A Method of Water Body Extraction Based on Multiscale Feature and Global Context Information

Ru Miao, Tongcan Ren, Ke Zhou, Yanna Zhang, Jia Song, Guangyu Zhang, and Jiaqian Wang

***Abstract*—Water body extraction is an essential mission in the field of semantic segmentation of remote sensing images. It plays a significant role in natural disaster prevention, water resources utilization, hydrological monitoring, and other territories. In practice, the background of the majority of water remote sensing images is complicated. Owing to insufficient semantic mining and rough water body boundary extraction, traditional segmentation methods may be unable to adequately distinguish water bodies. We put forward a multiscale feature and global context fusion network (MSGFNet). In addition, multiscale feature extraction and fusion (MSEF) module based on UperNet and global context enhancement block (GCE Block) are designed. The MSEF module is capable of handling complex scenes by dynamically capturing multiscale semantic information and fusing different layers of features. The GCE Block can help the network to infer the location, shape and contextual information of the water bodies. The GF-1 dataset and Sentinel-2 dataset are used for model training simultaneously. The experimental results indicate that the extraction accuracy of the MSGFNet proposed are superior than other methods on GF-1 dataset and Sentinel-2 dataset, with overall accuracy of 98.60% and 98.22%, respectively. Compared to UperNet, the overall accuracy increases by 1.28% and 0.85%, respectively. In conclusion, the learning method build upon multiscale features and global context information can effectively prohibit noise, heighten the extraction accuracy of water bodies under intricate background, as well as ameliorate the matter of inaccurate water edge segmentation.**

***Index Terms*—Global context information, multiscale feature, remote sensing image, water body extraction.**

I. Introduction

WATER resource is one of the uppermost crucial natural resources on the Earth and an important part of the terrestrial ecosystem. The investigation of water body extraction from remote sensing images is helpful to enhance the comprehension of water body features. In water body extraction research, novel algorithms and methodologies need to be explored continuously, which can contribute to the advancement of remote sensing technology, image processing technology, and other related technologies. In addition, water body extraction research also holds significant practical implications. The water area and water level changes of water bodies in remote sensing images can be accurately obtained through water body extraction technology, which can provide reference for hydrological monitoring, water resources management, urban planning and construction, flood disaster monitoring, and other applications.

Early water body extraction methods mainly relied on visual interpretation, which was intuitive but inefficient and easily affected by subjective factors. As the remote sensing technology progress, especially the emergence of high-resolution remote sensing satellite technology, a new extraction solution for water bodies has been provided. Remote sensing technology can quickly and accurately obtain image data of a wide range of areas, providing abundant data sources for water body extraction. Moreover, remote sensing images are beginning to be widely used in natural resources investigation, water body extraction, change detection, and other important fields, and remote sensing water bodies information extraction methods based on spectral features have emerged. These methods are mainly classified into single-band methods [1], spectral relationship methods [2], and water index methods, among which the water index methods are universally applied in water body extraction research, such as multiband water index [3], normalized difference multiband water index [4], and other methods. As the machine learning rapidly progress, researchers have gradually explored a series of machine learning classification approaches, such as semisupervised learning [5], supervised learning, and unsupervised learning, by combining the spectral, spatial, and textural characteristics of water bodies in remote sensing images. Common supervised learning methods primarily contain support vector machine [6], decision trees, and random forest. While the unsupervised learning refers to learning on samples without categorical labels, with the aim of mining the intrinsic structure of the data, mainly including principal component analysis [7] and clustering. For instance, Zhang et al. [8] put forward a water body multifeature automatic extraction method, which used an unsupervised approach to extract water bodies from GF-1 images by integrating spectral and spatial characteristics, and clustered the features using k-means.

Manuscript received 11 December 2023; revised 1 May 2024; accepted 7 June 2024. Date of publication 19 June 2024; date of current version 12 July 2024. This work was supported in part by the National Science and Technology Major Project of High-Resolution Earth Observation System under Grant 80-Y50G19-9001-22/23, and in part by the Science and Technology Project of Henan Province under Grant 222102210061. *(Corresponding author: Ke Zhou.)*

Ru Miao is with the School of Computer and Information Engineering, Henan Engineering Research Center of Spatial Information Processing, Henan University, Kaifeng 475004, China (e-mail: mr1015@henu.edu.cn).

Tongcan Ren, Ke Zhou, Guangyu Zhang, and Jiaqian Wang are with the School of Computer and Information Engineering, Henan Engineering Research Center of Spatial Information Processing, Henan Technology Innovation Center of Spatio-Temporal Big Data, Henan University, Kaifeng 475004, China (e-mail: Rentc@henu.edu.cn; zhouke@radi.ac.cn; zhangguangyu@henu.edu.cn; wangjiaqian@henu.edu.cn).

Yanna Zhang is with the School of Computer and Information Engineering, Henan University, Kaifeng 475004, China (e-mail: zyn@henu.edu.cn).

Jia Song is with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China (e-mail: songj@igsnrr.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3416623

As a novel research direction in machine learning filed, deep learning can provide efficient algorithms to process and analyze massive data. In recent years, the development of deep learning technology has made significant achievements in the fields of image classification [9], object detection [10], and semantic segmentation [11]. In semantic segmentation tasks, as a typical representative of deep learning methods, the rapid development of convolutional neural networks (CNNs) provides technical support for water body extraction from remote sensing images [12], which is capable of capturing the location relationship of shallow networks and the feature representation of deep networks [13]. In the development progress of semantic segmentation networks, the encoder–decoder has gradually become a representative architecture configuration. The encoder is primarily employed to extract features, while the decoder is employed to restore spatial dimensions and fuse features. U-Net [14] was an approach based on CNN structure. In the decoder part, it established the spatial correlation of the coding stage by adopting skip connections, so as to refine the feature loss problem generated by the encoder in the process of feature transmission. SegNet [15] was a semantic segmentation network with encoder–decoder architecture. As the performance of computer equipment improves, some deeper networks have also been applied to semantic segmentation tasks. However, with the deepening of network layers, the phenomenon of gradient disappearance may occur [16]. Different scholars have proposed different semantic segmentation algorithms. DeepLabV3+ [17], based on DeepLabV3 [18], introduced a decoder module to fuse features of different levels, so as to extract multiscale features and repair segmentation edge information.

Although encoder–decoder architectures are effective in learning and distinguishing spatial features, there are still two challenges in network training and inference: the limited ability of encoders to capture spatial details and the insufficient ability of decoders to fuse multiscale features [19]. In order to capture sufficient water body features and enhance the extraction effect of water bodies, some researchers have started to ameliorate network with encoder–decoder architectures. For instance, Ge et al. [20] and Qin et al. [21] significantly improved the water body extraction accuracy by modifying the network depth of the U-Net. For the sake of aggregating the multiscale feature of water bodies, Wang et al. [19] put forward a network with multiscale dense connections, named MSLWENet, which preserved small lakes information to a certain extent. In terms of extracting slender water bodies and identifying shadow effects of building, the dense local feature compression network put forward by Li et al. [22] was superior to traditional methods and semantic segmentation networks. MSAFNet proposed by Lyu et al. [23], which integrated multiscale features with attention mechanism, and enhanced the interaction capability of semantic interaction. This method achieved optimality in both QTPL and LoveDA datasets, and ameliorated the segmentation effects of various water bodies. Dai et al. [24] put forward a multiscale location attention network, which concerned the details of tributaries of rivers through location channels to decrease the issues of feature map loss caused by downsampling operations, and constructed a two-branch network structure to acquire multiscale semantic information. These approaches have demonstrated that heightening the capability of encoder–decoder to capture multiscale features is feasible to improve the extraction accuracy of water bodies.

CNN has become an indispensable tool in the field of semantic segmentation, showing apparent advantages in spatial position representation [25]. However, the acceptance filed of CNN is easily limited by the size of convolutional kernel, which leads to more attention to local features [26]. Remote sensing images are characterized by tight context information and strong noise information. For instance, SegNet adopted common convolution kernel to extract water body, but the acceptance filed of convolution kernel is small, so the extracted features could not adequately describe the context information of water bodies [27]. Unlike CNN, transformer [28], as a new deep learning method, mainly captured location relationships by introducing the location embedding, avoiding convolution and pooling operations. Based on vision transformer [29], swin transformer designed by Liu et al. [30] achieved cross-window information interaction by introducing a sliding window mechanism, which opened up a new research idea for global relationship modeling. Since then, many scholars have begun to apply transformer in the research of water body extraction. Xu et al. [31] put forward a swin transformer model for lightweight edge classification, which further diminished the computational complexity when refining water body boundaries. Transformer cannot accurately identify the detailed texture of the water body boundaries because of the weak inductive bias capability. In recent years, many hybrid networks of transformer and CNN have been proposed to take full advantage of both. For instance, MST-UNet, a water body extraction network with symmetric encoder–decoder structure proposed by Li et al. [32], extracted features from input images and captured the interdependence between different pixels by using CNN and swin transformer, thereby heightening the capture ability of global image information. Zhang et al. [33] put forward a water body extraction network based on CNN and swin transformer. The network combined the ability of CNNs to capture local details and the capability of transformers to gather global context information, showing certain advantages in water body extraction accuracy. To make up for the disadvantages of CNN in feature extraction, Chen et al. [34] designed a double-branch parallel interactive network, using Rensnet50 and swin transformer to obtain contextual and spatial information. DEDNet proposed by Wang et al. [35], extracted the local and global information through parallel structure of CNN and swin transformer in the encoding part, while in the decoder stage, it enhanced the recognition capability of the target boundary by building cross-stage fusion modules. In addition, considering the limitations of using CNN and transformer alone for water body extraction of remote sensing images, Dong et al. [36] designed a DSCTs frame. Channel-weighted attention-guided feature distillation was used for learning spatial dependencies, and a target-nontarget KD module was put forward to enhance the flexibility of classification logic.

