AEDNet: An Attention-Based Encoder–Decoder Network for Urban Water Extraction From High Spatial Resolution Remote Sensing Images

Yanjiao Song, Xiaoping Rui, and Junjie Li

***Abstract*—Accurate water extraction from urban remote sensing images holds great significance in assisting the formulation of river and lake management policies and ensuring the sustainable development of urban water resources. However, urban high-resolution remote sensing images encompass complex spatial and semantic information, which leads to disparities between the extracted water body features based on local and global information, consequently affecting the accuracy of urban water extraction. To tackle this issue, an attention-based encoder–decoder network was proposed. In this network, the backbone employing atrous convolution (AC) facilitated the acquisition of low-level and high-level features of urban remote sensing images at various scales. Integrated with the attention mechanism, the encoder–decoder structure extracted global features in both the spatial and channel domains. Subsequently, these two types of features were merged to yield the urban water segmentation. Moreover, considering both intersection over union and class weights, a joint loss function (JLF) was introduced to further enhance the accuracy of urban water extraction. Experimental results demonstrated the strong performance of the proposed method on both GID and LoveDA datasets.**

***Index Terms*—Atrous convolution (AC), attention mechanism, joint loss function (JLF), remote sensing, urban water extraction.**

I. Introduction

WATER serves as the fundamental source of life, playing a crucial role in supporting all known life forms on the earth. Its impact extends to climate, biodiversity, and the well-being of humans [1]. Notably, urban water resources assume a vital role in preserving the ecological equilibrium of cities and fostering the robust development of urban economies [2]. However, owing to the influence of climate change, human activities and other factors, the distribution of urban water bodies is highly heterogeneous [3]. Therefore, obtaining an accurate depiction of urban water body distribution holds immense significance in assisting governmental efforts to formulate effective river and lake management policies and ensure the sustainable development of urban water resources.

Due to its significant advantages, such as wide coverage, long time series, high efficiency, and cost effectiveness in terms of manpower and resources, remote sensing has gained increasing importance in water extraction efforts [4]. The extraction of water bodies primarily relied on the spectral differences between water and other objects in various bands of remote sensing images [5]. Among these traditional methods, the widely used approach is the water index method. This kind of method calculated a specific index that reflects the water characteristics by considering multiple bands of remote sensing images. Subsequently, threshold segmentation was performed on this index to classify the image into water and nonwater areas. Over the years, scholars have proposed different water body indices, such as the normalized difference water index (NDWI) [6], modified normalized difference water index (MNDWI) [7], automated water extraction index (AWEI) [8], and weighted normalized difference water index (WNDWI) [9]. However, due to the complex semantic information present in urban high spatial resolution remote sensing images, roads, buildings, and building shadows are prone to misclassification as water bodies. Furthermore, a major challenge with the water index method lies in determining the appropriate threshold, and improper selection of the threshold can significantly impact the accuracy of water extraction results [10], [11]. Automatic binarization algorithms like OTSU [12] is commonly used in image thresholding, but it is inappropriate to apply a local optimal threshold to remote sensing images of different regions at different times [13].

Classical machine learning methods, such as support vector machine [14], [15], decision tree [16], [17], and random forest [18], [19], have been applied to water body extraction tasks. These algorithms effectively address the problem of water misclassification by taking manually labeled training data and teaching the computer to detect similar features in the data [20]. However, feature construction in these methods can be time consuming, leading to relatively lower efficiency in water extraction using machine learning [21].

Deep learning methods offer advantages by automatically extracting features from raw images through multiple convolutional layers, eliminating the need for intricate feature engineering and significantly improving efficiency [22], [23], [24]. Among them, semantic segmentation models based on convolutional neural networks (CNNs) can extract semantic features

Manuscript received 8 August 2023; revised 2 November 2023; accepted 27 November 2023. Date of publication 1 December 2023; date of current version 14 December 2023. This work was supported by Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China under Grant KLSMNR-G202212 and in part by the National Natural Science Foundation of China under Grant 42376180. *(Corresponding author: Xiaoping Rui.)*

Yanjiao Song and Junjie Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: songyanjiao2000@163.com; junjieli@whu.edu.cn).

