Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **9**, Number **3**, August **2024**

Automatic Extraction Model of Lake Remote Sensing Images in Cold-Dry Regions Using Improved U-net Fusion CRFs

Honghui Li*, Tianming Zhao

Department of Computer and Information Engineering
Inner Mongolia Agricultural University, Hohhot 010018, China
lihh_chf@163.com, zhaotm0609@163.com

Jiangyin Fu

School of Electronics and Informaiton Engineering
Harbin Institute of Technology, Harbin 150006, China
JYF021@outlook.com

Xueliang Fu

Department of Computer and Information Engineering
Inner Mongolia Agricultural University, Hohhot 010018, China
fuxl@imau.edu.cn

*Corresponding author: Honghui Li

Received September 14, 2023, revised December 22, 2023, accepted February 20, 2024.

Abstract. *Lakes in cold and arid regions play an important role in the daily lives of people. However, the health of the lake ecosystem is seriously affected by climate change, especially human activities. Remote sensing satellites provide effective means for lake ecological health monitoring. How to extract lakes in cold and arid regions from remote sensing images has become a research hotspot. There exist some issues in most of the traditional methods regarding extraction accuracy as well as susceptibility to interference. Therefore, this paper proposes a novel method named AU-net, which is based on the improved U-net model and Fully Connected Conditional Random Field (FCCRF).* Firstly, semantic segmentation datasets are established using Landsat 8 OLI remote-sensing images for the intended research objects, i.e., Wulliangsuhai Lake and Hulunhu Lake. Then, the U-net model is constructed. To improve the model's adaptive ability, the Convolutional Block Attention Module is added after the feature map is extracted from the model. Finally, post-processing is conducted based on the FCCRF. The experimental results depict that the AU-net extraction accuracy is superior to the Normalized Difference Water Index, Deeplab v3+ model, and the initial U-net model. Moreover, AU-net can eliminate the interference of complex backgrounds and small holes. Furthermore, AU-net has achieved good extraction results for Wuliangsuhai Lake and Hulunhu Lake in different seasons.
Keywords: Lakes in cold and dry regions, Water body extraction, U-net, Convolutional block attention module, Fully connected conditional random field

1. **Introduction.** Lakes in cold and dry regions are significant freshwater resources on Earth, which play an essential role in biological reproduction, farmland irrigation, and ecological environment regulation in surrounding areas. Recently, it has become an urgent problem of the lake-area reduction and water pollution in the cold-dry region due to

climate change and human activity. Waterbody change can be monitored through remote sensing images, which are of great importance in pollution prevention and control [1], rational water resource development, and acceleration of ecological civilization construction. Therefore, how to efficiently and thoroughly extract water body from lake images in cold-dry regions becomes one of the research hotspots.

For the purpose of withdrawing water body, the researchers have suggested many automatic or semi-automatic methods lately. These extraction methods consist of four types. The first type is the Single-Band Threshold (SBT) method, which is the earliest method used. The second type is called the Multi-band Spectral Relationship Method (MBSR) method. The third type commonly used is the Water Index method (WI). The fourth one is the classifier method. Considering the water's low reflectivity in the mid-infrared and near-infrared bands, SBT [2] distinguishes straightforwardly the water part and the background part by taking advantage of these two bands. Lira [3] proposed an innovative scheme to extract water part using an improved principal component analysis. Through the combination of multiple bands, this framework probes the gray value features of the water part. On the basis of these features, the water body in the images can be segmented. Zhang et al. [4] propose a mechanical MultiFeature Water body Extraction (MFWE) framework, which combines spectral with spatial characteristics. With MFWE, water masks can be achieved for water body abstraction from a compound background. With the intention of picking water body out of their surroundings, WI main idea aims to choose compound bands for the proportion process. Currently, the popular method is called the Normalized Difference Water Index (NDWI) method [5], where the near-infrared band is employed besides the green band. Xu [6] enhanced NDWI. There, two bands, i.e., the green band and the mid-infrared band, were selected. Experiments show that this method can extract water body precisely from both vegetated areas and urban areas. Zhang et al. [7] proposed an algorithm to identify human motion in a virtual reality environment based on Support Vector Machine (SVM). Experimental results show this algorithm achieves better performance than other methods. The classifier method first trains the classifier [8] through the selection of representative images as training samples and then produces a categorization model. Based on this model, the water body can be withdrawn in accordance with the differentiations in spectral characteristics of the water body with the backdrop. The classifier method can get more polished water extraction in comparison with SVM [9], decision tree, SBT, MBIR as well as WI method. Saghafi et al. [10] proposed a water body extraction framework based on fusing remote sensing images, which come from multiple remote sensing satellites. The integration of water indices improved the surface water extraction accuracy.

Although the traditional extraction methods are relatively simple, it is usually required to set parameters manually. This results in a low degree of automation and weak versatility and is not suitable for large-scale extraction. Moreover, with a large amount of information about the ground objects in the remote sensing image, it is easy to have the phenomenon of wrong mention and omission. Therefore, the extraction accuracy has a great space for improvement. Deep learning method offers an efficient way to tackle the problems mentioned above. Krizhevsky et al. [11] present the Convolutional Neural Network (CNN) framework, named as AlexNet. With the purpose of image recognition, this framework acquires supplementary abstract image characteristics through a great quantity of data samples. However, CNN only can achieve the image abstract characteristics, which is useful for image categorization. Moreover, it is tough for CNN to identify the targets in the figure in detail. Based on the CNN, Long et al. [12] suggest Fully Convolutional Networks (FCN) architecture to fulfill semantic segmentation. In the proposed architecture, Long et al replace a convolutional layer with a fully connected layer, and

employ deconvolution to rebuild the image abstract characteristics to categorize the image pixels. The enhancement [13] was presented of false positive reduction in the detection of lymph nodes based on CNN and sparse coding. The simulation results show that this enhancement efficiency is superior to the previous processes. For the purpose of scene classification, the GM-SMO model [14] was suggested to select features in order to deal with the overfitting issues in the CNN models. This model extracts features from the augmented images in combination with AlexNet and VGG19. And then these features are fed into LSTM for scene classification. Zhao et al. [15] come up with a mechanism FCN-OPT for image segmentation of Wuliangsuhai Lake. On the basis of the improved FCN, the suggested framework can accomplish enhanced performance for water body extraction than MBSR and NDWI. In order to withdraw water body, especially from Sentinel-2 images, Parajuli et al. [16] suggests a novel method AD-CNN. On the basis of CNN, AD-CNN employs dense assemblies and a novel attention module to reach the ideal performance on water body extraction. Tambe et al. [17] proposed a CNN-based end-to-end multi-featured network model named as W-Net. The W-Net model uses shrinkage networks to capture semantic information of image contexts, while localization is achieved by extending the network and using an asymmetric network structure to reduce training parameters, which can ultimately be trained on fewer image data to achieve accurate extraction of water body. Li et al. [18] came up with a FCN approach for the purpose of segmenting water body out of remote sensing images. With this framework, the improved results can be accomplished in terms of automation and accuracy.Wang et al. [19] propose an AFD neural network in order to segment CT images with high variability. The performances of this network outweigh the existing methods shown in the experimental results. A semi-supervised CNN model DNetUnet is suggested [20] for the segmentation of medical images, and its advantage is of processing large images.