Deep learning methods require a mass of data samples, and training data are necessary condition to ensure that the network can accurately extract water bodies. For the sake of exploring

the influence of cloud shadow samples on the accuracy of water body extraction, Song et al. [37] conducted research by dividing training datasets with different proportions of cloud shadows and applied them to the swin transformer. The experimental results indicated that the network could better deal with misclassification of cloud shadows. Wieland et al. [38] produced a global reference dataset (S1S2-Water) for training, verifying and testing networks, and compared the water body extraction effects with different CNN architectures, respectively, which provided data support for water research.

The progress of deep learning technology promotes the wide application of semantic segmentation network in water body extraction research. With the enhancement of remote sensing image resolution, water bodies hold rich texture and spectral features. There are plentiful surface coverings similar with water bodies in remote sensing images, which may have similar shape size and spectral characteristics. Considering the complexity of image background information, abundant contextual information and detailed features are served as clues for water body extraction [39], [40], [41]. Meanwhile, the coverings in complex scenes possess multiscale features. In pixel-level semantic segmentation task, single-scale features alone cannot adequately identify each pixel, so the introduction of multiscale information needs to be considered to achieve accurate semantic segmentation.

A novel segmentation network named multiscale feature and global context fusion network (MSGFNet) is put forward for improving the extraction accuracy of water bodies. The overall network framework consists of three parts. Swin transformer is selected as backbone network for extracting multiscale features of water bodies. In addition, a multiscale feature extraction and fusion (MSEF) module based on UperNet is designed. The MSEF module consists of four dynamic convolution modules (DCMs) with different kernel sizes [42] and the feature pyramid network (FPN) [43]. These dynamic convolutions of different kernel sizes can help the network capture water body information of different scales, thereby adapting to water changes in different scenarios and conditions. The FPN can effectively fuse shallow and deep features through top-down approach and lateral connection. The combination of these two effective mechanisms endows the MSEF module with better multiscale feature extraction capability. Moreover, a global context enhancement block (GCE Block) is designed for strengthening the capability of network to attend to global information.

The main contributions of this article are summarized as follows.

1) A novel water body extraction network MSGFNet is put forward in this article for extracting water bodies from high-resolution remote sensing images. The proposed method simultaneously takes into account multiscale variations of the water bodies and global context, and better balances model performance and computational resources while maintaining a certain water body extraction accuracy.
2) The MSEF module and GCE Block are designed in this article. The MSEF module is capable of handling complex scenes by dynamically capturing multiscale semantic information and fusing different layers of features. The GCE Block can help the network to infer the location, shape and contextual information of the water bodies, thereby reducing the possible localized errors in water body extraction.
3) A water body classification dataset named GF-1 dataset composed of GF-1 images and corresponding labels is constructed. Simultaneously, the validation experiment based on the public sentinel-2 dataset is conducted for testing the generalization ability and adaptability of the proposed method. Compared to other approaches, the experimental results indicate that the proposed method can effectively enhance the extraction accuracy of small water bodies and large waters, and ensures the integrity of lake boundary extraction. (Sentinel-2 dataset address: https://link.zhihu.com/?target=https%3A//doi.org/10.5281/zenodo.5205674).

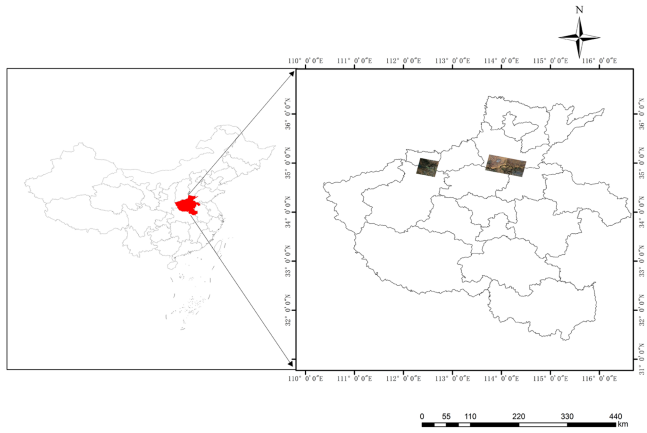

Fig. 1. Study area location map.

II. Study Area and Data

A. Study Area

Henan (31°23–36°22, 110°21–116°39) is situated in the middle and lower reaches of the Yellow River basin. The Yellow River flows through several terrains, including mountains, hills and plains in Henan province, with complex and changeable topography, forming a rich natural landscape and ecological environment. The study area is situated in the Henan section of the Yellow River mainstream, which primarily contains part of the Zhengzhou section, part of the Kaifeng section, Xiaolangdi Reservoir and Xixiayuan Reservoir, as shown in Fig. 1. The watershed characteristics of the Henan section of the Yellow River mainstream are very remarkable.

1) There are relatively few tributaries in this section, and almost no tributaries converge in the downstream, thus forming a relatively independent water system characteristic.
2) The channel morphology of the Yellow River in the Henan section is varied, and complex riverbed landforms have been formed due to multiple channel changes. In some sections, the Yellow River shows the characteristics of a

wandering channel, with a wide and shallow riverbed and scattered currents. In canyons and mountain areas, narrow and fast-flowing channels may be formed.

3) Henan is located at the boundary between the second and third terraces of the Yellow River Basin, with a large difference in topography, making the middle reaches of the Yellow River rich in hydropower resources. However, due to the sparse soil and less vegetation on the Loess Plateau, the Yellow River carries a large amount of sediment when it flows through this section, so it is known as one of the richest rivers in the world. This section is the most concentrated section of the middle reaches of the Yellow River with the layout of water conservancy hubs, and it has rich water types, which contain both large river reservoirs and small water flow ponds. In addition, the study area includes a variety of surface covers, such as mountains, bridges, and construction areas, which provides data guarantee for the research of extracting water bodies under different background conditions.

B. Data

1) GF-1 Dataset: In this study, three GF-1 domestic high-resolution remote sensing images were adopted. The application of GF-1 data for water body extraction task has obvious advantages. 1) The high-resolution data can provide more details, which contributes to identify and extract water body boundaries. 2) The GF-1 satellite has three visible light bands of blue, green and red, as well as a near-infrared band. These different spectral bands have different sensitivities to water bodies. By utilizing rich spectral information, water bodies, and other coverings can be finely identified. The water bodies show strong absorbency peculiarities in the near-infrared band, while the vegetation and soil hold high reflectivity. Therefore, the near-infrared band can better separate water bodies from other coverings. In this study, three bands of green, red and near-infrared are selected for the water body extraction study.

In this article, the selected GF-1 data were preprocessed, including radiometric calibration, atmospheric correction, geometric correction, image fusion, and image cropping. First, the water bodies were labeled by manual visual interpretation, and the original images and corresponding water body labels were randomly cropped using Python program to produce water body classification datasets. Then, the diversity of data samples is enriched in this article through data augmentation, which mainly includes rotating, flipping, random cropping, adding Gaussian noise and other expansion methods. Finally, a total of 11 270 images with the size of 512 × 512 pixels and corresponding labels were obtained in the GF-1 dataset. The dataset was divided according to the ratio of 6:2:2. The training dataset contained 6762 images, the validation dataset and test dataset each contained 2254 images. The training dataset and validation dataset were primarily used to train and inference model, while the test dataset was used for evaluating model performance. Some of the water body samples are illustrated in Fig. 2, and each image contains water and nonwater categories, where blue represents water body and black represents nonwater body.

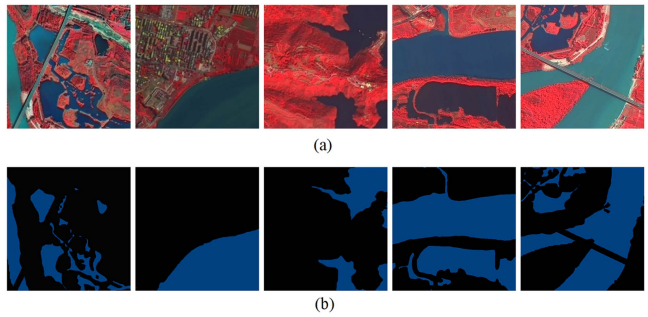

Fig. 2. Training samples of GF-1 dataset. (a) Is the real water body images. (b) Is the labeled water body images.

