frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 04 December 2023
DOI 10.3389/frsen.2023.1283615

OPEN ACCESS

EDITED BY
Xiang Zhang,
China University of Geosciences Wuhan, China

REVIEWED BY
Weitao Chen,
China University of Geosciences Wuhan, China
Xiaolei W.,
Zhengzhou University, China

*CORRESPONDENCE
Xiaomei Yi,
yxm@zafu.edu.cn
Hao Liang,
lhao@zafu.edu.cn

†These authors have contributed equally to this work.

RECEIVED 26 August 2023
ACCEPTED 01 November 2023
PUBLISHED 04 December 2023

CITATION
Wu P, Fu J, Yi X, Wang G, Mo L, Maponde BT, Liang H, Tao C, Ge W, Jiang T and Ren Z (2023), Research on water extraction from high resolution remote sensing images based on deep learning.
Front. Remote Sens. 4:1283615.
doi: 10.3389/frsen.2023.1283615

Research on water extraction from high resolution remote sensing images based on deep learning

Peng Wu[1†], Junjie Fu[1†], Xiaomei Yi[1,2*], Guoying Wang[1], Lufeng Mo[1], Brian Tapiwanashe Maponde[1], Hao Liang[1*], Chunling Tao[1], WenYing Ge[1], TengTeng Jiang[1] and Zhen Ren[1]

[1]College of Mathematics and Computer Science, Zhejiang A and F University, Hangzhou, China, [2]School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China

Introduction: Monitoring surface water through the extraction of water bodies from high-resolution remote sensing images is of significant importance. With the advancements in deep learning, deep neural networks have been increasingly applied to high-resolution remote sensing image segmentation. However, conventional convolutional models face challenges in water body extraction, including issues like unclear water boundaries and a high number of training parameters.

Methods: In this study, we employed the DeeplabV3+ network for water body extraction in high-resolution remote sensing images. However, the traditional DeeplabV3+ network exhibited limitations in segmentation accuracy for high-resolution remote sensing images and incurred high training costs due to a large number of parameters. To address these issues, we made several improvements to the traditional DeeplabV3+ network: Replaced the backbone network with MobileNetV2. Added a Channel Attention (CA) module to the MobileNetV2 feature extraction network. Introduced an Atrous Spatial Pyramid Pooling (ASPP) module. Implemented Focal loss for balanced loss computation.

Results: Our proposed method yielded significant enhancements. It not only improved the segmentation accuracy of water bodies in high-resolution remote sensing images but also effectively reduced the number of network parameters and training time. Experimental results on the Water dataset demonstrated superior performance compared to other networks: Outperformed the U-Net network by 3.06% in terms of mean Intersection over Union (mIoU). Outperformed the MACU-Net network by 1.03%. Outperformed the traditional DeeplabV3+ network by 2.05%. The proposed method surpassed not only the traditional DeeplabV3+ but also U-Net, PSP-Net, and MACU-Net networks.

Discussion: These results highlight the effectiveness of our modified DeeplabV3+ network with MobileNetV2 backbone, CA module, ASPP module, and Focal loss for water body extraction in high-resolution remote sensing images. The

reduction in training time and parameters makes our approach a promising solution for accurate and efficient water body segmentation in remote sensing applications.

KEYWORDS

remote sensing image, semantic segmentation, deep learning, water extraction, DeepLabV3+

1 Introduction

The task of extracting water bodies from remote sensing images involves the segmentation of surface water from these images. Surface water is a general term for static and dynamic water on the land surface, including river channels, lakes, reservoirs and other water bodies. It is a crucial component of Earth's ecosystem, constituting only 1.75% of the global total water storage. Surface water plays a significant role in biodiversity, climate change, and the global water cycle. However, due to the continuous development of human society and industry, surface water pollution and depletion have become increasingly severe. Consequently, it is vital to study the extraction and accurate delineation of surface water bodies to foster sustainable human development, environmental protection, and urban planning. Despite the paramount importance of surface water to Earth's ecosystem and human survival, our research on its trends and area changes has been limited.

This limitation arises from our heavy reliance on manual investigation and annotation to understand surface water. Although this approach yields highly accurate results, it is time-consuming, lacks real-time data availability, and incurs high labor costs. It fails to meet the growing demand for water body extraction from various regions. In recent years, the advent of remote sensing satellite technology has transformed the field of remote sensing research. Satellite systems such as Landsat and Sentinel-I have been widely deployed, providing low-cost and reliable remote sensing images. Equipped with high-resolution microwave sensors unaffected by day-night variations and capable of penetrating thick clouds, these satellites offer imaging capabilities across diverse terrains.

During the initial stages of remote sensing image water body extraction research, various algorithms were developed to leverage the disparities in spectral reflectance between land and water. These included techniques like spectral unmixing, single-band thresholding, and the spectral moisture index method. The variance in spectral reflectance arises because water predominantly absorbs energy at near-infrared and mid-infrared wavelengths, while vegetation, soil, and impermeable surfaces exhibit higher reflectance at these wavelengths. The spectral moisture index, which accounts for the correlation between different bands, emerged as a widely utilized method due to its high accuracy and cost-effectiveness (Xie et al., 2014).

Numerous other water indices were proposed in the early stages as well. In 2011, researchers suggested combining the NDVI-NDWI difference with SLOPE and near-infrared bands (Lu et al., 2011). This combination proved more effective than using the NDVI or NDWI indices alone, enhancing the contrast between water bodies and other surface features. Additionally, scholars introduced the Automatic Water Extraction Index (AWEI) to enhance the classification accuracy of shadows and dark surfaces that other methods often struggle to identify correctly (Feyisa et al., 2014). However, these methods generally encountered challenges in complex scenes like shadows and mountainous areas, necessitating manual adjustment of suitable thresholds. Determining the optimal threshold for achieving the highest possible accuracy proved to be a daunting task, as it varied with the image acquisition time and location.

Several commonly employed water body extraction methods have also been established. Some researchers employed radial basis functions (RBF) to construct machine learning support vector machine (SVM) classification models, yielding promising outcomes in water body segmentation (Li et al., 2013). Moreover, the concept of object-based classification was proposed. Huang et al. (Huang et al., 2015) developed a two-level machine learning framework that employed geometric and texture features to identify water bodies at the object level in high-resolution remote sensing images of urban areas. Traditional water body extraction methods, on the whole, provide effective water body information but are susceptible to the influence of complex environments and involve significant workload.

