Received September 2, 2019, accepted October 9, 2019, date of publication October 25, 2019, date of current version November 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949635

Multiscale Features Supported DeepLabV3+ Optimization Scheme for Accurate Water Semantic Segmentation

ZIYAO LI, RUI WANG, WEN ZHANG, FENGMIN HU, AND LINGKUI MENG
School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

Corresponding author: Lingkui Meng (lkmeng@whu.edu.cn)

This work was supported by the National Key Research and Development Program of China. (No. 2017YFC0405806).

ABSTRACT In the task of using deep learning semantic segmentation model to extract water from high-resolution remote sensing images, multiscale feature sensing and extraction have become critical factors that affect the accuracy of image classification tasks. A single-scale training mode will cause one-sided extraction results, which can lead to ''reverse'' errors and imprecise detail expression. Therefore, fusing multiscale features for pixel-level classification is the key to achieving accurate image segmentation. Based on this concept, this paper proposes a deep learning scheme to achieve fine extraction of image water bodies. The process includes multiscale feature perception splitting of images, a restructured deep learning network model, multiscale joint prediction, and postprocessing optimization performed by a fully connected conditional random field (CRF). According to the scale space concept of remote sensing, we apply hierarchical multiscale splitting processing to images. Then, we improve the structure of the image semantic segmentation model DeepLabV3+, an advanced image semantic segmentation model, and adjust the feature output layer of the model to multiscale features after weighted fusion. At the back end of the deep learning model, the water boundary details are optimized with the fully connected CRF. The proposed multiscale training method is well adapted to feature extraction for the different scale images in the model. In the multiscale output fusion, assigning different weights to the output features of each scale controls the influence of the various scale features on the water extraction results. We carried out a large number of water extraction experiments on GF1 remote sensing images. The results show that the method significantly improves the accuracy of water extraction and demonstrates the effectiveness of the method.

INDEX TERMS Remote sensing, deep learning, semantic segmentation, water information extraction, multiscales, DeepLabV3+.

I. INTRODUCTION

Accurate extraction of water information from remote-sensing images has always been an important research topic in the field of remote sensing image analysis because it plays a vital role in national land/water resource monitoring and environmental protection. The traditional remote sensing image water body extraction methods are divided into two main types. One is based on the spectral characteristics of water bodies and involves setting different thresholds and using a water body index to classify and extract water information [1]–[3]. The other is based on mathematical morphology [4] and segments water body edges [5], [6] to extract their features. Among the two types of methods, the index method (NDWI), which is a spectrum-photometric method, is widely used in the task of water extraction because of its simplicity and universality. In actual appliances, especially in the large-scale daily monitoring of water body nationwide, it is inevitable to process massive remote sensing images. The traditional methods with more manual intervention are not able to guarantee the quality of data products. In particular, there is substantial random interference in images, such as clouds, shadows, fog, etc. Even if some models have good generality, there will also be an extraction loss on the details of water. These will greatly affect our monitoring and use efficiency. Therefore, it is of far-reaching significance to study a water extraction model that can meet the requirements of high precision and high generalization

The associate editor coordinating the review of this manuscript and approving it for publication was Jingchang Huang.

ability and effectively reduce the need for human intervention. As deep learning in the field of computer vision has developed and improved, intelligent pattern recognition based on image spatial features has increasingly been applied to remote sensing image target detection and pixel semantic segmentation tasks [7]–[9]. Compared with traditional methods, semantic segmentation methods based on deep learning can map pixels to semantics [10] and mine spectral features of deep remote sensing images that traditional methods cannot extract. Many researchers use semantic segmentation models to achieve water body recognition and extraction tasks [11], [12]. Under the continuous optimization of complex neural network models, target feature mining and learning occurs at deeper levels; thus, image classification accuracy has improved continually. Moreover, in the remote sensing field, with the continuous improvement of image resolution, the textural details of surface features have also been revealed, and the accuracy requirements for water extraction have also continually improved. Nonetheless, the complexity of water background information combined with interference from complex homologous and heterogeneous (the same water body with different spectra) phenomena must also be considered. Consequently, adopting complex deep learning networks is very important to learn and identify the features of high-resolution images.

Compared with the commonly used images for semantic segmentation [13], [14], high-resolution satellite images have large differences in the number of spectral bands, the extent of the image range, and the target scale. In particular, the target range and shape of water bodies are uncertain: a large area of water can cover the entire image segment, while a small water body may be expressed by only one or two pixels in the image. In [11], a large-scale block size of 512 × 512 is used for the experiments because the author believes that large-scale analysis has a good effect on the remote sensing image segmentation model, allowing it to maximally perceive the acceptance domain. However, in our study of water extraction, we found that the model's generalization ability and extraction accuracy did not achieve the expected results when using a single-scale sample for training. A large sample scale may reduce the precision of the details extracted during the water extraction process, while small-scale samples are insufficient for identifying features clearly at the information macro level. Thus, scale selection and processing has become a primary problem that affects the deep learning water extraction process. Consequently, determining how to design a reasonable model that can not only perceive the remote sensing image water body background information and completely extract macroscopic large-area water bodies but also maintain the extraction accuracy of small water bodies has become a difficult research problem.

The main contributions of this paper are as follows: we provide a new technology flow for the use of semantic segmentation techniques to extract water features from high-resolution remote sensing images. We execute the study according to the three aspects of the image multiscale feature sampling principle, the neural network model structure optimization and multifeature fusion and form a set of standardized production processes with application and promotion value. The experimental results show that the accuracy index of this method is markedly improved and has good generalizability. This scheme focuses on multiscale features and uses a hierarchical multiscale splitting method combined with a stretching algorithm to enrich the detailed features of the samples, thereby preserving the multiscale features of the samples. We control both the scale of the sample data used for model training and the input and output scales of the image data during inference. Simultaneously, by changing the DeepLabV3+ network structure, the model can better perceive the global background context characteristics of the image block. Finally, the extraction results are optimized with a conditional random field (CRF) to meet the accuracy requirements of microscopic and macroscopic features of water bodies.

The remaining sections of this article are organized as follows: The second section introduces the problems and reviews related works from the literature. The third section introduces the DeepLabV3+ model application optimization scheme, including the image scale division specification, the application details of the optimized DeepLabV3+ model network structure, and the fully connected CRF. In the fourth section, we describe experiments and results comparison using a GF1 image as the research object. The experimental results and revealed problems are described in the fifth section, and we present conclusions in the sixth section.

II. RELATED WORKS

A. LIMITATIONS OF THE SPECTRUM-PHOTOMETRIC METHOD

In this section, we discuss the limitations of traditional methods. The spectrum-photometric method combined with the NDWI index is widely used as the main method for water body extraction [15]. The principle is to compute a suitable segmentation threshold based on a normalized ratio index, which is calculated using the green band and the near-infrared band. Many researchers have conducted extensive research on the selection of thresholds [16]–[18]. Some researchers have also used unsupervised classification methods (Isodata or K-means) for water classification based on the NDWI index [19]–[21].

Although this type of method performs well in local areas and involves simple calculations, some subtypes of the method cannot be applied to large regions or to entire images. Segmentation based only on information from two bands will result in a misclassification of similar nonwater pixels as water, especially in areas rich in complex background environments (buildings, etc.). Furthermore, due to the limitation of the spectral reflectance physical mechanism and the lack of correlation information between pixels in the analysis process, the phenomenon of "the same water body with different spectra" exists, which causes the spectrum-photometric method to be unable to achieve highly precise classification. Figure 1 shows the incorrect classifications of the NDWI index method for water segmentation.

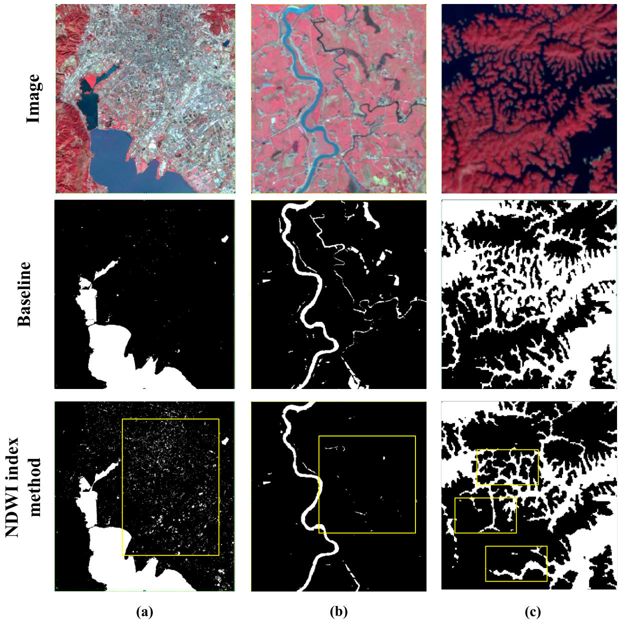

FIGURE 1. Limitations of the NDWI index method for water segmentation under different scenarios. Column (a) shows the errors in areas rich in complex buildings; Column (b) shows the errors extracted when the spectral differences of water bodies are large; Column (c) shows the losses in small water bodies and near edges.

Therefore, deep learning is introduced to supplement the user's need for highly precise water body classification.

B. REVERSE ERRORS

Semantic segmentation is a hot topic in deep learning areas and has been widely applied in computer vision and remote sensing image classification. However, the computational mode of a convolutional neural network may lead to insufficient extraction of background information [22]. Figure 2 shows the water extraction effect from a single remote sensing image using different types of semantic segmentation models and different image block sizes. We compared the results of water extraction at different scales using four different deep neural networks that are favorable for scale sensing (a fully convolutional neural (FCN) network [10], a feature pyramid network (FPN) [23], InceptionV3 [24], and DeepLabV3+ [25]. Through experiments and comparisons, we found more water porphyroclasts in the results of FCN network extraction and found that FPN and DeepLabV3+ have better classification effects than do FCN and InceptionV3, because their network structures consider scale induction. After training on the same data set, the DeepLabV3+ network is more conducive for extracting multiscale features due to its atrous convolution. DeepLabV3+ not only performs well for classification results at different scales but also completely extracts water body details.

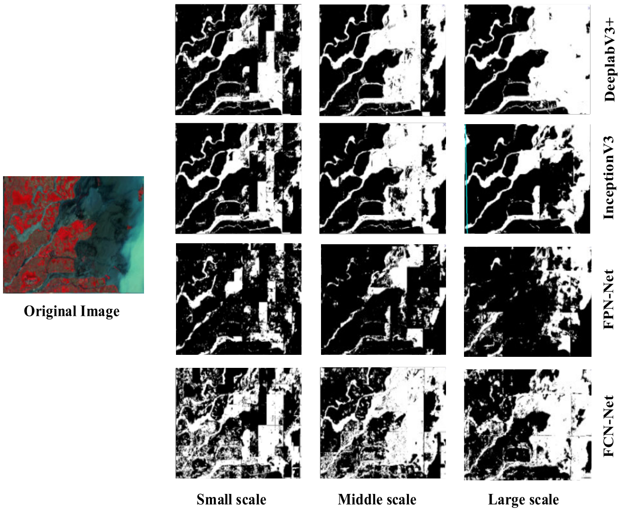

FIGURE 2. The effects of different network structure models on water extraction at different scales from the same image (DeepLabV3+, InceptionV3, FPN-Net, and FCN-Net).