The surrounding geomorphology of lakes varies greatly with the seasons in cold and dry regions and presents relatively complex features on remote sensing images. Also, there are distracting factors such as mountains, shadows, grasses, and aquatic plants. All of these make it tough to withdraw the lake waters entirely and perfectly with the traditional methods. Therefore, this paper studies how to extract water body out of remote sensing images accurately. For this purpose, a novel method is presented through taking advantage of the improved U-net architecture. First, it is introduced into the U-net core feature extraction part of the Convolutional Block Attention Module (CBAM). Then, Fully Connected Conditional Random Field (FCCRF) is utilized for back-end processing to optimize the extraction results. Finally, precise water body extraction is achieved in cold and dry regions.

2. **Study Area Profiles and Image Sources.**

2.1. **Overview of the Study Area.**

Two typical cold-dry region lakes, WuliangSuhai Lake and Hulunhu Lake in Inner Mongolia Autonomous Region, are selected as the research objects. Wuliangsuhai Lake lies in the former Ulat banner in Bayannur City, Inner Mongolia Autonomous Region. It is a large grassland lake that is very uncommon in the global desert as well as semi-desert areas, with a total area of 300 km^2 now. It ranges in latitude and longitude between 108°43′-108°57′E and 40°36′-41°03′N, and the entire lake area is crescent-shaped as shown in Figure 1(a). Due to the massive discharge into the lake of urban residential sewage and industrial effluent, the lake area has been drastically reduced, the water quality has been seriously polluted and the water body is seriously eutrophic. These factors accelerate the aging process of the lake and cause severe damage to the balance of the local ecosystem [21].

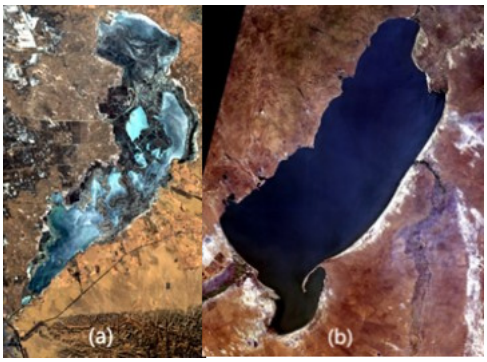

FIGURE 1. Example of the Wuliangsuhai Lake and the Hulun Lake.

Hulun Lake is situated on western Hulunbuir Prairie, between the two banners of Manzhou City and Jalainur, with geographical coordinates of 117°00′10″-117°41′40″E and 48°30′40″-49°20′40″N. It is the greatest lake in the Inner Mongolia Autonomous Region of the People's Republic of China. The lake area is irregularly shaped in a diagonal rectangle, as shown in Figure 1(b). The area of Hulunhu Lake has declined significantly due to river breakage, drought, and sanding of the grasslands, and has now been maintained at 2,339 km^2 after recent years of treatment [22].

2.2. **Data Source.** The Landsat 8 satellite was launched by NASA in February 2013. It has a 16-day revisit period and a 30m spatial resolution. The carried OLI land imager by Landsat8 satellite contains 9 bands, among which Band 5 ranges from 0.845 to 0.855um, which can prohibit the steam absorption effect at 0.825um on Band 8. In comparison to Landsat 7, the panchromatic band is slimmer than Landsat 7. As a result, the distinction in the image between vegetated and non-vegetated areas is more pronounced. Also, it is added for coastal zone as well as water consumption monitor that of Band 1 (Blue Band) [23] and Band 9 (short-wave infrared band). From the USGS website, this paper took the whole images utilized, and those with less than 10% cloudiness were selected for their clarity.

3. **AU-net Method.** Figure 2 depicts the process of the proposed AU-net method for water body extraction out of the remote-sensing images of the Wuliangsuhai Lake and Hulunhu Lake. This method takes advantage of an improved U-net architecture, and is comprised of three key steps: (1) acquiring Landsat 8 OLI remote sensing images of the Wulianghai Lake and Hulunhu Lake region, and building a dataset; (2) constructing a U-net network model, then inserting the convolutional attention module CBAM into the U-net model and training it; (3) using the improved model to initially extract the water body regions in the Wulianghai Lake and Hulunhu Lake images and optimizing them with a fully connected conditional random field.

3.1. **Create Dataset.**

3.1.1. *Data preprocessing.* Firstly, it was downloaded of a total of 141 Landsat 8 remote sensing images in the Wuliangsuhai Lake and Hulunhu Lake areas from 2015 to 2021 from the website http://www.usgs.gov/. Then, ENVI 5.3 was exploited for the image radiation measurement calibration and atmospherical correction in order to frame the Wuliangsuhai Lake and Hulunhu Lake areas. These two areas were cropped out of the remote sensing image and the redundant parts were removed. In order to facilitate future water extraction, ENVI's Change RGB Bands tool was used for the band combination

