MRSE-Net: Multiscale Residuals and SE-Attention Network for Water Body Segmentation From Satellite Images

Xinyu Zhang, Jinjiang Li, and Zhen Hua

***Abstract*—Automatic extraction of water bodies from various satellite images containing complex targets is a very important and challenging task in remote sensing and image interpretation. In recent years, convolutional neural networks (CNNs) have become an important choice in the field of semantic segmentation of remote sensing images. However, generic CNN models present many problems when performing water body segmentation, such as: 1) blurred water body boundaries; 2) difficulty in accommodating different scales of rivers, often losing information about many small-scale rivers; and 3) a large number of trainable parameters. This article proposes an end-to-end CNN structure based on multiscale residuals and squeeze-and-excitation (SE)-attention for water segmentation, called MRSE-Net. MRSE-Net consists of an encoder–decoder and a skip connection, which captures contextual information at different scales using the encoder, and then passes the encoder feature mapping through the improved skip connection, while localization is achieved by the decoder is implemented. With the multiscale residual module, the number of parameters in our model can be significantly reduced and water pixels can be extracted accurately. The SE-attention module is used to enhance the prediction results, mitigate the blurring effect, and make the segmented water boundaries more continuous. Landsat-8 images are used to train our model and validate our proposed method's performance and effectiveness. In addition, we evaluate our method on Landsat-7 and Sentinel-2 images and obtain the best water segmentation results. Preliminary results on Sentinel-2 images show that the cross-sensor generalization capability of our model is beyond the range of the Landsat sensor family.**

***Index Terms*—Convolutional neural network (CNN), deep learning, multiscale residual, SE-attention, satellite image analysis, water body extraction.**

I. Introduction

SURFACE water refers to the general term for static and dynamic water on the land surface and contains various solid and liquid water bodies, such as rivers, lakes, and glaciers. Surface water is an important ecosystem on Earth, and although it only accounts for 1.75% of the total global water storage, it plays an important role in biological survival [1], climate change [2], and the global water cycle [3]. With the increasing urbanization and industrialization, the pollution and decline of surface water are becoming more and more serious; therefore, studying the spatial and temporal distribution of surface water and accurately identifying its boundaries are of great significance for sustainable human development [4], environmental protection [5], and urban planning [6].

Even though surface water is so important for the Earth's ecosystem and human survival, our research on surface water trends and area changes is very poor [7]. This is mainly because our understanding of it relies mainly on manual surveys and annotations, and although the results obtained are highly accurate, the inspection period is long, real-time data are often not available, and the labor cost is extremely high, which cannot meet our growing practical needs for regional extraction of water bodies. In recent years, with the development of the space industry, the large-scale popularity of remote sensing satellites, such as Landsat and Sentinel-I has provided us with low-cost and reliable global remote sensing images, which can be imaged in various terrain conditions by their being equipped with high-resolution microwave sensors that are not disturbed by day and night alternations and penetrate through thick clouds.

Remote sensing is a rapidly developing and widely used technology that is time-sensitive and periodic [8], [9], its access to information is less restricted by conditions, and it allows simultaneous observation of large areas, and the land satellite images we obtain through remote sensing can cover an area of more than 30 000 km2, and many algorithms have been developed to extract water bodies using remote sensing images. In the early days, people mainly extracted water bodies by the threshold method, in which the single-band threshold method uses the low reflectance of water bodies in the near-infrared band to determine a grayscale threshold that distinguishes water bodies from other objects for water body extraction, which is the simplest method for water body identification. However, this method has limitations as it cannot distinguish water bodies from mountain shadows when used in mountainous areas. The spectral water index, on the other hand, fully considers the correlation between different bands, and the mapping accuracy is relatively high, plus the lower cost [10] makes it more widely used, among

Manuscript received 13 March 2022; revised 13 June 2022 and 17 June 2022; accepted 19 June 2022. Date of publication 22 June 2022; date of current version 30 June 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61772319, Grant 62002200, Grant 61972235, and Grant 12001327, in part by the Shandong Natural Science Foundation of China under Grant ZR2021QF134 and Grant ZR2021MF068, and in part by the Yantai science and technology innovation development plan under Grant 2022JCYJ031. *(Corresponding author: Jinjiang Li.)*

Jinjiang Li is with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: lijinjiang@gmail.com).

Xinyu Zhang and Zhen Hua are with the School of Information and Electronic Engineering, Shandong Technology and Business University, Institute of Network Technology, Yantai 264005, China (e-mail: 1170752813@qq.com; huazhen66@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3185245

which, the normalized difference water index (NDWI) proposed by McFeeters [11] is the first water body index, and the subsequent. There is also the modified normalized difference water index (MNDWI) proposed by Xu [12] for the poor performance of NDWI in the face of buildings. In the past decades, many other water indices have been proposed [13]–[15], but they perform generally in the face of complex scenes that include shadows, buildings, and thin clouds at the same time, and they all require manual adjustment of thresholds, which can easily fall into local optimal thresholds and not express the best results.

In addition, there are some more commonly used methods for water body extraction, such as support vector machines [16], [17], the active contour model [18]–[20], the Markov random field (MRF)-based model [21]–[23], and object-based classification [24], [25]. However, traditional machine learning methods, such as support vector machines and decision trees are based on single-pixel points for recognition, which do not take into account the connection between individual pixel points, and the recognition accuracy is not very high. The active contour model is sensitive to the initial position, and it is difficult to obtain the initial position automatically. MRF-based methods are computationally intensive and difficult to apply to large-area images, especially in remote sensing directions, and often result in many small-scale image objects (pretzel noise) during segmentation. Although object-based classification methods utilize texture and spectral features in remotely sensed images, the determination of the optimal scale for water body extraction and the selection of features directly affect the accuracy of the final segmentation. Overall, although these traditional water body extraction methods can effectively obtain water body information, the extraction results have more serious pretzel noise and are susceptible to complex environments, making them difficult to be applied to large-scale automatic water body extraction globally.

In recent years, the update of hardware devices and the emergence of large-scale datasets have driven the development of deep learning, especially convolutional neural networks (CNNs), which have improved the prediction accuracy by introducing the correlation between adjacent pixels in images into the content recognition process through their unique perceptual field mechanism, making them the mainstream methods in the field of semantic segmentation. Among them, the full CNN-based [26]–[28] and the encoder–decoder structure (U-type) [29]–[31] have become two representative network structures in the field of semantic segmentation of remote sensing images, especially the U-Net, which has greatly outperformed the traditional methods for water body segmentation [32], [33]. Dai *et al.* [34] proposed a new edge-based loss function for the problem of boundary detail loss in segmentation of the bilateral segmentation network (BiSeNet) [35], which improved BiSeNet and improved the segmentation accuracy. Li *et al.* [36] added spatial pyramid pooling (SPP) modules and attention modules to build a more robust PA-U-Net water extraction network [37], reducing the probability of false segmentation. Dirscherl *et al.* [38] used the atrous spatial pyramid pooling (ASPP) module [39] extracted multiscale features and improved the U-Net by combining shallow and deep features using jump connections to improve the accuracy of water extraction for Sentinel-1 and Sentinel-2. Ren *et al.* [40] proposed a dual-attention U-Net model (DAU-Net), which uses a position attention module (PAM) and a channel attention module (CAM) to improve the model's characterization capability and improves the accuracy of water body segmentation by 1% compared to the original U-Net network.

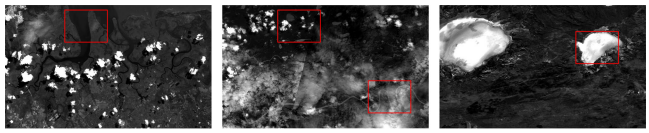

Fig. 1. Scale variation of water bodies in remote sensing images. (a)–(c) Examples of wider rivers, narrow rivers, and oval-like lakes captured by the Landsat-8 remote sensing satellite, respectively.

In this article, we treat water body extraction as a binary semantic segmentation task and use it to generate pixel-level water body annotations by labeling each pixel point in our adopted remote sensing images as water bodies and nonwater bodies to generate binary masks, a step that is implemented by Google Earth Engine (GEE) [41] and water body labels in the Global Surface Water Dataset [42]. From this perspective, water body extraction and semantic segmentation have the same goal of assigning a label to each pixel point of the input image, which belongs to a pixel-level object classification task, and semantic segmentation techniques based on CNNs have been successfully applied to many fields, such as medical image processing and autonomous driving. Therefore, it is possible to use CNN-based methods, which perform well in various semantic segmentation tasks, for the water body extraction problem and obtain a good result, which gives us a key basis for using CNN-based methods for water body extraction, and the increasing number of CNN-based water body extraction methods [43]–[45] proposed in recent years validates our view.

However, if generic CNN structures [fully convolutional networks (FCNs) [26] or U-Net [29]] are used directly for surface water extraction without modification, poor prediction accuracy and blurred water body boundaries often occur, accompanied by visual degradation. Therefore, how to improve the prediction accuracy and maintaining the accurate segmentation of water body boundaries are two important problems faced by water body segmentation of remote sensing images. In the remote sensing image water body extraction task, we are interested in segmenting various water bodies (lakes, rivers, glaciers, etc.) from various forms of remote sensing images, however, these objects of interest often have irregular and different proportions. As shown in Fig. 1, we find that the size of water bodies in remote sensing images may vary greatly, from narrow or wide rivers to oval-like lakes, which are very common in the global remote sensing image water body extraction tasks.