2) Sentinel-2 Dataset: The dataset [44] was constructed from 31 Sentinel-2 remote sensing images collected worldwide by the Sentinel-2 satellite. All Sentinel-2 images were acquired and downloaded from the Sentinel Science Data Center, which contained all seasons of the year. The dataset was composed of 95 complex scenes, which contained a wealth of water body types, such as rivers, lakes, reservoirs, and oceans, as well as surface coverings, such as built-up areas, roads, and ice. In addition, the dataset also comprehensively considered the influence of different terrain types such as plains, hills and mountains on water body extraction. The diversity of terrain types can provide richer water body features. Moreover, different weather types, such as cloudy and cloudless, are also incorporated to the dataset. Due to the highly similar spectral features of cloud shadows and water bodies, learning about cloud shadows can diminish the misclassification between water bodies and cloud shadows to a certain extent. Therefore, the diversity and richness of the dataset can provide comprehensive and high-quality data guarantee for the deep learning network to learn surface water body features.

The original dataset images consisted of four 10 m resolution bands and two 20 m resolution bands. To standardize the format of the dataset, the Python program was firstly utilized to crop the band of 95 images in this study. The same three 10 m bands of green, red and near-infrared were selected for extracting water bodies from remote sensing images. Then the same cropping method and expansion method were conducted on Sentinel-2 dataset. Finally, a total of 10 000 images with the size of 512 × 512 pixels and corresponding labels were obtained form Sentinel-2 dataset. Similarly, the Sentinel-2 dataset was divided according to the ratio of 6:2:2, where 6000 images were served as the training dataset, 2000 images were served as the validation dataset and the remaining 2000 images were served as the test dataset. Part of the water body samples are illustrated in Fig. 3.

III. Methodology

A. Overview

The MSGFNet proposed adopts a representative encoder–decoder architecture. Swin transformer is used as the feature extractor in the encoder, while the decoder mainly consists of the MSEF module and GCE Block. The overall architecture is illustrated in Fig. 4.

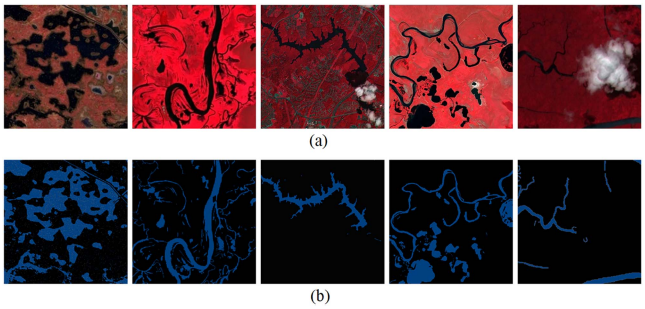

Fig. 3. Training samples of Sentinel-2 dataset. (a) Real water body images. (b) Labeled water body images.

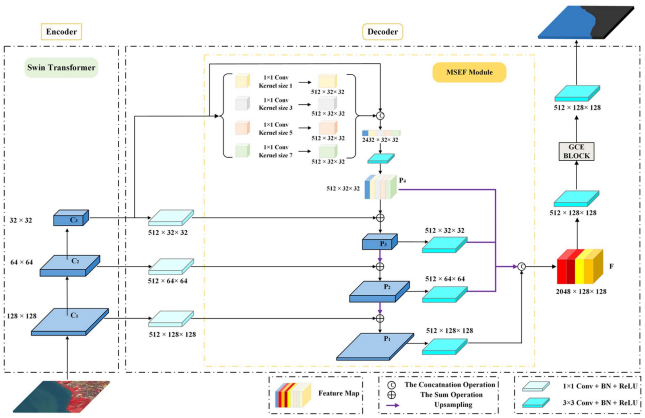

Fig. 4. Overall framework.

In this article, the features C_1, C_2, and C_3 extracted from swin transformer are selected as the input features of the MSEF module. Since DCMs with different convolution kernel sizes are embedded inside the MSEF module, which enables it to process the semantic information inside the image and enhance the high-level feature representations, thereby capturing multiscale features of the water bodies. Meanwhile, FPN is used in MSEF module to integrate deep and shallow features, so that the network contains abundant semantics for features at different levels. Before input to the MSEF module, a 1×1 convolution module is used to adjust the number of channels for input features, while enhancing nonlinear capability. The MSEF module can output the feature F with multiscale information. Then a 3×3 convolution module is utilized to capture the local features of the image. Eventually, the GCE Block is used for improving the utilization of global information, and enhancing the processing of image details and edge information. Next, the structure of different modules of the MSGFNet proposed is described in detail.

B. Swin Transformer

As a new-fashioned deep learning method, transformer was initially applied in the field of natural language processing. Since the efficient modeling abilities of transformer, it has been applied to semantic segmentation domain. Swin transformer was a deep learning network with transformer structure, which adopted the windows multihead self-attention (W-MSA) mechanism to

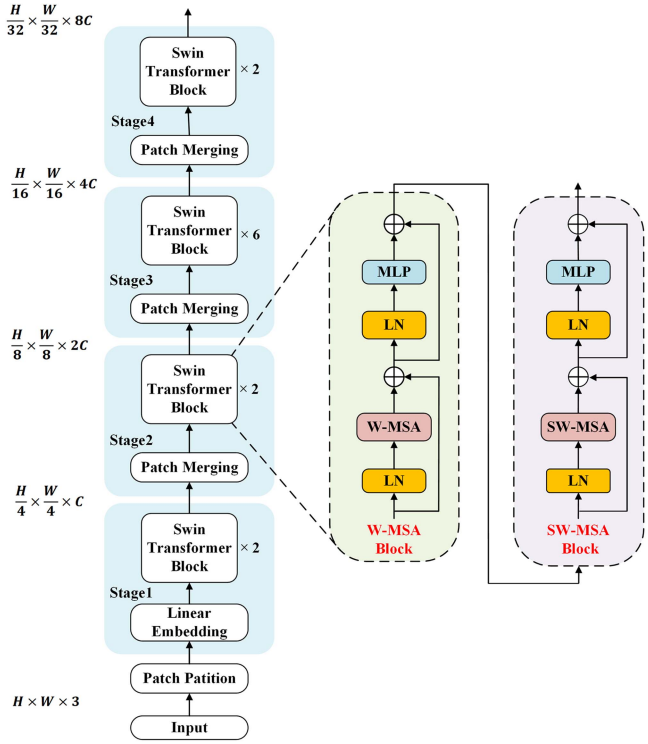

Fig. 5. Network structure of swin transformer.

divide the feature graph into several fixed-size window, and computed the self-attention within each window. This method significantly reduced the computational complexity of global attention while maintaining the perceptual range of global features. In order to enable features to interact across different windows, the sliding W-MSA mechanism was introduced to help swin transformer capture long-range dependencies in the image.

Swin transformer adopted a representative hierarchical structure, including patch partition module and four stages, as shown in Fig. 5. Since each stage of swin transformer will diminish the resolution of the input feature, making it easy to create multilevel features, so swin transformer is served as the backbone network of MSGFNet in this article.

C. UperNet

UperNet was put forward by Xiao et al. [45], whose main purpose was to solve a recognition task called unified perceptual parsing. The task required identifying as many visual concepts as possible from a single image, which included various scene labels, objects, textures, and materials. The UperNet performs feature fusion at each layer of the FPN, and inputs the fused features into two structurally equivalent detection heads for image segmentation tasks. Since UperNet is able to unify the structural visual knowledge and predicts the layer output effectively. Therefore, during the training process, UperNet can learn the differentiated data in different image datasets to explore the rich visual knowledge in the image, which enables UperNet to better understand the image content and enhance the segmentation accuracy. The network structure of the UperNet is illustrated in Fig. 6.

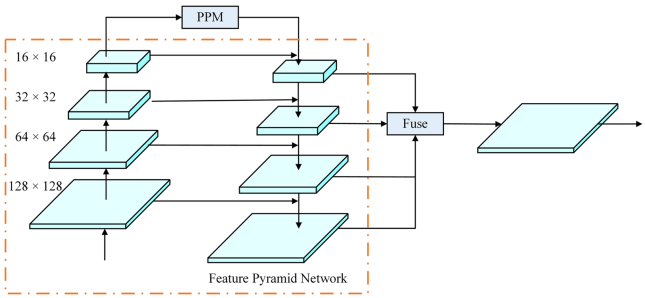

Fig. 6. Network structure of UperNet.

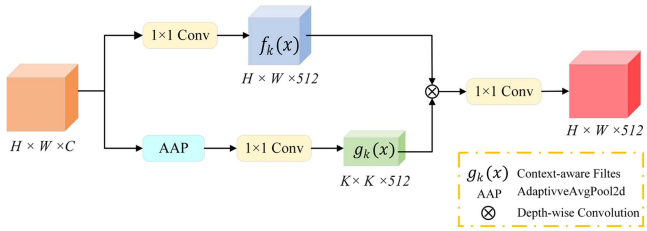

Fig. 7. Network structure of DCM.

The PPM [46] in UperNet utilizes multiple pooling operations of different sizes to capture multiscale features. However, these pooling operations usually uniformly downsampling the input feature maps, which make UperNet easy to lose fine detail during the pooling process. In order to improve the extraction accuracy of remote sensing water body information, an MSEF module is designed for extracting and fusing multiscale features based on UperNet. Next, the MSEF module is described in detail in this article.