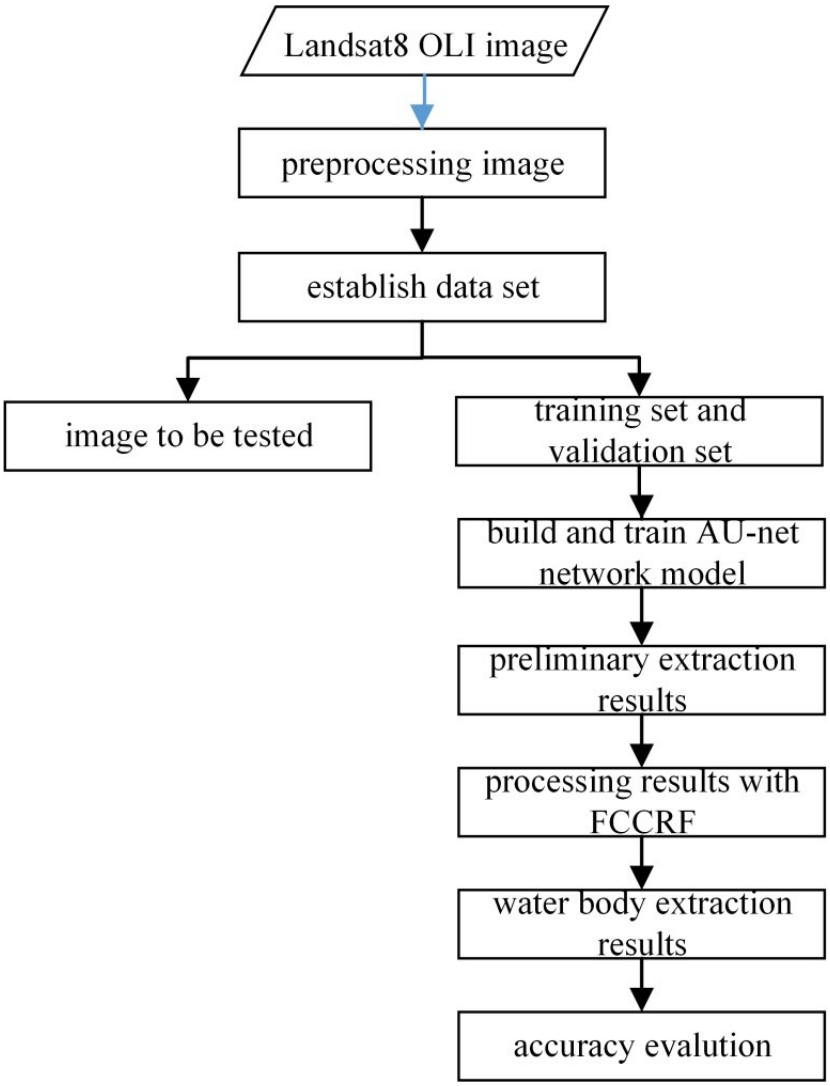

FIGURE 2. Work flow of the proposed AU-net method

of the cropped image so as to emphasize the disparities between the water part and the background part. Explicitly, the SWIR2, NIR, and Red bands of the Landsat8 OLI remote sensing image were assigned respectively to the R, G, and B wavebands. Following this, a 2% linear enhancement was conducted to get original images of the Wuliangsuhai Lake and Hulunhu Lake regions clearly.

A total of 141 images were obtained after preprocessing. There are 57 images of the Hulunhu Lake area, each with a size of 2464×3136. This paper selected 28 images of Wuliangsuhai Lake and 15 images of Hulunhu Lake to make the semantic segmentation data set. The remaining images were used to test the accuracy of the proposed framework.

3.1.2. *Data annotation.* The method is adopted of regular grid segmentation. After the combination processing of each band used for training, each image downloaded is divided into 224×24 pixel subgraphs. These segmentation results were screened, and the images were maintained of both water and land parts. After that, with the help of LabelMe software, the water body outline in every single sub images was indicated manually. The contour inner part was marked as the forefront, and the surplus part as the surroundings. These tagged files were saved as those images in PNG format. Following this, standard normalization was done. The image pixel values were set to 0 for the background section

and 1 for the lake water section. Some labeling results are shown in Figure 3, converting all the labeled images to 8-bit single-channel images.

For the sake of increasing data diversity and avoiding overfitting problem in the process of cultivating the model, data enhancement is performed here on the dataset. The specific operations are brightness enhancement, contrast adjustment, image rotation, flip, and sharpness adjustment. The same operations are applied to the image downloaded as well as the corresponding image labeled in data enhancement. In total, 8388 pairs of image samples are obtained finally. In a ratio of 8 to 2, 6710 samples are randomly split into the training image dataset and the validation dataset. As a result, the validation dataset consists of 1678 image pairs.

FIGURE 3. Label the image

3.2. **Water Body Extraction Method Using Improved U-net Model.**

3.2.1. *Introduction to U-net Model.* CNN is mainly applied to address image classification [24]. It is composed of four layers. These four layers are in order the convolution layer, the excitation layer, the pooling layer, and the full connection layer. The image is processed in these four layers sequentially. After that, the image can be categorized properly. However, it is hard for CNN to classify the inner object of the image. To deal with this issue, the full convolutional network replaces the CNN fourth layer (i.e., the the full connection layer) with the convolution layer. Processed by this multi-layer convolution layer, the FCN output is a rough feature map rather than the image category probability. Afterward, in order to obtain a grayscale one without scaling the input image, the deconvolution layer is developed to up-sample the feature image. Each pixel classification of the input image is determined, and the semantic level is segmented of the input image [25]. In such a way, specific areas in the image is extracted and a marked picture is achieved.

The u-net model belongs to the full convolutional network. The segmentation of medical images first employed the U-net model. Now U-net is extensively utilized in the remote sensing domain. It is named U-net because its network structure is like the letter U.

3.2.2. *Establish the Initial U-Net Framework.* The U-net framework is typically established on the VGG technique, which is presented by Oxford University. In the VGG network model, multiple small convolution kernels (3×3) are stacked to replace large convolution kernels to extract features. In this way, the parameter quantity and the network depths can be decreased, and more abstract and advanced features can be gained. The VGG16 model is a VGG variant, which incorporates 13 convolution layers, and 3 fully connected layers.

Herein the VGG16 network is used as the backbone network. Those are retained of 13 convolutional layers of Original architecture. On the other hand, the following three

ones are replaced with three convolution layers. The weight parameters achieved by the VGG16 network were utilized to construct the initial U-net model. Figure 4 shows the U-net architecture.