In recent years, the field of image segmentation has witnessed significant advancements due to the rapid progress of deep learning techniques. Deep learning networks, capable of achieving pixel-level classification, have found widespread application in semantic segmentation. Zhao et al. (Zhao et al., 2017) enhanced the FCN (Long et al., 2017) model by incorporating a Pyramid Pooling Module, resulting in the proposed PSP-Net network model. This novel approach effectively aggregates contextual information from various regions, thereby improving the model's ability to capture global information. Zhou et al. (2023) put forward EG-UNet model to solve the problems of feature loss and limited interpretation accuracy in mining land cover classification. EG-UNet improves the feature representation ability by extracting boundary information using Sobel operator and capturing remote features using graph convolutional network. Wenkuana and Shicai (2023) proposed an improved segmentation model based on deeplabv3. By combining image level adjustment and attention mechanism, they solved the problems of lower segmentation accuracy and difficult segmentation of dense fog in fog-day image, and achieved better results. Li et al. (2022), in 2021, addressed the issue of insufficient feature utilization in the U-Net (Ronneberger et al., 2015) model by introducing asymmetric convolution to enhance the feature representation and extraction capabilities of the convolutional layer. The resultant MACU-Net network outperformed the U-Net network when tested on WHDLD and GID datasets.

Dai et al. (2020) proposed an enhanced water body segmentation network based on the bilateral segmentation network (BiSeNet) and utilized the loss function of the edge region to improve the network's segmentation capability. Wang et al. (2022) introduced an intelligent water body extraction method

known as SADA Net, designed specifically for high-resolution remote sensing images. This network framework integrates three key components: Shape Feature Optimization (SFO), Hollow Space Pyramid Pooling, and Dual Attention Module. Yang et al. (2020) presented a model based on MASK R-CNN for automatic detection and segmentation of water bodies in remote sensing images, eliminating the need for manual feature extraction. Zhang et al. (2022) proposed MRSE Net, a multi-scale residual network structure for water segmentation. Similar to U-Net in structure, MRSE Net comprises an encoder-decoder and hop connections, enabling it to capture context information of different scales.

Dirscherl et al. (2021) proposed the use of an improved U-Net architecture for semantic segmentation of lakes in Sentinel-1 images, which yielded favorable segmentation results. The enhanced U-Net network consisted of four downsampling blocks in the encoder and four upsampling blocks with convolution in the decoder, with both blocks implemented as ResNet blocks. Chen et al. (2015) introduced the DeepLab network, utilizing hole convolution based on FCN to mitigate pooled information loss. They further incorporated a conditional random field (CRF) module into the feature extraction network output. Building upon DeepLab, Chen L. C. et al. (2017) developed the DeeplabV2 network, enhancing the receptive field without increasing the number of parameters through the introduction of the Atrous Spatial Pyramid Pooling (ASPP) module, thus improving network performance.

Continuing the evolution, Chen LC. et al. (2017) proposed the DeeplabV3 network, which eliminated the conditional random field while optimizing ASPP modules and leveraging multiple empty convolution cores, resulting in superior segmentation outcomes compared to DeeplabV2. Recognizing the limited inclusion of shallow features in the DeeplabV3 network, Chen et al. (2018) introduced the DeeplabV3+ network. Considered a breakthrough in semantic segmentation, the DeeplabV3+ network integrated shallow features with deep features using a classic encoding and decoding structure based on DeeplabV3, significantly improving segmentation accuracy. While DeeplabV3+ has demonstrated a leading edge on multiple publicly available data sets in the field of semantic segmentation, there are still some problems in high-resolution remote sensing image segmentation. One of the problems is the lack of accuracy of semantic information. Since remote sensing images usually have higher resolution and more complex scenes, it makes it difficult for networks to capture details and boundaries. This can lead to a lack of accuracy and detail in segmentation results due to ambiguous semantic information. Another problem is the computational complexity of the network. Due to the large volume of high-resolution remote sensing images, this can cause the training and inference process to become more time-consuming and may require more powerful hardware support.

This article acknowledges the challenges faced by the DeeplabV3+ network when dealing with the complex features in high-resolution remote sensing image segmentation of water bodies. To address these challenges, the proposed approach utilizes the MobilenetV2 network as the feature extraction network, incorporating an attention mechanism module into the network structure. Focal loss balancing is introduced, and the algorithm is tested and validated using high-resolution remote sensing image datasets. Comparative analysis is conducted with traditional DeeplabV3+ networks to evaluate the performance of the proposed method.

2 Related research

2.1 Deeplabv3+network model

The DeepLabV3+ algorithm has gained significant popularity as a network model structure in the realm of image semantic segmentation. Since its inception, it has been extensively employed for achieving high-precision image segmentation. In recent years, DeepLabV3+ networks have found wide application in remote sensing image segmentation tasks (Li et al., 2019; da Cruz et al., 2022; Du et al., 2021). These networks excel in leveraging multi-scale contextual information and employ spatial information reconstruction techniques to delineate object boundaries. The ASPP (Atrous Spatial Pyramid Pooling) module within the network structure takes an input feature map x and produces an output feature map y, as expressed by Equation:

$$\boldsymbol{y} = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{K}} \boldsymbol{x}[\boldsymbol{i} + \boldsymbol{r} \times \boldsymbol{k}]\boldsymbol{w}[\boldsymbol{k}] \tag{1}$$

Here, i denotes the input signal, $w[k]$ represents the filter value, r corresponds to the expansion rate, and K signifies the convolution length of the cavity.

The DeeplabV3+ network introduces numerous dilated convolutions within the encoder module, enabling the network to expand its receptive field without sacrificing valuable information. The calculation of the receptive field for dilated convolutions aligns with that of standard convolution kernels, where the value of K, denoting the receptive field of the dilated convolution, is determined by Equation below:

$$\boldsymbol{K} = \boldsymbol{k} + (\boldsymbol{k} - \boldsymbol{1})(\boldsymbol{r} - \boldsymbol{1}) \tag{2}$$

In the formula, k is the size of the original convolutional kernel; r is the size of the actual convolution kernel for empty convolutions.

DeepLabV3+ is based on the structure of DeepLabV3, adding a simple and efficient decoding module to refine feature information and improve segmentation performance. The encoder is used for feature processing. Firstly, an improved Xception model is used, using Modified Aligned Xception as the backbone network. Through the Modified Aligned Xception feature extraction network, the depth of different channels in the network can be separated by convolution operations. Finally, two effective feature layers are generated, namely, shallow features and deep features. Afterwards, utilizing the spatial pyramid pooling module 1×1 Parallel convolution with 3 void rates of 6,12,18, respectively 3×3 after processing deep features through hollow convolution and global average pooling operations 1×1 convolutional channel compression results in multi-scale features with 256 channel bits. In the decoder section, the shallow features extracted from the input layer of the backbone network are first utilized 1×1 Convolutional dimensionality reduction, followed by fusion with deep features obtained through encoder upsampling, and then utilizing several 3×3 The spatial information in the feature map is restored through convolution, and the final prediction result is obtained through the Softmax function. Figure 1 shows the network structure of DeeplabV3+.