D. MSEF Module

Water bodies in remote sensing images present different morphological features at different scales. Since the influence of a variety of factors, such as noise, light variations, and shadows, single-scale features may not be able to completely overcome these interferences. On the contrary, multiscale feature fusion can synthesize different scale information, so as to improve the extraction accuracy of water bodies. The MSEF module proposed mainly consists of four DCMs with different kernel sizes, FPN and four convolutional modules. DCM is able to generate convolutional kernel dynamically according to the characteristics of input data, thereby strengthening the ability of the MSEF module to perceive the local information of input data. The use of multiple DCMs enables the MSEF module to adaptively adjust the convolution kernel according to different dynamic convolution, thus enhancing its adaptability. Since the MSEF module can flexibly process dynamical changing data, it can be well applied to river identification tasks under complex backgrounds.

The MSEF module structure diagram is illustrated in Fig. 4. In this article, the features C_1, C_2, and C_3 extracted from swin transformer are selected as inputs for the MSEF module, where one branch of the feature C_3 is through the DCM inside the MSEF module. The interior of the DCM consists of two parallel branches. One path performs 1×1 convolution on the input feature to acquire the feature x_k. The other path generates a filter $y_k \in R^{512\times k\times k}$ with a kernel size of k through an adaptive average pooling operation and a 1×1 convolution operation. This filter is called context-aware filter. The x_k and y_k are used as the output of the DCM after fusing features by deep convolution and adjusting channels using 1×1 convolution, the calculation formula is as follows:

$$D_k = \mathrm{Conv}_{1\times1}(x_k \otimes y_k) \tag{1}$$

$$x_k = \mathrm{Conv}_{1\times1}(C_3) \tag{2}$$

where $\otimes$ represents depthwise convolution.

Different DCMs have context-aware filters with different kernel sizes. Multiple feature maps obtained from DCM are spliced with original feature C_3 to get feature map P_4, which has rich and detailed multiscale feature information. The specific calculation method of the multiscale special diagnosis extraction process is as follows. The structure of DCM is illustrated in Fig. 7

$$x_k = \mathrm{Conv}_{3\times3}(\mathrm{Cat}([C_3, D_k])), k = 1, 3, 5, 7. \tag{3}$$

Then P_4 is used as the top feature in the top-down branch of FPN, and links it horizontally with the feature map of the previous level. In this way, the high-level feature information is strengthened, and the feature expression capability of the whole network is enhanced. Horizontal connection mainly means that the channel number of the feature map of the previous layer is adjusted by 1×1 convolution operation, and is consistent with the feature map sampled on the later for feature fusion. Then, four feature maps of different scales can be obtained from the top down, namely P_4, P_3, P_2, and P_1, which contain abundant positional information and semantic information. Finally, the 3×3 convolution operation and dimension splicing of P_4, P_3, P_2, and P_1 are conducted to output a feature map F with plentiful multiscale information. The specific calculation formula is as follows. The purpose of the 3×3 convolution module is to eliminate aliasing effect of upsampling

$$P_i = \mathrm{Conv}_{1\times1}(C_i) \otimes \mathrm{Up}(P_{i+1}), i = 3, 2, 1 \tag{4}$$

$$\mathrm{Out}_{\mathrm{MSEF}} = \mathrm{Cat}[(\mathrm{Conv}_{3\times3}(P_i)), P_1], i = 4, 3, 2. \tag{5}$$

E. GCE Block

Nonlocal networks [47] was put forward for dealing with the essential matter of capturing remote dependence in deep learning networks and enhancing the efficiency of network information transmission. It discards the concept of distance and coalesces global information by calculating the correlation between different locations. This structure is called nonlocal block, while GCNet [48] further simplifies and diminishes the calculated quantity while reinforcing context understanding.

As shown in Fig. 8, a GCE Block is designed to reinforce the network's capability to concern global information. The GCE Block consists of a GC Block and elementwise add operation, where the GC Block is divided into three parallel branches. The first route aggregates the global position features through a 1×1 convolution operation and a softmax function. The

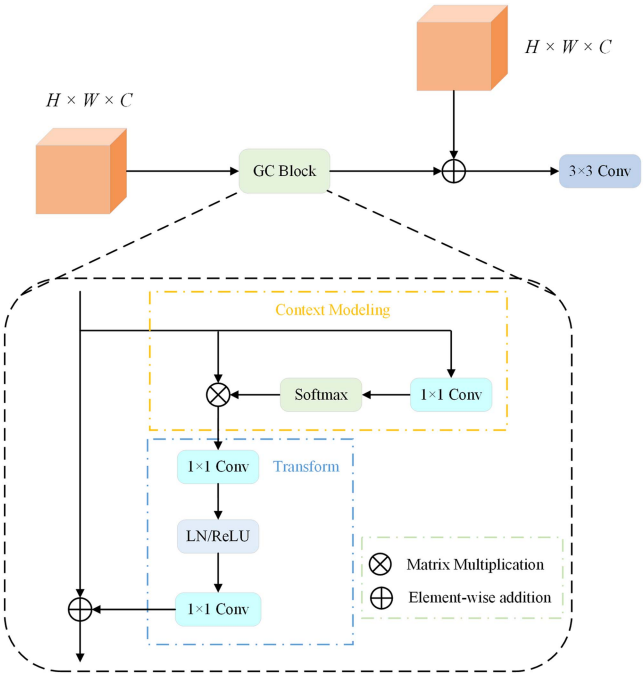

Fig. 8. Network structure of GCE Block.

second route performs feature transformation on the aggregated features to capture the interchannel dependencies, which mainly consists of two 1 × 1 convolutions and an LN/ReLU layer. The third way will transform features with the input feature by aggregating global information to each position, enhancing the global representation. Finally, the feature output by GC Block and the input feature are added element-by-element. The calculation formula is as follows:

$$\text{Out} = \text{Conv}_{3\times3}(\text{Add}(\text{GC Block}(F_1), F_1)). \tag{6}$$

IV. Results

A. Experimental Environment

All experiments in this article were conducted under the PyTorch framework, with CentOS 7.9.200 as the operating system, Montage Jintide(R) C12301 as the CPU and GeForce RTX 3080 Ti as the image processor. The programming language was Python 3. In terms of the neural network design, to ensure efficiently training, the Adamw was selected as the network training optimizer, the initial learning rate was set to 0.00006, the total number of iterations was set to 169 000, and the weight decay was 0.01. Based on the performance of the computer, the batch size of the training was set to 4, and the loss function adopted CrossEntropyLoss for calculating the loss value between the prediction plot and the label plot. The experimental environment is exhibited in Table I.

B. Evaluation Metrics

In this article, the mean intersection over union (MIoU), precision, recall, F1 score (F1), and overall accuracy (aAcc) are adopted to evaluate the extraction results with quantitative accuracy. The formula of all indicators as follows:

$$\text{MIoU} = \frac{1}{\text{K}+1}\sum_{\text{i}=0}^{\text{k}}\frac{\text{TP}}{\text{FN}+\text{TP}+\text{FP}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

$$\text{aAcc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{11}$$

where TP represents the truly predicted water body pixel, TN represents the truly predicted nonwater body pixel, FP represents the incorrectly predicted water body pixel, and FN represents the incorrectly predicted nonwater body pixel.

TABLE I
Experimental Environment Configuration Information

Surrounding	Name
Operating system	CentOS 7.9.2009
Deep learning framework	Pytorch 1.13
Programming language	Python 3.9.13
Image processing software	ENVI 15.3 / ArcGIS 10.8
CPU	Montage Jintide (R) C12301 (64GB)
GPU	GeForce RTX 3080 Ti (12GB)

TABLE II
Extraction Accuracy of Different Modules of the MSGFNet Proposed on the GF-1 Dataset

Method	MIoU (%)	F1 (%)	aAcc (%)
Swin transformer	89.05	91.66	95.79
Swin trasnformer + FPN	90.02	92.42	96.26
Swin transformer + DCM	93.48	95.10	97.87
Swin transformer + MSEF	94.70	96.04	98.41
Swin transformer + MSEF + GCE Block	**95.12**	**97.49**	**98.60**

C. Experimental Results

In this section, water body extraction studies are mainly conducted on GF-1 dataset and Sentinel-2 dataset through ablation experiments and comparison experiments. Where the bold numbers in the table indicate the best results. In addition, in the extraction result plots, the blue areas and red areas represent water bodies, the black areas represent nonwater bodies, as well as the green box indicate the areas that need to be focused.

1) Ablation Experiment on GF-1 Dataset: The quantitative results of the evaluation metrics are exhibited in Table II, and the visualization results of the ablation experiment are illustrated in Fig. 9. In this study, only swin transformer is initially used for water body extraction, with the overall accuracy of 95.79%. In the subsequent experiments, whether FPN with accuracy of 96.26% or DCM with accuracy of 97.87% is added alone, the extraction result is significantly higher than swin transformer. Then the MSEF module designed is introduced in this article, and the accuracy is improved by 2.62%. The experimental results

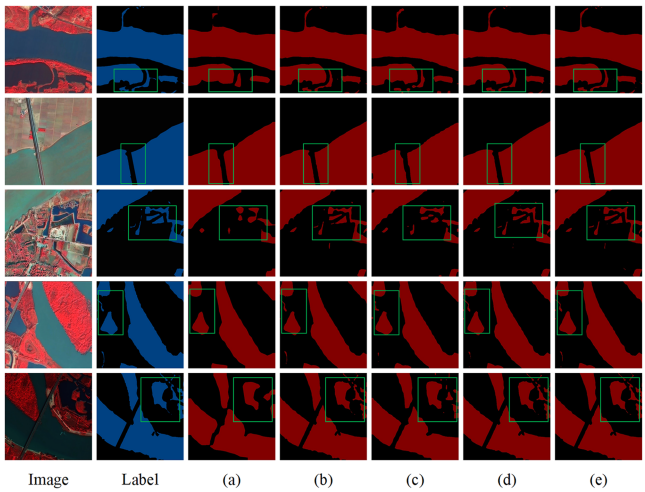

Fig. 9. Visualization results of ablation experiments on the test dataset of GF-1 dataset. (a) The extraction effect of the Swin Transformer. (b) The extraction effect of the Swin Transformer+FPN. (c) The extraction effect of the Swin Transformer+DCM. (d) The extraction effect of the Swin Transformer+MSEF. (e) The extraction effect of the Swin Transformer+MSEF+GCE Block.

