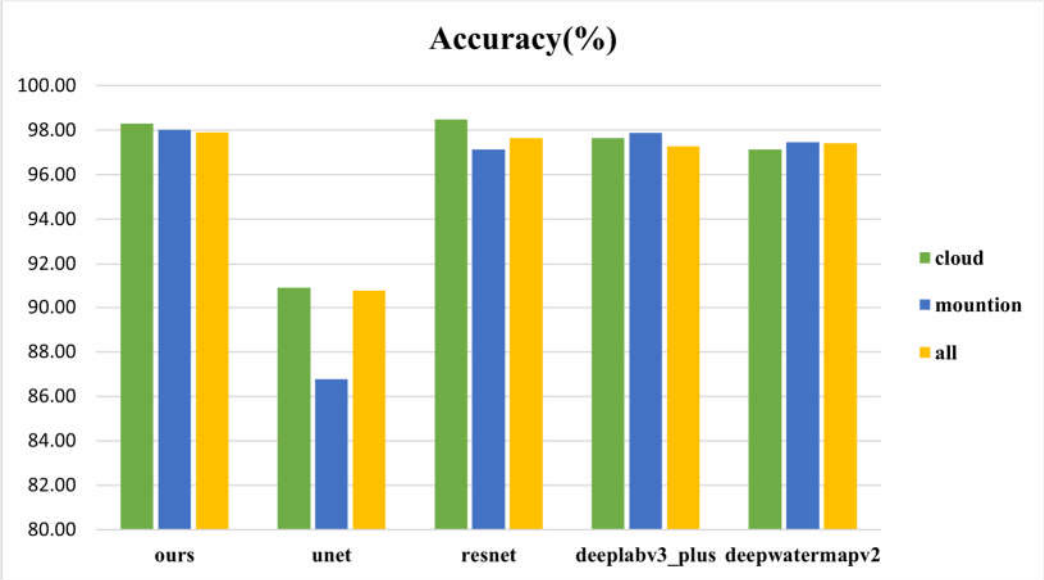


Figure 11. Comparison of the accuracy of several methods in various geographical areas.

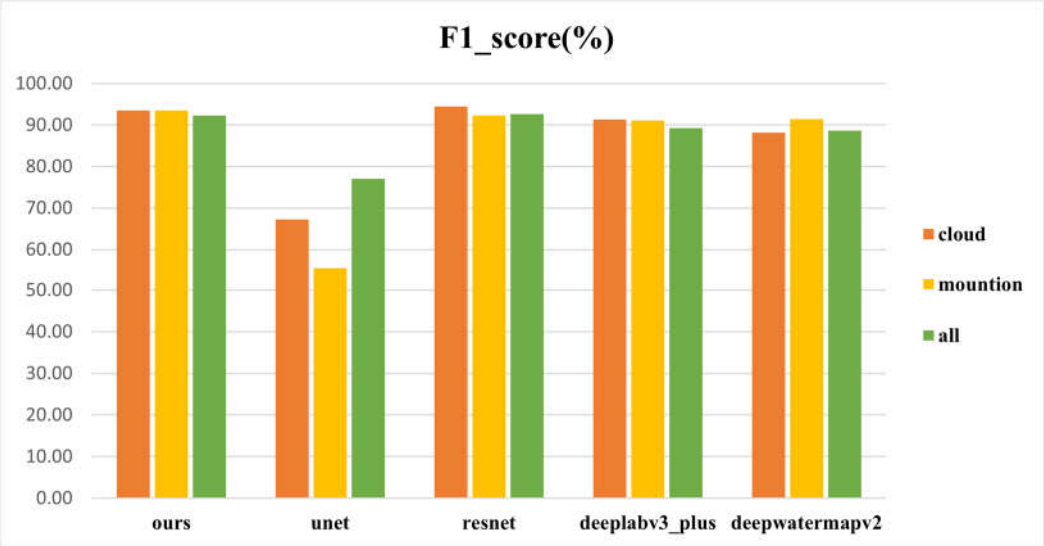


Figure 12. Comparison of the F1 score of several methods in various geographical areas.

Observing Table 3, we can see that our proposed model performs best in mountainous regions, with accuracy, recall, and F1 score values of 98.03%, 96.61%, and 93.52%, respectively. In the extraction of water bodies in mountainous regions, except for U-Net, all other methods achieved satisfactory results, with accuracy above 97%. Although U-Net's performance in mountainous regions is not ideal, it achieved good results in handling elongated and small water bodies in cloudy and urban areas.