After considering the above findings, if we want to get a satisfactory prediction result, the CNN we use should be designed to be able to analyze water bodies at different scales, and to our knowledge, this problem has been solved in some well-developed computer vision fields, such as medical image segmentation, target detection, etc., but in the field of remote sensing

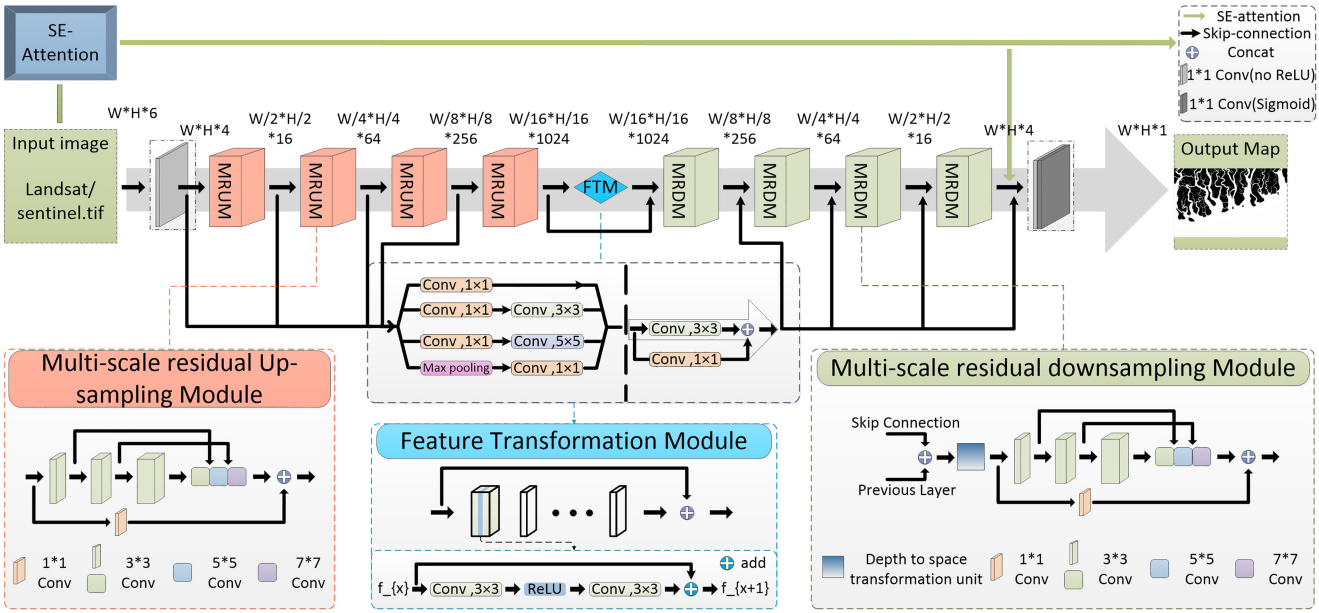

Fig. 2. Our proposed MRSE-Net network structure. We replace the convolutional blocks at each level of the original U-Net with the proposed multiscale residual module. In addition, we propose two methods to improve the skip connection and incorporate the SE-attention mechanism.

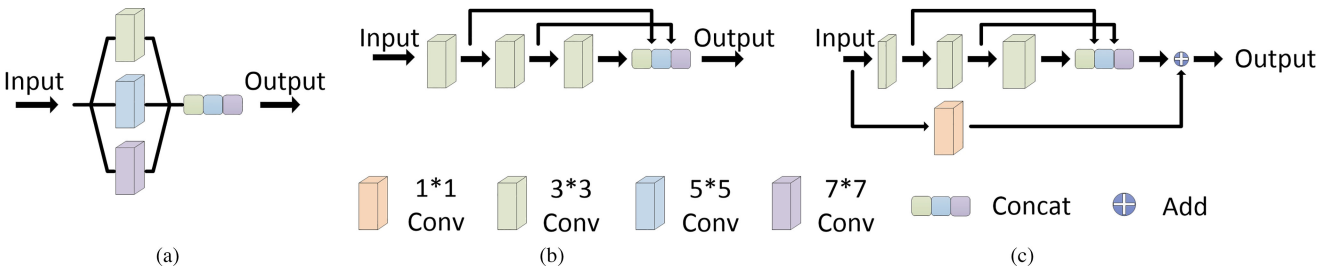

Fig. 3. We propose a multiscale residual module. (a) Starting from an inception-like structure, 3×3, 5×5, and 7×7 convolutions are used in parallel and connected to generate the output. This allows our network to obtain spatial features at different scales. However, instead of using the 3×3, 5×5, and 7×7 filters in parallel, (b) we decompose the 5×5 and 7×7 filters into a series of smaller 3×3 filters. (c) shows our multiscale residual module. To reduce the number of parameters, we gradually increased the number of consecutive three 3×3 filters and added a residual connection using a 1×1 filter for dimensionality maintenance.

image water body extraction, this problem is not well solved. The problem has not been well solved in the field of remote sensing image water extraction. In the early days, Serre *et al.* [46] were inspired by the visual modeling of human and primate vision by the visual cortex to solve the multiscale variation in images by a series of Gabor filters of different sizes. Later, Szegedy *et al.* [47] innovatively proposed the inception network structure, in which the inception module obtains information at different scales in an image by using convolutional kernels of different sizes in parallel, and later combines this feature information containing different scales together and passes it deep into the network. And Chen *et al.* [39] were inspired by SPP and proposed ASPP, which similarly obtained different scales of perceptual fields by using convolutional layers with different atrous rates. We found that in the general semantic segmentation model U-Net, there are two consecutive 3×3 convolutional layers, and it was demonstrated in [48] that these two consecutive 3×3 convolutional operations are equivalent to one 5×5 convolutional operation, so we use a series of consecutive 3×3 convolutional layers instead of 5×5 and 7×7 convolutional layers and increase the number of filters in each convolutional layer in turn, which can greatly reduce the memory requirement and also get the multiscale information we need in the image. Using larger convolutional kernels or pooling layers with larger strides to obtain larger perceptual fields will lead to too much computation and too much resolution loss, respectively, and too much resolution loss will result in feature information at the image boundaries not being utilized. Our approach can be said to be the easiest way to obtain multiscale spatial features from U-Net and ensure that the resolution does not drop too much, which we name the multiscale residual module (Fig. 3), and unlike the inception module, we obtain multiscale information by concatenating convolutional layers in series, rather than in parallel.

To obtain accurate and continuous water body boundaries as well as to improve the prediction accuracy, we propose an improved multiscale residual encoder–decoder network to

effectively extract water bodies from images captured by remote sensing satellites. Since the remote sensing images captured by satellites contain rich feature information and scenes, there are many redundant features, which greatly affect the model's feature recognition ability for water bodies. To improve the ability of the network to extract the most important key features for the segmentation task based on the relationships between individual features without increasing the model complexity and introducing new spatial dimensions, we incorporate a lightweight squeeze-and-excitation (SE) attention mechanism [49], which explicitly models the dependencies between individual channels through a "feature recalibration" strategy. We believe that the SE attention mechanism plays an important role for MRSE-Net in the global remote sensing image water extraction task, and the experiments in Section IV also demonstrate that the SE attention mechanism improves the prediction accuracy. The bottleneck layer with the highest number of channels in the network uses a feature transformation module of our design instead of the normal convolution layer to recover the missing information through the residual block and enable the network to learn feature information at the abstraction level, and also prevent the gradient disappearance or explosion problem in the deep network. The inception module or residual block is introduced into the skip connection to alleviate the semantic gap between the high-level encoder–decoder features at both ends of the skip connection, and to help upsampling recover images with feature information at a closer semantic level. In summary, the main contributions of our work are summarized as follows.

1) We propose an end-to-end CNN-based water segmentation network based on multiscale residuals and SE attention. Unlike previous conventional methods, our approach does not need to use the handcrafted features provided by the domain knowledge, but automatically extracts features through convolutional kernels at different scales, and this part of the features is more robust compared to the feature information obtained from a single convolutional layer, and is more robust in terms of image size and context.
2) The multiscale residual module can extract deeper high-level feature information and can analyze targets at different scales. Its combined application with the attention mechanism can not only obtain spectral features, shape, and texture features of water bodies but also capture the differences between water bodies and nonwater bodies at the pixel level.
3) Using the feature transformation module at the bottleneck layer allows the network to learn more feature information at the abstraction level, improving skip connections to alleviate the semantic gap between high-level encoder–decoder features. The water extraction results can be generated end-to-end by simply putting the remote sensing images into our trained network, which is more practical than the traditional methods with tedious steps.

The rest of the work in this article is organized as follows. Section II reviews the related work, Section III introduces our proposed method (MRSE-Net), and Section IV gives the details of the experiments. Finally, Section V concludes this article.

II. Related Work

A. Water Body Extraction Method

In the past decades, many methods have been proposed to extract water bodies from remotely sensed images, among which the most commonly used is still based on water spectral indices. The NDWI proposed by McFeeters *et al.* [11] initiated this method and played an inspirational role in a series of later methods based on water spectral indices. It has the drawback of being influenced by thin clouds and mountain shadows and is difficult to be applied to complex environments for water extraction. Xu *et al.* [12] replaced the near-infrared band in NDWI with the mid-infrared band and proposed the MNDWI, which better highlighted the water features and suppressed the noise of soil and vegetation, but it was also not very effective in the face of shadows. To address the problem that the optimal threshold in each image varies with time and geographical location, Feyisa [14] proposed the automated water extraction index (AWEI) to provide stable thresholds, but it does not apply to targets with high reflectance. The manual selection of the optimal threshold by commission and omission error rate is tedious and inefficient, but it is faced by almost every water spectral index-based method, and to overcome this limitation, Guo *et al.* [50] proposed weighted NDWI (WNDWI), which improves the overall accuracy (OA). The method based on the water spectral index tends to misclassify turbid water bodies and small water bodies in the shaded area, making it difficult to perform large-scale high accurate water extraction, and it is also a difficult task to choose the best threshold value.

Subsequent studies have mostly addressed the problems associated with water-based spectral indices, and Huang *et al.* [24] proposed a two-stage machine learning-based water extraction method for urban high-resolution remote sensing images, first extracting water bodies from the pixel level and then further identifying them from the object level, using both water/shadow/vegetation indices as well as geometric and textural features. Essa *et al.* [51] proposed a new hyperspectral image feature extraction method that considered not only local neighboring pixels in hyperspectral images but also neighboring pixels in three consecutive bands, thus extracting rich contextual information. Yao *et al.* [52] proposed an automatic urban water extraction method (UWEM) that addressed the building shadow effect in cities. However, although these newly proposed methods have higher accuracy and better water extraction, they are only applicable to urban water extraction and are difficult to be extended to scenes with different spectral and spatial features on a global scale.

B. Encoder–Decoder Structure

In recent years, the emergence of graphics processing units and large-scale datasets has driven the development of deep learning, especially CNNs, and has been successfully applied to various fields of remote sensing [53]. In the field of semantic segmentation of remote sensing images, FCNs [54] and encoder–decoder structures [43], [45], [55], [56] are the two of the most

representative network structures. Lin *et al.* [54] improved the FCN by localizing in the shallow layer of the network and detecting in the deep layer, achieving a compromise between accuracy and feature representation capability in the FCN. Feng *et al.* [43] combined superpixel segmentation and conditional random field (CRF) to propose an enhanced deep convolutional encoder–decoder network (DCED) for water body extraction from remote sensing images, which is slightly more parametric than other CNN-based water body extraction methods, although it effectively suppresses the pepper noise and ensures the continuity of the water body segmentation boundary. Tambe *et al.* [45] proposed a W-Net based on the encoder–decoder structure for water segmentation, where contextual information is obtained in the encoder part and image recovery is achieved in the decoder part, and the network parameters are reduced using the inception module and the blurring effect is reduced using the refinement module. Li *et al.* [55] proposed a novel deep CNN DeepUNet for sea-land segmentation, replacing the convolutional layers in the encoder and decoder with DownBlock and UpBlock, and obtaining more accurate segmentation results.