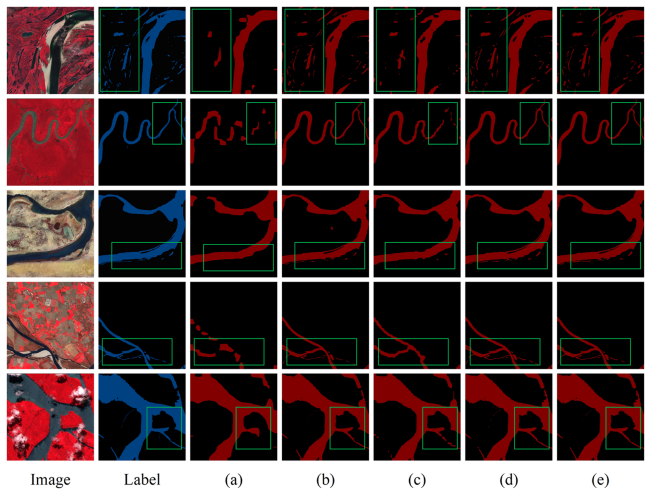

Fig. 10. Visualization results of ablation experiments on the test dataset of Sentinel-2 dataset. (a) The extraction effect of the Swin Transformer. (b) The extraction effect of the Swin Transformer+FPN. (c) The extraction effect of the Swin Transformer+DCM. (d) The extraction effect of the Swin Transformer+MSEF. (e) The extraction effect of the Swin Transformer+MSEF+GCE Block.

TABLE III
EXTRACTION ACCURACY OF DIFFERENT MODULES OF THE MSGFNET PROPOSED ON THE SENTINEL-2 DATASET

Method	MIoU (%)	F1 (%)	aAcc (%)
Swin transformer	81.29	83.25	93.41
Swin trasnformer + FPN	87.96	89.67	96.20
Swin transformer + DCM	88.47	90.14	96.42
Swin transformer + MSEF	92.67	93.85	98.05
Swin transformer + MSEF + GCE Block	**93.12**	**94.25**	**98.22**

TABLE IV
EXTRACTION ACCURACY OF THE MSGFNET PROPOSED WITH OTHER METHODS ON THE TEST DATASET OF GF-1 DATASET

Method	Index				
	MIoU (%)	Precision (%)	Recall (%)	F1 (%)	aAcc (%)
PSPNet	90.30	92.56	92.90	92.73	95.74
U-Net	80.16	80.79	88.53	84.49	90.50
DeepLabV3+	93.08	95.16	94.61	94.89	97.02
HRNet	86.86	88.67	91.35	89.99	94.06
SegFormer	89.52	92.11	91.91	92.01	96.03
APCNet	91.64	91.92	95.48	93.66	97.01
ANNNet	91.02	93.16	93.31	93.23	96.73
ENCNet	90.28	91.85	93.37	92.61	96.38
PSANet	92.11	94.22	93.95	94.08	97.23
UperNet	92.30	93.41	95.00	94.20	97.32
MSGFNet	**95.12**	**95.86**	**96.81**	**96.33**	**98.60**

indicate that the MSEF module holds significant effect in water body recognition, which is due to the fine image features captured inside the MSEF module. Finally, GCE block is introduced into the network, and the overall network performance has reached a good state, with the overall accuracy increased by 2.81%.

2) Ablation Experiment on Sentinel-2 Dataset: Similarly, ablation experiments are conducted on the Sentinel-2 dataset. The quantitative results of the evaluation metrics and the visualization results are shown in Table III and Fig. 10. All index values of the MSGFNet proposed reach the highest, with aAcc of 98.22%. Compared with the swin transformer alone, the extraction accuracy is increased by 4.81%.

3) Comparison Experiment on GF-1 Dataset: In this section, the MSGFNet proposed is compared with other approaches, including PSPNet, U-Net, DeepLabV3+, HRNet [49], SegFormer [50], APCNet [51], ANNNet [52], ENCNet [53], PSANet [54], and UperNet. In addition, the MIoU, precision, recall, F1, and aAcc are adopted as evaluation metrics. The quantitative results of evaluation metrics are exhibited in Table IV, and the visualization results are illustrated in Fig. 11.

The extraction accuracy of different approaches is exhibited on the test dataset in Table IV. The MSGFNet proposed accomplishes significant improvement in all evaluation indexes, with MIoU, precision, recall, F1, and aAcc reaching 95.12%, 95.86%, 96.81%, 96.33%, and 98.60%, respectively. The index values of the MSGFNet proposed increased by 2.82%, 2.45%, 1.81%, 2.13%, and 1.28%. This improvement is attributed to the loss of plentiful refined semantic information of UperNet during pooling operations, leading to erroneous judgements in the recognition of intricate water bodies. Fig. 11 illustrates that U-Net holds the poorest extraction performance, with a higher number of misclassified pixels and noticeable of misjudgment and omission, falling to fully identify water bodies. While DeepLabV3+ improves boundary extraction by introducing the decoder module to fuse underlying features and advanced semantics, it still has error detection issues. Compared to DeepLabV3+, the aAcc of the MSGFNet proposed is improved by 1.58%, with more precise handling of local details.

Fig. 11(d) and (f) demonstrates that these methods can well distinguish the water body contour to a certain extent. But when there are some complex ground objects, such as bridges. Owing to the similarity between the characteristics of the bridges and water bodies, the adjacent regions of the water bodies and bridges cannot be well recognized. For instance, PSPNet

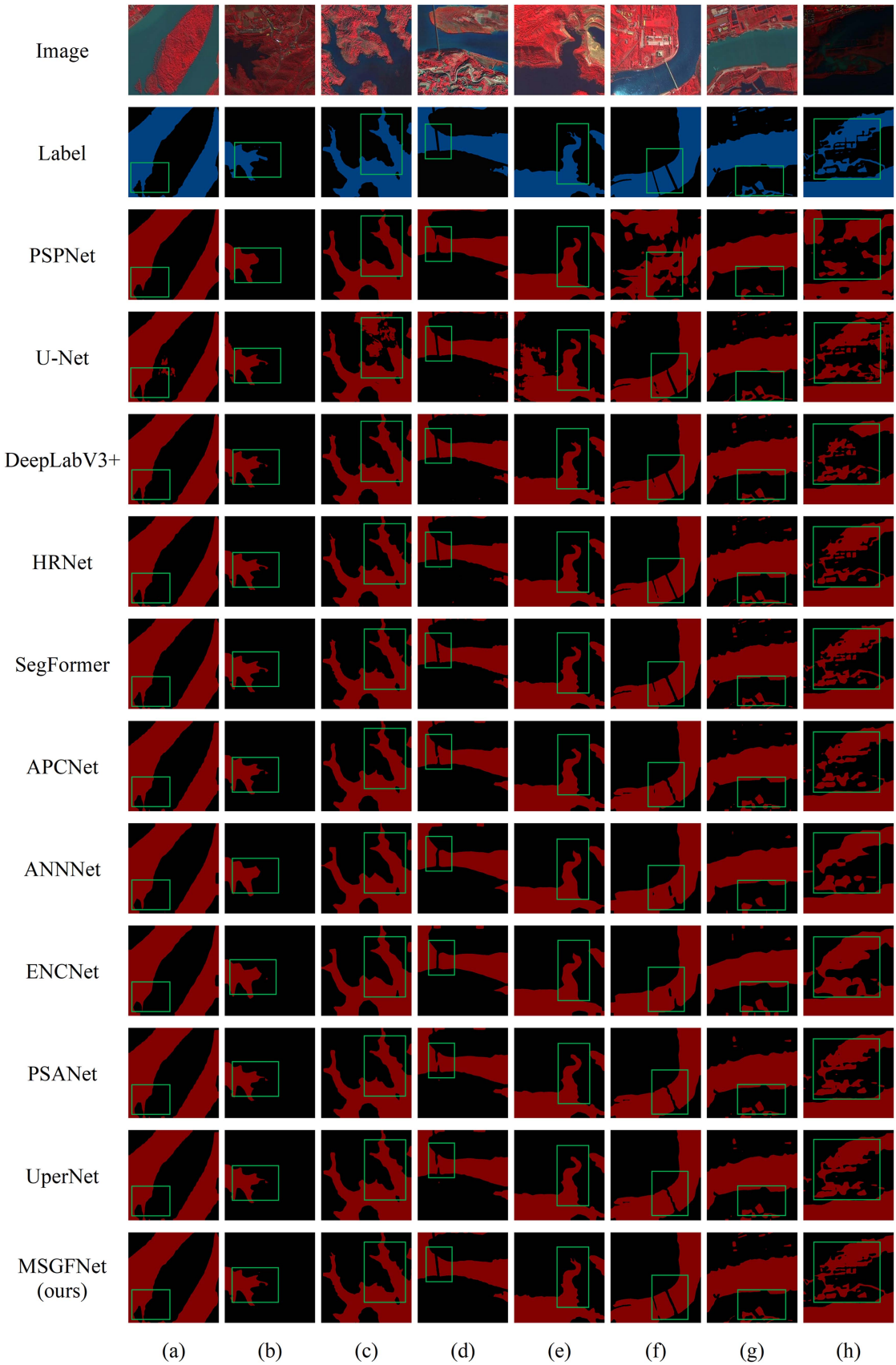

Fig. 11. Visualization results of comparative experiments on the test dataset of GF-1 dataset. (a) Image 1. (b) Image 2. (c) Image 3. (d) Image 4. (e) Image 5. (f) Image 6. (g) Image 7. (h) Image 8.