Xiaoping Rui is with the School of Earth Science and Engineering, Hohai University, Nanjing 211100, China (e-mail: ruixp@hhu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3338484

from images and associate them with specific class labels. In recent years, substantial progress has been made in semantic segmentation models such as FCN [25], U-Net [26], PSPNet [27], and DeepLab [28], [29], [30], [31], achieving high accuracy in water extraction tasks [32], [33]. However, the aforementioned semantic segmentation models can only extract water features with a fixed receptive field size or extract multiscale water body features with multiple receptive field sizes [34], [35], [36], [37]. Nevertheless, the intricate semantics present in urban high-resolution remote sensing images often cause water features extracted based on local information to deviate from global information, thereby affecting the accuracy of urban water semantic segmentation [38].

In recent years, the attention mechanism has emerged as a significant research focus in deep learning area, and many related neural networks have been proposed, such as nonlocal neural networks [39], SENet [40], CBAM [41], DANet [42], and so on. It enables the calculation of weight distributions based on global information and applies these weights to emphasize specific features, thereby facilitating the extraction of global features. Many scholars have applied attention mechanism into water extraction tasks. Wang et al. [43] adopted the spatial and channel squeeze and excitation attention mechanism in flood extraction from SAR image, and achieved better accuracy of the model prediction. Zhang et al. [44] involved SE-attention to an end-to-end CNN structure, in which the SE-attention module can enhance the prediction results, mitigate the blurring effect, and make the segmented water boundaries more continuous. Yu et al. [45] presented a hierarchical attentive high-resolution network to export semantic-discriminative, target-oriented feature representations for precise water body segmentation. Dai et al. [46] proposed a multiscale location attention network, which focused on location-spatial information and channel information of water and improved the boundary extraction of water bodies.

In this article, to address the challenges in urban water extraction, an attention-based encoder—decoder network (AEDNet) was proposed. AEDNet tackles these challenges by incorporating several key components. First, the backbone of AEDNet utilizes atrous convolution (AC) to capture both low-level and high-level features of urban remote sensing images across multiple scales. Subsequently, within the encoder–decoder architecture, the dual attention module facilitates the extraction of global features in both the spatial and channel domains. Ultimately, features from both domains are fused, yielding accurate water extraction results. In addition, a joint loss function (JLF) is proposed to address the challenge of imbalanced positive and negative samples arising from the uneven distribution of urban water bodies. This JLF combines the intersection over union (IoU)-based loss function, specifically Lovász hinge [47], with a weighted cross-entropy loss function. The main contributions of this article can be summarized as follows.

1) A backbone incorporating AC was designed, allowing for the extraction of both low-level and high-level water features at various scales.
2) A dual attention module within the encoder–decoder structure was proposed. This module applies a spatial attention mechanism to the low-level features and a channel attention mechanism to the high-level features. By combining these two types of attention, AEDNet can effectively extract global water features, taking into account both spatial and semantic information.
3) A JLF was designed to address the issue of imbalanced urban water samples and the importance of IoU in accuracy evaluation. This loss function combines the Lovász hinge loss based on IoU with a weighted cross-entropy loss.

II. Proposed Method

The overall structure of the proposed AEDNet is shown in Fig. 1. The baseline is an encoder–decoder network, while AC, channel attention module (CAM), and position attention module (PAM) are incorporated to enhance feature extraction. ResNet [48] with AC is employed to extract both low-level and high-level features from the input images at different scales. CAM is utilized to extract global features in the channel domain, capturing the interdependencies between different channels. PAM focuses on extracting global features in the spatial domain, capturing spatial relationships within the image. In the decoder section, the features extracted by CAM and PAM are concatenated. Further feature extraction is then performed to refine the combined features and generate the final segmentation result. In addition, a JLF for urban water semantic segmentation is proposed to ensure the accuracy and stability during model training.

A. Backbone With Multiscale AC

Water bodies exhibit significant spatiotemporal variability, and remote sensing images often contain large-scale water features. To capture the dependencies of such water features, a larger receptive field is required in the backbone of the model. AC, also known as dilated convolution, addresses this need. AC introduces holes in the filters of standard convolution to increase the receptive field. By adjusting the dilation rate, the receptive field of AC can be dynamically modified, allowing the model to capture feature information at different scales. This mechanism enables the extraction of contextual information over a larger area, enhancing the understanding of complex water features in remote sensing images. Fig. 2 illustrates the concept of AC and its effect on the receptive field. With the same number of parameters, AC expands the effective field of view, enabling the model to incorporate information from a wider range of neighboring pixels.