In addition, these encoder–decoder-based network structures have been widely used in other directions in remote sensings, such as road segmentation [57], [58], classification [59], [60], and building detection [61], [62]. This U-shaped encoder–decoder structure can perform data enhancement by applying random elastic deformation to the training images, which improves the invariance and stability of the network and gives good results even under the condition of small training samples. Second, U-Net can identify irregular boundaries well, so U-Net is widely used in various semantic segmentation. In this article, we use U-Net for remote sensing image water body extraction, but directly using it for remote sensing images with such a large image size (even up to 300 million pixels) often does not get ideal results. We find that it is difficult to capture the size variation of different water bodies in remote sensing images with only 3×3 convolution in the network, so we design a multiscale residual module to obtain multiscale information in the images. Then we add the SE-attention module [49], which is used to improve the classification accuracy and accelerate the convergence of the network, and the feature transformation module, which helps the network to obtain more abstract information at the bottleneck layer, and finally, the skip connection is deepened and widened using the inception module or the residual module, which helps the decoder to recover the image better and finally obtain more accurate water body extraction results.

III. Proposed Method

The images captured by remote sensing satellites contain rich feature information and a large number of redundant features, which greatly affect the model's ability to recognize the features of water bodies, and the size and shape of water bodies in remote sensing images may vary greatly, which is very common in the task of water body extraction from global remote sensing images. To enable the model to better handle the scale variation of water bodies in remote sensing images and to obtain accurate and continuous water body boundaries as well as to improve the prediction accuracy, we propose an improved multiscale residual encoder–decoder network to efficiently extract water bodies from images captured by remote sensing satellites. In this section, we detail the architecture of our proposed network (Fig. 2 and Table I) and explore the multiscale residual module that can improve the prediction accuracy and reduce the computational complexity of the network.

TABLE I
MRSE-Net Architecture Details

MRSE-Net			
	Unit Level	Layer(filter size)	Output size
Input			512*512*6
Encoding	Level 1	Conv2D(1*1)	512*512*4
	Level 2	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	256*256*16
	Level 3	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	128*128*64
	Level 4	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	64*64*256
	Level 5	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	32*32*1024
Bridge	Level 6	Conv2D(3*3) Conv2D(3*3)	32*32*1024
Decoding	Level 7	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	64*64*256
	Level 8	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	128*128*64
	Level 9	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	256*256*16
	Level 10	Conv2D(3*3) Conv2D(3*3) Conv2D(3*3) Conv2D(1*1)	512*512*4
Output	Level 11	Conv2D(1*1) Conv2D(1*1)	512*512*1

A. MRSE-Net Architecture

The main body of our network consists of four parts: 1) multiscale residual downsampling module, 2) feature transformation module, 3) multiscale residual upsampling module, and 4) an improved skip connection. To keep the size of the feature maps in the network consistent, we make the number of input and output channels in each upsampling and downsampling module multiply or divide by 4. In each multiscale residual module, we use strided convolution to reduce the size of the input image instead of max pooling, Springenberg *et al.* [63] stated that using pooling works best when the network is shallow in-depth, and conversely, using a convolution layer with a step size of 2 works best and will reduce the memory usage, where we use convolution (step size 2) to save some of the convolution and

pooling operations compared to the case of using convolution (step size 1) plus a max-pooling layer. A constant feature map size in the network is ensured by the above two steps, e.g., the feature map sizes of the first and bottleneck layers are $W \times H \times 4$ and $W/16 \times H/16 \times 1024$, respectively. In each multiscale residual downsampling module, the number of input channels is first multiplied by 4, followed by a multiscale residual module to extract features and reduce the image size. The multiscale residual upsampling module first consolidates the size $W \times H \times C$ feature information passed from the previous layer and sibling skip connections into $2W \times 2H \times C/4$ by a depth-to-space transformation unit, followed by a multiscale residual module to reduce the number of output channels and recover the image.

The bottleneck layer, as the location with the highest number of network channels, is often needed to recover the missing information after the downsampling operation, to learn the feature information at many abstraction levels, and prevent the gradient disappearance or explosion problem in deep networks, we use residual blocks, a method proven effective in image enhancement [64]. To maintain a balance between computational efficiency and memory usage, we define a feature transformation module according to the setting of [65], which includes three sets of residual modules (one set includes three residual blocks) to perform feature learning.

In the input part of the network, we use one 1*1 convolutional layer to accept input images of arbitrary size. Our input image is a remote sensing satellite image with six bands, corresponding to 6 channels of the input, but after a series of operations of downsampling, the number of channels will reach 1536 in the bottleneck layer, which greatly affects the memory usage. Therefore, we reduce the output channels to 4 by convolutional layers and place most of the trainable parameters in the layers with high feature levels. Compared with a channel number of 6, we reduce the memory usage by half by doing so, and we do not use the ReLU activation function after it to prevent excessive loss of feature information to prevent the effect of the early channel number reduction on the network. The final 1*1 convolutional output layer of the network has a Sigmoid activation function and is responsible for outputting the probability that each pixel point in the final feature map is a body of water. The remaining convolutional layers are followed uniformly by batch normalization and the ReLU activation function.

B. Structure of Multiscale Residual Module

Following AlexNet [66] winning the ImageNet competition by a large margin in 2012, CNNs began to emerge, and gradually CNN models, such as GoogLeNet [47], ResNet [67], and DenseNet [68] emerged, which greatly influenced the image segmentation field and are currently the mainstream image segmentation methods. Surface water segmentation, as a part of conventional remote sensing image segmentation, also suffers from large image size (images captured by Landsat are several orders of magnitude larger than those in conventional segmentation direction) and different sizes of segmentation targets (see Fig. 1). To extract deeper high-level feature information and to analyze targets at different scales, we propose a multiscale residual module (as in Fig. 3), through a series of 3*3 convolutional layers which are concatenated to achieve the effect of simulating convolutional layers of different scales. In this module, the input of the latter layer is convolved by the previous layer operation, which greatly reduces the number of parameters, and to some extent avoids the memory computation overload and gradient disappearance or explosion caused by the network being too deep.

In surface water segmentation, we are interested in segmenting water bodies, such as lakes and rivers from waveband images taken by remote sensing satellites under different topographic and environmental conditions. However, in most cases, these water bodies tend to have irregular and different scales, as shown in Fig. 1, where we have demonstrated that the size and shape of rivers and lakes vary greatly from region to region. Therefore, the network should be able to analyze water bodies at different scales, and although these problems have been well addressed in other computer vision fields, to our knowledge, this problem has not been well addressed in the direction of surface water segmentation of remotely sensed images, where generic network frameworks tend to focus only on improving accuracy by increasing the depth of the network; until GoogLeNet revolutionarily introduced the inception module, which obtains feature information at different scales by using a series of convolutional layers with different Kernel sizes in parallel and combining them to increase the width of the network. Following inception's approach, the simplest way to get multiscale feature information in U-Net is to use convolutional layers of 3*3, 5*5, and 7*7 kernel sizes in parallel at each layer, as in Fig. 3(a). However, using 5*5 and 7*7 kernel size convolutional layers in parallel in the network will greatly increase the memory requirement, as mentioned in [48]; the feature information extracted from adjacent layers of the neural network is correlated, so we follow its idea by using multiple 3*3 convolutional layers with smaller memory requirement instead of 5*5 and 7*7 kernel size convolutional layers with higher memory requirement, and then combine them by the concatenation operation, as in Fig. 3(b). The outputs of two consecutive 3*3 and three 3*3 convolutional layers are effectively close to those of 5*5 and 7*7 convolutional operations, respectively, which reduce the parameters while enhancing the nonlinearity, and the performance will be even better.

However, the number of filters in the first convolutional layer has a square effect on the memory requirement when multiple convolutional layers are connected in series in this way [47], so we reduced the number of filters in the first convolutional layer (1/6, 1/3, and 1/2 of the number of output channels from the first to the third, respectively), which greatly reduces the memory requirement and prevents the network from computational collapse at the early stage of training. In addition, inspired by [69], we introduced residual connectivity and used a 1*1 convolution to provide feature information at other scales, which we define as a multiscale residual module, as shown in Fig. 3.

C. Structure of SE-Attention Module

Convolutional kernels obtain global information by learning spatial and channel information in the local perceptual field; however, it is very difficult to learn a large amount of feature

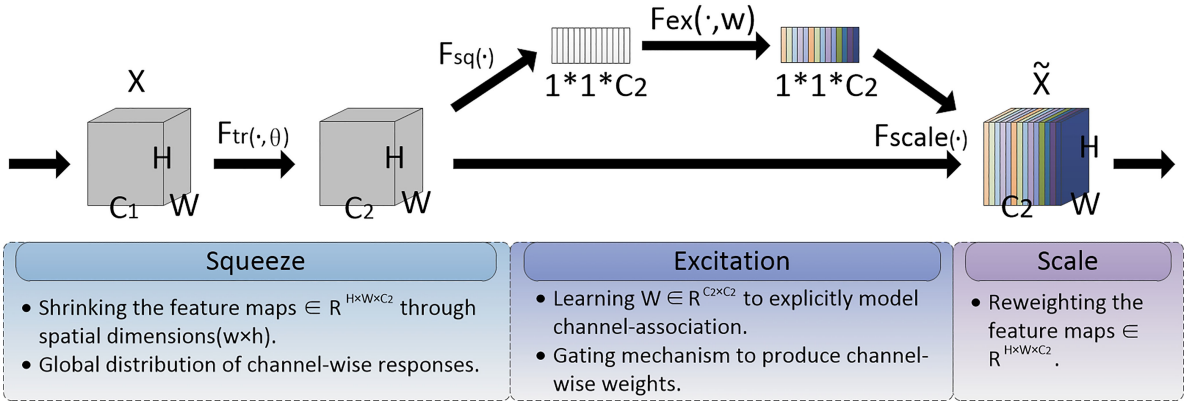

Fig. 4. Specific network structure of the squeeze-and-excitation (SE) attention module.

information and perform well, and most of the earlier studies have focused on improving spatial coding ability to improve network performance, in this context, SE-attention [49] starts from the perspective of feature channel information by explicitly modeling the interdependence of its channels. Ultimately, the SE-attention module does not increase the depth or width of the network, but only leads to a slight increase in the number of parameters, and the feature channel recalibration strategy it employs can adjust the importance weights of each channel according to the different levels of contribution of each channel to the network; in general, the network can automatically obtain new weights for each feature channel from the learned global information to enhance useful and suppress the redundant feature channels. As shown in Fig. 2, we introduce the SE-attention module in the last layer sampled on the network to make the output feature channels more directional. The function of this module is implemented in three main steps, the first step is to obtain the global compressed feature amount of each feature channel through the global average pooling operation, and the second step obtains a new weight value corresponding to each feature channel from 0 to 1 through the two fully connected layers, and the last step. The feature channel recalibration function of the SE-attention module is implemented by multiplying the new weight value with the 2-D matrix of the corresponding channel of the initial feature map and using it as the next input, see Fig. 4.