We describe the results of water extraction at different scales. When small-scale image blocks are used to train a neural network model and extract water bodies, the resulting prediction accuracy for small rivers and discrete water bodies is high. However, for large-area water body recognition, the model is unable to perceive sufficient feature differences from the small image blocks. This can cause a complete misclassification of the larger water bodies. We call this type of error "reverse" error. In contrast, if the model is trained for water extraction with large-scale image blocks, the recognition accuracy for large-area water information improves but causes a loss of the fine water details, resulting in incomplete water extraction. To quote Chinese poetry: "The true face of Lushan is lost to my sight, for it is right in this mountain that I reside." The goal is to achieve a method that can perceive the background information about the water body in remote sensing images, allowing it to extract macroscopic large-area water bodies while also maintaining high extraction precision for small water bodies.

C. OPTIMIZATION METHOD

Various methods have been proposed to improve semantic segmentation model accuracy. Among these methods, optimizing their own network structures and adding auxiliary postprocessing modules are the two main approaches. The first approach, optimizing the network structure essentially simulates a more complex signal transmission path to achieve more accurate fitting of feature characteristics [26]. Among these, the use of multiscale features plays an important role in the success of semantic segmentation. Therefore, most researchers use the following two strategies [27] to optimize neural network models by exploiting multiscale features. The first strategy is to learn the multiscale features through the changes that occur in the internal network structure under the same image input scale. The features in these

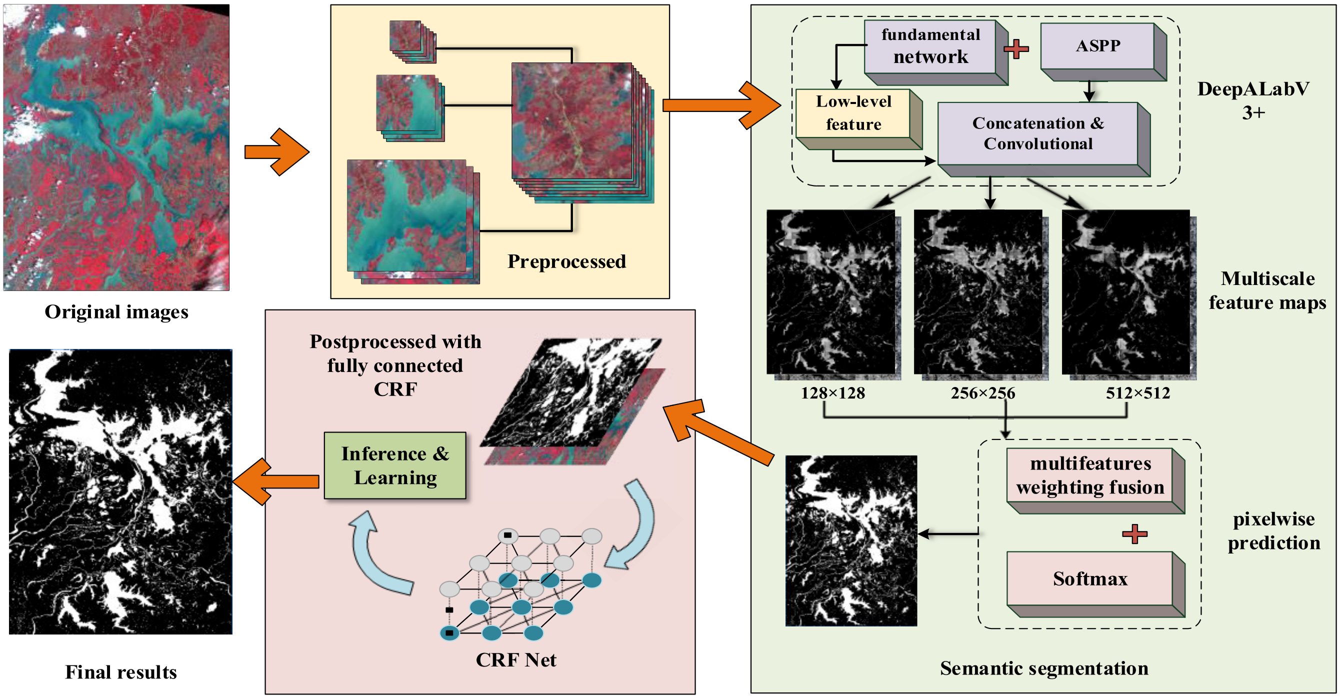

FIGURE 3. The outline of our methodology: first, we propose the multiscale image-splitting method, which focuses on background information; then, we design a DeepLabV3+ optimization model with multiscale feature fusion (the detailed network structure is shown in Figure 4). Finally, we apply a fully connected CRF postprocessing module for edge feature refinement.

networks are multiscale in nature because they use different receptive field and deconvolution sizes [28]–[30]. The representative model for the first strategy is the DeepLab series of neural network models [31]. The second strategy is the multiscale input learning strategy, in which after the input image has been resized at several scales, it is input to the same deep network and the resulting multiscale features are fused [32]–[34]. The FPN is the typical network structure used in this approach.

Another approach for improving accuracy is to add a postprocessing module (e.g., an SVM [35], GPM [36], or attention model [26]) after the deep network model. Combining a semantic segmentation model with a probabilistic graphical model (PGM) has become an application trend and is widely used in image classification tasks. Neural network models generally lack the ability to classify pixels along the edge contours of objects [22]. The postprocessing module optimizes the prediction results by analyzing the distribution probabilities of pixel characteristics based on the neural network model's prediction. Thus, the role of the postprocessing module is mainly that of optimizing the details and improving the discriminative ability of edge-contour pixels.

III. METHODS

In this study, we used a multiscale input specification for the image data. Correspondingly, we improved the multiscale fusion output method of the DeepLabV3+ neural network architecture and further optimized the prediction results by adding a CRF postprocessing module. Figure 2 shows the details of our method.

A. A MULTISCALE IMAGE FEATURE COLLECTION METHOD FOCUSING ON BACKGROUND INFORMATION

In this section, we discuss how to obtain multiscale feature information from RS images in detail. In remote sensing images, the distribution of objects is unbalanced and cannot be described by regular features. For example, a river water body may have a minimum width of 3–5 pixels in an image, but lake and reservoir water bodies will have widths of hundreds of pixels. This imbalanced object distribution must be considered during the neural network training process. To overcome the feature distribution imbalance, we mainly use two strategies, including obtaining context information of different scales from the original image and collecting the multiscale features of an image block.

1) THE HIERARCHICAL EXPANSION SPLITTING METHOD

With the first strategy, in deep learning, the context perception is often an important factor affecting convolution kernel sampling and is also a key link in multiscale feature extraction from remote sensing images. We define the maximum range for which a convolution kernel can carry out convolution calculation of an image as the context perception domain, that is, the size of the image block. The schematic is shown in Figure 4(a). When the context perception domain is too small, the context information of image blocks is more easily lost, which can lead to ‘‘reverse’’ errors, and will result in

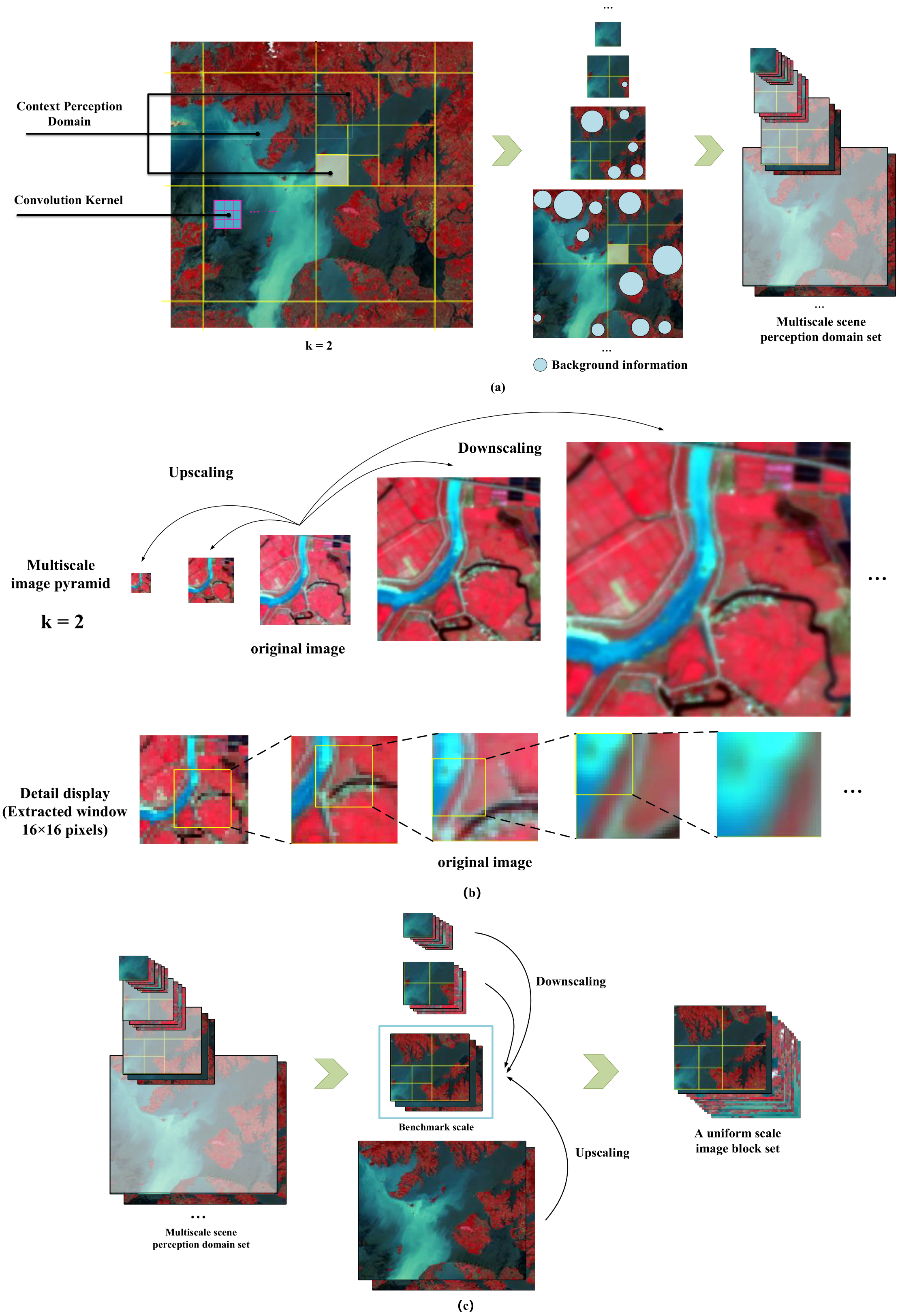
Context Perception Domain
Convolution Kernel
k = 2
Background information
Multiscale scene perception domain set
(a)
Upscaling
Downscaling
Multiscale image pyramid
k = 2
original image
Detail display (Extracted window 16×16 pixels)
original image
(b)
Downscaling
Benchmark scale
Upscaling
A uniform scale image block set
Multiscale scene perception domain set
(c)

FIGURE 4. Schematic of the multiscale image feature collection method. (a) The hierarchical expansion image-splitting steps; (b) the "Scale space" of a remote sensing image with local enlargements to display the details; (c) a scheme that considers the richness of the context perception domain and enhances multiscale detailed features.

a convolutional sampling number that is too low to extract sufficient detailed features. By contrast, a context perception domain that is too large may provide abundant global reference features but will also increase the computation load. Since the richness of the scene contextual information is related to the size of image splitting, we propose a hierarchical expansion splitting method to cut the image multiple times.