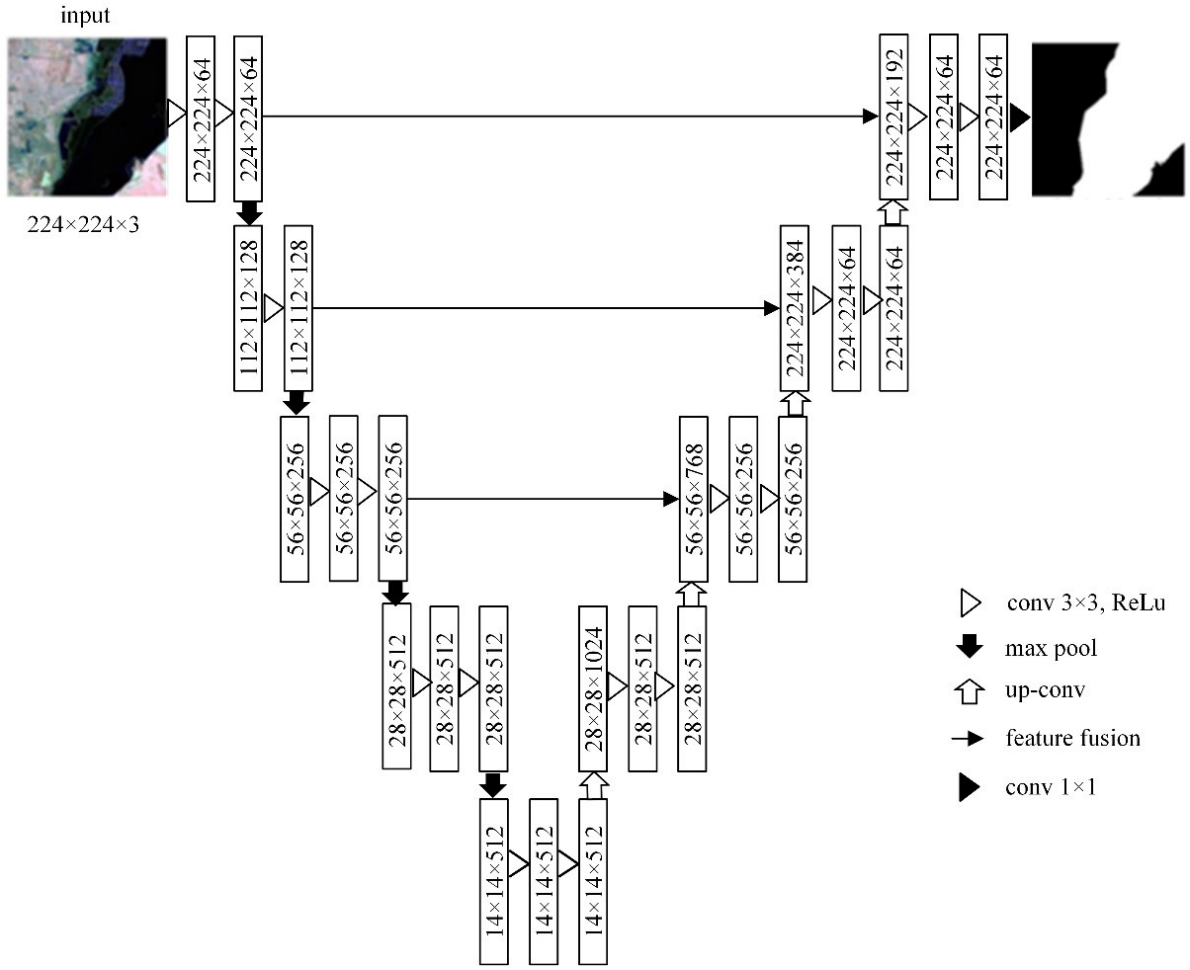

FIGURE 4. The U-net framework

The U-net framework mainly consists of the following parts:

(1) Convolution layer: Through convolution operations, diverse input image characteristics can be achieved in the convolution layer. And then these features are saved in multiple convolution kernels. After that, with the backpropagation algorithm, the convolution kernel is optimized in order to achieve closer to the image features. Table 1 presents each convolution layer parameters in the U-net architecture constructed here, part of which is identical to the values shown in literature [15]. (2) Activation layer: The

TABLE 1. Summary statistics

Convolution layer	**kernel_size**	**kernel_number**
Conv1_1 − conv1_2	3×3	64
Conv2_1 − conv2_2	3×3	128
Conv3_1 − conv3_3	3×3	256
Conv4_1 − conv4_3	3×3	512
Conv5_1 − conv5_3	3×3	512

nonlinear activation function is employed in the activation layer to fulfill the network nonlinearization, remove the jobless features and preserve the valuable characteristics. Herein, the ReLU function [26] is exploited in order to relieve the issue of network gradient vanishing as well as quicken the training process. After each convolution layer, the Formula (1) below is added.

$$F(X) = max(0, X) \tag{1}$$

Therein, X is variates, which indicates the activation function input.

(3) Pooling layer: The input image in the pooling layer can be condensed, and its dimensionality can be decreased. In the meantime, those main characteristics can be preserved. Also, the training process can be accelerated. It can be enhanced by the generalization ability of the model. Overfitting can also be inhibited. A maximum pooling layer is exploited with a 2×2 pooling window as well as 2 step size. After pooling the input image, the image size is reduced to half of the original. The pooling layer formula is described as follows.

$$F_{ij} = \frac{1}{c^2}\left(\sum_{i=1}^{c}\sum_{j=1}^{c} G_{ij}\right) + b_1 \tag{2}$$

Therein, F_{ij} indicates this layer output, G_{ij} represents the characteristic map matrix achieved from the first layer. Regarding a pooling window, c represents its slide step size, b_1 implies the offset amount.

(4) Up-sampling: The images are subjected to a series of convolution and pooling operations to obtain five feature maps with resolutions of 1/2, 1/4, 1/8, 1/16, and 1/16 of the original image resolution. For the purpose of obtaining the pixel category from the original remote sensing image, a bilinear interpolation upsampling method is employed such that the feature map size is shrunken to the original.

If the feature graph is directly up-sampled 16 times with the 1/16 size of the original fifth graph, only the convolution kernel features can be reinstated in the fifth convolution layer. The results obtained are crude and not refined enough. Therefore, U-net adopts a hierarchical structure, which fuses the features distilled from the previous several convolution layers in order to enhance the image details and increase extraction accuracy. Firstly, the bilinear interpolation method is used to double up-sample the last feature graph. Then the feature is fused with the previous feature graph by stacking. After four times of up-sampling, the feature map dimension is back to the original size. Finally, a 1×1 convolution layer is used to adjust the final feature graph channel number to the classification number. In this paper, the feature graph is not clipped like the initial U-net, but directly sampled twice. In this way, the final output result of the model is restored to the input image size. Thereby, the model's applicability is increased.

(5) Softmax layer: After the upsampling is completed, in order to categorize each pixel, this paper makes use of the Softmax function to evaluate each pixel possibility of belonging to each category. The Softmax function is calculated as Formula (3).

$$\text{Soft}\max(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}} \tag{3}$$

Where z_i represents the i-th node output, C indicates the category amount.

3.2.3. *Improved U-net Model.* Regarding water bodies in the remote sensing images of Wuliangsuhai Lake as well as Hulunhu Lake, their colors are complicated and mixed with

the background. For these reasons, a new network model AU-net is put forward here. AU-net introduces an attention mechanism to the U-net architecture to increase its adaptive ability. As a result, AU-net can focus more on the water body features in the image.

The Attention Mechanism (AM) is built by draw lessons from the human visual system [27]. When observing an image, people usually ignore the unimportant parts and focus their attention on its main body. When employment of deep learning model for purpose of processing images, it is also desired that the model can keep an eye on the key features required by the task. In this case, it is needed to add the AM to the deep learning model. This paper exploits CBAM. An input feature map will be handled in turn through channel AM as well as spatial AM, and then the weights of the channels and feature points are assigned. As a result, CBAM focuses on both channel and spatial features of images. The CBAM working flow is illustrated in Figure 5.

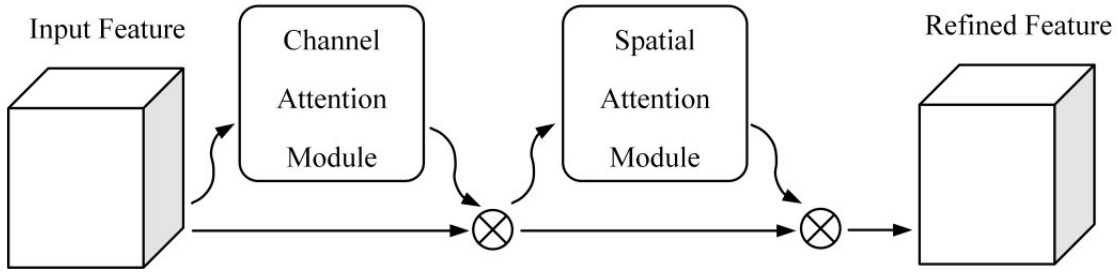

FIGURE 5. CBAM Working Flow

Theoretically, as a plug-and-play AM module, CBAM is suitable to be placed anywhere in the network. However, if it is placed in the backbone network, the pre-training weight of the model cannot be used, so this paper inserts the CBAM module behind each feature layer extracted by U-net. Figure 6 depicts the improved network architecture.

Firstly, each feature map is processed from the channel dimension through the channel attention mechanism module in CBAM. After an average and a maximum pooling operation on the feature graph F, two feature graphs of $1 \times 1 \times C$ are obtained. Then, these two feature graphs are fed into a shared Multi-Layer Perceptron (MLP), where the neuron amount is set to C/r in the hidden layer, and r denotes the decay ratio. Finally, the channel attention weight M_c is achieved through adding the two output results of the MLP element-by-element and activated by Sigmod function.