D. Structure of Improved Skip Connection

One of the highlights of the U-Net architecture is the skip connection introduced between the codecs, which passes the missing information due to the downsampling operation and improves the segmentation accuracy. However, we speculate that although this makes use of the missing information, the semantic level of the feature information of the codec connected by the first skip connection is very different, when the feature information from the encoder is at an earlier layer of the network and has undergone only a little processing, while the feature information of the corresponding decoder is deeper in the network and has undergone more processing, and its semantic level is higher and more abstract. Therefore, we believe that there may be a semantic gap between them, and directly splicing them may hurt network optimization.

To alleviate the semantic gap between them, we need to increase the semantic level of the feature information from the encoder by passing them through more processing before splicing them at the decoder side, and instead of using the usual convolutional layers, we use residual modules because they will make learning simpler, while not reducing the network efficiency [70]. As shown in Fig. 5, instead of directly splicing the codecs through the skip connection, we introduce a residual connection in the skip connection to pass the processed feature information from the encoder, and we define the improved skip connection as the multiresidual skip connection (Multi-Res path). It is worth noting that the semantic level of feature information from the encoder and decoder gradually increases or decreases in the next skip connections, and the difference between them gradually decreases, so we reduce the number of residual modules in the skip connections step by step, where the semantic level of feature information on both sides of the bottleneck layer is not much different, so we do not improve the skip connections here to avoid introducing too many parameters. In particular, we use 4, 3, 2, and 1 residual modules in each of the four multiresidual skip connections.

However, this inevitably introduces a certain amount of parameters to the network, so we also provide another way to improve the skip connection by adding the inception module to it instead of the residuals module. As shown in Fig. 2, we add the inception module to the skip connection except for the last level, which also shortens the semantic gap of feature information on both sides of the skip connection and does not introduce too many parameters, and defines the improved skip connection as inception skip connection (inception path). Under our tests, although the average performance of the network incorporating the Inception path is lower than that of the network using multiresidual skip connections (which may be because the residual module improves the efficiency of network learning), both are stronger than the normal skip connections, and it introduces a very small number of parameters. If the requirement for the number of parameters is not high, it is recommended to use the multiresidual skip connection, and vice versa with the Inception module.

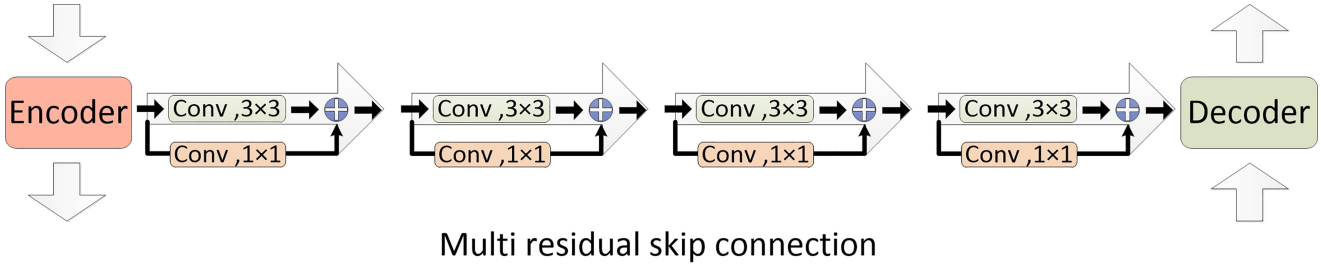

Fig. 5. Proposed multiresidual skip connection. Instead of passing the encoder feature mapping directly to the decoder, we use many residual connections to pass the encoder features. These nonlinear operations reduce the semantic gap between encoder and decoder features, make learning easier, and learn much more abstract semantic information through residual connections.

E. Loss Function

In water body extraction, our network only needs to determine whether the pixel points of the input image are water or nonwater pixels, and the training image only contains labels for water and nonwater pixels, but we also have pixels that approximate water bodies, such as clouds, shadows, snow, and ice, etc. It is difficult to classify them accurately if only generic loss functions (e.g., Mse, Softmax) are used because they will place the easier-to-learn pixels and harder-to-learn pixels in the same hierarchy guide as the network training. Therefore, we use a combination of max-pooled adaptive loss [71] and focal loss functions [72] to help the network to enhance the learning of this part of the pixels that are harder to classify. Among them, the max-pooled adaptive loss function can automatically shift the focus of the network learning to the most difficult part of the input image to better handle river crossings and coastlines, thus improving the prediction accuracy. And the focal loss functions can reduce the weight of simple negative samples, thus focusing the learning on the more difficult to classify water-like pixels.

F. Implementation

Our proposed MRSE-Net is trained and tested using the publicly available Tensorflow. The model parameters are: input image size (512*512), ground truth (512*512), batch size (16) with several epochs (150), and optimizer (SGD with the momentum of 0.9) with the learning rate (0.1). First, we performed horizontal or vertical rotation of each patch for data enhancement and later trained MRSE-Net with the enhanced data. all our experiments were performed on Intel(R) Core(TM) i7-10700 CPU at 3.80 GHz with 16 GB RAM and NVIDIA TITAN RTX with 24 GB GDDR5X memory, CUDA 11.0 edition.

IV. Experiment

To demonstrate the effectiveness of our proposed MRSE-Net, we performed qualitative and quantitative evaluations with remotely sensed images taken by Landsat-8 and validated the generalization capability of our method by performing cross-dataset evaluations on Landsat-7 and Sentinel-2 images in different bands. In addition, we perform ablation experiments to investigate the effect of the proposed module on water segmentation of remotely sensed images.

A. Data

The dataset we used is a dataset consisting of multispectral images and their corresponding pixel-level labels collated through GEE [41]. To maximize compatibility between different sensors, we used images in bands 2–7 of Landsat-8, which have equivalence with images taken by Sentinel-2 and some Landsat remote sensing satellites, as input, and did not use other less common bands of this remote sensing satellite, ground truth output was obtained using images from the global surface water dataset [42] in the water body labels, and the synthetic images are generated by calculating the median pixel values of the input images with the ground-truth labels at the same moment to form the dataset needed for our model, which does not remove the clouds considering the presence of more or fewer clouds in the real images, but keeps the clouds to let the network learn how to process them, thus improving the accuracy of water body segmentation in the real images.

The quality of the dataset has a significant impact on the segmentation accuracy of the network; too little quantity or too low quality can lead to degradation of segmentation accuracy and problems, such as artifacts. The DeepWaterMap [73] designed by Isikdogan *et al.* confused water bodies and nonwater bodies because of inaccurate definitions of the labels of clouds and shadows in the dataset, which led to false positives for these categories. Zhou *et al.* 's work [74] used a dataset that was corrected to increase the classification accuracy from 83% to 92%, all of which demonstrate the importance of high-quality datasets. Since most of the images in the synthetic dataset contain almost exclusively land or water bodies, this class-imbalanced dataset leads the network to learn shortcuts (with little use of nonlinear transformations) and to situations where accuracy is high during training but very low during testing. To avoid the above problem, we balanced the proportion of different categories by removing images with more than 99% of land or water bodies in the dataset, and then after data enhancement (random horizontal or vertical flip) of them, we finally obtained a synthetic dataset of more than 400 000 images (as in Fig. 6) and used 80% of them for training and 20% for validation and testing.

To improve the model's ability to generalize across datasets and to make the model robust to different sensors, environmental conditions and calibration methods, we use dynamic perturbation and normalization to preprocess the images of the dataset.

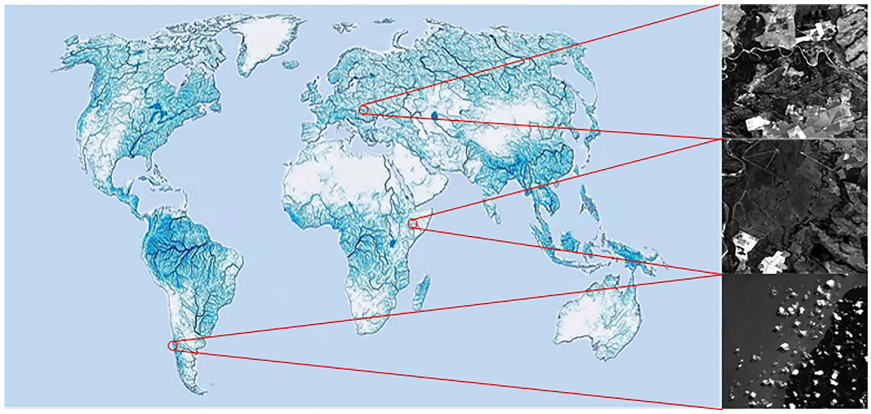

Fig. 6. We balance the proportion of different categories by removing the tiles in our dataset where more than 99% of the pixels have the same category label and ensuring that at least 1% of the pixels are water. The blue area shows the position of each tile in our dataset.

First, in dynamic perturbation, we randomly crop a patch of 512*512 size among all samples of each batch, and then multiply the bands of this patch with a smoothing unit matrix to obtain $P_{512*512*6} * G_{1*3}(I_6)$ to randomly leak information in the continuous bands to achieve the simulation the spectral response of sensors in remote sensing satellites in different observation missions, where P is the input and G is a Gaussian filter . In addition, we distorted these patches using additive Gaussian noise of random magnitude, and finally used min-max scaling for the input, i.e., $P_{\text{Final}} = (P - \min(P))/\max(\max(P) - \min(P), 1)$, where the 1 can stabilize the normalization when the input scene is full water or land.

B. Evaluation Metrics

We evaluated the performance of the proposed MRSE-Net with the optimized dataset obtained above, for which we selected three commonly used metrics for semantic segmentation: precision (user accuracy), recall (producer), and F1-score. Precision is the ratio of pixels correctly classified as water bodies to all pixels classified as water bodies (both correct and incorrect), and recall denotes the ratio of detected water body pixels to all water body pixels in the ground truth label, and F1-score is the summed average of precision and recall, which is used to measure the overall performance, and its value ranges from 0 to 1, with 1 representing the best output of the model and 0 representing the worst output of the model. The following are the formulas for precision, recall, and F1-score:

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive + false positive}} \tag{1}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive + false negative}} \tag{2}$$

$$\text{F1-score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

where true positive indicates a correct positive case, i.e., water pixels are correctly classified as water pixels, false positive indicates a wrong positive case, i.e., nonwater pixels are incorrectly classified as water pixels, and false negative indicates a wrong negative case, i.e., water pixels are incorrectly classified as nonwater pixels.