Let the minimum context perception domain (splitting size) of the primary splitting operation $S_0(x, y)$ be set to n × n pixels. Each expanded multiple is set to k, and the number of expansion operations is set to m. When performing the m-th expansion image-splitting step, the context perception domain will be expanded k^{2m} times; that is, the expanded split image blocks provide k^{2m-1} times the neighborhood scene contextual information of the minimum-size image block. The relationship between the size S of the context perception domain after expansion splitting and the values of m and k is expressed as equation (1). Then, we can obtain a set of image blocks with different scene contextual information richness levels. The image-splitting steps are shown in Figure 4(a).

$$S(x, y) = k^{2m} \cdot S_0(x, y) \quad k \in (0, +\infty) \tag{1}$$

2) SCALE SPACE OF A REMOTE SENSING IMAGE

In the remote sensing field, scale conversion is the main method used to extract different scale features of a single image block and transfer information from one scale to another [37], including upscaling (spatial resolution reduction) and downscaling (spatial resolution improvement) [38], [39]. We introduce the concept of "scale space" from the field of image processing to remote sensing image classification and incorporate the invariable single-scale image information processing technology into the variable dynamic analysis framework [40]. A "scale space" set can be constructed for each remote sensing image block, and the transferred image $L(x, y, k, \sigma)$ can be expressed as equation (2), where $f(k, \pm\sigma)$ denotes the interpolation operators, σ is a spatial scale factor indicating the number of scaling operations,k is a scaling multiple, and n is the pixel number of the rows and columns of the original image. $I(x, y)$ represents the original image block; that is, based on the spatial resolution of the current original image block, a multiscale image pyramid is constructed by the image interpolation algorithm. The details are shown in Figure 4(b).

$$L(x, y, k, \sigma) = \begin{cases} f(k, \sigma)\, I(x, y) & \sigma \in (0, +\infty) \quad \textit{upscaling} \\ f(k, -\sigma)\, I(x, y) & \sigma \in \left(0, \lceil \log_k n \rceil\right] \quad \textit{downscaling} \end{cases} \tag{2}$$

In the process of upscaling calculation, the spatial resolution of the image block is reduced by pixel-based fusion. This process will lead to the appearance of mixed pixels in the edge between the different categories but will reduce the spectral heterogeneity in the same ground object [41], which is advantageous for the extraction of large-area water bodies but harmful to the accuracy of small water body and water boundary distinction. The phenomenon of image mixed pixels becomes more serious with the composition σ, and when $\sigma = \lceil \log_k n \rceil$, the image is merged into a single pixel.

In contrast, downscaling is an operation that decomposes information at one scale into its constituents at smaller scales, which improves image resolution. After interpolation and enlargement, the number of pixels of local objects becomes larger, more pixels are classified into the same object, the number of mixed pixels at the edges of different categories is reduced, and the boundary of ground objects is smoother. During the m-th magnification k-multiples operation, one pixel is theoretically interpolated to $k^{2\sigma}$ pixels, which accommodates the sampling of a large receptive field convolution kernel and increases the extraction probability of detailed features. However, with an increase in σ, the spectral heterogeneity in the same category becomes larger, resulting in lower inter-class separability, which wastes computing resources and causes internal classification errors for the same object. The local enlargement details are shown in Figure 4(b).

3) THE INTEGRATED METHOD

To obtain a scheme that considers the richness of the context perception domain and enhances multiscale detailed features, we perform the hierarchical expansion splitting operation on an original image to obtain image block sets with different scene contextual information richness levels. An optimal scale in this set is used as the benchmark scale, and the image blocks of other scales are upscaled or downscaled to the benchmark scale according to the method of the spatial scale of remote sensing images. As shown in Figure 4(c), the image blocks smaller than the benchmark size are subjected to upsampling interpolation to downscale to the benchmark scale, while the image blocks larger than the benchmark size are subjected to downsampling interpolation to upscale to the benchmark scale. Accordingly, a uniform scale image block set can be obtained.

In fact, we maintain the scale diversity not only during model training but also during inference: the target image of the input model is also split at multiple scales. In this way, our model extracts multiscale features of the same image area multiple times. By combining the features extracted at each scale, the accuracy of the final model prediction results is guaranteed. After stretching and enlarging the small-scale images, the feature details will be clearly expressed. The large-scale image blocks contain rich background information of ground objectsčwhich overcomes the "reverse" error caused by the lack of local background features that is problematic in small-scale image classifications. However, overstretching the small features (such as rivers with two-pixel widths) in small-scale image blocks causes feature distortion, which in turn can cause cognitive errors in the model. To prevent such distortion, the medium-scale image block can be regarded as a transitional scale for the model. Fusing these three scale features makes it possible to achieve both comprehensive and fine extraction of water body information.

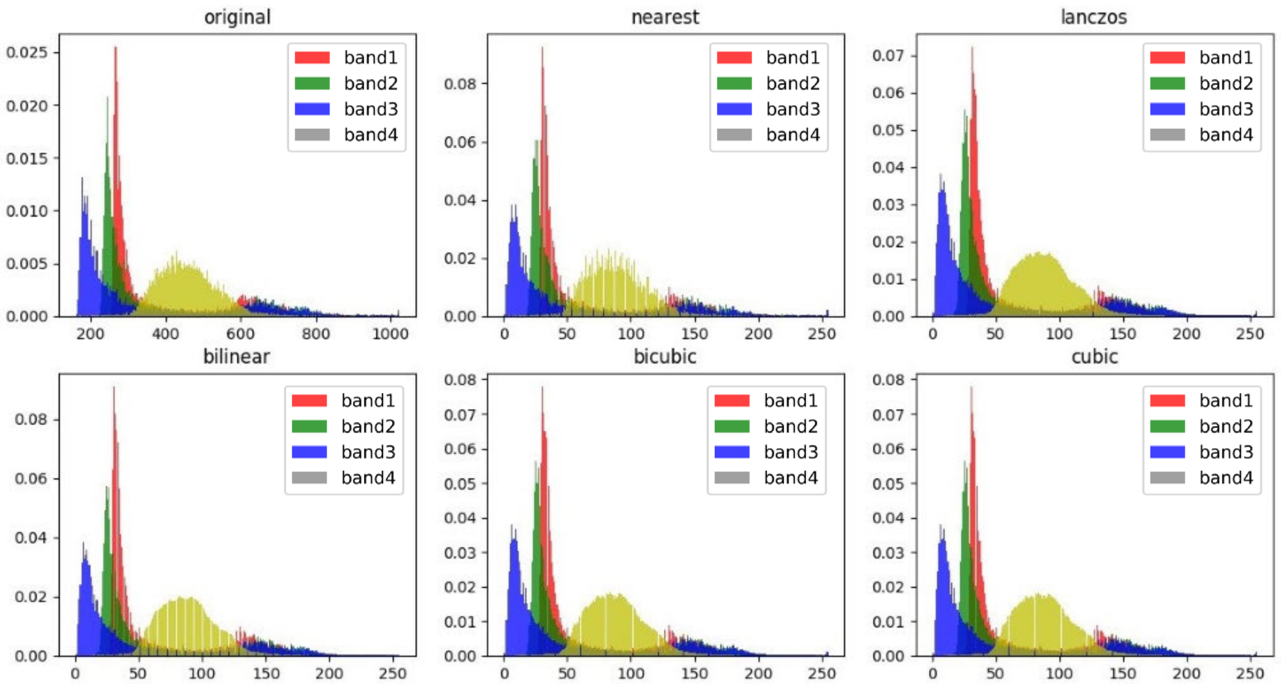

FIGURE 5. The pixel frequency distribution histogram of different interpolation algorithms. The Lanczos method achieves the most consistent pixel distribution.

To prevent abnormal features caused by the interpolation algorithms, we also evaluated the interpolation algorithms commonly used in the amplification process, which mainly include nearest neighbor interpolation [42], Lanczos interpolation [43], bilinear interpolation [44], bicubic interpolation [45], and cubic interpolation [45], [46]. The pixel distributions of remote sensing images should be continuous. However, if this continuity is destroyed during the process of image interpolation and enlargement, some image features may be lost, which leads to cognitive bias in the model. Figure 5 shows gray histograms of the studied remote sensing image amplified using different interpolation algorithms. Obviously, the gray histogram image resulting from the Lanczos interpolation is smoother. Thus, Lanczos interpolation better maintains the continuity of the image pixel values.

B. MULTISCALE FUSION PREDICTION MODEL BASED ON DEEPLABV3+

The DeepLabV3+ model is a variant of the typical FCN network [10, 47] that has achieved good performance in semantic segmentation using contextual information. DeepLabV3+ is the latest improved version of the DeepLab series of networks. The model inherits the atrous spatial pyramid pooling (ASPP) [48] module based on spatial pyramid pooling (SPP) [49] from DeepLabV3. On one hand, the model performs convolution operations by employing parallel atrous convolutions at various rates to capture contextual features at multiple scales [25]. On the other hand, the model uses an encoder-decoder structure [50]. Through its effective decoder module, the model can recover detailed object boundaries. Furthermore, it reduces the boundary loss problem found with traditional CNNs. Thus, the DeepLabV3+ model can perform a deeper analysis of the contextual features in a single image. After splitting the input images at multiple scales. we developed a modified version of the DeepLabV3+ architecture that can achieve feature fusion from multiscale feature maps and obtain primary water prediction results.