Our model achieved an accuracy and F1 score of 98.30% and 93.46%, respectively, in cloudy areas. Compared to the best-performing ResNet, the differences are only 0.19% and 1%, respectively. Compared to other methods, our recall is highest at 92.22%, indicating that when predicting cloudy areas, we sacrificed some accuracy to reduce the omission of water pixels, enabling fine extraction of water bodies. This is also evidenced by our predicted images. Meanwhile, by comparing the labeled image with the predicted image, we observed that our model exhibits a higher similarity in extracting small water bodies' boundaries. This indicates that our experiment has certain advantages in refining the processing of small water bodies.

 doi:10.20944/preprints202406.0084.v1

Table 3. Accuracy, Precision, Recall, and F1 score results of water mapping in different regions for the five methods.

Method	Ours	Unet	Resnet	Deeplabv3+	Deepwatermapv2
Mountainous Area					
Accuracy	**98.03%**	86.77%	97.14%	97.88%	97.47%
Precision	95.99%	89.24%	97.97%	95.42%	**99.34%**
Recall	**91.61%**	49.18%	87.91%	89.18%	85.65%
f1_score	**93.52%**	55.56%	92.32%	91.08%	91.46%
Cloud Area					
Accuracy	98.30%	90.93%	**98.49%**	97.64%	97.14%
Precision	94.97%	93.66%	97.12%	93.81%	**98.61%**
Recall	**92.22%**	63.05%	92.12%	89.27%	84.85%
f1_score	93.46%	67.29%	**94.46%**	91.33%	88.19%

4. Discussion

In this part, we performed ablation tests to evaluate the significance of various components, along with discussing the impact of the Swin Transformer and Deformable Convolution on the network's performance. The model used the same training set, test set, and validation set in the ablation experiments.

Firstly, to validate the effectiveness of Swin Transformer in land water mapping tasks, we first compared the combined Swin Transformer and DeepLabV3+ model (see Model 1 in Table 4) with the DeepLabV3+ model using the original Xception backbone (see Model 3 in Table 4). Accuracy, Precision, Recall, and f1_score were used as evaluation metrics, the results showed that using Swin Transformer improved Accuracy and f1_score by 1.20% and 3.93% respectively. Precision and Recall also increased by 6.82% and 0.28% respectively. The analysis indicated that using Swin Transformer increased the accuracy of model predictions and improved its reliability. Additionally, there was a 6.82% increase in Precision which suggests that after incorporating Swin Transformer into the model, it became more accurate at predicting water pixels in instances where it is necessary for classification tasks while reducing classification errors caused by misjudgments of water pixels.

Table 4. The results of the ablation experiments.

Model	Swin Transform	Deformable Conv	Accuracy(%)	Precision(%)	Recall(%)	f1_score(%)
1	✓	✓	**97.89**	**94.98**	**90.05**	**92.33**
2	✓		97.67	94.96	89.04	91.73
3		✓	96.70	88.17	89.77	88.40

Secondly, to validate the effectiveness of Deformable Convolutional Networks (Deformable Conv), we compared a DeepLabV3+ model with Swin Transformer as its backbone network using regular 3×3 convolutions with another one using Deformable Conv (refer to Model 2 in Table 4). The results showed that after incorporating Deformable Convolutional Networks into our model's architecture led to improvements in Accuracy (+0.22%), Precision (+0.03%), Recall (+1.01%), and f1_score (+0.6%). This indicates an enhancement in both accuracy and reliability after integrating Deformable Conv into our model's architecture while also improving its ability to identify water pixels through an increase of +1 .01 % recall rate.

Through experimental analysis, we can confirm that both Swin Transformer and Deformable Conv are beneficial for water extraction tasks. The combined use of these two components has also enhanced the precision of land water mapping to achieve fine-grainedextraction of water bodies.

5. Conclusions

To achieve fine-grained extraction of water bodies in remote sensing images, this study explored a novel combination of deep learning networks, employing an encoder-decoder structure for

 doi:10.20944/preprints202406.0084.v1

extracting multiscale information. The model was developed in a two-step process: (1) the most advanced image classification and segmentation model is integrated into the encoder-decoder network structure; and (2) utilizing deformable convolution to adaptively adjust the receptive field for fine-grained extraction of water body features. More precisely, the model employed the Swin Transformer as the encoder to capture multiscale information, and leveraged deformable convolution to dynamically adjust the receptive field for the precise extraction of water body characteristics.