C. Evaluation

We compare the proposed MRSE-Net with NDWI [11], MNDWI [12], MLP [73], U-Net [29], DeepWaterMap-3, i.e., DWM-3 [73], Deepcrack [75], FCN8s [76], DeepWaterMapV2, i.e., DWMV2 [77], and DeepUNet [55] were compared. We used Landsat-8 tiles randomly selected from the global surface water dataset and their corresponding labels as our test set and quantitatively evaluated these methods using accuracy, recall, and F1-score as our evaluation metrics.

Fig. 7 shows the comparison between our proposed MRSE-Net and the state-of-the-art water body extraction methods. The water body images extracted by the traditional water body index-based methods (NDWI and MNDWI) are very noisy, and since our test image is not calibrated at the top of the atmosphere, it is particularly susceptible to interference from dense clouds [Fig. 7 D, E(b) and (c)], as seen in Table II, the simple MNDWI classifier obtains the highest recall, but the classification accuracy is not very high and produces many false positives. the boundary of water bodies segmented by MLP and U-Net is blurred and a part of the fine river information is ignored [Fig. 7(F)(d) and (e)], and a high number of artifacts are produced. DeepWaterMap-3 classifies a part of the rivers obscured by clouds as nonwater bodies [Fig. 7B(f)] Deepcrack was originally used to identify cracks, and because the structure of cracks and rivers is relatively similar, we use it here to extract water bodies, but it performs poorly in identifying scenes with nonriver structures (including large areas of land and water bodies with nonriver structures). The water segmentation results of FCN8s show many artifacts and blurred boundaries, probably due to the influence of thick cloud cover. DeepWaterMapV2 and Deepunet both use a U-shaped structure, and they perform better on images with nonriver structures, but also have more pixel classification errors. These indicate that in scenes with nonriver structures, the

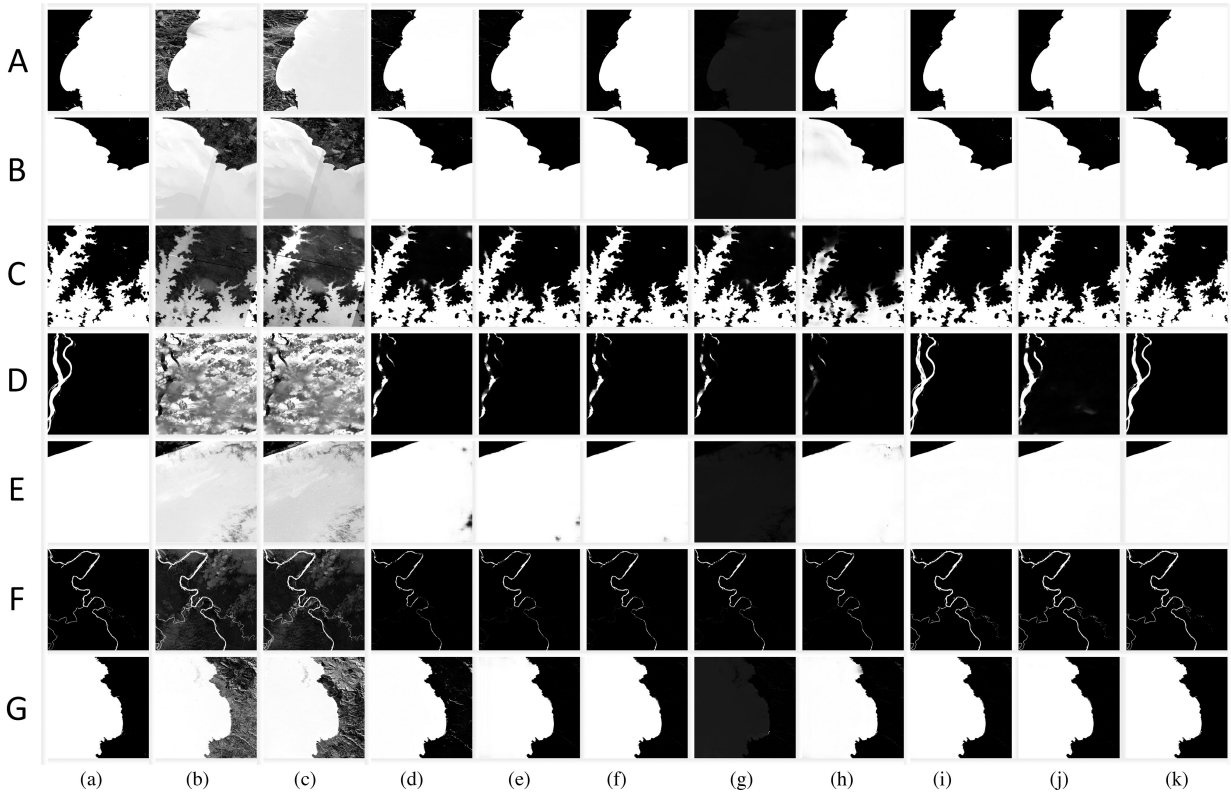

Fig. 7. Landsat-8 test images are compared with different methods. Each column presents: (a) ground truth, (b) NDWI [11], (c) MNDWI [12], (d) MLP [73], (e) U-Net [29], (f) DeepWaterMap-3 [73], (g) Deepcrack [75], (h) FCN8s [76], (i) DeepWaterMapV2 [77], (j) DeepUNet [55], and (k) MRSE-Net.

TABLE II
VALUES OF THE QUANTITATIVE ASSESSMENT INDICATORS

Method	NDWI [11]	MNDWI [12]	MLP [73]	UNet [29]	DWM-3 [73]	Deepcrack [75]	FCN8s [76]	DWMV2 [77]	DeepUNet [55]	MRSE-Net
Precision	0.5325	0.6094	0.8108	0.8527	0.8908	0.4854	0.8624	0.9518	0.9586	**0.9891**
Recall	0.9502	**0.9830**	0.6575	0.7354	0.8575	0.4656	0.7122	0.8706	0.8402	0.8955
F1-score	0.6818	0.7515	0.7253	0.7897	0.8730	0.4753	0.7801	0.9085	0.8955	**0.9400**

The best of them are shown in bold.

lower-level convolutional layers have local features with smaller receptive fields, while the receptive fields get larger as the depth of the network increases, which increases the false positive (nonwater pixels). Our proposed method fuses information from different scales and passes the receptive fields of convolutional layers of different sizes to the network, obtaining the closest results to the ground truth, surpassing existing SOTA methods in accuracy, recall, and F1-score.

Table II gives the values of the quantitative evaluation metrics of the different methods, except for MNDWI [12], which has the highest recall; our method is higher than the other SOTA methods in terms of accuracy and F1-score, DeepWaterMapV2 [77] is second, DeepUNet [55] is located third, and Deepcrack [75] has a very poor performance, because it is designed to detect cracks and performs poorly in scenes that contain nonriver structures, while only capturing crack-like river structures. As can be seen in Table III, our proposed method has fewer convolutional layers, but slightly larger trainable parameters due to the large image size in the dataset used (hundreds of millions of pixels for each scene of Landsat-8), but only more than DeepWaterMap-3 and less than the number of parameters of any other water body extraction method.

TABLE III
COMPARISON OF THE NUMBER OF LAYERS, TRAINABLE PARAMETERS, AND RUNTIME FOR VARIOUS METHODS

Network	Layers	Trainable parameters	Running time
U-Net [29]	23	31	265
DWM-3 [73]	20	15	167
Deepcrack [75]	23	15	176
FCN8s [76]	50	135	579
DWMV2 [77]	20	37	283
DeepUNet [55]	32	124	512
MRSE-Net	20	29	243

Trainable parameters are measured in millions and runtime in milliseconds.

D. Cross-Dataset Evaluation

We also evaluate the cross-dataset generalization ability of our model on a set of Landsat-7 and Sentinel-2 images without

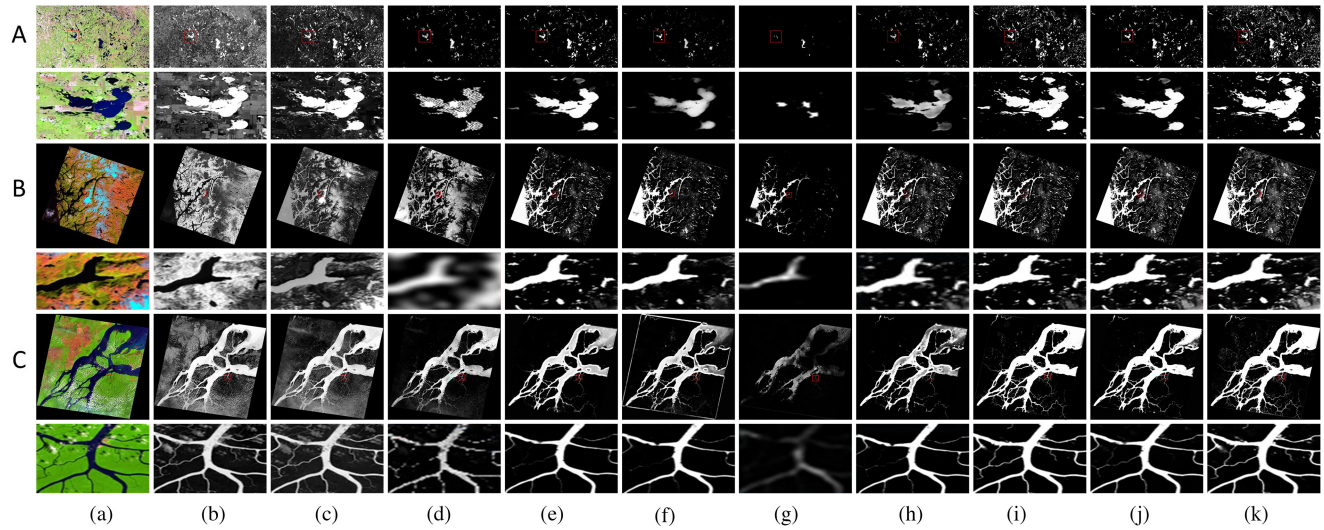

Fig. 8. Different methods were compared for the top-of-atmosphere calibrated Landsat-7 test set. (A,a), (B,a), and (C,a) denote RGB color composite images of the input images. (b) NDWI [11]. (c) MNDWI [12]. (d) MLP [73]. (e) U-Net [29]. (f) DeepWaterMap-3 [73]. (g) Deepcrack [75]. (h) FCN8s [76]. (i) DeepWaterMapV2 [77]. (j) DeepUNet [55]. (k) MRSE-Net.

ground truth labels and exclude any possible message leakage in the validation set. We use the original Landsat-8 images for training, while the test set images used in the cross-dataset test are top-of-atmosphere calibrated, and we use the model trained on Landsat-8 images without fine-tuning directly for water extraction on Landsat-7 and Sentinel-2 images equipped with different sensors since the test set. Since there is no corresponding ground truth in the test set, we introduce RGB color composite images of the input images to subjectively compare the extraction results of the different methods.