The overall process can be expressed by Equation (4).

$$\begin{aligned} M_c(F) &= \sigma(\mathrm{MLP}(\mathrm{Avg\,Pool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F_{\mathrm{crg}}^{C}\right)\right) + W_1\left(W_0\left(F_{\max}^{c}\right)\right)\right) \end{aligned} \tag{4}$$

In Formula (4), σ represents the Sigmoid activation function, $W_0 \in R^{C/r \times C}$ stands for the hidden layer weight of the multi-layer perceptron, and $W_1 \in R^{C \times C/r}$ represents the output layer weight of the multi-layer perceptron.

After the channel attention feature weight is acquired, it is multiplied by the original feature graph element by element. Each channel's corresponding weight is assigned to obtain a new feature graph. It can be expressed by Formula (5).

$$F' = M_c(F) \otimes F \tag{5}$$

In Formula (5), $\otimes$ stands for multiplying element by element.

The input to the spatial AM module is the feature map F' handled by channel attention mechanism. Firstly, maximum pooling and average pooling are accomplished on each feature point's channel in the feature figure F' to obtain two $1 \times H \times W$ feature pictures.

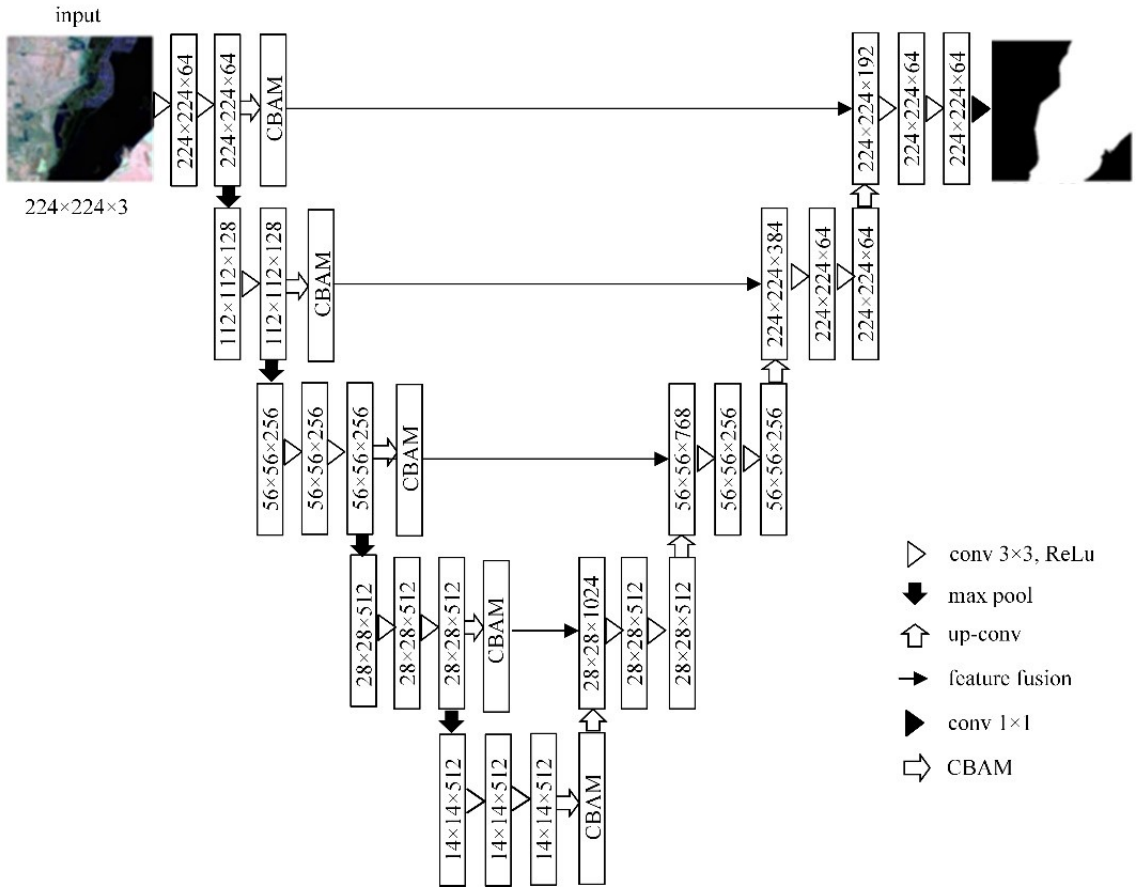

FIGURE 6. Structure of improved U-net model

Then, the two feature pictures are spliced together according to the channel, and a 7×7 convolution layer is got through to adapt channel number. Finally, spatial attention feature weight M_s was obtained by Sigmoid function activation.

The overall process can be expressed by Formula (6).

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}([\text{AvgPool}(F);\text{MaxPool}(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{\text{avg}};F^s_{\text{max}}\right]\right)\right) \end{aligned} \tag{6}$$

Hereon, σ denotes the Sigmoid activation function, $f^{7\times7}$ symbolizes the convolution operation with 7×7 convolution kernel size.

After obtaining the spatial attention feature weight M_s, multiply it with the feature graph F' element by element to gain the final feature graph F''. It can be expressed by Formula (7).

$$F'' = M_s(F') \otimes F' \tag{7}$$

3.2.4. *Optimization of the Fully Connected Conditional Random Field.* Although U-net uses a hierarchical structure to fuse shallow and deep features, it does not consider the affiliation between image pixels. In addition, in the feature extraction process, some features of the original image could inevitably be lost, which will lead to the blurred boundary of the extracted water body and errors. Therefore, FCCRF is exploited here as post-processing in order to optimize water extraction results.

Define the original image to be segmented as an arbitrary area $X = \{X_1, X_2, ..., X_N\}$. In the arbitrary field, N denotes the quantity of all pixels in the image. X_i denotes the

pixel i label value in the image. The label value 1 represents the water body target, and the label value 0 represents the background. At the same time, another random field $I = \{I_1, I_2, ..., I_N\}$ is defined according to the image to be segmented. I^i represents the pixel i color feature vector, and (X, I) constitutes a conditional random field model. The expression is defined as Formula (8).

$$P(X \mid I) = \frac{1}{Z(I)} \exp\left(-\sum_{c \in C} \Phi_c \left(X_c \mid I\right)\right) \tag{8}$$

Therein, $Z(I)$ implies the normalization function, C denotes the set of pixel blocks of the same category, Φ_c stands for potential function. The potential function can be expressed as Formula (9), which conforms to the Gibbs distribution.

$$E(x \mid I) = \sum_i \psi_u \left(x_i\right) + \sum_{i,j} \psi_p \left(x_i, y_i\right) \tag{9}$$

where $\psi_u \left(x_i\right)$ is a univariate potential function. This paper uses the preliminary extraction results of the improved U-net to construct the FCCRF univariate potential. The formula is defined as follows:

$$\psi_u \left(x_i\right) = -\log P(x_i) \tag{10}$$

In Formula (10), $P(x_i)$ indicates the pixel probability predicted by AU-net that belongs to the water body or the background.