Figure 6 shows the model structure used for the experiments in this paper. The model still consists of two parts: an encoder and a decoder. The encoder module uses ResNet-50 [51] as the fundamental network. The network used in this paper consists of 5 convolutional layers (Con1–Con5), each of which contains a different number of bottleneck building blocks (for details of the structure, see [51].) A lower-level convolution result corresponds to a richer and larger context with higher resolution. As the convolutional layer deepens, the output features of convolutional computation reduce the spatial resolution of the image due to downsampling and pooling operations, resulting in loss of the initial global details. Therefore, to preserve a sufficient amount of initial global information, we selected the feature map from the first convolutional layer after pooling operations as low-level feature 1 and the feature map from the third bottleneck building block in the 2nd convolutional layer as low-level feature 2, which is then stretched via interpolation to the same size of low-level feature 1. Then, these two low-level feature maps are concatenated into one feature map that subsequently passes through a 1∗1 convolution with 64 channels. Thus, the number of output channels of the

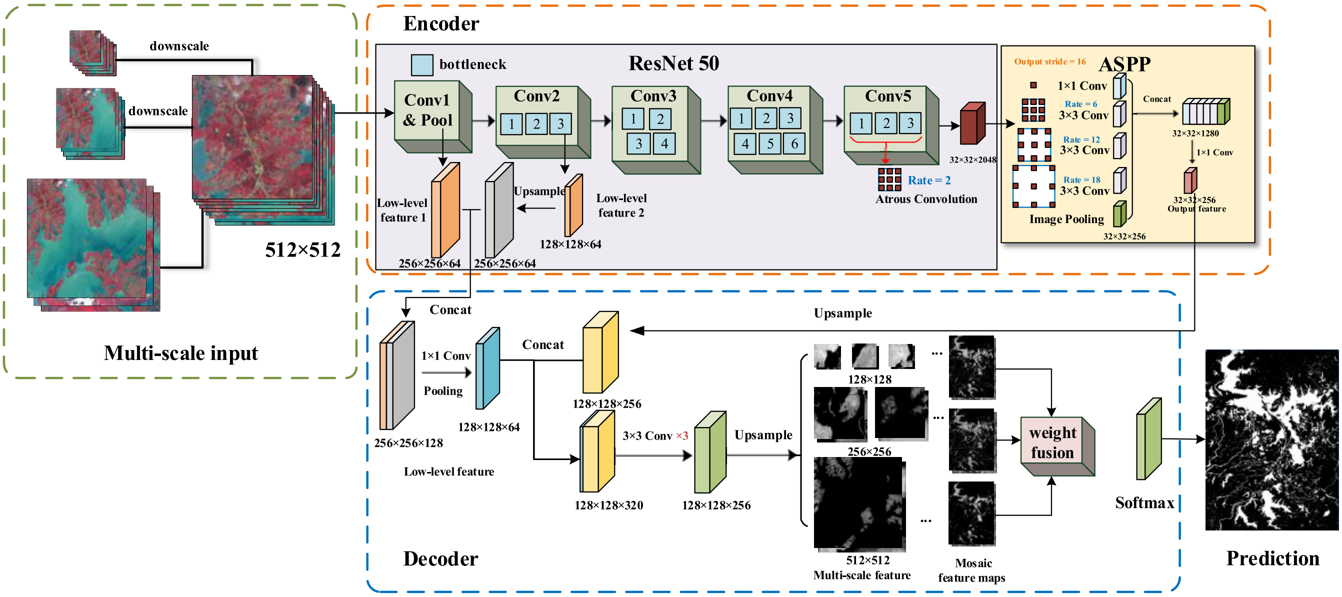

FIGURE 6. The deep learning network model structure, including the encoder (the fundamental Resnet network and the ASPP module) and the decoder (the upsampling steps and the weighted feature fusion) structure.

final low-level feature map is 64, which does not exceed the output channel of the ASPP module. After performing downsampling in the five convolutional layers, the ASPP module is connected. The output feature map from the fifth convolutional layer is regarded as the input feature map to the ASPP module. For more details of the ASPP module, see [48]. The five parallel computed feature maps in ASPP are concatenated into one feature map with 256 channels which then passes through a 1 × 1 convolutional layer before upsampling to the same size of the corresponding low-level feature map.

In the decoder module, to recover the water body segmentation details accurately, the output feature map of the ASPP module is concatenated with the corresponding low-level feature map and passed through a 1 × 1 convolution. A convolution calculation is performed on the connected features three times to obtain a new feature map with two channels. Then, the obtained features are upsampled to the different sizes of the corresponding input image blocks. This step is performed to adjust the segmentation logic size of feature map, making it possible to recover detailed image features. Through the decoder module calculation, we obtain the feature maps with 2 channels from all 3 input image blocks scales and mosaically splice the upsampled feature maps of the same size. We thus obtain three feature maps for the remote sensing images of the same size.

To obtain a more accurate classification result, the mosaic features of different scales must be fused to form a new feature map. In existing studies [32], [33], [52], extracted features from different scales are always merged and input to one classifier for classification [22], [35]. This approach leads to equal treatment of the features of each scale but does not highlight the advantages of the feature expression at each scale. Thus, according to the different impacts of different scale features on the final prediction result, we use the weighted fusion method to set different weights for different scale feature maps to perform feature fusion. This method can adjust the influence factors of the different scale feature maps to control the degree of refinement of water body extraction. The calculation is as follows:

$$F_k\left(X, k\right) = \sum_{s=1}^{s} \left(\omega_s \times D\left(X_s\right)\right) \tag{3}$$

where $F_k\left(X, k\right)$ denotes the resulting feature map, X is the prediction feature map of different scales, k is the number of classifications (this model is the binary classification model, thus, $k = 2$), s is the number of multiple scales in this model, and $s = 3$, ω_s represents the weights assigned to the prediction feature map for each scale, and $\sum_{s=1}^{S} \omega_s = 1$; $D(X_s)$ represents the feature map matrix of the s-th scale prediction, which is a two-dimensional matrix. We test and discuss setting the weights ω_s, in Section 4 Part C. Finally, a softmax normalization (inormalized exponential function) is performed on the output feature map after the weighted fusion operation, which guarantee that the output is a probability distribution and obtains the pseudoprobabilities of the class labels. Finally, we use a classification rule to determine the final label for each pixel—that is, pixels with the greatest probability of being a water class are labeled as water.

C. POSTPROCESSING OPTIMIZATION METHOD BASED ON A FULLY CONNECTED CRF

In this section, the coarse classification map obtained after the prediction of the multiscale convolutional neural network model is used as the prior probability to calculate the maximum posteriori probability of the remote sensing image

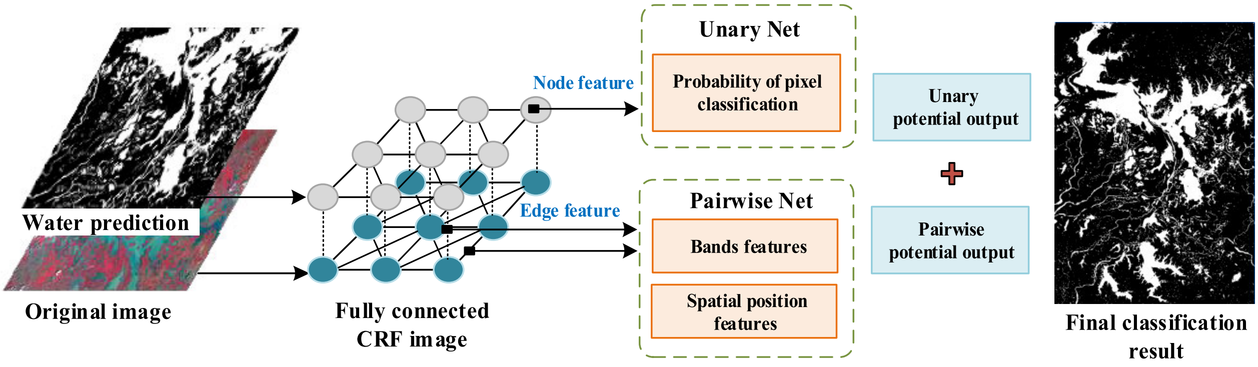

FIGURE 7. An illustration of the fully connected CRF. First, we construct a CRF graph model based on the classification results after the deep learning network and then apply a unary potential network to each pair of nodes and a pairwise potential network to each edge between node pairs to finally obtain the potential function outputs.

classification of each pixel using the fully connected CRF method [53]. Because the atrous convolution and pooling operations will lose some features in the calculation process and the feature map calculated by the neural network model is stretched due to upsampling, the water body details are not finely expressed. Moreover, no spatial regularization is used in the conventional segmentation method based on pixel classification, which leads to a lack of spatial consistency in the semantic segmentation results [54], [55]. The fully connected CRF can synthetically utilize the spatial information of remote sensing images to obtain spatially consistent results [56] and refine the upsampled coarse prediction results to sharpen the water body boundaries and details. Through this postprocessing module, we established a unary potential function and a pairwise potential function for all pairs of pixels in the image to achieve the maximum fine segmentation. The process is illustrated in Figure 7.

We define X as the original input image to the model, $x_i \in X$; and Y is the prediction label mask based on the result of the deep learning network model, where $y_i \in Y$. Here, y_i and x_i have a one-to-one correspondence, forming each node in the CRF graph model. Thus, the joint conditional probability for one image is expressed as follows:

$$P(Y, X) = \frac{1}{Z} \exp\{-E(Y, X)\} \tag{4}$$

where Z is the partition function, and $E(Y, X)$ is the energy function of the fully connected CRF. In this formula, the energy function under the fully connected nodes condition is defined as the sum of a unary term and a pairwise term.

Similar to most inference methods used in fully connected CRF, we apply the mean field approximation algorithm [52] to perform inference. During fully connected CRF training, we apply piecewise training [57] in the postprocessing module to iteratively train each parameters, until the value of the likelihood function is maximized. Eventually, the pixels that have the greatest probability of belonging to water classes are marked as water. Thus, we are able to obtain precise water extraction results from remote sensing images.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL FRAMEWORK DESIGN

To verify the generalizability and effectiveness of the proposed method, we designed two types of experiments to show the method's performance in the water extraction task. First, the study area is classified into three types, including a large-area water body region, a small-area water body region, and a mixed complex water body region, to test the effect of this scheme on water body extraction at various scales. Second, to ensure the accuracy of the comparison experiments, a parallel experiment and a self-step optimization experiment are adopted. The design of the experimental process is shown in Figure 8.