2.2 Modified Aligned Xception network model

The Modified Aligned Xception network represents an enhanced iteration of the original Xception network, which was introduced by the

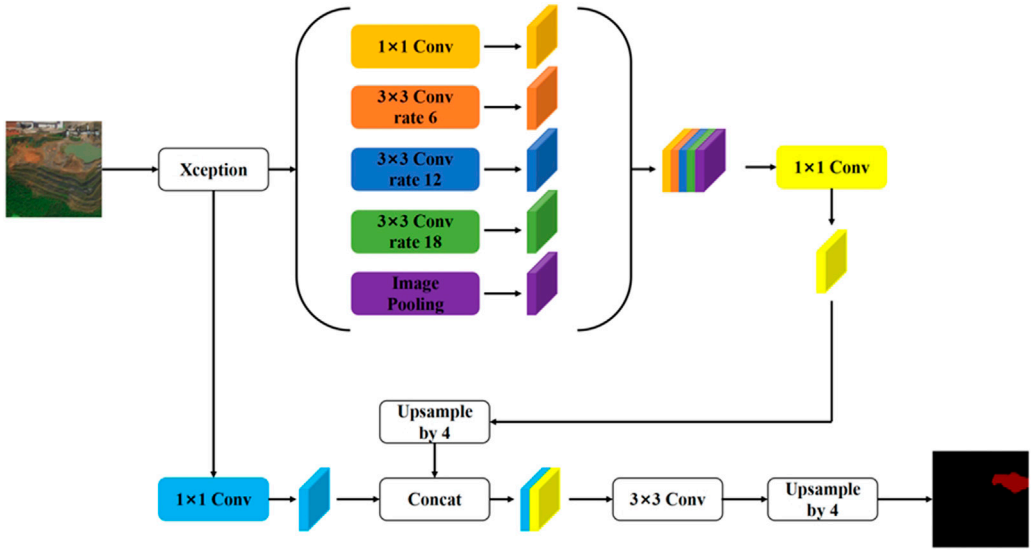

FIGURE 1
DeeplabV3+Structure diagram. Maps Data:Google, ©2023 CNES / Airbus, Maxar Technologies. Reproduced with permission.

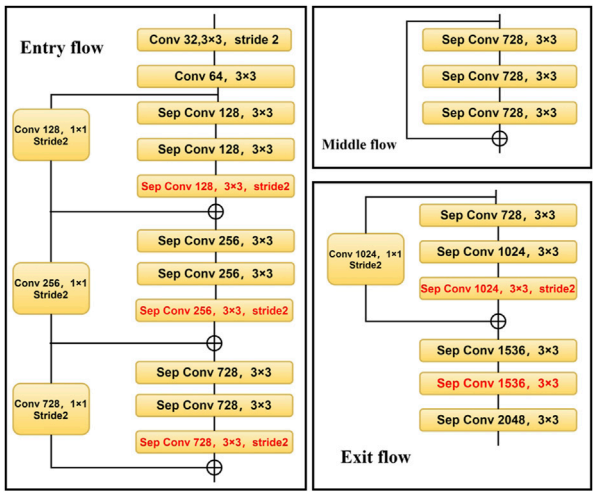

FIGURE 2
Structure diagram of Modified Aligned Xception.

Google team in 2016 as a network model. The Xception network structure comprises three distinct flows: the entry flow, middle flow, and exit flow, collectively incorporating a total of 36 convolutional layers. Notably, the Xception network introduces deep separable convolution, which involves performing three separate convolutions on each channel using 3 × 3 filters, followed by a 1 × 1 convolution operation, and ultimately merging the results. This method introduces more nonlinear transformation into the network and improves the capability of feature representation.

In addition, the Xception network structure incorporates the concept of residual learning, which significantly contributes to the model's effectiveness. Assuming the input is denoted as X, and the desired output is represented by Y, traditional linear network structures aim to learn the mapping F(X) = Y. In contrast, the residual structure in the Xception network directly transmits the input X to the output as the initial result, thereby shifting the learning objective from F(X) = Y to F(X) = Y - X, which represents the difference between the output and the input.

Figure 2 provides a visual representation of the network structure diagram for the Modified Aligned Xception. Upon observation, it becomes apparent that the Entry flow remains fixed, with the addition of multiple Middle flows. Furthermore, the Max pooling operation is substituted with depthwise separable convolutions. Additionally, in each 3 × 3 Normalization and ReLU activation functions were added after the convolution operation.

3 Materials and methods

3.1 Experimental dataset

The first step to build a water dataset is to collect high-resolution remote sensing images of different areas of Hangzhou through Google remote sensing images. Subsequently, these remote sensing images are cropped to suitable dimensions, followed by a filtering process to refine the selection. Finally, manual annotation is performed to identify and label the water bodies present in the images. The resulting Water dataset comprises 10,000 remote sensing images, each measuring 512 sheets × 512 pixels, with a spatial resolution of 0.3 m. This comprehensive dataset encompasses remote sensing images depicting rivers, reservoirs, and lakes. For visual reference, Figure 3 displays an example image from the Water dataset, along with its corresponding label image indicating the water bodies.

3.2 Improved DeepLabV3+ network model

The traditional DeeplabV3+network model has a large number of parameters and complex calculations. Therefore, in the

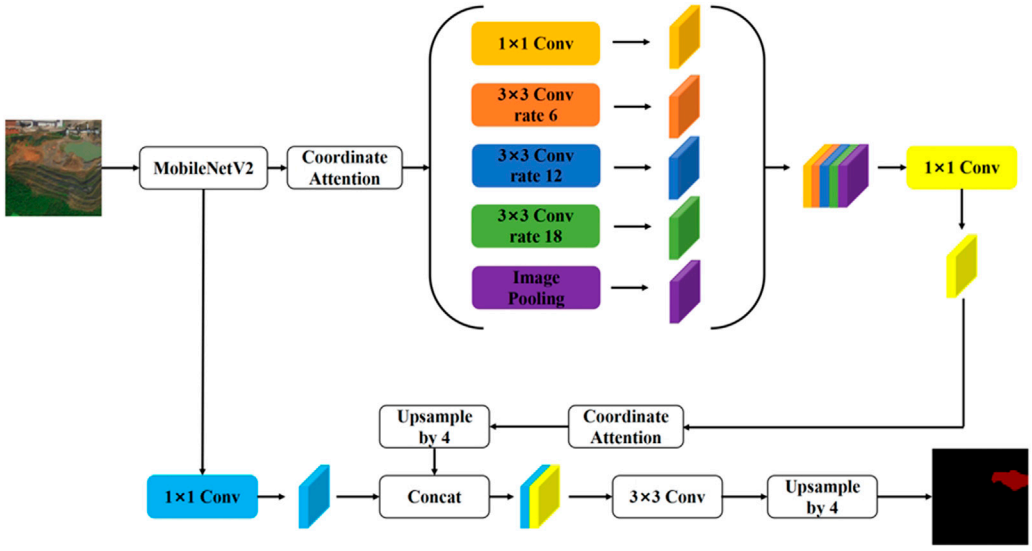

FIGURE 3
Water data set image and label map. Maps Data:Google, ©2023 CNES / Airbus, Maxar Technologies. Reproduced with permission.

implementation process of DeepLabV3+algorithm, this article proposes to replace the Modified Aligned Xception network model with a lighter weight MobileNetV2 network to reduce the number of model parameters. At the same time, the CA (Coordinated Attention) module is added after the ASPP module in the MobileNetV2 feature extraction network to improve the segmentation accuracy of the model. Finally, optimize the loss function and introduce Focal loss to equalize the loss. This article aims to achieve a faster and stronger encoding decoding network by making three improvements to the DeeplabV3+network. Figure 4 shows the improved DeeplabV3+network structure diagram.