Additionally, to assess the efficacy of our proposed method, we trained the novel integrated model on the ESWKB dataset and subsequently utilized it for mapping surface water in diverse geographical regions. Comparing with various existing methods, our proposed network combination outperformed them, especially in the fine-grained water body segmentation. This provides strong assistance for accurate water extraction and offers a strategy for fine-grained water body segmentation in remote sensing images.

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

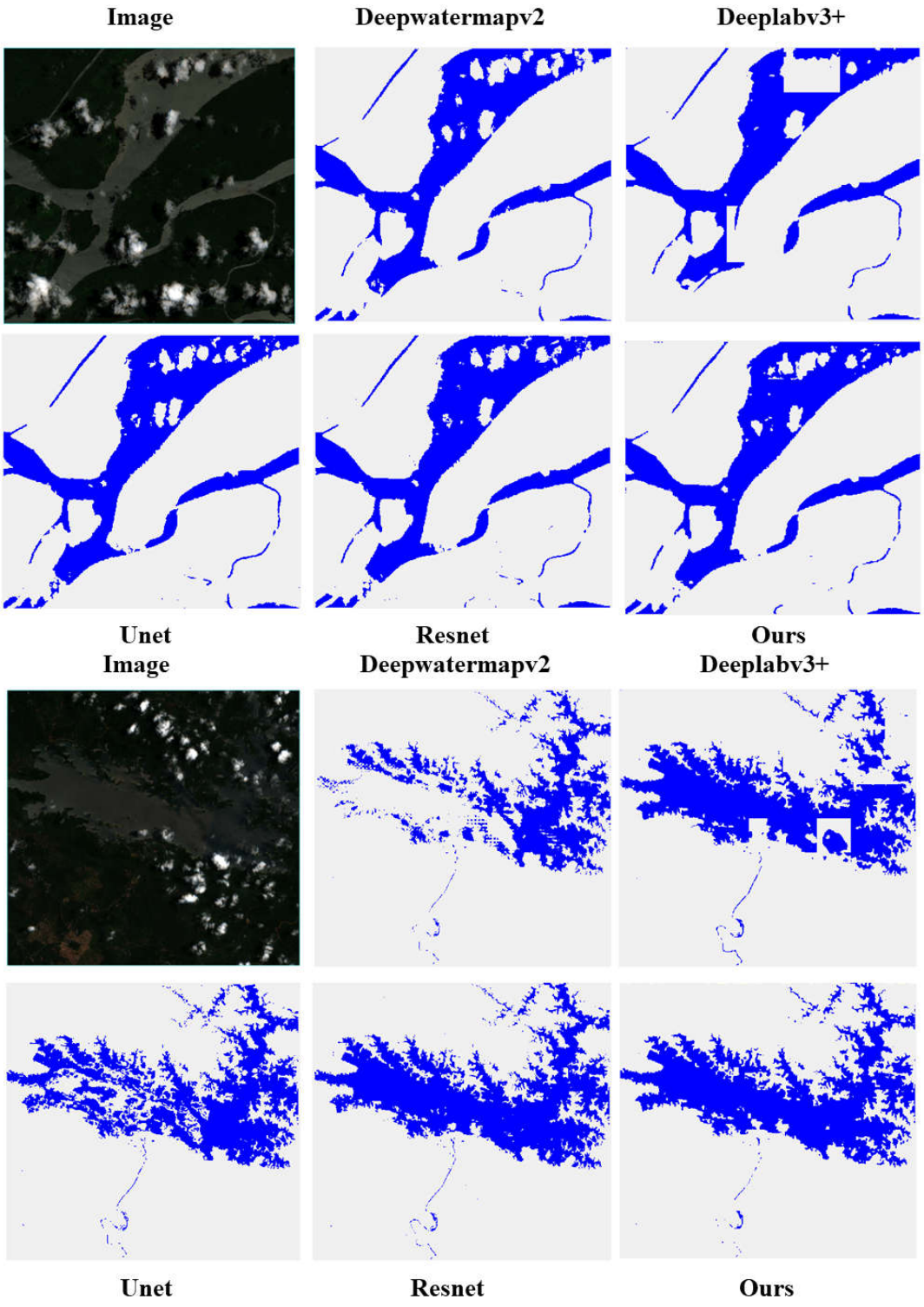
Funding: This research was funded by National Key Research and Development Programs, grant number 2022YFC3004405; National Natural Science Foundation of China, grant number: 42061073; Natural Science Foundation of Guizhou Province, grant number: [2020]1Z056.

Acknowledgments: We would like to thank the reviewers for their valuable comments and suggestions. In addition, the authors would like to thank X Luo for providing the Earth surface water knowledge base.