Let $\psi_p \left(x_i, y_i\right)$ be a binary potential function, which is used to describe the relationship between one pixel and another pixel in the image. The discrimination of the relationship is related to the color value and relative distance of the pixel in the original remote-sensing image. If the relationship between two pixels is close, they are appointed the same category label; If the relationship between two pixels is quite different, they are assigned different category labels. Thus, the water part of the image and the background are disconnected at the edge as far as possible, and the water extraction results with smoother boundaries are obtained. The expression of the binary potential function is defined as Formula (11).

$$\psi_p \left(x_i, y_i\right) = \mu \left(x_i, y_i\right) \Sigma \omega^m K_G^m \left(f_i, f_j\right) \tag{11}$$

Here, $\psi_p \left(x_i, y_i\right)$ denotes the label compatibility item, $K_G^m \left(f_i, f_j\right)$ stands for the Gaussian kernel function, ω^m denotes the weight parameter in each kernel. The Gaussian kernel function is defined as the Formula (12).

$$K_G^m \left(f_i, f_j\right) = W^{(1)} e^{\frac{-\left|p_i - p_j\right|^2}{2\theta_\alpha^2} - \frac{\left|I_i - I_j\right|^2}{2\theta_\beta^2}} + W^{(2)} e^{\frac{-\left|p_i - p_j\right|^2}{2\theta_\gamma^2}} \tag{12}$$

Where, p_i and p_j denote the position vectors of pixel i and pixel j respectively, I_i and I_j denote color vectors of pixel i and pixel j respectively. The first term of the formula is the surface kernel function. Pixels with more similar features in adjacent areas are prone to be organized into the same category. θ_α and θ_β are used respectively to control the area size and the color similarity range. The second term of the formula is the smoothing kernel function, which is used to eliminate the isolated area in the image, and θ_γ is used to control the area size. After several iterations, the energy function value reaches the minimum and the final segmentation result is obtained.

4. **Experiment and Result Analysis.**

4.1. **Training Model.** The experiments have been conducted with the Pytorch 1.2 deep learning framework under Centos 7.6. It is programmed with Python language. Computer used is with Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz, and its memory size is 32GB. Moreover, the NVIDIA Tesla V100 32GB GPU is used for training acceleration.

The parameter values exploited during training are assigned as follows. The class number is set to 2, the batch size 16, and the epoch 120. In the freezing stage, the model backbone network is frozen, its parameters are fine-tuned. The Initial Learning Rate (ILR) is set to the value of 0.0001. And the last 60 epochs are the unfreezing phase. In this phase, the network parameters will be changed, and ILR is 0.00001. Figure 7 depicts the train-loss curve. Figure 7 illustrates that the loss value declines rapidly among the

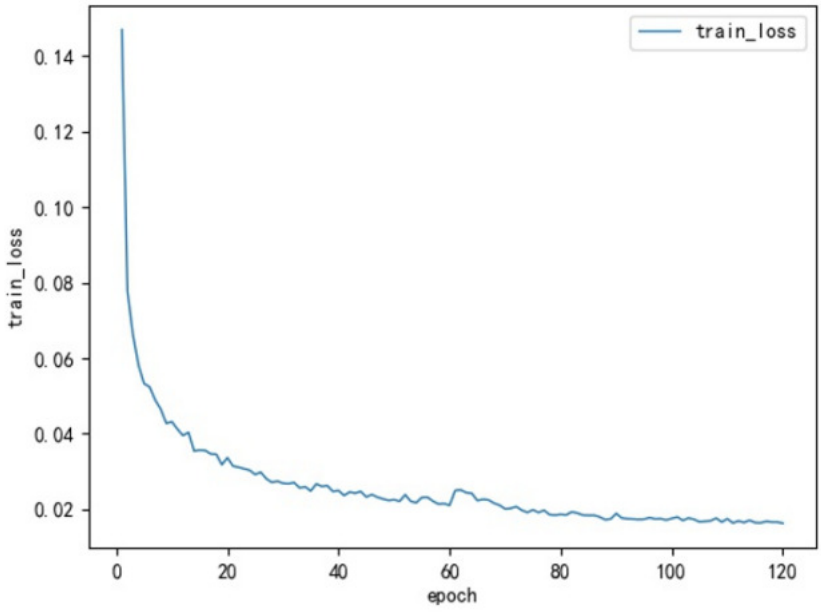

FIGURE 7. The change of loss function value during improved U-net model training

freezing training of the first 60 epochs. At the beginning of thawing training, the loss value increases slightly. Then it gradually declines and becomes stable. At this time, the model converges.

4.2. **Evaluation Index.** For the purpose of evaluating the AU-net performance, three indexes are adopted in this paper, i.e., Pixel Accuracy rate (PA), Average Pixel Accuracy rate (MPA) as well as Mean Intersection ratio (MIoU). Among these three indexes, MIoU acts as a typical marker in the image extraction area [28]. Therefore, MIoU herein is taken as the primary assessment indicator.

(1) PA

PA is calculated as the proportion of the amount of those pixel points with correct categorization to the amount of all pixel points, which is defined as Formula (13):

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{13}$$

Where, k denotes the total category number, p_{ij} denotes the pixel number of category i that are wrongly classified into category j, p_{ii} denotes the pixel number of category i that are correctly classified.

(2) MPA

The mean pixel accuracy is calculated separately for each category and then averaged over all categories. The formula is:

$$\mathrm{MPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{14}$$

Where, k, p_{ij} and p_{ii} are with definitions identical to the counterpart in Formula (13).

(3) MIoU

Firstly, in each group, the ratio is needed to be calculated of the intersection and union of the true and projected values. Following this, MIoU is defined as the average ratio of all categories is evaluated. Ideally, MIoU is 1. MIoU is defined as Formula (15).

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{15}$$

Where, p_{ii} denotes the set intersection of correctly classified, p_{ji} denotes the pixel number being affiliated with category j is misclassified as category i.

4.3. **Experimental Results and Analysis.**

4.3.1. *Comparisons and Analysis.* For the sake of evaluation of the AU-net performance proposed, extensive experiments have been carried out. AU-net is in contrast with the traditional NDWI method, Deeplab v3+ model, and the initial U-net model. Test images are randomly selected. The extraction results are depicted in Figure 8.

For accurate comparisons of the water-body extraction methods which are shown in Figure 8, three performance indicators, PA, MPA, and MIoU, are exploited. The corresponding results are presented in Table 2. It can be observed that the proposed AU-net method outweighs NDWI, Deeplab v3+ model, and initial U-net in three evaluation indexes. The PA value of the AU-net method is 99.29%, which is improved by 10.27%, 5.01% and 1.41% compared with NDWI method, Deeplab v3+ and initial U-net model, respectively. The MPA value of the AU-net method is 99.01%, which is 16.79%, 4.33%, and 1.19% higher than that of the NDWI method, Deeplab v3+, and initial U-net correspondingly. The MIoU value of the AU-net method reaches 98.31%, which is improved by 22.19%, 10.41%, and 3.06% compared with the NDWI method, Deeplab v3+ and original U-net model, respectively.