The water body index-based methods (NDWI and MNDWI) appear quite noisy with black spots and produce many false positives. MLP produces fuzzier water body boundaries, and MLP is lower than U-Net in both metrics, but both lose a lot of information about small-scale rivers [Fig. 8C(d) and (e)]. Deepcrack was originally designed to identify slender fissures, thus making it difficult to identify water bodies with nonriver structures and resulting in poorer results. The FCN8s were able to identify most of the water bodies, but the prediction of the water body boundaries was still unsatisfactory, with significant blurring as well as artifacts. DeepWaterMap-3 and its improved version still perform well in most water body scenes but tend to classify mountain shadows as water bodies when they appear in the image [Fig. 8 B(f) and (h)]. As shown in Table II, the accuracy of DeepUNet is high, but it is missing a lot of information about small-scale water bodies [Fig. 8 A(i)]. Our proposed method outperforms other methods in terms of objective evaluation metrics, and the resulting water body boundaries are more accurate and continuous for segmentation of large- and small-scale water bodies (narrow rivers and lakes).

As seen in Fig. 9, the MNDWI method based on the water body index has a large range of shadows, and DeepWaterMapV2 and DeepUNet are difficult to segment many fine river branches despite the elimination of shadows. Our proposed method fuses the information of different sensory fields through the multiscale residual module and handles the river information of different scales well. Preliminary experimental results on Sentinel-2 show that our proposed method can be used to predict not only Landsat satellite images but also can be well applied to Sentinel satellite images, showing the good cross-sensor generalization ability of our model.

E. Ablation Study

We propose the MRSE-Net containing feature transformation module, multiscale residual module, SE-attention, and optionally Multi-Res path or inception path. Our network is based on the U-Net structure to improve it, so we compare the five variants of the MRSE-Net architecture with the U-Net to verify the effectiveness of each of our proposed modules. Table IV and Fig. 10 show the module configuration and model framework for each of our models, respectively.

1) *U-Net**: We added a feature transformation module to the bottleneck layer part of the original U-Net to enable the network to learn more abstraction-level feature information. (No Multi-Res path, inception path, multiscale residual module, and SE-attention).
2) *U-Net**/U-Net****: We add the Multi-Res path or inception path to the skip connection based on U-Net*, thus alleviating the semantic gap between the high-level encoder–decoder features at both ends of the skip connection and helping to upsample the recovered images by using feature information with closer semantic levels. (No Multi-scale residual module and SE-attention).
3) *MRU-Net*: We use a multiscale residual module instead of the convolutional layer in the encoder–decoder based on U-Net**, which helps the network to extract deeper high-level feature information and can analyze targets at different scales. The Multi-Res path is used at the skip connection in MRU-Net, because in our experiments to

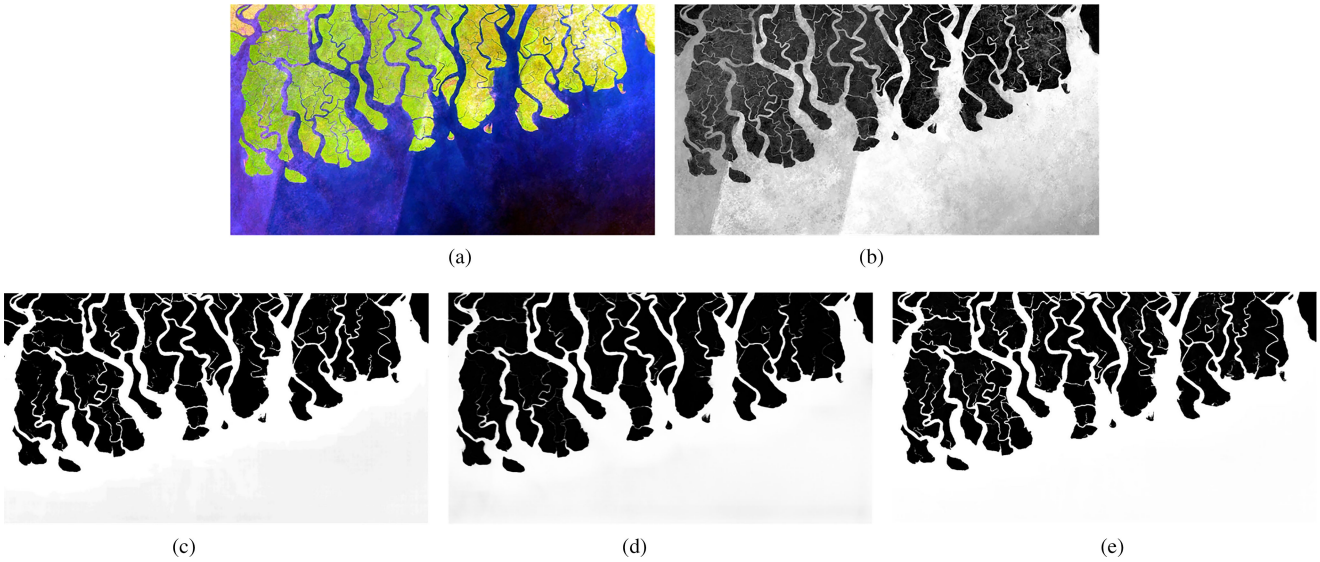

Fig. 9. Prediction of Sentinel-2 input images using different water body extraction methods. (a) RGB color composite image. (b) MNDWI [12]. (c) DeepWaterMapV2 [77]. (d) DeepUNet [55]. (e) MRSE-Net.

TABLE IV
FIVE DIFFERENT MODELS FOR ABLATION STUDIES ARE PROPOSED

Network	feature transformation module	Multi-Res path	Inception path	Multi-scale residual module	SE-attention
U-Net [29]	-	-	-	-	-
U-Net*	✓	-	-	-	-
U-Net**	✓	✓	-	-	-
U-Net***	✓	-	✓	-	-
MRU-Net	✓	✓	-	✓	-
MRSE-Net	✓	✓	-	✓	✓

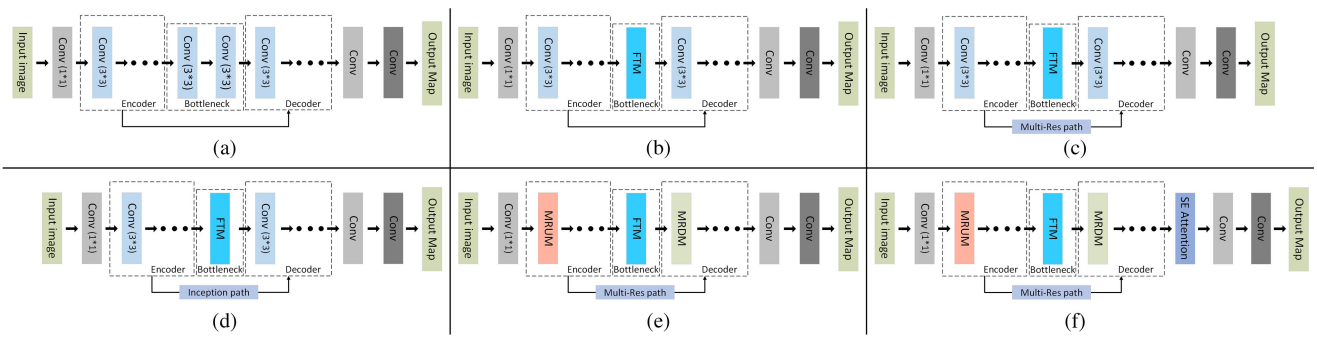

Fig. 10. Model framework in ablation experiments. (a) U-Net. (b) U-Net*. (c) U-Net**. (d) U-Net***. (e) MRU-Net. (f) MRSE-Net.

get the best water extraction results, the Multi-Res path with slightly higher performance than the Inception path is chosen here as our MRU-Net, and if there is a higher requirement for the number of model parameters, the inception path can be used. (No SE-attention).

4) *MRSE-Net*: We introduce the SE attention module in the last layer of upsampling on the network to make the output channels more directional, and explicitly model the dependencies between channels through a "feature recalibration" strategy to increase the weights of the feature channels that contribute to the network optimization.

The Landsat-8 images we used in the test set were not top-of-atmosphere calibrated, and the original U-Net misclassified some clouds as water bodies when performing water body extraction, and many water bodies were not extracted [Fig. 11A, B (b)], so the accuracy was not very high. After adding the feature transformation module, the network learns a lot of high-level semantic information and most of the water bodies are recognized, but more artifacts are generated [Fig. 11 A(c)]. The network that added the Multi-Res path or Inception path alleviated the problem of appearing artifacts [Fig. 11 A(d) and (e)], and as seen in Table V, the former was slightly better than the latter.