TABLE V
EXTRACTION ACCURACY OF THE MSGFNET PROPOSED WITH OTHER METHODS ON THE TEST DATASET OF SENTINEL-2 DATASET

Method	Index				
	MIoU (%)	Precision (%)	Recall (%)	F1 (%)	aAcc (%)
PSPNet	81.92	85.83	81.97	83.85	93.73
U-Net	83.20	83.74	86.86	85.27	93.55
DeepLabV3+	91.34	93.72	91.84	92.77	96.92
HRNet	88.81	89.53	91.60	90.55	95.89
SegFormer	87.55	89.71	88.86	89.28	96.05
APCNet	86.47	87.40	89.19	88.28	95.59
ANNNet	86.70	89.11	87.89	88.50	95.71
ENCNet	81.77	84.49	83.06	83.77	93.60
PSANet	86.87	89.05	88.34	88.70	95.76
UperNet	90.88	92.45	92.14	92.29	97.37
MSGFNet	**93.12**	**93.62**	**94.88**	**94.25**	**98.22**

and APCNet mistakenly identify bridges as water bodies. In Fig. 11(c), (e), and (h), some small water bodies are missing. This may be due to the insufficient features extracted by the networks, so it unable to deal with the water bodies and nonwater bodies in intricate regions well, leading to a mass of misclassifications. In the adjacent areas of water bodies and mountains, Fig. 11(b) indicates that there are missing water bodies in PSPNet, ANNNet, and ENCNet. As shown in Fig. 11, the prediction results acquired by the MSGFNet proposed are resembled to the real label, reaching the highest of prediction accuracy among all networks.

4) Validation Experiment on Sentinel-2 Dataset: Table V displays the extraction accuracy of different approaches in the same experimental conditions. Similarly, the MSGFNet proposed accomplishes the highest accuracy in the test region, with MIoU, precision, recall, F1, and aAcc reaching 93.12%, 93.62%, 94.88%, 94.25%, and 98.22%, respectively. The aAcc of the MSGFNet proposed is increased by 0.85% compared to UperNet. On the contrary, the extraction effects of PSPNet and ENCNet are the worst, with aAcc of 93.73% and 93.60%. As shown in Fig. 12, regions corresponding to cloudy (a, b, c), urban (d), ice (e, f), mountain (g), and other intricate scenes are selected for analysis. In cloudy and ice-containing regions, such as Fig. 12(a) and (e), PSPNet and ENCNet do not perform well in distinguishing cloud shadows and ice. Some ice bodies are misclassified as water bodies, and some water bodies are misidentified as cloud shadows. This is because the limited distinction between cloud shadows, ice and water bodies in certain spectral bands, leading to PSPNet and ENCNet unable to adequately learn water body features. There are elongated water bodies in Fig. 12(a)–(c) and (f). The proposed method indicates a more complete recognition of water body regions and outperformed other networks in terms of extraction completeness. In urban areas, there are numerous small water bodies distributed, with interference form building shadows, bridges and other ground objects, the extraction process becomes more complicated. In Fig. 12(d), (g), and (h), it can be observed that PSPNet, U-Net, HRNet, SegFormer, APCNet, ANNNet, ENCNet, and PSANet have noticeable noise in identifying small water bodies, and there are omissions and misrepresentation, especially in the processing of the edge details. This also validates the MSGFNet proposed has a certain generalization ability.

5) Negative Sample Experimental Results: In the research of water body extraction, data acquisition is a necessary condition. The data used for water body extraction should contain water bodies and nonwater bodies, in which the negative samples represent nonwater areas, including cloud shadows, buildings, roads, and other surface coverings. Negative samples can help the model better understand the background information, so that the model has strong robustness in extracting water bodies and is not easily interfered by nonwater body features. Owing to the high similarity between cloud shadows and water bodies, it is easy to be misidentified as water bodies. Therefore, cloud shadow is an indispensable negative sample in water body extraction. In this section, images containing cloud shadows in different scenes in the test dataset are selected, and experimental results are analyzed by comparing MSGFNet with other approaches. The visualization results of different approaches are illustrated in Fig. 13.

As shown in Fig. 13, MSGFNet exhibits a better extraction effect. In elongated water bodies containing cloud shadows, such as in Fig. 13(a), (c), (e), and (g), the extraction effect of PSPNet and ENCNet on elongated water bodies is poor, which may be due to the limitations of the model in recognizing water bodies of this morphology, where interfering factors such as cloud shadows in the image obscure the features of the water bodies, thus affecting the capability of model to accurately recognize elongated water bodies. In Fig. 13(d) and (f), there are plentiful cloud shadows distributed in the image, and the complex texture of cloud shadows brings a certain challenge to water body extraction. MSGFNet can better extract the water bodies from cloud shadows without being easily interfered with by the nonwater body features. This is because the global context provides rich spatial relationship and environmental information, and the integration of global context information can help the model adequately understand the relationship between water bodies and cloud shadows, and reduce misclassification and omission, thus strengthening the robustness and anti-interference ability of the model.

V. DISCUSSION

A. MSGFNet

UperNet is a classical segmentation network with encoder–decoder construction, composed of FPN and PPM. It aims to heighten segmentation accuracy by fusing multiscale features. However, since PPM adopts immobile pooling operations, leading to the receptive domain size cannot be adjusted. This limitation makes it difficult for UperNet to precisely capture the dynamic variations of water bodies inside the image. In this article, an MSEF module is designed based on UperNet, which consists of four dynamic convolutional modules, FPN, and four convolutional modules. The MSEF module is able to adapt to different scales and dynamically changing features through the filters with different kernel sizes inside the DCM, which enables the MSEF module to concern the internal changes of the input image, so as to flexibly perceive local information of the input data, and enhance the adaptability of the module in processing features at different locations. The convolution