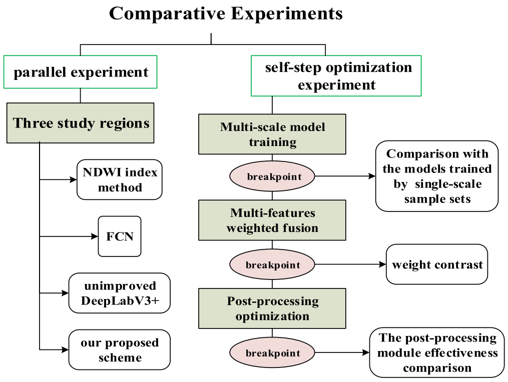
Comparative Experiments
parallel experiment
self-step optimization experiment
Three study regions
NDWI index method
FCN
unimproved DeepLabV3+
our proposed scheme
Multi-scale model training
breakpoint
Comparison with the models trained by single-scale sample sets
Multi-features weighted fusion
breakpoint
weight contrast
Post-processing optimization
breakpoint
The post-processing module effectiveness comparison

FIGURE 8. The experimental process design.

Here, in the parallel experiment, water bodies from three regions of China are arranged for method versatility verification. In the self-step optimization comparison experiment, we set experimental breakpoints, including comparison of the models trained by the multiscale and single-scale sample sets, the contrast experiment of feature weighted fusion and the postprocessing module effectiveness comparison experiment. Through this process, the effectiveness of the optimization strategy proposed in this paper is shown.

1) STUDY REGIONS (ROIS) AND EXPERIMENTAL DATA

A large-area region denotes a region where the water surface area is regular and large, and water features are predominantly extracted from large lakes and reservoirs. We select Kunming City in southwestern China as the study area, which includes three large lakes: the Dian Lake, Fuxian Lake and Yangzonghai Lake.

A small-area region usually denotes a region with small rivers, which is covered by only a few pixels in an image. In this paper, part of the Lancang River and the Heihui River basin is selected as the study area for extracting small water bodies.

A mixed complex region is an area with complex water systems, rich backgrounds and diverse types of water bodies. To test the comprehensive extraction ability of the model, we select the Poyang Lake in southeastern China as the study area.

The data sources are high-resolution remote sensing images with 16-m resolution acquired by the GF1 satellite WFV sensor, which includes 4 bands, including blue, green, red and near-infrared, and covers 3 study areas. Based on the polygon vector data of rivers and lakes in the First National Census for Water, this study used an automatic extraction method combined with manual verification to complete the water sample labeling. The image label maps after manual visual interpretation were sufficiently accurate; therefore, we regarded the water labeling results as the mapping labels (ground truth) for the sample image and as the baseline for the test image.

In this paper, 512 pixels are taken as the benchmark scale according to the resolution of the GF1 images and the various factors mentioned in [58]. This scale image block is rich in sufficient background scene information and is the most suitable scale for recognition of differences by the human eye. Therefore, the complete sample and test remote sensing images are split into 3 levels according to Section 3 Part A, where $m = 3$, $k = 2$, the minimum image block pixel is 128×128, and three scale image blocks (128×128, 256×256, and 512×512) were obtained. Thus, we obtained the initial sample set and test set. Then, for the sample set, we filter the initial sample set by first skipping the "empty" image blocks in which more than 95% of the pixels are labeled as nonwater. Next, to enhance the water body expression details and extract the complex features from the raw pixels more efficiently, we adjusted the proportions of the sample patches at the different scales such that the quantitative ratio between the three scales of the sample patches (128 × 128, 256 × 256, and 512 × 512) was approximately 5:3:2. In the sample set, a special case of a "wholly water" image block in which each pixel is labeled as water will be included. We define such patches as "whole water" samples. We retained the "whole water" samples in the sample set to increase the sample richness. Then, we performed an upsampling interpolation calculation (downscaling operation) twice on the small-scale patches (128 × 128) and once on the mesoscale patches (256 × 256) to obtain the benchmark scale (512 × 512). Therefore, the sample set and test set, which are rich in multi-level scale features, are ready.

In addition, we designed a contrast test to compare the effects of a multiscale sample set training model and a single-scale sample training model. We used small-scale image blocks (128 × 128 pixels) to construct a single-scale sample set. For convenience, the 128 × 128 pixel patches in the multiscale sample set are used as a single-scale sample set, which also contains the same number of "wholly water" patches and is upsampled to the benchmark scale. The composition of the experimental data set is shown in Table 1.

TABLE 1. Sample set composition.

Sample Set	*Scale*	*Num*	*"Wholly Water" Num*	*Patches Num*
Multiscale sample set	128×128 256×256 512×512	5,430 3,285 2,201	452	10,916
Single-scale sample set	128×128	5430	452	5,430

2) ACCURACY EVALUATION INDEXES

Background nonwater on images covers a large proportion of all pixels. The number of pixels predicted to be water is much smaller than the number of pixels predicted to be nonwater, reflecting an imbalanced classification model [59]. Therefore, we select three precision indexes to evaluate the performance of the method: Pixel Accuracy (PA), Recall, and Intersection Over Union (IOU), with IOU as the main index used to measure accuracy.

PA is used to calculate the ratio of correctly classified pixels of water to the total number of water pixels in the baseline image; we always refer to this metric as precision.

Recall is a metric that calculates the ratio of correctly classified pixels of water to the total number of water pixels in the predicted image.

IOU is a standard measure in the semantic segmentation field that calculates the ratio between the intersection and the union of two sets in each class.

$$\begin{aligned} \mathrm{PA} &= \frac{T_W}{T_W + F_W} \\ \mathrm{Recall} &= \frac{T_W}{T_W + F_N} \\ \mathrm{IOU} &= \frac{T_W}{T_W + F_N + F_W} \end{aligned} \tag{5}$$

where T_W (true water) denotes the number of pixels correctly classified as water; T_N (true nonwater) denotes the number of pixels correctly classified as nonwater; F_W (false water) denotes the number of pixels of the nonwater classes labeled as water; F_N (false nonwater) denotes the number of pixels of water classes classified incorrectly as nonwater pixels.

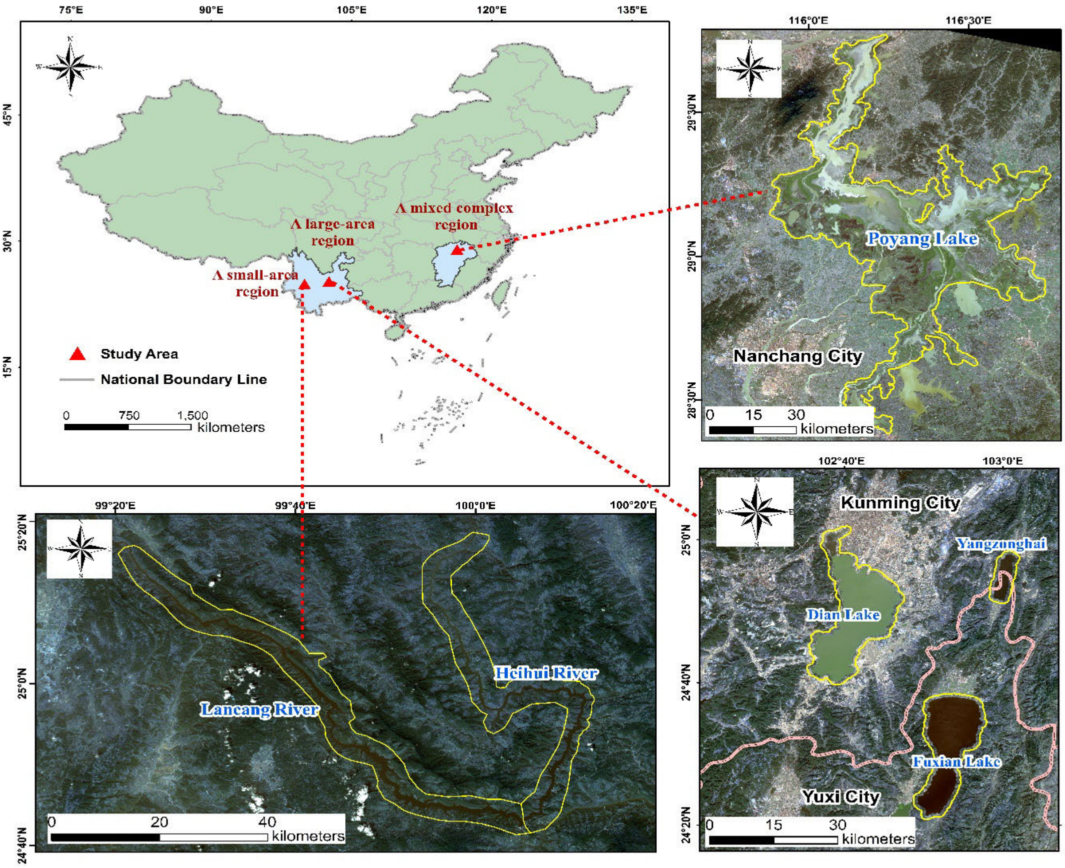

FIGURE 9. Three ROIs: A large-area region, a small-area region, and a mixed complex region.

B. PARALLEL EXPERIMENTAL RESULTS

We entered the test images of the three ROIs into the model trained in this paper and compared the results with the predictions of the traditional spectral-photometric method (NDWI), the classical neural network model FCN, and the unimproved DeepLabV3+ model. Additionally, we evaluate accuracy according to a manually verified baseline. The accuracy scores are shown in Table 2. Figure 10 shows the details of the different methods for water extraction.

The experimental results show that our proposed scheme outperforms the other models in multiple aspects, and improves the accuracy of the results. Compared with the unoptimized deep learning classification method, our method is more intelligent, and it improves the classification accuracy of the details of water bodies. Compared with the traditional method, the classification accuracy has also been greatly improved, especially in areas with rich complex background information and large-area water body regions. Moreover, the method is shown to have good generalizability through its application to different study areas.

C. SELF-STEP OPTIMIZATION EXPERIMENTAL RESULTS

1) COMPARISON OF THE MODELS TRAINED BY THE MULTISCALE AND SINGLE-SCALE SAMPLE SETS

To verify the effectiveness of the multiscale sample hybrid training strategy, we use two prepared sample sets and trained the same network configuration to obtain two models, a multiscale model and a single-scale model. During inference, the prepared multiscale test set is fed into the two models for prediction, and the prediction process stops before the feature weighted fusion. Here, we execute the normalized exponential function (softmax) on the three feature maps of different scales. Then, we mosaic the water prediction

TABLE 2. The accuracy scores of four methods for PA, Recall, and IOU.