3.2.1 Optimize feature extraction module

At the coding layer, the original network used for feature extraction has been replaced with the more lightweight MobileNetV2 network, which was introduced by the Google team in 2018. The MobileNetV1 network, a classic lightweight CNN neural network, was initially proposed by Google in 2017 with a focus on embedded devices. The MobileNetV2 network is an improvement over MobileNetV1, maintaining its lightweight nature. The MobileNetV2 network incorporates a linear bottleneck structure and a reverse residual structure.

During feature extraction operations, neural networks extract valuable target-related information, which can be embedded in low-dimensional subspaces. However, when mapping from low-dimensional to high-dimensional spaces and then back to low-dimensional spaces after applying the ReLU activation function, some features inevitably get lost. If the final mapped dimension is relatively high, the loss of information during the transformation back to low-dimensional space is relatively small. Conversely, if the mapped dimension is relatively low, a significant amount of information is lost. To address this, the linear bottleneck layer replaces the ReLU activation function of the penultimate layer with a linear function, thereby reducing the loss of useful information within the network.

The reverse residual structure of MobileNetV2 network application consists of three parts. Firstly, 1×1 convolution is used to increase the dimensionality of input features, followed by feature extraction using 3×3 depth separable convolution, and then 1×1 convolution is used to reduce the dimensionality. The specific network structure is shown in Table 1.

3.2.2 Add CA module

The attention mechanism is a computational approach that calculates the weighted sum of different weights assigned to feature vectors. During the model training process, varying weights are assigned to different regions of the input image, reflecting the varying importance of feature information. By assigning different weights, attention mechanisms enable the model to focus on significant information and reduce interference. Thus, incorporating attention mechanisms with convolutional networks can effectively enhance the performance of image segmentation tasks. Two commonly used attention mechanisms are the channel attention mechanism (Zhang et al., 2018) and the spatial attention mechanism (Chu et al., 2017). By introducing the attention mechanism in the encoder end, features can be extracted from the input more efficiently and the attention representation of the input can be obtained. Adding the attention mechanism to the input side of the decoder can solve the limitation that the output of the encoder is only a certain length tensor and can not store a lot of information, so as to realize the full use of information. In this article, coordinated attention (CA) module is added after the ASPP module in the MobileNetV2 feature extraction network.

The CA module employs two one-dimensional global pooling operations to aggregate vertical and horizontal input features into two separate directional perceptual feature maps. These feature maps, embedded with directional-specific information, are then encoded into attention maps capturing long-distance dependency relationships within the input feature map's spatial direction. This encoding process ensures that location information is preserved within the generated

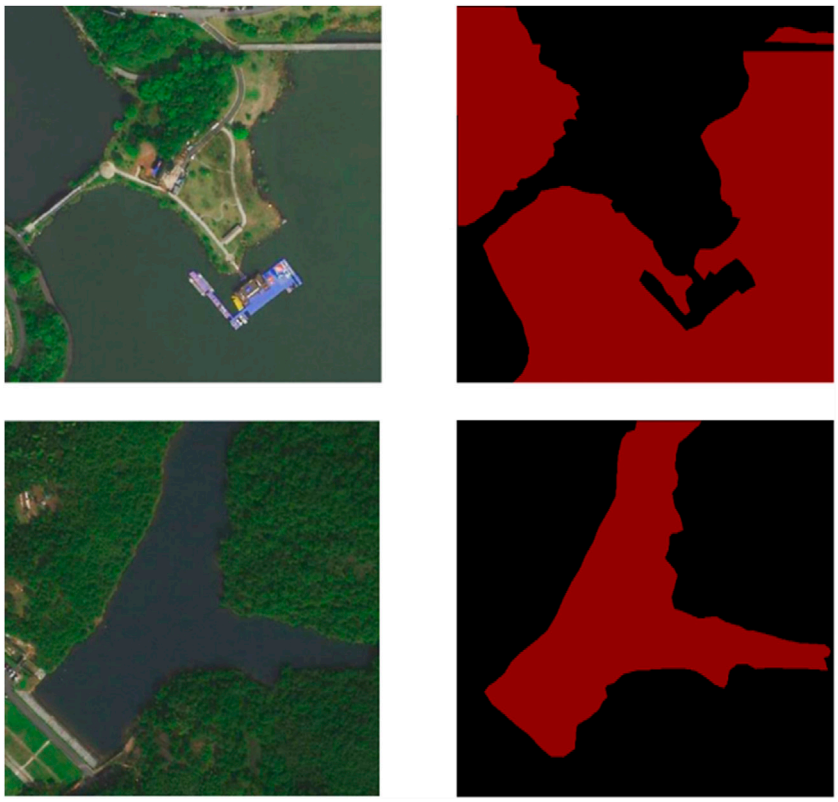

FIGURE 4
Structure diagram of improved DeeplabV3+. Maps Data:Google, ©2023 CNES / Airbus, Maxar Technologies. Reproduced with permission.

TABLE 1 MobileNetV2 network structure.

Input	Network	Expansion factor of input channel	Number of output channels	Module repetitions	Step
256 × 256 × 3	Conv2d	-	32	1	2
128 × 128 × 32	Bottleneck	1	16	1	1
128 × 128 × 16	Bottleneck	6	24	2	2
64 × 64 × 24	Bottleneck	6	32	3	2
32 × 32 × 32	Bottleneck	6	64	4	2
32 × 32 × 64	Bottleneck	6	96	3	1
16 × 16 × 96	Bottleneck	6	160	3	2
8 × 8 × 160	Bottleneck	6	320	1	1

attention maps. Finally, the two attention maps are applied to the input feature map through element-wise multiplication, emphasizing the representation of relevant features of interest.