In 1-D AC, for the input signal $x(i)$, the output $y(i)$ can be calculated using the following equation:

$$y(i) = \sum_{k=1}^{K} x(i + r \cdot k) \cdot \omega(k) \tag{1}$$

where ω refers to a filter of size K, and r is the dilation rate.

ResNet is a widely recognized deep CNN architecture, and its key idea of addressing overfitting through residual blocks has been incorporated into many contemporary CNN models. In this article, ResNet-101 with AC is employed as the backbone of AEDNet. The last three residual blocks of ResNet-101 are

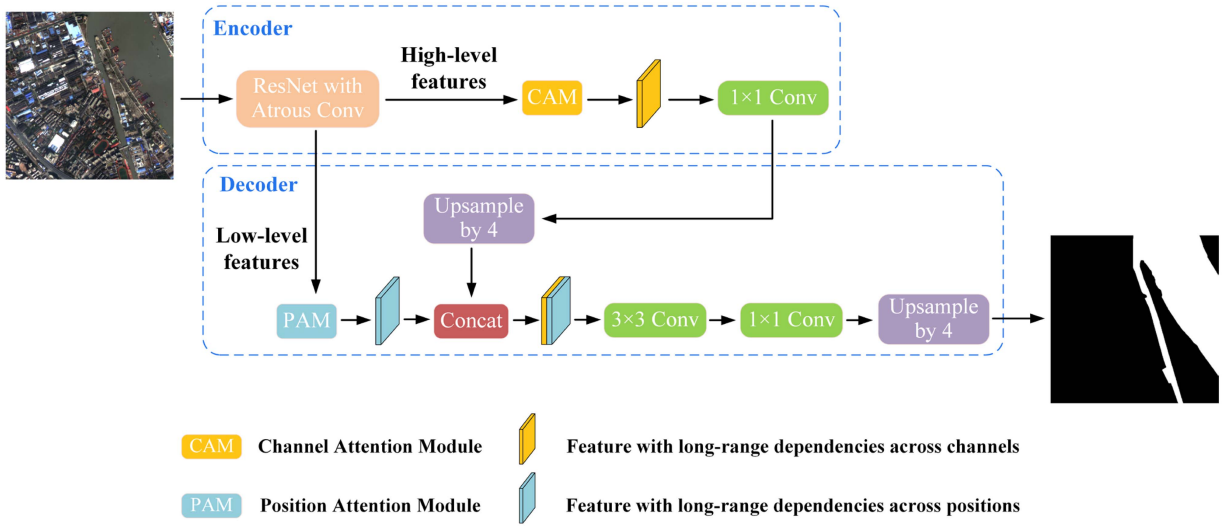

Fig. 1. Overall structure of AEDNet. The baseline is an encoder–decoder network, while AC, CAM, and PAM are incorporated to enhance feature extraction.

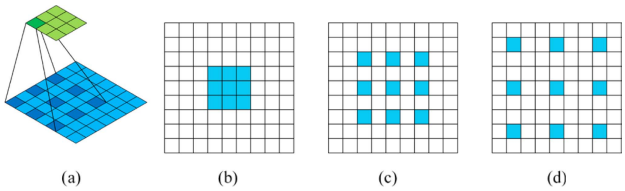

Fig. 2. AC with different dilation rate. (a) Atrous conv. (b) Rate = 1. (c) Rate = 2. (d) Rate = 3.

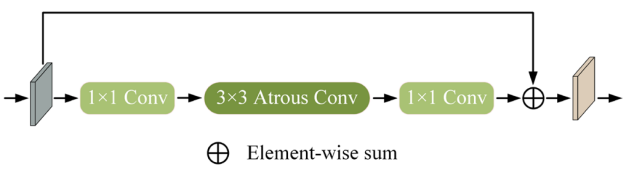

Fig. 3. Residual block with AC used in this article. The dilation rates are set to 2, 4, and 8, respectively.