Table 2. Extraction results of the four methods

Extraction method	**PA/%**	**MPA/%**	**MIoU/%**
NDWI	89.02	82.22	76.12
Deeplab v3+	94.28	94.68	87.90
The initial U-net	97.88	97.82	95.25
AU-net	99.29	99.01	98.31

4.3.2. *The extraction results of images in diverse seasons.* The water body spectral characteristics differ greatly from the seasons of the remote sensing images in the cold-arid regions. In order to assess the AU-net effectiveness, the images of Wuliangsuhai Lake and Hulunhu Lake in different periods, which include spring, summer, autumn and winter freeze-up period, are chosen at random from the image set. The AU-net extraction results are revealed in Figure 9 and Figure 10 in turn.

As shown from Figure 9 and Figure 10, the AU-net method can exclude the interference of complex backgrounds, and achieve ideal results on the images of both Wuliangsuhai

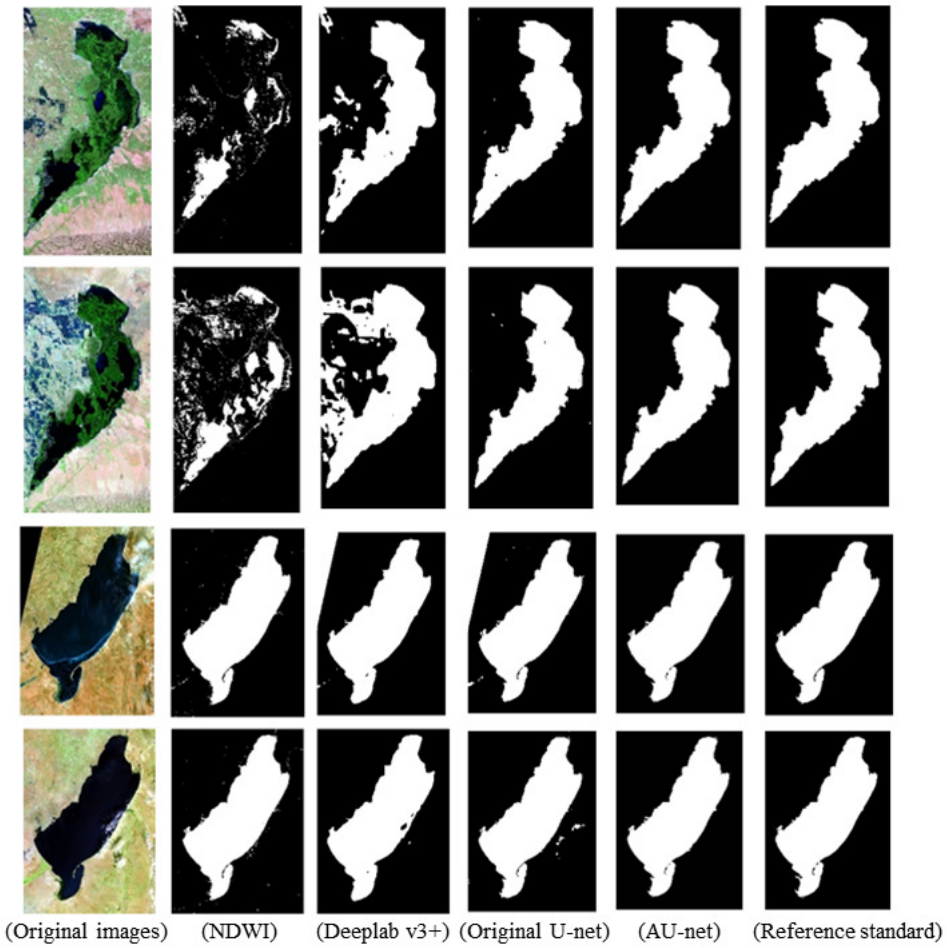

FIGURE 8. Result comparison of different methods.

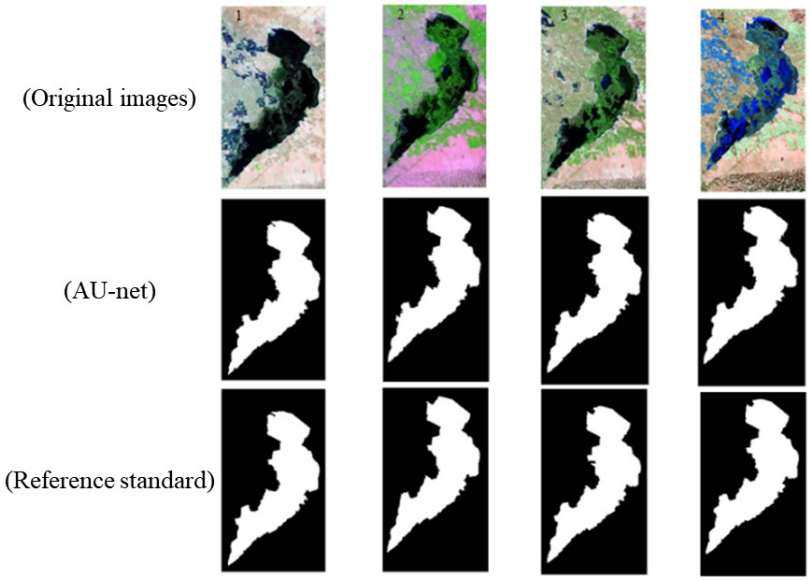

FIGURE 9. Extraction results of different seasonal images of Wuliangsuhai.

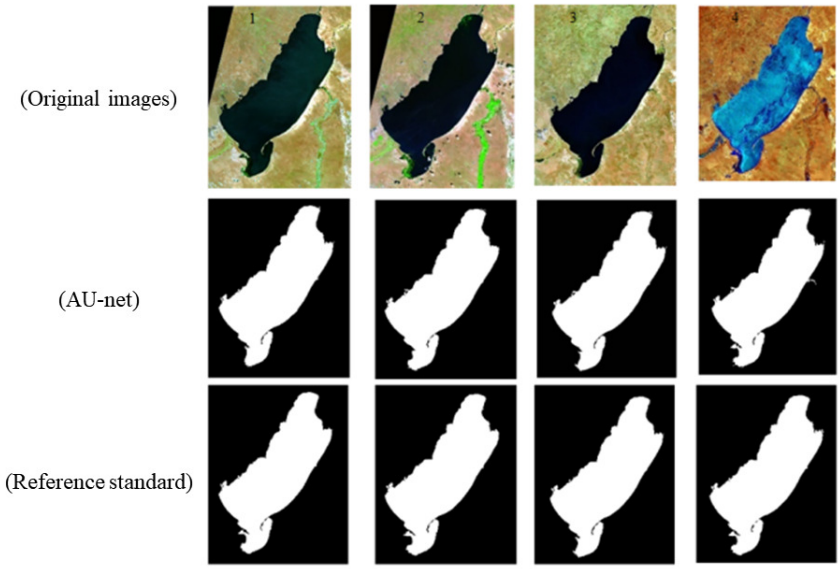

FIGURE 10. Extraction results of different seasonal images of Hulunhu Lake.