While global pooling allows for the global encoding of channel attention with spatial information, it struggles to preserve positional information, which is crucial for capturing spatial structure in visual tasks. To facilitate the attention module in capturing long-range spatial interactions with precise positional information, global pooling is decomposed, and paired one-dimensional feature encoding is performed as defined by Equation below:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{3}$$

Encode each channel along the horizontal and vertical coordinates using pooling kernels of size (H, 1) or (1, W) for the given input x. Therefore, the output of channel c with a height of h can be expressed as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(j, w) \tag{4}$$

Similarly, the output with a width of w for channel c can be represented as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le i < H} x_c(j, w) \tag{5}$$

Through the above transformation, two aggregated feature maps are obtained by aggregating features in two directions. This section will stack the generated two aggregated feature maps and then use 1×1 convolutional transformation function F_1. Transform it:

$$f = \delta\left(F_1\left[z^h, z^w\right]\right) \tag{6}$$

Among them, $[\cdot,\cdot]$ represents the concatenation operation along the spatial dimension, δ is a nonlinear activation function, and f is an intermediate feature map encoding spatial information in the horizontal and vertical directions, γ is the reduction ratio of the control module size. Decompose f into two independent tensors $f^h \in R^{C/\gamma \times H}$, $f^w \in R^{C/\gamma \times w}$ along the spatial dimension. Using the other two 1×1 convolutional transformations F_h and F_w, respectively, f^h and f^w are transformed into tensors of the same number of channels into input X, and obtain:

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{7}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{8}$$

σ is the sigmoid activation function. To reduce the computational cost and complexity of the model, appropriate reduction ratios are used γ to reduce the number of channels for f. Then expand the output g^h and g^w to use as attention weights, respectively. Finally, the output y of the CA module can be represented as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

3.2.3 Optimize loss function

The loss function plays a crucial role in defining the training effectiveness of a neural network model by optimizing its parameters and objectives. In the case of high-resolution remote sensing images, the dataset encompasses a vast ground range, with water bodies occupying a relatively small proportion. Consequently, during network training, larger targets tend to dominate, causing the classifier to misclassify other target categories as larger targets. This imbalance negatively impacts the performance of the segmentation network. To address this issue and balance the losses, the approach employed in this article involves utilizing Focal loss.

Focal loss is a technique used to tackle the problem of imbalanced classification target proportions. Its calculation formula is presented in Equation:

$$FL = -a_c(1 - p_c)^{\gamma} \log(p_c) \tag{10}$$

Within this equation, weight a_c is employed to balance the uneven sample proportions across different categories. The parameter γ serves as a hyperparameter, and pc represents the prediction probability for various categories. For samples that are simple and easily distinguishable, their corresponding weight diminishes as their prediction probability increases. Conversely, for samples that are complex and difficult to distinguish, their weight increases as their predicted probability decreases. By employing Focal loss, the aim is to effectively address the imbalanced nature of the classification targets and enhance the segmentation network's performance.

4 Experiments and result analysis

4.1 Experimental environment and evaluation standards

The experiment conducted in this study employed an Intel i7-10700 CPU, running on the Windows 10 operating system. The GPU utilized was the NVIDIA GeForce RTX 2060S with 8 GB of graphics memory. The development environment employed for the experiment consisted of Python 1.9.0 and Python 3.8.

The primary objective of this study was to evaluate the segmentation performance of high-resolution remote sensing image datasets. To assess the performance of the network segmentation, several indicators were selected, including the average intersection to union ratio (mIoU), average pixel accuracy (mPA), and average recall (mRecall). The formulas for calculating these indicators are as follows:

$$\mathbf{mIoU} = \frac{1}{n} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n} p_{ji} - p_{ii}} \tag{11}$$

$$\mathbf{mPA} = \frac{1}{n} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} p_{ij}} \tag{12}$$

$$\boldsymbol{mRecall} = \frac{1}{n} \sum_{i=0}^{n} \frac{p_{ii}}{p_{ii} + \sum_{i=0}^{n} p_{ji}} \tag{13}$$

In these formulas, n represents the total number of categories, p_{ij} denotes the number of pixels that predict class i as class j, p_{ji} represents the number of pixels that predict class j as class i, and p_{ii} signifies the number of pixels predicted to be of class i as class i.

4.2 Experimental process

In the improved DeepLabV3+ network, the initial step involves loading the MobileNetV2 pre-training weights before commencing model training. To optimize the network's performance, the Focal loss function is employed, and the Focal loss parameter is adjusted accordingly. After conducting multiple experiments with consistent control variables, it has been observed that the model achieves the highest segmentation accuracy when γ is set to 3 and a_c is set to 0.2.

When working with the Water dataset, the training set randomly selects 60% of the images in the dataset, the validation set is 20% of the images in the dataset, and the test set is the remaining 20% of the images.

4.3 Ablation experiment

To evaluate the segmentation accuracy of models with different structures on the Water dataset, an ablation experiment was conducted on high-resolution remote sensing image water extraction. The experiment aimed to test the effectiveness of modifying the network used for feature extraction to the MobileNetV2 network, along with the

addition of CA module after the feature extraction module and Atrous Spatial Pyramid Pooling (ASPP) module. Four distinct schemes were employed during the experiment:

Scheme 1: This scheme involved replacing the feature extraction network of the traditional DeeplabV3+ network structure with the MobileNetV2 network.
Scheme 2: Building upon Scheme 1, CA module was added to the MobileNetV2 network.
Scheme 3: In this scheme, the CA module was incorporated after the ASPP module, while maintaining the modifications introduced in Scheme 1.
Scheme 4: Based on Scheme 1, the ASPP module was integrated into the CA module within the MobileNetV2 network.
Scheme 5: Add CA module to backbone network based on traditional DeeplabV3+ network structure.
Scheme 6: Based on the traditional DeeplabV3+ network structure, the ASPP module is added to the CA module.
Scheme 7: Based on the traditional DeeplabV3+ network structure, the CA module is added to the ASPP module in the backbone network.

These seven schemes were investigated to assess their impact on the segmentation accuracy of high-resolution remote sensing image water extraction, providing valuable insights into the performance of different model configurations.

The initial learning rate of this experiment is 0.0005, and the training iteration is 200 steps. Table 2 shows the experimental results of different schemes on the Water dataset, among which Scheme 4 has the best performance on multiple indexes. The experimental result of scheme 4 network is 2.05% higher than that of traditional DeeplabV3+ network, and the experimental result of Scheme 7 network is 0.61% higher than that of traditional DeeplabV3+ network. By comparing the experimental results, it is found that in the traditional DeeplabV3+ network structure, adding CA module to the network structure can effectively improve the segmentation accuracy of the network. At the same time, the experiment proves that adding CA module to the backbone network and ASPP module can more effectively indicate the segmentation effect. The results of this ablation experiment are shown in Figure 5. It can also be seen from the figure that the model segmentation effect of improved Scheme 4 is the best. This shows that DeeplabV3+ network can complete the water segmentation task on the water dataset. At the same time, in Scheme 4, CA module is added after feature extraction module and ASPP module to improve the accuracy of model segmentation.

TABLE 2 Water Results of ablation experiment.

Method	mPA(%)	mRecall (%)	mIou(%)
Traditional DeeplabV3+	92.64	92.03	87.84
Scheme 1	92.75	92.25	88.22
Scheme 2	92.88	92.36	88.84
Scheme 3	93.04	92.89	89.24
Scheme 4	93.23	93.02	89.89
Scheme 5	92.70	92.16	88.09
Scheme 6	92.73	92.23	88.14
Scheme 7	92.81	92.31	88.45

4.4 Comparison of segmentation performance between different methods

To extensively evaluate the segmentation performance of the enhanced DeeplabV3+ model in high-resolution remote sensing image water extraction tasks, this section of the experiment employed a comparative analysis. The experiment included the utilization of PSP-Net, MACU Net, U-Net, DeeplabV3+, and the improved DeeplabV3+ model proposed within this article for conducting water extraction experiments. The aim was to comprehensively assess and analyze the experimental outcomes, thereby providing valuable insights into the effectiveness of the different models in the context of water extraction from high-resolution remote sensing images.