Study Area	*Model*	*PA*	*Recall*	*IOU*
large-area region	NDWI	0.93764	0.94051	0.91260
	FCN	0.93939	0.90196	0.86664
	Unimproved DeepLabV3+	0.94324	0.95805	0.93378
	our scheme	0.97094	0.97725	0.96887
small-area region	NDWI	0.92833	0.97099	0.90325
	FCN	0.93898	0.97756	0.91966
	Unimproved DeepLabV3+	0.94963	0.97412	0.93606
	our scheme	0.94933	0.97655	0.93749
mixed complex region	NDWI	0.90663	0.86670	0.85241
	FCN	0.90844	0.87905	0.85420
	Unimproved DeepLabV3+	0.92077	0.92437	0.87641
	our scheme	0.93132	0.92230	0.88501

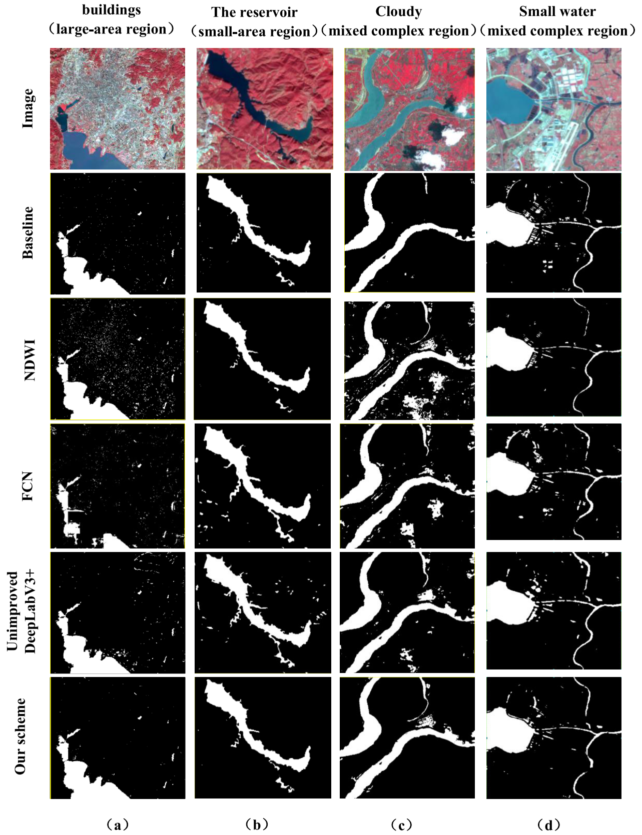

FIGURE 10. Some typical examples showing the details of the different methods for water extraction: (a) the water in background areas with buildings (from the large-area study region); (b) reservoir extraction (from the small-area region); (c) extraction under clouds (from the mixed complex region); (d) small water bodies (from the mixed complex region).

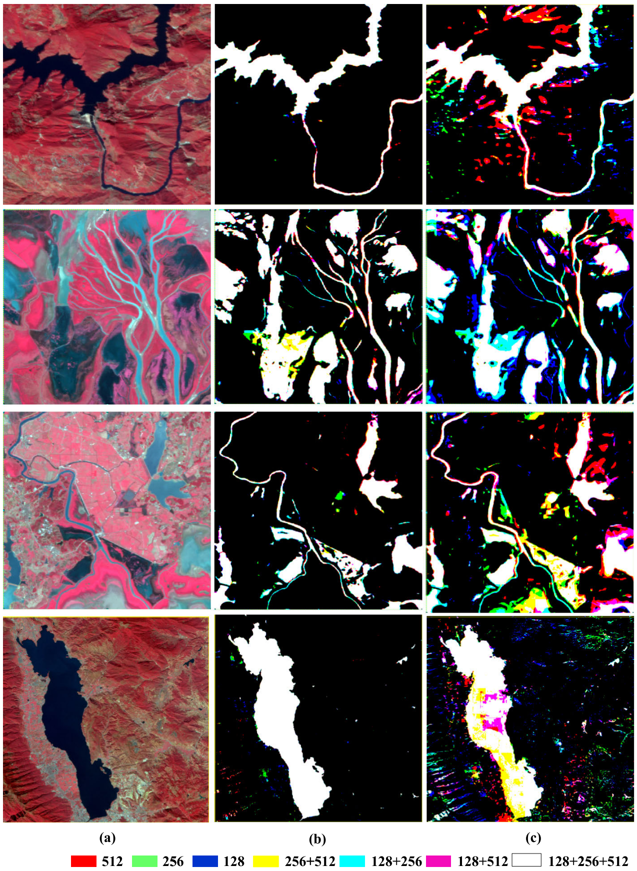

FIGURE 11. Details of the extraction results for different models on the Poyang Lake image: (a) the original image in false color; (b) multiscale model prediction results in pseudo color; (c) single-scale model prediction results in pseudo color.

TABLE 3. Water extraction accuracy scores of two models (a multiscale model and a single-scale model) for PA, Recall, and IOU.

Scale	*Model*	*PA*	*Recall*	*IOU*
512	multiscale	0.95752	0.96455	0.92372
	single-scale	0.92926	0.94593	0.89890
256	multiscale	0.95603	0.97302	0.93278
	single-scale	0.90791	0.92368	0.88570
128	multiscale	0.95237	0.97169	0.92798
	single-scale	0.90191	0.90475	0.87025

results of the same-scale image blocks and obtain three water prediction images of the same size. We stack the three-scaled predicted results into a pseudo color image by assigning different colors to each of the scale-classified images. In this image, the water pixels in the 128-scale prediction images are colored blue, the 256-scale prediction images are colored green, and the 512-scale prediction images are colored red. Using this approach, when combined, pixels that are all predicted to be water in the 128-scale and 256-scale prediction images are cyan, and the pixels that are all predicted to be water in the 128-scale and 512-scale prediction images are magenta, the pixels that are all predicted to be water in the 256-scale and 512-scale prediction images are yellow, and the pixels that are predicted to be water in all three scale-prediction images are white. The resulting pseudo color images are shown in Figure 11. Through band composition, we can clearly see the differences between the prediction results of the two models for different scale image blocks. To quantitatively describe the performance of the multiscale model and the single-scale model, we calculated the accuracy of the predicted results against the baseline, and the result is shown in Table 3.

Figure 11 presents a comparison of the water extraction details between the two models. We conducted a qualitative analysis based on these results. First, from Table 3, the accuracy of the multiscale models is higher than that of the single-scale models under any scale-prediction size. Second, from Figure 11, we find that the single-scale models performed poorly for feature extraction of large-scale image blocks. The extraction of water details relies primarily on the 128-scale prediction image, while the results of the other two scale images are coarse: many water body details and boundary extractions are incomplete; thus, they do not perform as well as supplementary classification images. In contrast, using the multiscale model, the prediction images at the three different scales are relatively consistent, and the fitting degree is high. The model trained by a multiscale sample set is able to extract water bodies satisfactorily at various scales, and the extracted water bodies are complete and continuous. From another viewpoint, comparing the prediction results at different scales, we find that a few ‘‘reverse’’ errors still occur in the 128-scale prediction results. but this problem is greatly reduced in the 256-scale and 512-scale prediction results.

2) EXPERIMENTAL RESULTS OF WEIGHTED FUSION

The previous section explored the influences of different sample sets on the prediction results of different scale test images. Multiscale model prediction results provide different feature angles for water extraction. In Section 3 Part B, we proposed a weighted fusion algorithm that assigns different weights to the feature maps of the three scales, fuses them into a new feature map, and then classifies them on this final feature map basis. In this section, we test the impact of different weight combinations on classification accuracy to obtain the weight combination most suitable for precise water body segmentation.

In this process, we used a weight stride of 0.1. We assess the accuracy of the classification results obtained by feature fusion against the baseline labeled image for each weight combination. The calculated accuracy scores are shown in Table 4, and the distribution of each indicator is shown in Figure 12.

The experimental results show the influence of different weight assignments on the classification results and show that the weights of small-scale features are not ‘‘the higher, the better’’ when considering the effect of water extraction. When the weights of the 128-scale feature maps increase, the models extract large amounts of water porphyroclasts, which reduces the accuracy of image classification. The highest precision (PA) is achieved at the weight combination where 512/256/128 = 0.4/0.3/0.3, and the IOU value is also the highest. The classification result using the weight combination of 512/256/128 = 0.4/0.3/0.3 is closest to the baseline, and the effect is optimal. Therefore, the method proposed in this paper uses this weight combination as the final weight setting.

TABLE 4. Comparison of different weight combinations.

Combination of Weight			*Evaluation Indexes*		
512×512	*256×256*	*128×128*	*PA*	*Recall*	*IOU*
0.1	0.1	0.8	0.92941	0.96612	0.90642
0.1	0.2	0.7	0.92966	0.97181	0.91118
0.1	0.3	0.6	0.93184	0.97643	0.91676
0.1	0.4	0.5	0.93342	0.97906	0.92025
0.1	0.5	0.4	0.93517	0.98237	0.92451
0.1	0.6	0.3	0.93517	0.98444	0.92619
0.1	0.7	0.2	0.93499	0.98414	0.92581
0.1	0.8	0.1	0.93837	0.98213	0.92713
0.2	0.1	0.7	0.93000	0.97171	0.91135
0.2	0.2	0.6	0.93468	0.97643	0.91922
a0.2	0.3	0.5	0.93539	0.97916	0.92207
0.2	0.4	0.4	0.93956	0.98054	0.92687
0.2	0.5	0.3	0.93690	0.98486	0.92811
0.2	0.6	0.2	0.94003	0.98253	0.92895
0.2	0.7	0.1	0.93832	0.98280	0.92760
0.3	0.1	0.6	0.93432	0.97538	0.91804
0.3	0.2	0.5	0.93721	0.97861	0.92318
0.3	0.3	0.4	0.93822	0.98240	0.92721
0.3	0.4	0.3	0.94171	0.98282	0.93065
0.3	0.5	0.2	0.93972	0.98382	0.92972
0.3	0.6	0.1	0.93932	0.98344	0.92899
0.4	0.1	0.5	0.93692	0.97837	0.92269
0.4	0.2	0.4	0.94206	0.97896	0.92764
0.4	0.3	0.3	**0.94325**	**0.98805**	**0.93378**
0.4	0.4	0.2	0.94312	0.98132	0.93056
0.4	0.5	0.1	0.93907	0.98237	0.92787
0.5	0.1	0.4	0.93988	0.98022	0.92671
0.5	0.2	0.3	0.94063	0.98327	0.92993
0.5	0.3	0.2	0.94078	0.98185	0.92888
0.5	0.4	0.1	0.93990	0.98099	0.92738
0.6	0.1	0.3	0.93842	0.98187	0.92673
0.6	0.2	0.2	0.94121	0.98197	0.92933
0.6	0.3	0.1	0.93857	0.98252	0.92741
0.7	0.1	0.2	0.93642	0.98207	0.92506
0.7	0.2	0.1	0.93659	0.98116	0.92447
0.8	0.1	0.1	0.93609	0.97940	0.92250

TABLE 5. Accuracy comparison before and after adding the fully connected CRF.