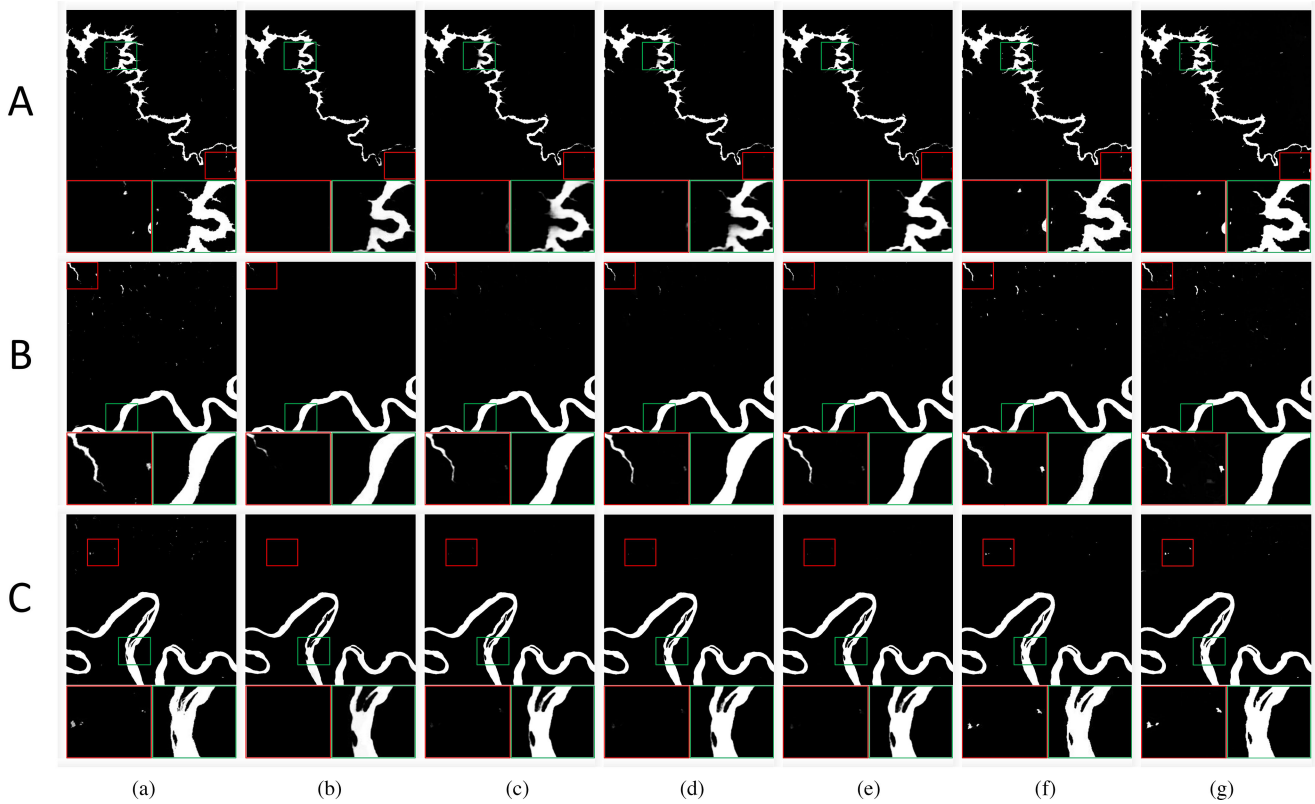

Fig. 11. Effects of adding each module are compared on Landsat-8 images. Each column presents: (a) Ground truth, (b) U-Net [29], (c) U-Net*, (d) U-Net**, (e) U-Net***, (f) MRU-Net, and (g) MRSE- Net. Every two rows represent an example, and the red and green boxes are enlargements of the corresponding details in the image, respectively.

TABLE V
VALUES OF THE QUANTITATIVE ASSESSMENT INDICATORS

Method	U-Net [29]	U-Net*	U-Net**	U-Net***	MRU-Net	MRSE-Net
Precision	0.8527	0.8754	0.8933	0.8912	0.9623	0.9891
Recall	0.7354	0.7567	0.7785	0.7723	0.8668	0.8955
F1-score	0.7897	0.8117	0.8320	0.8275	0.9121	0.9400

The best of them are shown in bold.

Although U-Net and its improved version obtained good results, there are still many problems (blurred water body boundaries, missing information of small details). To alleviate this problem, we obtained continuous and accurate water body boundaries by combining feature information from different scales through the multiscale residual module [Fig. 11 A, C(f)] and successfully predicted water bodies at different scales, and the accuracy, recall, and F1-score metrics of the method were improved. After adding SE-attention, the image details are preserved and the accuracy is improved, and the final prediction results produced are closest to the ground truth.

We also compared the effect of adding each module on Landsat-8 images with a relatively close proportion of water bodies and nonwater bodies. The original U-Net misclassified some of the clouds as water bodies when performing segmentation, which resulted in a lot of shadows in the final generated image [Fig. 12A(b)]. After adding the feature transformation module, the appearance of shadows was alleviated [Fig. 12 A(c)]. After adding the Multi-Res path or inception path, the image no longer has large shadows, but the detailed information is still blurred as seen in the magnified image, especially the boundary part of the water body. These problems were alleviated after we added the multiscale residual module [Fig. 12 A, B, C (f)], which, together with the attention mechanism, further improved the prediction results and obtained the water body extraction results closest to the ground truth.

Table VI gives the details of the number of layers and trainable parameters between the two variants of MRSE-Net and U-Net and their three variants, and it can be seen that all models and their variants have the same number of convolutional layers. However, since MRU-Net uses a lightweight multiscale residual module that requires less memory, its trainable parameters are reduced and the final number of proposed model parameters is slightly less than that of U-Net.

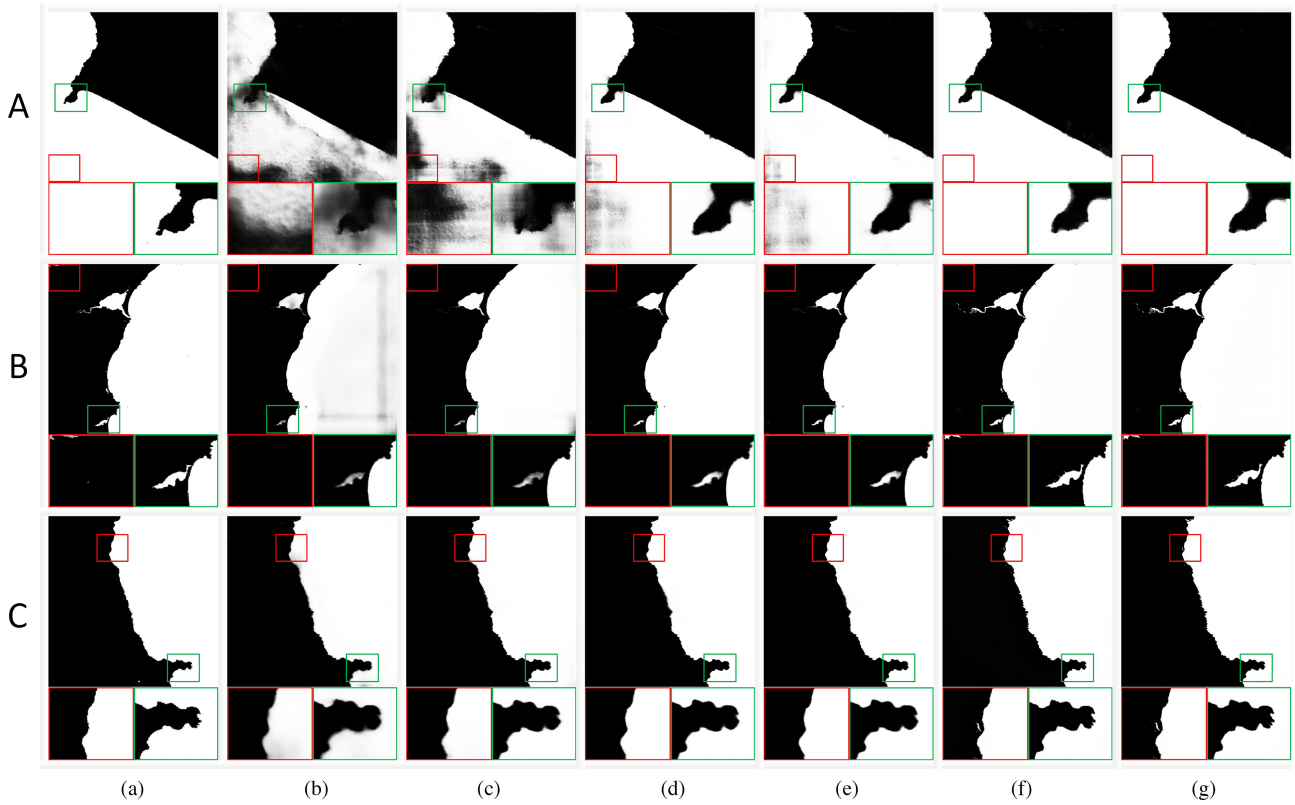

Fig. 12. Effects of adding each module are compared on Landsat-8 images. Each column presents: (a) Ground truth, (b) U-Net [29], (c) U-Net*, (d) U-Net**, (e) U-Net***, (f) MRU-Net, and (g) MRSE-Net. Every two rows represent an example, and the red and green boxes are enlargements of the corresponding details in the image, respectively.

TABLE VI
COMPARISON OF MRSE-NET AND U-NET TRAINABLE PARAMETERS AND RUNTIME

Network	Layers	Trainable parameters	Running time
U-Net [29]	20	31	265
U-Net*	20	33	296
U-Net**	20	37	342
U-Net***	20	34	309
MRU-Net	20	28	224
MRSE-Net	20	29	243

Trainable parameters are measured in millions, and runtime is measured in milliseconds.

V. CONCLUSION

In this article, propose an end-to-end multiscale residual and SE attention-based water segmentation method MRSE-Net for extracting water bodies from satellite images. MRSE-Net is implemented using an encoder and decoder to preserve contextual information and localization at different scales, respectively, and passing encoder feature mappings through an improved skip connection with semantic levels closer to the feature information to help the decoder to recover the image. The proposed multiscale residual module reduces the computational effort in the network so that the total number of trainable parameters of MRSE-Net is lower than that of U-Net. The SE-attention module is used to achieve "adaptive recalibration" of the feature channels, which enhances the water body prediction results and makes the segmented water body boundaries more continuous. Experiments were conducted on Landsat-8 images and compared with existing mainstream methods, and the experimental results demonstrate the superiority of our proposed method over existing methods using our validation set, reflecting the superiority of the MRSE-Net method over existing methods. In addition, we also evaluate our method on Sentinel-2 images, and the experimental results preliminarily show that the cross-sensor generalization capability of our model extends beyond the Landsat sensor family.

REFERENCES

[1] C. J. Vorosmarty, P. Green, J. Salisbury, and R. B. Lammers, "Global water resources: Vulnerability from climate change and population growth," *Science*, vol. 289, no. 5477, pp. 284–288, 2000.
[2] M. A. Holgerson and P. A. Raymond, "Large contribution to inland water CO2 and CH4 emissions from very small ponds," *Nature Geosci.*, vol. 9, no. 3, pp. 222–226, 2016.
[3] C. Prigent, F. Papa, F. Aires, W. B. Rossow, and E. Matthews, "Global inundation dynamics inferred from multiple satellite observations," 1993–2000, *J. Geophys. Res.: Atmospheres*, vol. 112, no. D12, pp. 1–13, 2007.
[4] A. I. Van Dijk *et al.*, "The millennium drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society," *Water Resour. Res.*, vol. 49, no. 2, pp. 1040–1057, 2013.
[5] C. Prigent, F. Papa, F. Aires, C. Jimenez, W. Rossow, and E. Matthews, "Changes in land surface water dynamics since the 1990 s and relation to population pressure," *Geophys. Res. Lett.*, vol. 39, no. 8, 2012, Art. no. L08403.