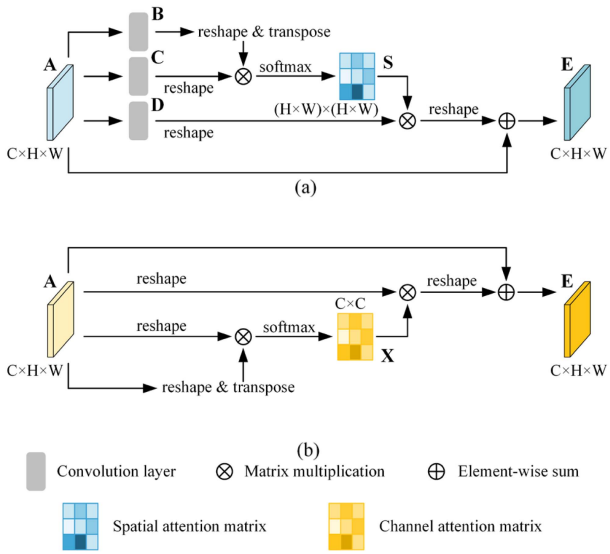

Fig. 4. Specific structure of PAM and CAM. (a) Position attention module. (b) Channel attention module.

modified to enhance the performance, as depicted in Fig. 3. First, a 1×1 convolutional layer is employed to reduce the number of feature map channels from 256 to 64. Subsequently, an atrous convolutional layer with varying dilation rates (2, 4, and 8 in this article) is applied to capture feature dependencies at different scales. Another 1×1 convolutional layer is used to restore the number of feature map channels from 64 to 256. Finally, the features before and after processing are element-wise added pixel by pixel to produce the output of the residual block. For feature extraction, the output of the first three residual blocks of ResNet-101 is utilized as low-level features, while the output of all residual blocks serves as the high-level features. These features are then fed into subsequent neural networks to further process.

B. Attention-Based Encoder–Decoder (AED) Structure

The low-level features generated by the backbone of AEDNet are rich in spatial information, whereas the high-level features contain more channel-related information. Building upon the success of DANet, similar PAM and CAM are incorporated into the encoder–decoder structure. This enables the processing of low-level and high-level features separately, allowing the network to capture global feature dependencies in both the spatial and channel domains. The specific structure of PAM and CAM is illustrated in Fig. 4.

In the PAM, the feature map $A \in \mathbb{R}^{C\times H\times W}$ is first subjected to a 1×1 convolution operation to obtain feature maps $\{B, C\} \in \mathbb{R}^{(C/8)\times H\times W}$, and reshaped to $\mathbb{R}^{(C/8)\times N}$ ($N = H \times W$, number of pixels). Then, the transposed B is multiplied by C, and the product is normalized by a softmax layer to obtain an attention map $S \in \mathbb{R}^{N\times N}$ as follows:

$$s_{ji} = \frac{\exp\left(B_i \cdot C_j\right)}{\sum_{i=1}^{N} \exp\left(B_i \cdot C_j\right)} \tag{2}$$

where s_{ji} represents the influence of pixel i on pixel j. At the same time, a 1 zed by a softmax layer to obtain an attention map features separately, allowimap $D \in \mathbb{R}^{C\times H\times W}$. D is first reshaped to $\mathbb{R}^{C\times N}$ and multiplied by the transpose of S, then the product is reshaped from $\mathbb{R}^{C\times N}$ to $\mathbb{R}^{C\times H\times W}$. Finally, using the idea of residuals, the product is weighted by the parameter α and added element-wise with the original feature map A to obtain PAM result $E \in \mathbb{R}^{C\times H\times W}$.

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{3}$$

where the initial value of the weight α is 0. In the feature map E output by PAM, each element is the weighted sum of features between all positions and the original features, so it can represent the global spatial dependence.

Similar to PAM, the CAM calculates the dependencies between different channels. In order to better preserve the relationship between channels, CAM do not perform convolution operations, but directly reshapes the feature map A to $\mathbb{R}^{C\times N}$, multiplies A by its transpose, and normalizes the product with a softmax layer to obtain an attention map $X \in \mathbb{R}^{C\times C}$ as follows:

$$x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C} \exp(A_i \cdot A_j)} \tag{4}$$

where x_{ji} represents the influence of pixel i on pixel j. Then, the transposed X is multiplied by A and the product is reshaped to $\mathbb{R}^{C\times H\times W}$. Finally, using the idea of residuals, the product is weighted by the parameter β and added element-wise with the original feature map A to obtain CAM result $E \in \mathbb{R}^{C\times H\times W}$ as follows:

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{5}$$

where the initial value of the weight β is 0. In the feature map E output by CAM, each element is the weighted sum of features between all channels and the original features, so it can represent the global channel dependence.