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

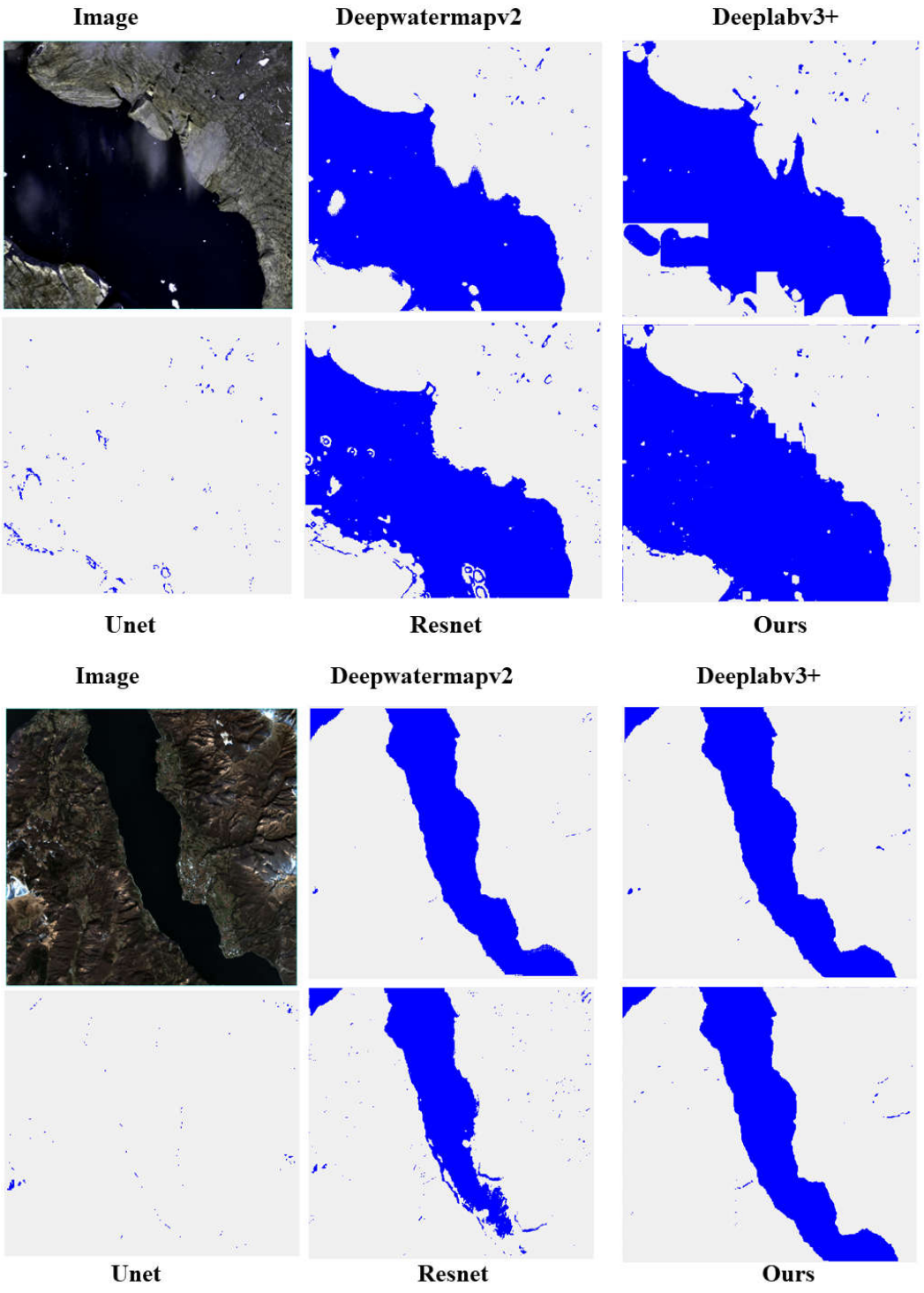
 doi:10.20944/preprints202406.0084.v1

Appendix A: Cloud Area



 doi:10.20944/preprints202406.0084.v1

Appendix B: Mountainous Area



References

1. Parajuli J, Fernandez-Beltran R, Kang J, et al. Attentional Dense Convolutional Neural Network for Water Body Extraction From Sentinel-2 Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15, 6804-6816.
2. Luo X, Hu Z, Liu L. Investigating the seasonal dynamics of surface water over the Qinghai–Tibet Plateau using Sentinel-1 imagery and a novel gated multiscale ConvNet. International Journal of Digital Earth, 2023, 16(1), 1372-1394.
3. Li H, Zech J, Ludwig C, et al. Automatic mapping of national surface water with OpenStreetMap and Sentinel-2 MSI data using deep learning. International Journal of Applied Earth Observation and Geoinformation, 2021, 104, 102571.