Classification	*PA*	*Recall*	*IOU*
result without CRF	0.943247	0.98605	0.93378
result with CRF	0.950733	0.97874	0.94063

3) EXPERIMENTAL RESULTS OF POSTPROCESSING OPTIMIZATION WITH A FULLY CONNECTED CRF

The prior experiments verified that the classification map accuracy after feature fusion is already sufficiently high, but the prediction results for very small water details (such as paddy fields, canals, etc.) still need to be refined. In particular, the models do not perform well for boundary segmentation of small water bodies under complex background environments. As described in Section 3 Part C, we treated the feature maps after multiscale feature weighted fusion as the prior probability and input them into a fully connected CRF model to obtain the final optimized classification results. The calculated accuracy. before and after including the fully connected CRF is shown in Table 5. The experimental result shows that the PA value and the IOU value are improved by this approach, but the Recall value decreases. We discuss this problem further in Section 5 Part C. Figure 13 clearly shows the detailed classification results of water bodies under a context with six types of objects: (1) water-body edges, (2) the water under building environments, (3) paddy fields, (4) shallow water beaches, (5) water under cloud shadows, (6) large-area water bodies and (7) small rivers.

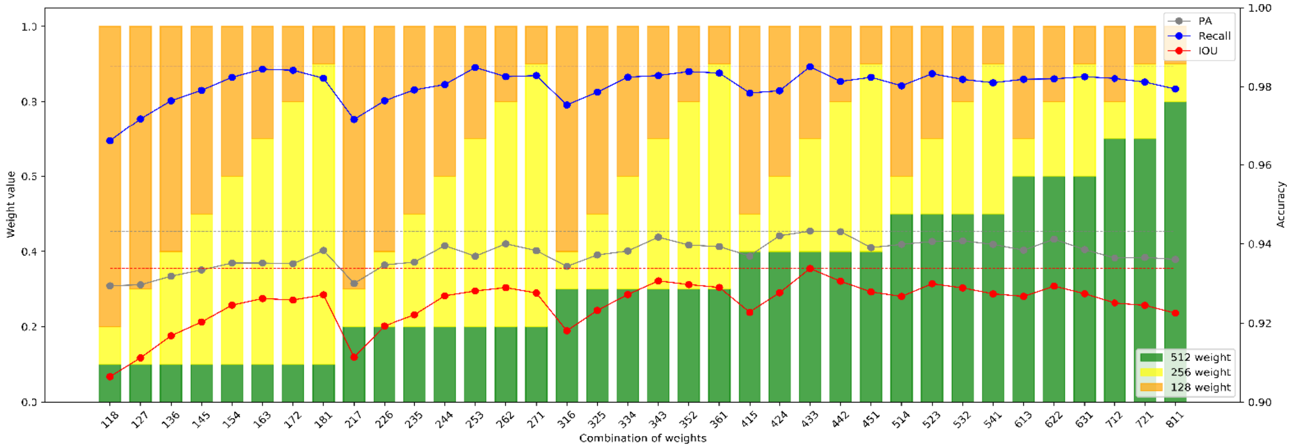

FIGURE 12. Classification results after feature fusion using different weight combinations.

As the comparison shows, the effect of the fully connected CRF for optimizing the details and for boundary refinement is quite obvious. The extracted water bodies are both complete and continuous. As shown in Figure 13 (1/3/5/7), our proposed method performs well for water body extraction at the edges of water bodies, under cloud shadows and in paddy fields. The segmentation effect exceeds even the visual interpretation of the manually checked baseline. Figure 13 (2/4) shows that water body extraction for shallow water beaches and under complex building environments also achieves a good performance. Using the CRF module, a large number of water body porphyroclasts are removed from the water extraction results, making them more complete. In particular, as shown in Figure 13 (6), the extraction result for the large-area water body is complete, and no "reverse" errors exist.

V. DISCUSSION

A. EFFECTIVENESS OF A MULTISCALE SCHEME FROM TRAINING TO PREDICTION

By comparing the expressions of the two models obtained by training the two types of sample sets, we find that the multiscale image input training method effectively improves model learning and expression ability for different scale image features, and the feature expression of the multiscale sample set training is more generalizable. The experimental results show that the prediction accuracy when using large-scale blocks and multiscale-trained models is higher than that from single-scale-trained models and that the fitting degree (overlap) of the three scales is high, which is beneficial in cross-assisted classification. By adding the scaled input prediction learning mode and using the ASPP module in the network architecture, this method can fully exploit the feature details of each sample at the different scales to obtain feature mappings of different scales. Then, the effects of the different scales can be adjusted through weighted fusion.

Regarding the influence of different weight settings on classification accuracy, we can conclude that the weight setting process is more reasonable and the precision of the classification results is greater. The fusion prediction result can effectively solve the "reverse" error problem and meet the requirements for greater precision. This experiment fully demonstrates the advantages of training the multiscale model, that is, it results in small-scale enhancement of detail segmentation, a large-scale perception of water body background information, and an improvement in large-area water body extraction results.

B. EFFECTIVENESS OF POSTPROCESSING WITH FULLY CONNECTED CRF

From the precision of experimental results, the PA value of water extraction accuracy predicted by neural network models has reached 94%. However, when assessing the details, the results of water body edge extraction in complex terrains is still insufficiently accurate because the boundaries become blurred due to upsampling operations. By adding the fully connected CRF optimization, the experimental results show that the method can well distinguish small water bodies—even those that are only one pixel wide—and the prediction results also eliminate a large number of misclassified nonwater body porphyroclasts. This is crucial for water extraction; nevertheless, the elimination of some patches inevitably leads to the losses when extracting small rivers. This problem is related to the potential energy function calculation of the CRF. The CRF model used in this study integrates only the pixel position and band information relationship of the image. Subsequent research should include more feature information and additional relationships in the CRF process to assist in the optimization of water edge classification

C. DISCUSSION ON CLASSIFICATION ACCURACY

The experimental area used in this study is complex and surrounded by paddy fields, urban areas and fishponds. There are many water-containing areas and complex forms that are very troublesome for water extraction algorithms. Through the comparison experiments described above, we found that even manually and accurately checked data labels include some classification errors; some number of incorrect labels

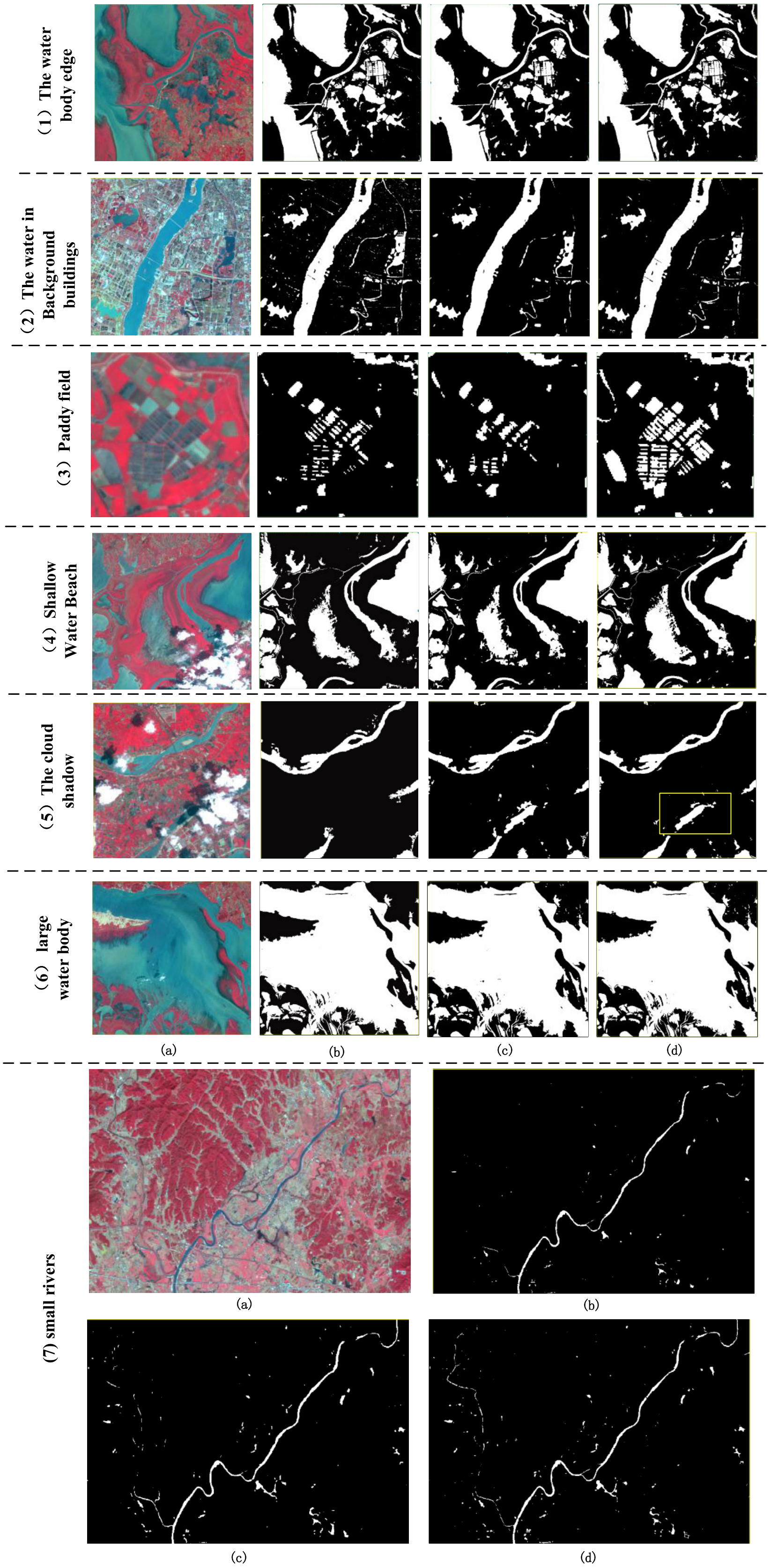

FIGURE 13. The before-and-after results of adding the fully connected CRF: (a) the original image in false color; (b) the manually checked baseline; (c) the classification results after weighted feature fusion; and (d) details of the postprocessing results after adding the fully connected CRF.

is inevitable. As shown in Figure 13 (2), some nonwater patches in the complex urban area are marked as water. In Figure 13 (5) the water body under the thin cloud in the baseline is not extracted. In Figure 13 (7), the smallest water body in the baseline is not marked, and the small river is disconnected. However, the method proposed in this paper accurately extracts the water bodies in these areas and eliminates some nonwater patches; these resulted in a reduction in the precision value of the quantitative algorithm evaluation when performing precision calculations against the baseline. Evaluations of remote sensing image water extraction results should not only be compared using a precision index but qualitative evaluations should also be performed that are based on the actual classification effect.