In the encoder part of AEDNet, a pointwise convolution followed by upsampling by a factor of 4 is employed. This sequential operation ensures that the output size of CAM matches that of the low-level features. Moving to the decoder part, the output of PAM is concatenated with the output of the encoder. This fusion of low-level and high-level features enhances the network's ability to capture both spatial and channel information. To complete the fusion process, a 3×3 convolution is applied. Finally, through a 1×1 convolution layer and upsampling by a factor of 4, the segmentation result with the same size as the input image is obtained.

C. JLF for Urban Water Extraction

In the field of image segmentation, compared to the commonly used cross-entropy loss function, IoU can better evaluate the performance of models [47]. The IoU score, also known as the Jaccard Index, can be expressed in the following form:

$$J_c(y^*, \tilde{y}) = \frac{|\{y^* = c\} \cap \{\tilde{y} = c\}|}{|\{y^* = c\} \cup \{\tilde{y} = c\}|} \tag{6}$$

where c is the category, y^* is the true value, and $\tilde{y}$ is the predicted value. In order to use the Jaccard index as a loss function, it can be transformed into

$$\Delta_{J_c}(y^*, \tilde{y}) = 1 - J_c(y^*, \tilde{y}). \tag{7}$$

The misclassified pixel set M_c can be represented as

$$M_c(y^*, \tilde{y}) = \{y^* = c, \tilde{y} \neq c\} \cup \{y^* \neq c, \tilde{y} = c\}. \tag{8}$$

Thus, (7) can be rewritten as

$$\Delta_{J_c} : M_c \in \{0,1\}^p \mapsto \frac{|M_c|}{|\{y^* = c\} \cup M_c|} \tag{9}$$

where p is the number of pixels in the image. Δ_{J_c} is discrete and nondifferentiable, so it cannot be used directly as a loss function. The Lovász hinge smooths Δ_{J_c} through the Lovász extension and transforms the input space from the discrete $\{0,1\}^p$ to the continuous $\mathbb{R}^p$. For a set function $\Delta : \{0,1\}^p \to \mathbb{R}$, its Lovász extension is shown as follows:

$$\overline{\Delta} : m \in \mathbb{R}^p \mapsto \sum_{i=1}^{p} m_i g_i(m) \tag{10}$$

$$\text{with } g_i(m) = \Delta(\{\pi_1, \ldots, \pi_i\}) - \Delta(\{\pi_1, \ldots, \pi_{i-1}\}) \tag{11}$$

where π represents the descending order of the elements in m. Although the input form of the function changes after the transformation, its output value remains unchanged and has convexity. For the Jaccard loss Δ_{J_1} of the positive class in binary classification, its Lovász hinge is shown as follows:

$$l(F) = \overline{\Delta_{J_1}}(m(F)) \tag{12}$$

where F is the model output, $\overline{\Delta_{J_1}}$ is the Lovász extension of Δ_{J_1}, and m is the hinge loss associated with the prediction.

The Lovász hinge has achieved the application of IoU as a loss function in CNNs and has shown good performance in the training and testing of semantic segmentation models. However, it has the problem of unstable training process, manifested by a fluctuating loss curve. In addition, in the task of urban water body semantic segmentation, the proportion of positive and negative samples usually differs greatly, but the Lovász hinge does not take into account this class imbalance problem. In contrast, although weighted cross entropy cannot bring higher IoU scores to the model, it has a stable training process and alleviates the problem of positive and negative sample imbalance through weighting. The equation for weighted cross entropy in binary classification is as follows:

$$l = -\frac{1}{N} \sum_{i=1}^{N} [(1-w) y_i \ln p_i + w(1-y_i) \ln(1-p_i)] \tag{13}$$

where N is the total number of samples, y_i is the label of sample i, p_i is the probability of sample i being predicted as positive, and w is the proportion of positive sample pixels in the total pixels of all data. Before training, w is calculated first. The fewer the

TABLE I
DATASETS AND DATA PREPROCESSING DETAILS

Datasets	Resolution	Band sequence	Base size	Crop size	Number of inputs
GID	1 m	NIR-R-G-B	7200×6800	512×512	training: 780 validation: 195
LoveDA	0.3 m	R-G-B	1024×1024	512×512	training: 1532 validation: 384

number of positive samples, the smaller the w, and the greater the weighting of the positive class in the loss function.