In this experiment, five different semantic segmentation networks were utilized, employing an initial learning rate of 0.0005 and conducting training iterations for a total of 200 steps. The experimental results on the Water dataset are presented in Table 3, indicating that the proposed method surpasses the performance of other networks. Specifically, compared to the U-Net network, the proposed method demonstrates a 3.06% increase in mIou (mean Intersection over Union); compared to the MACU Net network, there is a 1.03% increase in mIou, and compared to the traditional DeeplabV3+ network, there is a 2.05% increase in mIou.

Table 4 shows the comparison results of training time and number of parameters of different methods on the Water dataset. It can be observed that compared with the traditional DeeplabV3+ network, our method significantly reduces the training time and the number of parameters of the model. Compared with the traditional semantic segmentation networks PSP-Net and U-Net, our method has more training parameters, but it achieves higher segmentation accuracy. Therefore, the MobileNetV2 network proposed in this paper is adopted as the feature extraction network, and the CA module is added to the feature extraction module and ASPP module, which greatly reduces the number of parameters and training time of the model.

Figure 6 displays the experimental results, showcasing a comparison of the segmentation performance among different methods. The results indicate that semantic segmentation networks are effective in accurately segmenting water bodies, achieving high segmentation accuracy in the task of high-resolution remote sensing image water extraction. It is worth noting that the traditional DeeplabV3+ network outperforms U-Net and PSP-Net in the water body segmentation task in high-resolution remote sensing images, further affirming that the proposed improved method in this paper effectively accomplishes the water body extraction task in such images, yielding superior segmentation outcomes.

4.5 Comparison with existing high-resolution remote sensing image water extraction methods

At present, there are few research results on water extraction from high-resolution remote sensing images based on deep neural

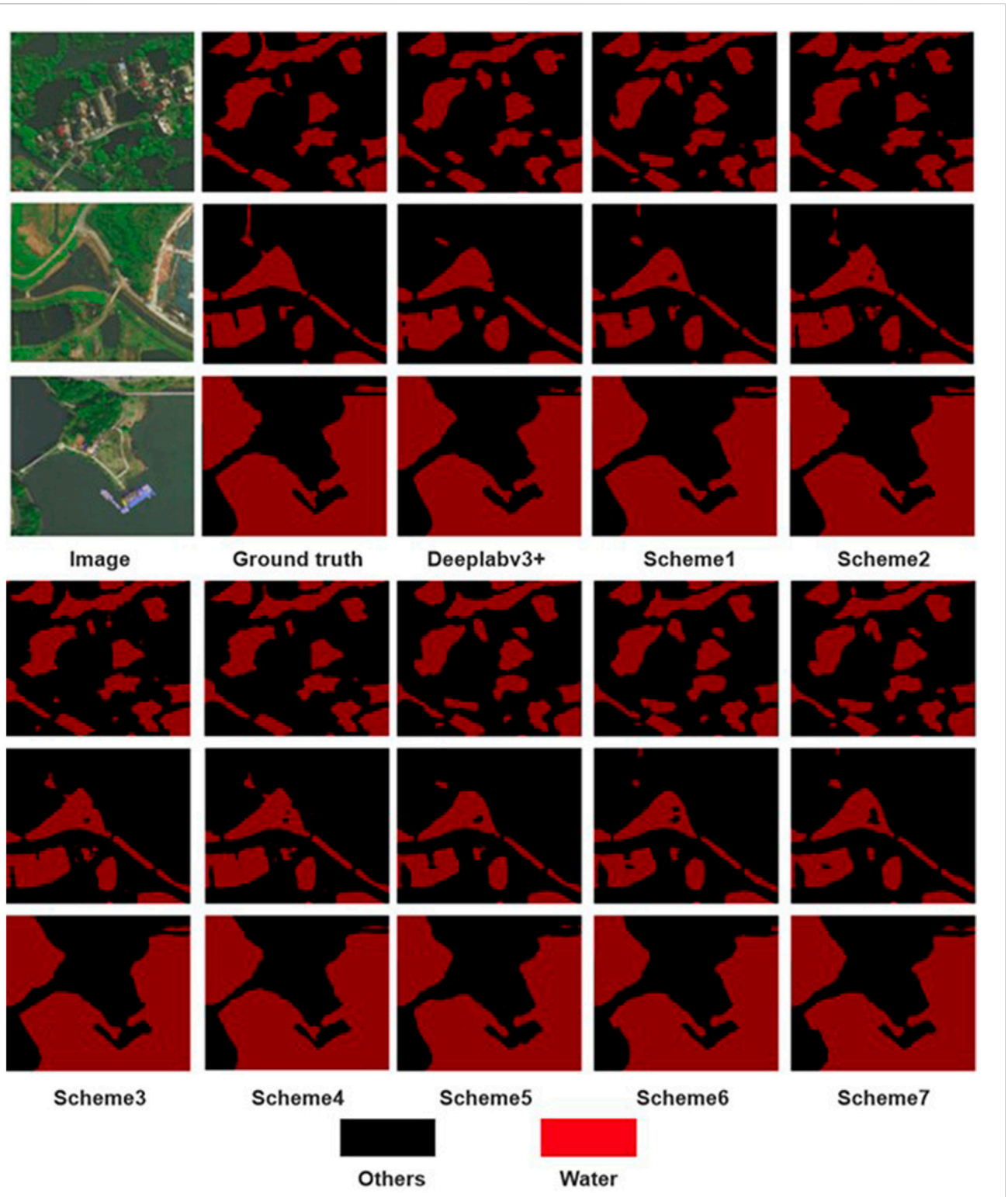

FIGURE 5
Water dataset ablation experiment results. Maps Data:Google, ©2023 CNES / Airbus, Maxar Technologies. Reproduced with permission.

networks. References (Ji et al., 2009; Sun et al., 2015; Ying et al., 2016 24, 25, 26) have proposed research on water extraction from high-resolution remote sensing images based on support vector machines, object-based classification, and NDWI index. The comparison between the water extraction method from high-resolution remote sensing images based on deep neural networks proposed in this paper and existing water extraction methods is shown in Table 5.

TABLE 3 Water comparison of segmentation methods.

Method	mPA(%)	mRecall(%)	mIou(%)
Traditional DeeplabV3+	92.64	92.03	87.84
U-Net	90.83	90.52	86.83
PSP-Net	89.64	89.32	86.27
MACU-Net	92.93	92.82	88.86
proposed method	93.23	93.02	89.89

The water extraction methods used in references (Ji et al., 2009; Sun et al., 2015; Ying et al., 2016) all require manual selection of spectral, texture, geometric, shadow, and background features of water in high-resolution remote sensing images, and the segmentation results depend heavily on the quality of image feature selection; This article utilizes deep neural networks to automatically extract features from high-resolution remote sensing images, fully leveraging the advantages of deep neural networks in image segmentation.