Lake and Hulunhu Lake in different seasons. The extracted water bodies of the lakes are clear and complete, and the agreement with the reference data is high.

The extraction results of Figure 9 and Figure 10 are evaluated using three evaluation metrics, PA, MPA and MIoU. Table 3 and Table 4 depict the metric values respectively.

The metric values are shown in Table 3 and Table 4 respectively.

TABLE 3. Image extraction results of Wuliangsuhai Lake in diverse periods

Seasons	**PA/%**	**MPA/%**	**MIoU/%**
1(Spring)	99.27	99.14	98.21
2(Summer)	99.46	99.51	98.65
3(Autumn)	99.34	99.28	98.36
4(Winter)	99.53	99.57	98.83

TABLE 4. Image extraction results of Hulunhu Lake in diverse periods

Seasons	**PA/%**	**MPA/%**	**MIoU/%**
1(Spring)	99.64	99.47	99.14
2(Summer)	99.68	99.53	99.25
3(Autumn)	99.80	99.83	99.53
4(Winter)	99.67	99.73	99.21

Table 3 and Table 4 show that PA and MPA reached over 99%, and the AU-net method is quite accurate on both images of the Wuliangsuhai Lake and Hulunhu Lake in different seasons. For the MIoU value, the image of the Hulunhu Lake area also reached more than 99%. However, in the image of Wuliangsuhai Lake with complex water color, it is slightly lower, but all of them are above 98%, and the overall accuracy is higher. All results above verify the effectiveness of the AU-net method.

5. **Conclusion.** A new water extraction method called AU-net is proposed in this paper in order to extract water body in cold-arid-zone lakes. Based on a revised U-net model and FCCRF, the AU-net method integrates the convolutional attention module CBAM into U-net to increase the adaptive capability of the network. And FCCRF is taken as the back-end optimization. In Comparison with NDWI, Deeplab v3+ and original U-net, as illustrated in the experiments, the AU-net method performs better in terms of elimination of complex background interference, removal of small area holes. Moreover, AU-net is superior to the other methods in terms of the three evaluation metrics, PA, MPA and MIoU. In the future, the AU-net method needs to be applied to the open datasets to explore its extraction efficiency. In addition, the data set exploited here remains to be extended so as to increase its diversity.

Acknowledgment. This work has been supported by the China National Science Foundation (6204121,61962047), China National Key Research and Development Program (2019YFC049205), Inner Mongolia Natural Science Foundation of China (2021MS06009, 2020MS06011), Basic research funds for universities directly under the Inner Mongolia Autonomous Region of China (BR22-14-05), and Inner Mongolia Autonomous Region science and technology plan project (2022YFHH0070).

REFERENCES

[1] R. Nagaraj and L. S. Kumar, "Multi scale feature extraction network with machine learning algorithms for water body extraction from remote sensing images," *International Journal of Remote Sensing*, vol. 43, no. 17, pp. 6349–6387, 2022.

[2] P. S. Frazier, K. J. Page *et al.*, "Water body detection and delineation with landsat TM data," *Photogrammetric Engineering and Remote Sensing*, vol. 66, no. 12, pp. 1461–1468, 2000.

[3] J. Lira, "Segmentation and morphology of open water bodies from multispectral images," *International Journal of Remote Sensing*, vol. 27, no. 18, pp. 4015–4038, 2006.

[4] Y. Zhang, X. Liu, Y. Zhang, X. Ling, and X. Huang, "Automatic and unsupervised water body extraction based on spectral-spatial features using GF-1 satellite imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 927–931, 2018.

[5] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[6] H. Xu, "A study on information extraction of water body with the modified normalized difference water index (MNDWI)," *National Remote Sensing Bulletin*, vol. 9, no. 5, pp. 589–595, 2005.

[7] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, pp. 1–15, 2019.

[8] J. Billson, M. S. Islam, X. Sun, and I. Cheng, "Water body extraction from Sentinel-2 imagery with deep convolutional networks and pixelwise category transplantation," *Remote Sensing*, vol. 15, no. 5, p. 1253, 2023.

[9] Z. Guo, L. Wu, Y. Huang, Z. Guo, J. Zhao, and N. Li, "Water-body segmentation for SAR images: past, current, and future," *Remote Sensing*, vol. 14, no. 7, p. 1752, 2022.

[10] M. Saghafi, A. Ahmadi, and B. Bigdeli, "Sentinel-1 and Sentinel-2 data fusion system for surface water extraction," *Journal of Applied Remote Sensing*, vol. 15, no. 1, pp. 014 521–014 521, 2021.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[13] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121–2133, 2022.

[14] A. L. H. P. Shaik, M. K. Manoharan, A. K. Pani, R. R. Avala, and C.-M. Chen, "Gaussian mutation–spider monkey optimization (GM-SMO) model for remote sensing scene classification," *Remote Sensing*, vol. 14, no. 24, p. 6279, 2022.

[15] T. Zhao, H. Li, H. Hu, and X. Fu, "Water body extraction from remote sensing image of Wuliangsuhai lake based on fully convolutional neural network," *Scientific Bulletin, Series C Electrical Engineering and Computer Science*, vol. 84, no. 2, pp. 185–202, 2022.

[16] J. Parajuli, R. Fernandez-Beltran, J. Kang, and F. Pla, "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*, vol. 15, pp. 6804–6816, 2022.

[17] R. G. Tambe, S. N. Talbar, and S. S. Chavan, "Deep multi-feature learning architecture for water body segmentation from satellite images," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103141, 2021.

[18] L. Li, Z. Yan, Q. Shen, G. Cheng, L. Gao, and B. Zhang, "Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks," *Remote Sensing*, vol. 11, no. 10, p. 1162, 2019.

[19] E. K. Wang, C.-M. Chen, M. M. Hassan, and A. Almogren, "A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain," *Future Generation Computer Systems*, vol. 108, pp. 135–144, 2020.

[20] K.-K. Tseng, R. Zhang, C.-M. Chen, and M. M. Hassan, "DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service," *The Journal of Supercomputing*, vol. 77, pp. 3594–3615, 2021.

[21] Z. Wang and B. Mei, "Current status and challenges of the ecological environment of wuliangsuhai basin in china," in *IOP Conference Series: Earth and Environmental Science*, vol. 829, no. 1. IOP Publishing, 2021, p. 012012.

[22] S. B. Z. S. L. Y. Z. M. YU Haifeng, SHI Xiaohong, "Analysis of water quality and eutrophication changes in Hulun lake from 2011 to 2020," *Arid Zone Research*, vol. 38, no. 6, p. 1534, 2021.

[23] L. H. D. L. QIAO Danyu, ZHENG Jinhui, "Application of water extraction methods from Landsat imagery for different environmental background," *Journal of Geo-information Science*, vol. 23, no. 4, p. 710, 2021.

[24] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, "Automatic water-body segmentation from high-resolution satellite images via deep networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 602–606, 2018.

[25] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[26] D. Wan and S. Yin, "Research on information extraction of the Dongting lake ecological wetland based on genetic algorithm optimized convolutional neural network," *Frontiers in Ecology and Evolution*, vol. 10, p. 944298, 2022.

[27] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[28] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.