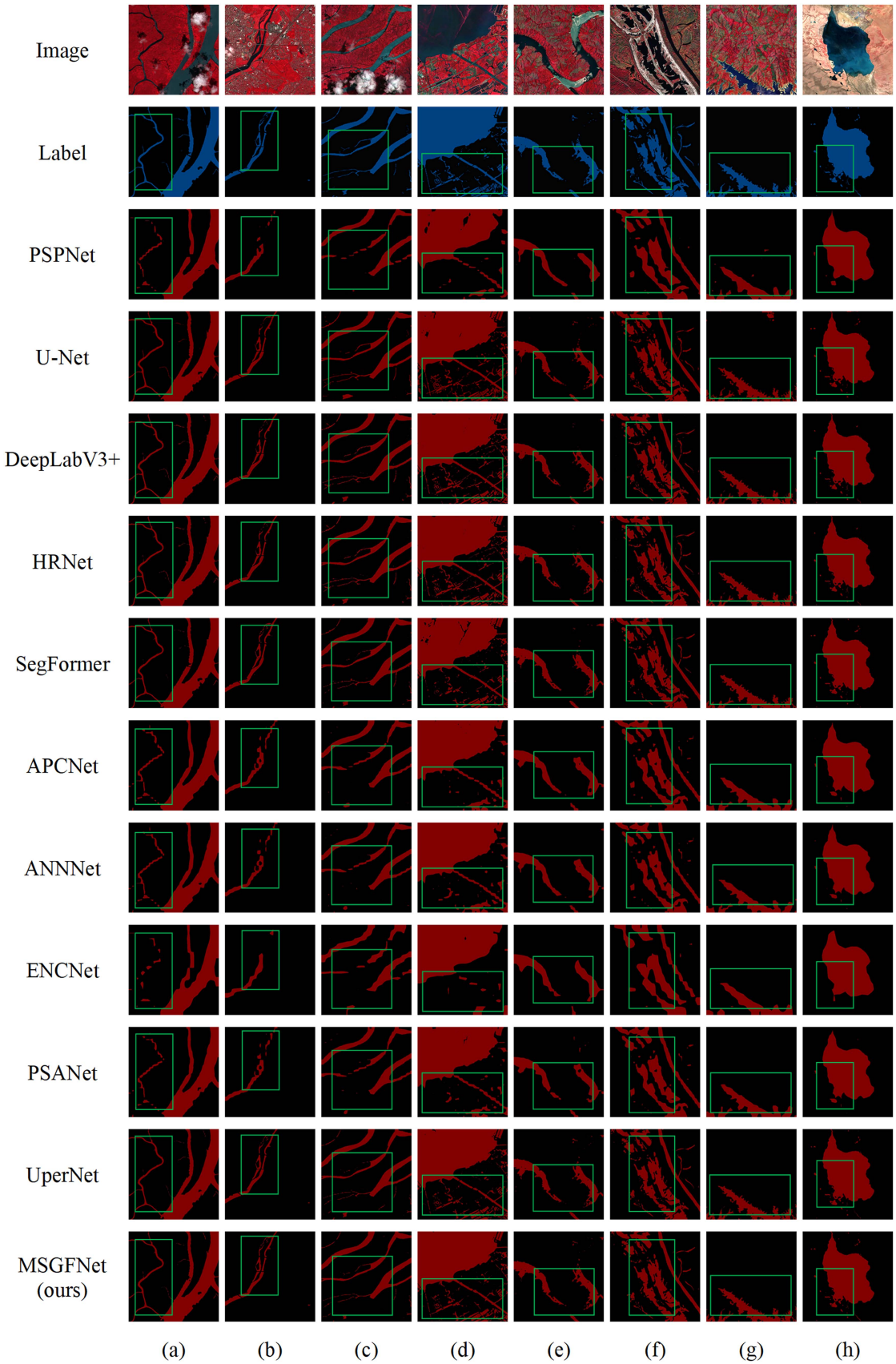

Fig. 12. Visualization results of comparative experiments on the test dataset of Sentinel-2 dataset. (a) Image 1. (b) Image 2. (c) Image 3. (d) Image 4. (e) Image 5. (f) Image 6. (g) Image 7. (h) Image 8.

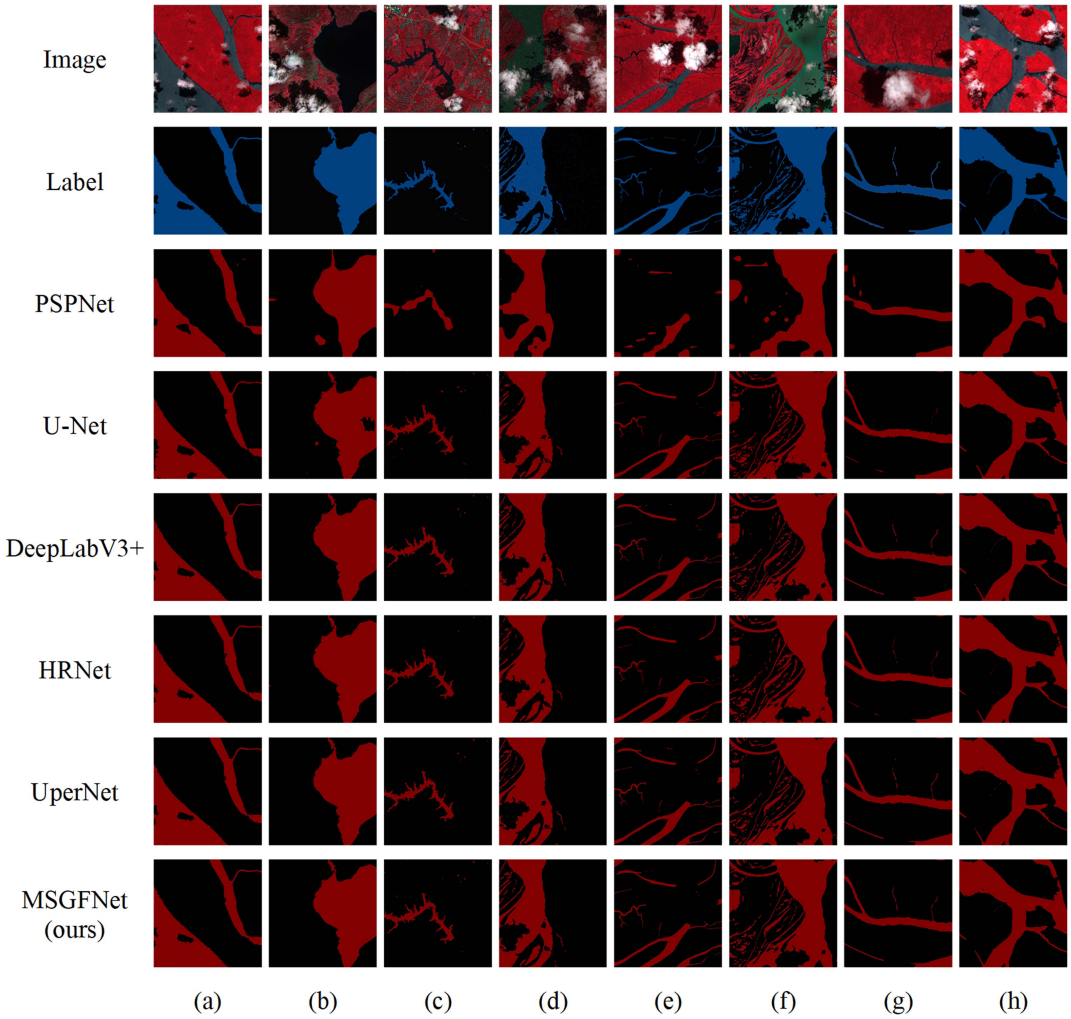

Fig. 13. Visualization results of negative sample experiments. (a) Negative sample image 1. (b) Negative sample image 2. (c) Negative sample image 3. (d) Negative sample image 4. (e) Negative sample image 5. (f) Negative sample image 6. (g) Negative sample image 7. (h) Negative sample image 8.

module mainly diminishes the dimension of feature map and eliminates aliasing effect caused by upsampling. In addition, a GCE Block is designed to help the network capture the global contextual information, and enhance the capability of network to recognize the semantics of the image.

B. MSGFNet Network Suitability Analysis

Both GF-1 and Sentinel-2 are high-resolution optical remote sensing satellites, which provide data support for resource monitoring, land management, and environmental protection. As the remote sensing technology progress, GF-1 images and Sentinel-2 images have been applied in water-related research.

In this article, the MSGFNet proposed accomplishes better result than PSPNet, U-Net, DeepLabV3+, HRNet, SegFormer, APCNet, ANNNet, ENCNet, PSANet, and UperNet on both GF-1 dataset and Sentinel-2 dataset. In Section IV, the water body extraction effect of different approaches is analyzed by metrics quantitative results and visualization results. Compared to other approaches, the MSGFNet proposed performs well in processing data with different sensors and resolutions, rather than just relying on a specific dataset. In addition, it also demonstrates the universality of the MSGFNet proposed to better accommodate the different data characteristics in the Gf-1 dataset and Sentinel-2 dataset. Compared with the GF-1 dataset, the Sentinel-2 dataset contains more abundant terrains, which can provide a more comprehensive scene understanding. Since different terrains require distinct extraction methods, learning the features of various terrains can help the network adapt to intricate topographical conditions, thereby reinforcing its adaptability. In addition, it can also help the proposed method to adequately understand diverse geomorphic features and enhance the generalization capability of the proposed method in some new or unseen geographic regions. Besides, the presence of negative samples, such as ice and cloud shadows, can increase the robustness of the proposed method. By learning how to deal with different interference factors, it can help the proposed method to make precise predictions when processing similar negative samples, consequently diminishing overfitting to noise and exceptional cases.

VI. Conclusion

In this article, part of the Henan section of the Yellow River main stream is taken as study area, and surface water bodies are extracted based on GF-1 remote sensing satellite data. To make up for the lack of some semantic information in UperNet in the process of feature fusion, an MSEF module is designed based on UperNet, and a water body extraction network named MSGFNet is proposed. The MSGFNet proposed can dynamically capture semantic information, according to the feature changes and contextual information of the input image. The optimal structure of the network is determined by training and testing based on GF-1 dataset, and the feasibility of the network is proved by ablation experiment and comparison experiment. In addition, a validation experiment is implemented to validate the learning capability of the network for different sensor data. In the experimental part, the network achieves the highest index values on different datasets compared to some existing methods. The experimental results demonstrate that the completeness and accuracy of the network for surface water body extraction are prominently superior to other approaches. In the study of water body extraction, the combined use of GF-1 and sentinel-2 datasets improves the unsatisfactory extraction results caused by data resolution, topography, climate, and other factors, which are difficult to be solved by single type of data to a certain extent.

Although the proposed method has accomplished better extraction effects on both GF-1 dataset and Sentinel-2 dataset, it still remain certain limitations and need to further improve in future work.

1) In subsequent study, we will focus on the integrated use of multisource data, adequately utilizing the advantages of various data to improve the accuracy and integrity of water body extraction.
2) Existing methods focus on the identification and utilization of positive samples, while ignore the importance of negative samples. The future study direction will further explore the application of negative samples in water extraction tasks.
3) As the continuous progress of deep learning technology, the advantages of CNN and swin transformer will be considered to combine and build a more comprehensive water body extraction model.
4) The practical application of the proposed method will be actively explored in the fields of water resources management, flood disaster monitoring, urban planning and construction, so as to provide true and accurate water body information.

References

[1] P.-S. Frazier and K.-J. Page, "Water body detection and delineation with landsat TM data," *Photogrammetric Eng. Remote Sens.*, vol. 66, no. 12, pp. 1461–1468, 2000.