[6] L. Li, P. Uyttenhove, and V. Van Eetvelde, "Planning green infrastructure to mitigate urban surface water flooding risk—A methodology to identify priority areas applied in the city of Ghent," *Landscape Urban Plan.*, vol. 194, 2020, Art. no. 103703.

[7] D. E. Alsdorf, E. Rodríguez, and D. P. Lettenmaier, "Measuring surface water from space," *Rev. Geophys.*, vol. 45, no. 2, 2007, Art. no. RG2002.

[8] A. M. Melesse, Q. Weng, P. S. Thenkabail, and G. B. Senay, "Remote sensing sensors and applications in environmental resources mapping and modelling," *Sensors*, vol. 7, no. 12, pp. 3209–3241, 2007.

[9] J. K. Lee, T. D. Acharya, and D. H. Lee, "Exploring land cover classification accuracy of Landsat 8 image using spectral index layer stacking in hilly region of South Korea," *Sensors Materials*, vol. 30, no. 12, pp. 2927–2941, 2018.

[10] H. Xie *et al.*, "New hyperspectral difference water index for the extraction of urban water bodies by the use of airborne hyperspectral images," *J. Appl. Remote Sens.*, vol. 8, no. 1, 2014, Art. no. 085098.

[11] S. K. McFeeters, "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.

[12] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *Int. J. remote Sens.*, vol. 27, no. 14, pp. 3025–3033, 2006.

[13] L. Ji, L. Zhang, and B. Wylie, "Analysis of dynamic thresholds for the normalized difference water index," *Photogrammetric Eng. Remote Sens.*, vol. 75, no. 11, pp. 1307–1317, 2009.

[14] G. L. Feyisa, H. Meilby, R. Fensholt, and S. R. Proud, "Automated water extraction index: A new technique for surface water mapping using landsat imagery," *Remote Sens. Environ.*, vol. 140, pp. 23–35, 2014.

[15] S. Lu, B. Wu, N. Yan, and H. Wang, "Water body mapping method with HJ-1A/B satellite imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 3, pp. 428–434, 2011.

[16] W. Li *et al.*, "Assessment for surface water quality in lake Taihu Tiaoxi river basin China based on support vector machine," *Stochastic Environ. Res. Risk Assessment*, vol. 27, no. 8, pp. 1861–1870, 2013.

[17] X. Sun, L. Li, B. Zhang, D. Chen, and L. Gao, "Soft urban water cover extraction using mixed training samples and support vector machines," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3331–3344, 2015.

[18] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3243–3254, Dec. 2010.

[19] B. Huang, H. Li, and X. Huang, "A level set method for oil slick segmentation in SAR images," *Int. J. Remote Sens.*, vol. 26, no. 6, pp. 1145–1156, 2005.

[20] M. Silveira and S. Heleno, "Separation between water and land in SAR images using region-based level sets," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 471–475, Jul. 2009.

[21] H. Deng and D. A. Clausi, "Unsupervised segmentation of synthetic aperture radar sea ice imagery using a novel Markov random field model," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 528–538, Mar. 2005.

[22] L. Gan, Y. Wu, F. Wang, P. Zhang, and Q. Zhang, "Unsupervised SAR image segmentation based on triplet Markov fields with graph cuts," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 4, pp. 853–857, Apr. 2014.

[23] F. Wang, Y. Wu, Q. Zhang, W. Zhao, M. Li, and G. Liao, "Unsupervised SAR image segmentation using higher order neighborhood-based triplet markov fields model," *IEEE Trans. Geosci. remote Sens.*, vol. 52, no. 8, pp. 5193–5205, Aug. 2014.

[24] X. Huang, C. Xie, X. Fang, and L. Zhang, "Combining pixel-and object-based machine learning for identification of water-body types from urban high-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2097–2110, May 2015.

[25] D. Ying, Z. Hong, W. Chao, and L. Meng, "An object-oriented water extraction method based on texture and polarimetric decomposition feature," *Remote Sens. Technol. Appl.*, vol. 31, no. 4, pp. 714–723, 2016.

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

[27] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 498.

[28] 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. 2018.

[29] 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.

[30] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields," *Remote Sens.*, vol. 9, no. 7, 2017, Art. no. 680.

[31] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite imagery feature detection using deep convolutional neural network: A Kaggle competition," 2017, *arXiv:1706.06169*.

[32] F. Chen, "Comparing methods for segmenting supra-glacial lakes and surface features in the mount everest region of the Himalayas using Chinese Gaofen-3 SAR images," *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2429.

[33] J. Wang *et al.*, "Flood inundation region extraction method based on sentinel-1 SAR data," *J. Catastrophol.*, vol. 36, pp. 214–220, 2021.

[34] M. Dai, X. Leng, B. Xiong, and K. Ji, "An efficient water segmentation method for SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1129–1132.

[35] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiseNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[36] 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. 2881–2890.

[37] J. Li, C. Wang, L. Xu, F. Wu, H. Zhang, and B. Zhang, "Multitemporal water extraction of dongting lake and Poyang lake based on an automatic water extraction and dynamic monitoring framework," *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 865.

[38] M. Dirscherl, A. J. Dietz, C. Kneisel, and C. Kuenzer, "A novel method for automated supraglacial lake mapping in Antarctica using Sentinel-1 SAR imagery and deep learning," *Remote Sens.*, vol. 13, no. 2, 2021, Art. no. 197.

[39] 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.

[40] Y. Ren, X. Li, X. Yang, and H. Xu, "Development of a dual-attention U-Net model for sea ice and open water classification on SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, Feb. 2021, Art. no. 4010205.

[41] Google Earth Engine Team, "Google Earth engine: A planetary-scale geospatial analysis platform," 2015. [Online]. Available: https://earthengine.google.com

[42] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418–422, 2016.

[43] W. Feng, H. Sui, W. Huang, C. Xu, and K. An, "Water body extraction from very high-resolution remote sensing imagery using deep U-Net and a superpixel-based conditional random field model," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 618–622, Apr. 2019.

[44] T. James, C. Schillaci, and A. Lipani, "Convolutional neural networks for water segmentation using sentinel-2 red, green, blue (RGB) composites and derived spectral indices," *Int. J. Remote Sens.*, vol. 42, no. 14, pp. 5338–5365, 2021.

[45] R. G. Tambe, S. N. Talbar, and S. S. Chavan, "Deep multi-feature learning architecture for water body segmentation from satellite images," *J. Vis. Commun. Image Representation*, vol. 77, 2021, Art. no. 103141.

[46] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[47] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

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

[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[50] Q. Guo, R. Pu, J. Li, and J. Cheng, "A weighted normalized difference water index for water extraction using Landsat imagery," *Int. J. Remote Sens.*, vol. 38, no. 19, pp. 5430–5445, 2017.

[51] A. Essa, P. Sidike, and V. Asari, "Volumetric directional pattern for spatial feature extraction in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1056–1060, Jul. 2017.

[52] F. Yao, C. Wang, D. Dong, J. Luo, Z. Shen, and K. Yang, "High-resolution mapping of urban surface water using ZY-3 multi-spectral imagery," *Remote Sens.*, vol. 7, no. 9, pp. 12336–12355, 2015.

[53] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
[54] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.
[55] R. Li *et al.*, "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.
[56] M. M. Pai, V. Mehrotra, S. Aiyar, U. Verma, and R. M. Pai, "Automatic segmentation of river and land in SAR images: A deep learning approach," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng.*, 2019, pp. 15–20.
[57] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.
[58] X. Yang, X. Li, Y. Ye, R. Y. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.
[59] J. H. Kim *et al.*, "Objects segmentation from high-resolution aerial images using U-Net with pyramid pooling layers," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 115–119, Jan. 2019.
[60] A. Rakhlin, A. Davydow, and S. Nikolenko, "Land cover classification from satellite imagery with U-Net and Lovász-softmax loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 262–266.
[61] V. Iglovikov, S. Seferbekov, A. Buslaev, and A. Shvets, "Ternausnetv2: Fully convolutional network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 233–237.
[62] V. Khryashchev, R. Larionov, A. Ostrovskaya, and A. Semenov, "Modification of U-Net neural network in the task of multichannel satellite images segmentation," in *Proc. IEEE East-West Des. Test Symp.*, 2019, pp. 1–4.
[63] J. T. Springenberg *et al.*, "Striving for simplicity: The all convolutional Net," 2014, *arXiv:1412.6806*.
[64] S. W. Zamir *et al.*, "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 492–511.
[65] K. Mei, A. Jiang, J. Li, and M. Wang, "Progressive feature fusion network for realistic image dehazing," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 203–215.
[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.
[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[68] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[69] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Berlin, Germany: Springer, 2016, pp. 179–187.
[70] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-V4, inception-Resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[71] F. Isikdogan, A. Bovik, and P. Passalacqua, "Learning a river network extractor using an adaptive loss function," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 6, pp. 813–817, Jun. 2018.
[72] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[73] 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.
[74] S. Zhou, P. Kan, J. Silbernagel, and J. Jin, "Application of image segmentation in surface water extraction of freshwater lakes using radar data," *ISPRS Int. J. Geo- Inf.*, vol. 9, no. 7, 2020, Art. no. 424.
[75] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "Deepcrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, 2019.
[76] L. Li, Z. Yan, Q. Shen, G. Cheng, L. Gao, and B. Zhang, "Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks," *Remote Sens.*, vol. 11, no. 10, 2019, Art. no. 1162.
[77] L. F. Isikdogan, A. Bovik, and P. Passalacqua, "Seeing through the clouds with deepwatermap," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1662–1666, Oct. 2020.

Xinyu Zhang received the bachelor's degree in discipline of communication engineering from the School of Science and Information Science, Qingdao Agricultural University, Qingdao, China, in 2020. He is currently working toward the master's degree in electronic information with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, China.

His research interests include computer graphics, computer vision, and image processing.

Jinjiang Li received the B.S. and M.S. degrees from the Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shandong University, Jinan, China, in 2010, all in computer science.

From 2004 to 2006, he was an Assistant Research Fellow with the Institute of Computer Science and Technology, Peking University, Beijing, China. From 2012 to 2014, he was a PostDoctoral Fellow with Tsinghua University, Beijing. He is currently a Professor with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China. His research interests include image processing, computer graphics, computer vision, and machine learning.

Zhen Hua received the B.S. and M.S. degrees in electrical automation from the Taiyuan University of Technology, Taiyuan, China, in 1989 and 1992, respectively, and the Ph.D. degree in electronic information engineering from the China University of Mining and Technology, Beijing, China, in 2008.

She is currently a Professor with the Shandong Technology and Business University, Yantai, China. Her research interests include computer-aided geometric design, information visualization, virtual reality, and image processing.