D. CONTRIBUTION TO ACTUAL PRODUCTION APPLICATIONS

We applied the DeepLabV3+ neural network architecture to the practical application of water extraction from remote sensing images and put the neural network model trained using this method into an actual production environment involving daily water monitoring. The water body was extracted from 188 GF1/GF2 images in 2017–2018. The extraction effect was stable, and the influences of clouds and mountain shadows were well controlled. After the optimization of the CRF postprocessing module, the water extraction details were accurate. The results of water extraction can be applied to the production of water samples again to expand the sample set of water bodies, thus forming a good sample expansion-model optimization production cycle that can reduce the labor and time costs of sample production.

VI. CONCLUSION

In this study, the research goal is the accurate extraction of complex water bodies from high-resolution remote sensing images. Considering the problems of inaccuracy exhibited by traditional neural network models due to the large changes in water body scale and the richly detailed features, we proposed an improved subscale model training method based on DeepLabV3+. This method combines the advantages of DeepLabV3+ multiscale feature extraction and fuses multiscale feature maps with appropriate weights. Finally, CRF is used for precise boundary segmentation of the pre-extraction results. We show the experimental results with different effects. The experimental method achieves PA, Recall and IOU accuracy scores of 95%, 97% and 94% respectively. Moreover, the model's detailed feature expression is even better than the manually labeled baseline water body segmentation result.

The multiscale model training method is highly adaptable to feature extraction from input images of different scales. Regarding the subscale feature mapping results, in a sense, this approach realizes multiple predictions of the same image. We adopt a controllable weighted fusion method to adjust the influence weights of different scale features for the overall prediction. The experiments show that this method not only correctly extracts large-area water bodies but also achieves accurate identification of water body details and improves the overall generalizability of the model. Adopting multiscale prediction to extract the deep features instead of the traditional single-scale training prediction produced satisfactory results in the identification of water features at different scales. The introduction of the fully connected CRF optimizes the water extraction boundary and reduces noise spots. The CNN+CRF training mode is highly suitable for semantic segmentation in the remote sensing field.

This study provides an idea for semantic segmentation of remote sensing images and achieved highly precise classification of water bodies. In the next step, we will focus on testing the adaptability of this scheme to the classification of other object features and improve the intelligence of the water extraction model. In addition, a learning transfer mechanism could be introduced to allow the model to independently learn to adapt to refined image water body segmentation of different resolutions and different phases.

ACKNOWLEDGMENT

(Ziyao Li and Rui Wang contributed equally to this work.)

REFERENCES

[1] H.-Q. Xu, "A study on information extraction of water body with the modified normalized difference water index (MNDWI)," *J. Remote Sens.*, vol. 9, pp. 589–595, Sep. 2005.

[2] S. W. Yang, C. S. Xue, T. Liu, and Y. K. Li, "A method of small water information automatic extraction from TM remote sensing images," *Acta Geodaetica Et Cartographica Sinica*, vol. 39, no. 6, pp. 611–617, 2010.

[3] S. K. Mcfeeter, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *Int. J. Remote Sens.*, vol. 17, no. 7, pp. 1425–1432, 1996.

[4] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 4, pp. 532–550, Jul. 1987.

[5] C. R. Dillabaugh, K. O. Niemann, and D. E. Richardson, "Semi-automated extraction of rivers from digital imagery," *Geoinformatica*, vol. 6, no. 3, pp. 263–284, 2002.

[6] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[7] L. Yu, Z. Wang, S. Tian, F. Ye, J. Ding, and J. Kong, "Convolutional neural networks for water body extraction from landsat imagery," *Int. J. Comput. Intell. Appl.*, vol. 16, p. 12, Mar. 2017.

[8] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[9] H. Bischof, W. Schneider, and A. J. Pinz, "Multispectral classification of Landsat-images using neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 3, pp. 482–490, May 1992.

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[11] F. Isikdogan, A. C. Bovik, and P. Passalacqua, "Surface water mapping by deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 4909–4918, Nov. 2017.

[12] R. Wang, Y. Meng, W. Zhang, Z. Li, F. Hu, and L. Meng, "Remote sensing semantic segregation for water information extraction: Optimization of samples via training error performance," *IEEE Access*, vol. 7, pp. 13383–13395, 2019.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[15] S. D. Jawak, K. Kulkarni, and A. J. Luis, "A review on extraction of lakes from remotely sensed optical satellite data with a special focus on cryospheric lakes," *Adv. Remote Sens.*, vol. 4, no. 3, pp. 196–213, 2015.

[16] X. J. Hou, L. Feng, H. Duan, X. Chen, D. Sun, and K. Shi, "Fifteen-year monitoring of the turbidity dynamics in large lakes and reservoirs in the middle and lower basin of the Yangtze River, China," *Remote Sens. Environ.*, vol. 190, pp. 107–121, Mar. 2017.

[17] J. Luo, Y. Sheng, Z. Shen, and J. Li, "High-precise water extraction based on spectral-spatial coupled remote sensing information," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2010, pp. 2840–2843.

[18] G. Q. Zhang, G. Zheng, Y. Gao, Y. Xiang, Y. Lei, and J. Li, "Automated water classification in the tibetan plateau using Chinese GF-1 WFV data," *Photogramm. Eng. Remote Sens.*, vol. 83, no. 7, pp. 509–519, 2017.

[19] J. M. McCarthy, T. Gumbricht, T. McCarthy, P. Frost, K. Wessels, and F. Seidel, "Flooding patterns of the Okavango wetland in Botswana between 1972 and 2000," *Ambio, J. Hum. Environ.*, vol. 32, no. 7, pp. 453–458, 2003.

[20] L. G. Olmanson, M. E. Bauer, and P. L. Brezonik, "A 20-year Landsat water clarity census of Minnesota's 10,000 lakes," *Remote Sens. Environ.*, vol. 112, no. 11, pp. 4086–4097, 2008.

[21] S. Reis and H. M. Yilmaz, "Temporal monitoring of water level changes in Seyfe Lake using remote sensing," *Hydrol. Processes, Int. J.*, vol. 22, pp. 4448–4454, Oct. 2008.

[22] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1134–1142.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[26] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3640–3649.

[27] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 3–18, Dec. 2017.

[28] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[29] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3376–3385.

[30] P. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. I-82–I-90.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[32] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[33] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.

[34] Y. Wang, Z. Dong, and Y. Zhu, "Multiscale block fusion object detection method for large-scale high-resolution remote sensing imagery," *IEEE Access*, vol. 7, pp. 99530–99539, 2019.

[35] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sens.*, vol. 11, no. 3, p. 227, 2019.

[36] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3194–3203.

[37] D. J. Marceau, P. J. Howarth, and D. J. Gratton, "Remote sensing and the measurement of geographical entities in a forested environment. 1. The scale and spatial aggregation problem," *Remote Sens. Environ.*, vol. 49, pp. 93–104, Aug. 1994.

[38] P. Aplin, "On scales and dynamics in observing the environment," *Int. J. Remote Sens.*, vol. 27, no. 11, pp. 2123–2140, 2006.

[39] D. J. Marceau, "The scale issue in the social and natural sciences," *Can. J. Remote Sens.*, vol. 25, no. 4, pp. 347–356, 1999.

[40] T. Lindeberg, *Scale-Space Theory in Computer Vision*, vol. 256. Norwell, MA, USA: Springer, 2013.

[41] L. Su, X. Li, and Y. Huang, "An review on scale in remote sensing," *Adv. Earth Sci.*, vol. 16, no. 4, 2001.

[42] O. Rukundo and H. Cao, "Nearest neighbor value interpolation," 2012, *arXiv:1211.1768*. [Online]. Available: https://arxiv.org/abs/1211.1768

[43] A. Poghosyan, "Asymptotic behavior of the Krylov–Lanczos interpolation," *Anal. Appl.*, vol. 7, no. 2, pp. 199–211, 2009.

[44] T. Blu, P. Thévenaz, and M. Unser, "Linear interpolation revitalized," *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 710–719, May 2004.

[45] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981.

[46] H. S. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 6, pp. 508–517, Dec. 1978.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[48] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[50] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[52] H. Lin, Z. Shi, and Z. Zou, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," *Remote Sens.*, vol. 9, no. 5, p. 480, 2017.

[53] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," 2015, *arXiv:1503.02351*. [Online]. Available: https://arxiv.org/abs/1503.02351

[54] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, and D. Du, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1529–1537.

[55] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[56] S. Qingsong, Z. Chao, C. Yu, W. Xingli, and Y. Xiaojun, "Road segmentation using full convolutional neural networks with conditional random fields," *J. Tsinghua Univ. (Sci. Technol.)*, vol. 58, no. 8, pp. 725–731, 2018.

[57] C. Sutton and A. McCallum, "Piecewise training for undirected models," 2012, *arXiv:1207.1409*. [Online]. Available: https://arxiv.org/abs/1207.1409

[58] J. T. Sample and E. Ioup, *Tile-Based Geospatial Information Systems: Principles and Practices*. Boston, MA, USA: Springer, 2010.

[59] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" in *Proc. BMVC*, 2013, pp. 1–11.

ZIYAO LI received the bachelor's degree from Northwest Agriculture and Forestry University, Yangling, China. She is currently pursuing the master's degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2017.

Her research interests include deep learning semantic segmentation and applications of deep learning techniques in remote sensing image water body extraction.

RUI WANG is currently pursuing the Ph.D. degree in remote sensing and information engineering with Wuhan University, Wuhan, China.

His research interests include uses of big data and remote sensing technology to time-series reservoir water changes, remote sensing data management methods for computational analysis, and applications of deep learning techniques in water monitoring.

Mr. Wang has participated in the construction of the Water Responsive Response Remote Sensing Intelligent Platform and won the First Prize of the Dayu Water Conservancy Science and Technology Award under Grant DYJ20150307-G15.

WEN ZHANG received the Ph.D. degree from Wuhan University, Wuhan, China, in 2009. She was a Visiting Scholar with the joint Centre of Cambridge–Cranfield for High Performance Computing, Cranfield University, Cranfield, U.K., from 2007 to 2008. She is currently a Lecturer with the School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than ten peer-reviewed articles. Her research interests include network GIS, remote-sensing applications, and spatial data analysis.

FENGMIN HU received the master's degree from Henan Polytechnic University, Jiaozuo, China, in 2017. She is currently pursuing the Ph.D. degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

Her research interests include scaling problem of quantitative remote sensing, soil moisture inversion, and deep learning.

LINGKUI MENG was a Visiting Professor with the joint Centre of Cambridge–Cranfield for High Performance Computing, Cranfield University, Cranfield, U.K., from 2006 to 2007. He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. His research interests include remote-sensing applications in hydrology, cloud computing, and big data analysis.