[2] K. Jia, W. Jiang, J. Li, and Z. Tang, "Spectral matching based on discrete particle swarm optimization: A new method for terrestrial water body extraction using multi-temporal landsat 8 images," *Remote Sens. Environ.*, vol. 209, pp. 1–18, 2018.

[3] X.-B. Wang et al., "A robust multi-band water index (MBWI) for automated extraction of surface water from landsat 8 OLI imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 68, pp. 73–91, 2018.

[4] K.-Y. Deng and C. Ren, "Water extraction model of multispectral optical remote sensing image," *Acta Geodaetica et Cartographica Sinica*, vol. 50, no. 10, pp. 1370–1379, 2021.

[5] Y.-X. Wu, Z.-M. Chen, and J.-J. Li, "Semi-supervised deep learning network based water body segmentation method," *J. Zhengzhou Univ. (Natural Sci. Ed.)*, vol. 55, no. 6, pp. 29–34, 2023.

[6] Q.-Y. Duan, L.-K. Meng, Z.-W. Fan, W.-G. Hu, and W.-J. Xie, "Applicability of the water information extraction method based on GF-1 image," *Remote Sens. Natural Resour.*, vol. 27, no. 4, pp. 79–84, 2015.

[7] D. Sharma, T. Sharma, and J. Singhai, "Extraction of water bodies from visible color satellite images using PCA feature map," in *Proc. IEEE Int. India Geosci. Remote Sens. Symp.*, 2021, pp. 1–4.

[8] 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 Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 927–931, Jun. 2019.

[9] Q.-Y. Li, Y.-S. Chen, X. He, and L.-B. Huang, "Co-training transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024, doi: 10.1109/TGRS.2024.3354783.

[10] J.-H. Lin, Y. Zhao, S.-G. Wang, and T. Yu, "YOLO-DA: An efficient YOLO-based detector for remote sensing object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[11] W. Long, Y.-J. Zhang, Z.-W. Cui, Y.-J. Xu, and X.-X. Zhang, "Threshold attention network for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, doi: 10.1109/TGRS.2022.3197334.

[12] G.-J. Wang et al., "Water identification form the GF-1 satellite image based on the deep convolutional neural networks," *Nation Remote Sens. Bull.*, vol. 26, no. 11, pp. 2304–2316, Nov. 2022.

[13] J.-J. Li et al., "Accurate water extraction using remote sensing imagery based on normalized difference water index and unsupervised deep learning," *J. Hydrol.*, vol. 612, Sep. 2022, Art. no. 128202, doi: 10.1016/j.jhydrol.2022.128202.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput., Assist Intervention*, 2015, pp. 234–241.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[16] Z. Dong, G. Wang, S. O. Y. Amankwah, X. Wei, Y. Hu, and A. Feng, "Monitoring the summer flooding in the Poyang Lake area of China in 2020 based on Sentinel-1 data and multiple convolutional neural networks," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, Oct. 2021, Art. no. 102400.

[17] 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.*, 2018, pp. 801–818.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," Jun. 2017, Accessed: Aug. 5, 2023, *arXiv:1706.05587*.

[19] Z.-B. Wang, X. Gao, Y.-N. Zhang, and G.-H. Zhao, "MSLWENet: A novel deep learning network for lake water body extraction of Google remote sensing images," *Remote Sens.*, vol. 12, no. 24, Dec. 2020, Art. no. 4140.

[20] C.-J. Ge, W.-J. Xie, and L.-K. Meng, "Extracting lakes and reservoirs from GF-1 satellite imagery over China using improved U-net," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[21] P. Qin, Y.-L. Cai, and X.-L. Wang, "Small waterbody extraction with improved U-net using Zhuhai-1 hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[22] M. Li, P. Wu, B. Wang, H. Park, H. Yang, and Y. Wu, "A deep learning method of water body extraction from high resolution remote sensing images with multisensors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3120–3132, 2021.

[23] X. Lyu, W. Jiang, X. Li, Y. Fang, Z. Xu, and X. Wang, "MSAFNet: Multiscale successive attention fusion network for water body extraction of remote sensing images," *Remote Sens.*, vol. 15, no. 12, Jun. 2023, Art. no. 3121.

[24] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multi-scale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023.

[25] J. Kim, J.-K. Lee, and K.-M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[26] D.-H. Ma, L.-G. Jiang, J. Li, and Y. Shi, "Water index and swin transformer ensemble (WISTE) for water body extraction from multispectral remote sensing images," *GIScience Remote Sens.*, vol. 60, no. 1, 2023, Art. no. 2251704, doi: 10.1080/15481603.2023.2251704.
[27] Q. Zhao, J.-H. Liu, Y.-W. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[28] A. Vaswani et al., "Attention is all you need," Jun. 2017, *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5999–6009, 2017.
[29] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–22.
[30] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[31] Z.-Y. Xu, W.-C. Zhang, T.-X. Zhang, Z.-F. Yang, and J.-Y. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3585.
[32] J.-K. Li, G.-G. Li, T. Xie, and Z.-B. Wu, "MST-UNet: A modified swin transformer for water bodies' mapping using Sentinel-2 images," *J. Appl. Remote Sens.*, vol. 17, no. 2, 2023, Art. no. 026527.
[33] Q. Zhang, X.-Y. Hu, and Y. Xiao, "A novel hybrid model based on CNN and multi-scale transformer for extracting water bodies from high resolution remote sensing images," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 10, pp. 889–894, 2023.
[34] J. Chen, M. Xia, D.-H. Wang, and H.-F. Lin, "Double branch parallel network for segmentation of buildings and waters in remote sensing images," *Remote Sens.*, vol. 15, no. 6, Mar. 2023, Art. no. 1536.
[35] Z.-C. Wang, M. Xia, L.-G. Weng, K. Hu, and H.-F. Lin, "Dual encoder-decoder network for land cover segmentation of remote sensing image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2372–2385, 2024.
[36] Z. Dong, G.-M. Gao, T.-Z. Liu, Y.-F. Gu, and X.-R. Zhang, "Distilling segmenters from CNNs and transformers for remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, Jul. 2023.
[37] J. Song and X.-B. Yan, "The effect of negative samples on the accuracy of water body extraction using deep learning networks," *Remote Sens.*, vol. 15, no. 2, Jan. 2023, Art. no. 514.
[38] M. Wieland et al., "S1S2-Water: A global dataset for semantic segmentation of water bodies from Sentinel-1 and Sentinel-2 satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1084–1099, 2024.
[39] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021, doi: 10.1109/TGRS.2020.2994150.
[40] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[41] H.-F. Zhong, Q. Sun, H.-M. Sun, and R.-S. Jia, "NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[42] J.-J. He, Z.-Y. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3561–3571.
[43] 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.*, 2017, pp. 936–944.
[44] X. Luo, X.-H. Tong, and Z.-H. Hu, "An applicable and automatic method for earth surface water mapping based on multispectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102472.
[45] T.-T. Xiao, Y.-C. Liu, B.-L. Zhou, Y.-N. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 432–448.
[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[47] X.-L. Wang, R. Girshick, A. Gupta, and K.-M. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[48] Y. Cao, J.-R. Xu, S. Lin, F.-Y. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 1971–1980.
[49] K. Sun, B. Xiao, D. Liu, and J.-D. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. 32nd IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.
[50] E.-Z. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[51] J.-J. He, Z.-Y. Deng, L. Zhou, Y.-L. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7511–7520.
[52] Z. Zhen, M.-D. Xu, S. Bai, T.-T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 593–602.
[53] H. Zhang et al., "Context encoding for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.
[54] H.-S. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 270–286.

Ru Miao received the M.S. degree in applied mathematics from the School of Computer and Information Engineering, Henan University, Kaifeng, China, in 2008, and the Ph.D. degree in cartography and geographical information system from the College of Geography and Environmental Science, Henan University, in 2014.

She is currently an Associate Professor with the School of Computer and Information Engineering, Henan University. Her research focuses on spatial data processing and visualization.

Tongcan Ren received the M.S. degree in computer technology from Henan University, Kaifeng, China.

His research focuses on image processing.

Ke Zhou received the M.S. degree in applied mathematics from the School of Computer and Information Engineering, Henan University, Kaifeng, China, in 2011, and the Ph.D. degree in signal and information processing from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2016.

He is currently an Associate Professor with the School of Computer and Information Engineering, Henan University. His research focuses on spatial information processing.

Yanna Zhang received the M.S. degree in applied mathematics from School of Computer and Information Engineering, Henan University, Kaifeng, China, in 2008.

She is currently with Advanced Laboratory, School of Computer and Information Engineering, Henan University. Her research focuses on image processing.

Jia Song received the M.S. degree in communication and information systems from the China University of Geosciences, Wuhan, China, in 2005, and the Ph.D. degree in cartography and geographical information system from the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently an Associate Professor with the Institute of Geographic Sciences and Natural Resources Research. His research interests include intelligent methods and applications of remote sensing and remote sensing geographic information cloud computing technology.

Guangyu Zhang received the B.S. degree in civil engineering from the North China Institute of Science and Technology, Langfang, China, in 2019. He is currently working toward the M.S. degree in computer technology from Henan University, Kaifeng, China.

His research interest is image processing.

Jiaqian Wang received the B.M. degree in electronic commerce in 2021 from Henan University, Kaifeng, China, where she is currently working toward the M.Sc. degree in software engineering.

Her research focuses on image processing.