Reference (Ji et al., 2009) used the Normalized Difference Water Index (NDWI) to delineate the characteristics of surface

TABLE 4 Comparison of training time and number of parameters in water.

Method	Training time/epoch(S)	Parameter Quantity(M)
Traditional DeeplabV3+	311	208.51
U-Net	235	12.56
PSP-Net	189	10.31
MACU-Net	300	6.23
proposed method	220	23.05

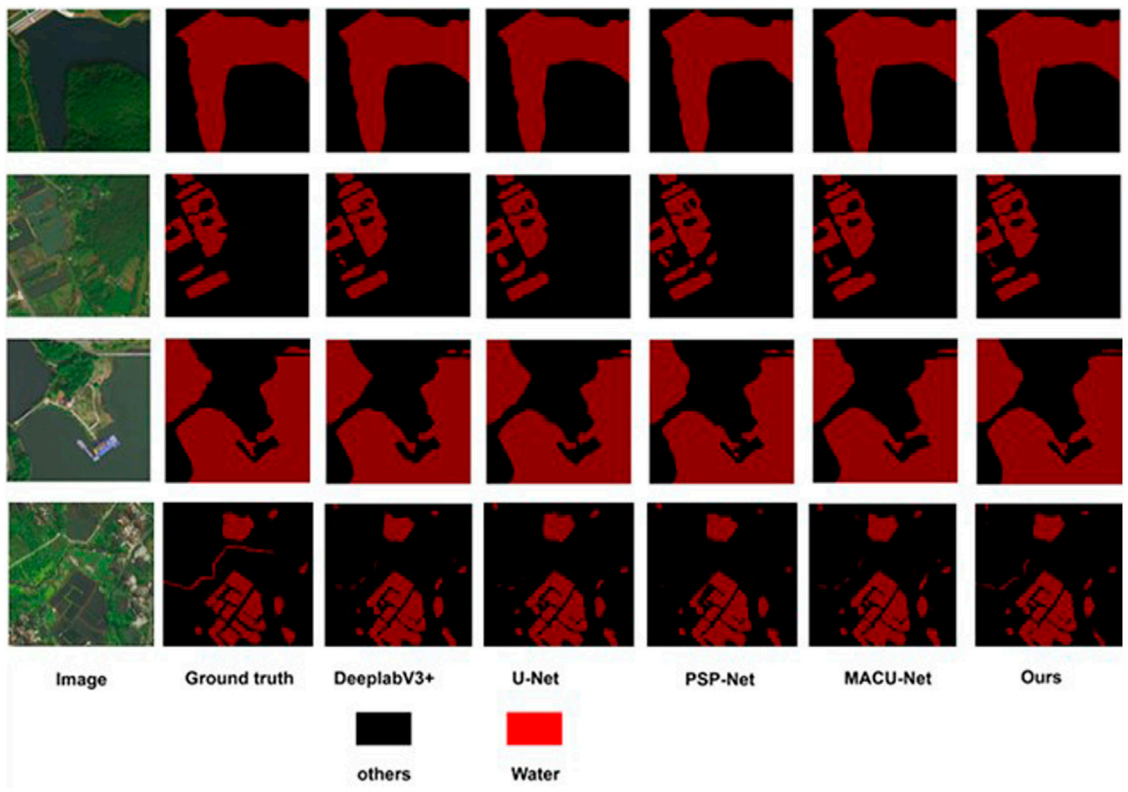

FIGURE 6
Water segmentation effect. Maps Data:Google, ©2023 CNES / Airbus, Maxar Technologies. Reproduced with permission.

From Table 5, it can be seen that compared with the surface coverage classification method in references (Ji et al., 2009; Sun et al., 2015; Ying et al., 2016), the improvement of this method lies in: water bodies. However, two main problems are often encountered: calculating NDWI from different band combinations yields different results; The NDWI threshold depends on the

TABLE 5 Comparison between this paper and the existing water object classification methods.

Reference	Dataset	Data type	Experimental methods	Feature extraction
Reference [24]	ASTER Spectral library	Text	NDWI	Manual extraction
Reference [25]	Custom dataset	Text	Support vector machine	Manual extraction
Reference [26]	WorldView-2 and Quick Bird	Text	Object based classification	Manual extraction
This paper	Water dataset	Image	Deep Neural Network	Automatic extraction

proportion of non-aqueous components. Therefore, literature (Ji et al., 2009) evaluates all NDWIs to determine the optimal performance index and establishes appropriate thresholds to clearly identify water characteristics. Reference (Sun et al., 2015) proposed a strategy for extracting urban water bodies based on mixed training data and SVM. This strategy was applied to the classification of Landsat 8 multispectral data in the Beijing area and validated through a large amount of manual surveying data. Reference (Ying et al., 2016) proposed a high-precision object-oriented water extraction scheme based on polarized SAR data. Through experiments, it has been proven that it can maintain accurate water edges, and the combination of texture features and decomposition components can distinguish between grasslands, wasteland, and shadows. Overall, traditional water body extraction methods can effectively obtain water body information, but the extraction results are easily affected by complex environments and have a large workload.

This article applies the attention mechanism to a high-resolution remote sensing image water extraction model based on deep neural networks and uses the lightweight MobilenetV2 network as the backbone feature extraction network to improve the accuracy of water extraction in the model, reduce the number of model parameters, and reduce the cost of training the model.

5 Conclusion

On the basis of traditional DeeplabV3+network, this article adopts MobilenetV2 network as the backbone feature extraction network, and adds a CA module after the feature extraction module and ASPP module, introducing Focal loss equalization loss for high-resolution remote sensing image water semantic segmentation. Experiments on the Water dataset have comprehensively confirmed the performance of our method over the benchmark method.

The main work of this article in the later stage is twofold: firstly, to improve the segmentation accuracy of the network. In the water extraction experiment, it was found that there is still interference in high-resolution remote sensing images, which affects the accuracy of network segmentation. Uncertain factors such as noise and shadows in images can interfere with the accuracy of network segmentation, so it is necessary to improve the segmentation performance of the network in complex scenes in the future. The second is to select more high-resolution remote sensing images and more complex water remote sensing images for further research and experiments.

Data availability statement

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

Author contributions

PW: Conceptualization, Methodology, Project administration, Resources, Supervision, Writing–review and editing, Writing–original draft. JF: Data curation, Writing–original draft. XY: Software, Supervision, Writing–review and editing. GW: Writing–review and editing, Formal Analysis. LM: Writing–review and editing, Resources. BM: Writing–review and editing. HL: Writing–review and editing, Investigation. CT: Writing–review and editing, Data curation. WG: Writing–review and editing, Methodology. TJ: Writing–review and editing, Data curation. ZR: Writing–review and editing, Project administration.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Key Research and Development Program of Zhejiang Province under grant number 2022C03039.

Acknowledgments

We are appreciative of the reviewers' valuable suggestions on this manuscript and the editor's efforts in processing the manuscript.