 doi:10.20944/preprints202406.0084.v1

4. Jiang C, Zhang H, Wang C, et al. Water Surface Mapping from Sentinel-1 Imagery Based on Attention-UNet3+: A Case Study of Poyang Lake Region. Remote Sensing, 2022, 14(19),4708.
5. Zhao B, Sui H, Liu J. Siam-DWENet: Flood inundation detection for SAR imagery using a cross-task transfer siamese network. International Journal of Applied Earth Observation and Geoinformation, 2023, 116, 103132.
6. Tong X, Luo X, Liu S. An approach for flood monitoring by the combined use of Landsat 8 optical imagery and COSMO-SkyMed radar imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 136, 144-153.
7. McFeeters S K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 1996, 17, 1425-1432.
8. Xu H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International Journal of Remote Sensing, 2006, 27, 3025-3033.
9. Wang X, Xie S, Zhang X, et al. A robust Multi-Band Water Index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery. International Journal of Applied Earth Observation and Geoinformation, 2018, 68, 73-91.
10. Li L, Su H, Du Q, et al. A novel surface water index using local background information for long term and large-scale Landsat images. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 172, 59-78.
11. Cai Y, Shi Q, Liu X. Spatiotemporal Mapping of Surface Water Using Landsat Images and Spectral Mixture Analysis on Google Earth Engine. Journal of Remote Sensing, 2024, 4 ,117.
12. Sun Q, Li J. A method for extracting small water bodies based on DEM and remote sensing images. Scientific reports, 2024, 14, 760.
13. Yan X, Song J, Liu Y, et al. A Transformer-based method to reduce cloud shadow interference in automatic lake water surface extraction from Sentinel-2 imagery. Journal of Hydrology, 2023, 620, 129561.
14. Isikdogan F, Bovik A C, Passalacqua P. Surface Water Mapping by Deep Learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017,10(11), 4909-4918.
15. Chen F, Chen X, de Voorde T, et al. Open water detection in urban environments using high spatial resolution remote sensing imagery. Remote Sensing of Environment, 2020, 242,111706.
16. Kang J, Guan H, Peng D, et al. Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images. International Journal of Applied Earth Observation and Geoinformation, 2021, 103, 102499.
17. Luo X, Tong X, Hu Z. An applicable and automatic method for earth surface water mapping based on multispectral images. International Journal of Applied Earth Observation and Geoinformation,2021, 103, 102472.
18. Li Z, Zhang X, Xiao P. Spectral index-driven FCN model training for water extraction from multispectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 192, 344-360.
19. Zhang X, Li J, Hua Z. MRSE-Net: Multiscale Residuals and SE-Attention Network for Water Body Segmentation From Satellite Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2022, 15, 5049-5064.
20. Yu Y, Huang L, Lu W, et al. WaterHRNet: A multibranch hierarchical attentive network for water body extraction with remote sensing images. International Journal of Applied Earth Observation and Geoinformation, 2022, 115, 103103.
21. Lyu X, Fang Y, Tong B, et al. Multiscale Normalization Attention Network for Water Body Extraction from Remote Sensing Imagery. Remote Sensing, 2020, 14, 4983.
22. Kang J, Guan H, Ma L, et al. WaterFormer: A coupled transformer and CNN network for waterbody detection in optical remotely-sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 206, 222-241.
23. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. IEEE International Conference on Computer Vision, 2021, 9992-10002.
24. Dai J, Qi H, Xiong Y, et al. Deformable Convolutional Networks. IEEE International Conference on Computer Vision, 2017, 764-773.
25. Seale C, Redfern T, Chatfield P, et al. Coastline detection in satellite imagery: A deep learning approach on new benchmark data. Remote Sensing of Environment, 2022, 278, 113044.
26. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015.
27. Chen L, Zhu Y, Papandreou G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. European Conference on Computer Vision, 2018.
28. Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network. IEEE Conference on Computer Vision and Pattern Recognition, 2017, 6230-6239.
29. Pan H, Chen H, Hong Z, et al. A Novel Boundary Enhancement Network for Surface Water Mapping Based on Sentinel-2 MSI Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16, 9207-9222.

 doi:10.20944/preprints202406.0084.v1

30. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2015, 770-778.