Conflict of interest

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

Publisher's note

References

Chen, L. C., Papandreou, G., Kokkinos, l., MurPhy, and K., and Yuille, A. L. (2015). *Semantic image segmentation with deep convolutional nets and fully connected CRFs[J/OL]. ICLR. arxiv.org/abs/1412.7062.*

Chen, L. C., Papandreou, G., Kokkinos, l., MurPhy, and K., and Yuille, A. L. (2017a). DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Analysis Mach. Intell.* 40, 834–848. doi:10.1109/TPAMI.2017.2699184

Chen, L. C., Papandreou, G., Schroff, F., and Adam, H. (2017b). *Rethinking atrous convolution for semantic image segmentation[J/OL].*Computer Vision and pattern recognition. *arxiv.org/abs/1706.05587.*

Chen, L. C., Zhu, Y. K., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. *Proc. Eur. Conf. Comput. Vis.*, 801–818. doi:10.48550/arXiv.1802.02611

Chu, Q., Ouyang, W. l., Li, H. S., Wang, X. G., Liu, B., and Yu, N. H. (2017). Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. *IEEE Int. Conf. Comput. Vis.*, 4846–4855. doi:10.1109/ICCV.2017.518

da Cruz, L. B., Júnior, D. A. D., Diniz, J. O. B., Silva, A. C., de Almeida, J. D. S., de Paiva, A. C., et al. (2022). Kidney tumor segmentation from computed tomography images using DeepLabv3+ 2.5 D model. *Expert Syst. Appl.* 192, 116270. doi:10.1016/j.eswa.2021.116270

Dai, M. C., Leng, X. G., Xiong, B. L., and Ji, K. F. (2020). An efficient water segmentation method for SAR images. *IGARSS 2020-2020 IEEE Int. Geoscience Remote Sens. Symposium*, 1129–1132. doi:10.1109/IGARSS39084.2020.9324113

Dirscherl, M., Dietz, A. J., Kneisel, C., and Kuenzer, C. (2021). A novel method for automated supraglacial lake mapping in Antarctica using sentinel-1 SAR imagery and deep learning. *Remote Sens.* 13 (2), 197. doi:10.3390/rs13020197

Du, S. J., Du, S. H., Liu, B., and Zhang, X. Y. (2021). Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Digit. Earth* 14, 357–378. doi:10.1080/17538947.2020.1831087

Feyisa, G. L., Meilby, H., Fensholt, R., and Proud, S. R. (2014). Automated Water Extraction Index: a new technique for surface water mapping using Landsat imagery. *Remote Sens. Environ.* 140, 23–35. doi:10.1016/j.rse.2013.08.029

Huang, X., Xie, C., Fang, X., and Zhang, L. P. (2015). Combining pixel-and object-based machine learning for identification of water-body types from urban high-resolution remote-sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 8(5), 2097–2110. doi:10.1109/JSTARS.2015.2420713

Ji, L., Zhang, L., and Wylie, B. K. (2009). Analysis of dynamic thresholds for the normalized difference water index. *Photogrammetric Eng. Remote Sens.* 75(11), 1307–1317. doi:10.14358/PERS.75.11.1307

Li, R., Duan, C. X., Zheng, S. Y., Zhang, C., and Atkinson, P. M. (2022). MACU-net for semantic segmentation of fine-resolution remotely sensed images. *IEEE Geoscience Remote Sens. Lett.*, 19, 1–5. doi:10.1109/LGRS.2021.3052886

Li, W., Yang, M. Y., Liang, Z. W., Zhu, Y., Mao, W., Shi, J. Y., et al. (2013). Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine. *Stoch. Environ. Res. risk Assess.* 27, 1861–1870. doi:10.1007/s00477-013-0720-3

Li, Z. Y., Wang, R., Zhang, W., Hu, F. M., and Meng, L. K. (2019). Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. *IEEE Access* 7, 155787–155804. doi:10.1109/ACCESS.2019.2949635

Long, J., Shelhamer, E., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. pattern analysis Mach. Intell.* 39, 640–651. doi:10.1109/tpami.2016.2572683

Lu, S. L., Wu, B., Yan, N. N., and Wang, H. (2011). Water body mapping method with HJ-1A/B satellite imagery. *Int. J. Appl. Earth Observation Geoinformation* 13(3), 428–434. doi:10.1016/j.jag.2010.09.006

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation[C]". in International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241. doi:10.48550/arXiv.1505.04597

Sun, X. X., Li, L. W., Zhang, B., Chen, D. M., and Gao, L. R. (2015). Soft urban water cover extraction using mixed training samples and support vector machines. *Int. J. Remote Sens.* 36(13), 3331–3344. doi:10.1080/01431161.2015.1042594

Wang, B., Chen, Z. L., Wu, L., Yang, X. H., and Zhou, Y. (2022). SADA-net: a shape feature Optimization and multiscale context information-based Water Body extraction method for high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 15, 1744–1759. doi:10.1109/JSTARS.2022.3146275

Wenkuana, D., and Shicai, G. (2023). Hazy images segmentation method based on improved DeeplabV3. *Acad. J. Comput. Inf. Sci.*, 6(5): 21–29. doi:10.25236/AJCIS.2023.060504

Xie, H., Luo, X., Xu, X., Tong, X. H., Jin, Y. M., Pan, H. Y., et al. (2014). New hyperspectral difference water index for the extraction of urban water bodies by the use of airborne hyperspectral images. *J. Appl. Remote Sens.* 8 (1), 085098. doi:10.1117/1.JRS.8.085098

Yang, F. Y., Feng, T., Xu, G. Y., and Chen, Y. (2020). Applied method for water-body segmentation based on mask R-CNN. *J. Appl. Remote Sens.* 14(1): 1. doi:10.1117/1.jrs.14.014502

Ying, D., Hong, Z., Chao, W., and Meng, L. (2016). An object-oriented water extraction method based on texture and polarimetric decomposition feature\. *Remote Sens. Technol. Appl.* 31(4), 714–723. doi:10.11873/J.ISSN.1004-0323.2016.4.0714

Zhang, X. Y., Li, J. J., and Hua, Z. (2022). MRSE-net: multiscale residuals and SE-attention network for water body segmentation from satellite images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 15, 5049–5064. doi:10.1109/JSTARS.2022.3185245

Zhang, Y. L., Li, K. P., Li, K., Wang, L. C., Zhong, B. N., and Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks[C]. In Proceedings of the European conference on computer vision, 286–301. doi:10.48550/arXiv.1807.02758

Zhao, H. S., Shi, J. P., Qi, X. J., Wang, X. G., and Jia, J. Y. (2017). "Pyramid scene parsing network". in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6230–6239. doi:10.1109/CVPR.2017.660

Zhou, G., Xu, J., Chen, W., Li, X., Li, J., and Wang, L. (2023). Deep feature enhancement method for land cover with irregular and sparse spatial distribution features: a case study on open-pit mining. *IEEE Trans. Geoscience Remote Sens.* 61, 1–20. doi:10.1109/tgrs.2023.3241331