Taking into account the importance of IoU, the stability of the training process and the handling of sample imbalance problems, a JLF that combines Lovász hinge and weighted cross entropy was proposed for urban water extraction. Its equation is as follows:

$$L = \gamma l_1 + (1-\gamma)\, l_2 \tag{14}$$

where $\gamma \in (0, 1)$ is a hyperparameter set before training, l_1 is the Lovász hinge in (12), and l_2 is the weighted cross-entropy loss function in (13).

III. EXPERIMENT

To evaluate the performance of our proposed method, experiments were conducted on two publicly available high-resolution remote sensing image semantic segmentation datasets: GID [49] and LoveDA [50].

A. Datasets

GID is a large-scale land-cover dataset containing 150 Gaofen-2 satellite images acquired from more than 60 cities in China. Two panchromatic (PAN) and multispectral (MS) sensors with effective spatial resolution of 1 m (PAN) / 4 m (MS) are onboard the Gaofen-2 satellite. After processing, each Gaofen-2 image in GID providing a spatial dimension of 7200×6800 pixels and 4 bands: near-infrared (NIR), red (R), green (G), and blue (B).

LoveDA dataset contains 5987 high spatial resolution images with 166 768 annotated objects from three cities in China, and encompasses urban and rural domains. Images in this dataset are obtained from the Google Earth platform with spatial resolution of 0.3 m, providing R, G, and B bands.

To prepare the data for urban water semantic segmentation, 34 images from urban areas were manually selected from the GID dataset, with R, G, and B bands selected. From the LoveDA dataset, images classified as urban were selected. Then, all images were cropped to samples of 512×512 pixels to ensure consistency. Next, certain criteria were applied to filter the samples. If a sample group did not contain any water, that group would be removed from further analysis. Similarly, if a sample group did not contain any urban buildings, such as houses or roads, that group would be randomly deleted with a 90% probability. Finally, the processed data would be divided into training and validation sets, with an 8:2 ratio for training and validation. This division allowed us to obtain the necessary data for training and evaluating the performance of our semantic segmentation model. The details of the dataset division can be found in Table I. During the training and validation process, data augmentation techniques were performed on the samples, containing random flipping (horizontal and vertical) and random rotation (90°, 180°, and 270°) operations. Not increasing the number of samples, these operations helped to enhance the diversity and robustness of the training data, improving the performance of the model.

B. Evaluation Metrics

The evaluation metrics used in this article are divided into accuracy evaluation metrics and efficiency evaluation metrics.

The efficiency evaluation metrics are the number of model parameters and floating-point operations (FLOPs). The accuracy evaluation metrics are precision, recall, and IoU, all of which are pixel-level metrics calculated from the confusion matrix. The calculation equations are as follows, where TP represents the number of true positive pixels, TN represents the number of true negative pixels, FP represents the number of false positive pixels, and FN represents the number of false negative pixels.

Precision refers to the proportion of pixels that are actually positive among those predicted to be positive. The equation is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{15}$$

Recall refers to the proportion of pixels predicted to be positive among those that are actually positive, as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{16}$$

IoU refers to the ratio of the intersection to the union of pixels predicted to be positive and pixels that are actually positive. The equation is

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{17}$$

C. Implementation Details

AEDNet was designed based on the Pytorch 1.13.0 framework, and the network model training was performed on a GPU server: containing 1 CPU (Intel Xeon E5-2640 v4) with a total of 64-GB RAM, 2 GPUs (NVIDIA Tesla V100 16 GB) with a total of 32-GB VRAM. The main parameter settings were:the data batch size was 4, the initial learning rate was 1e-4, the dynamic adjustment strategy of learning rate was StepLR, the optimizer was stochastic gradient descent, the normalization method was synchronized batch normalization [51], and the number of training epochs was 100.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE GID DATASET

Model	Backbone	Params (M)	FLOPs (G)	Precision (%)	Recall (%)	IoU (%)
U-Net	ResNet-101	31.03	873.83	93.86	93.39	88.01
PSPNet	ResNet-101	72.20	189.11	94.20	93.62	88.51
DeepLabv3+	ResNet-101	59.23	263.76	95.38	92.65	88.67
CBAM	ResNet-101	51.86	168.08	90.83	88.99	81.65
DANet	ResNet-101	66.31	187.49	92.98	92.62	86.57
MECNet	-	30.07	743.39	93.85	94.16	88.69
MSResNet	ResNet-101	69.55	137.53	86.21	88.41	77.45
AEDNet	ResNet-101	44.77	334.02	**95.52**	**95.66**	**91.55**

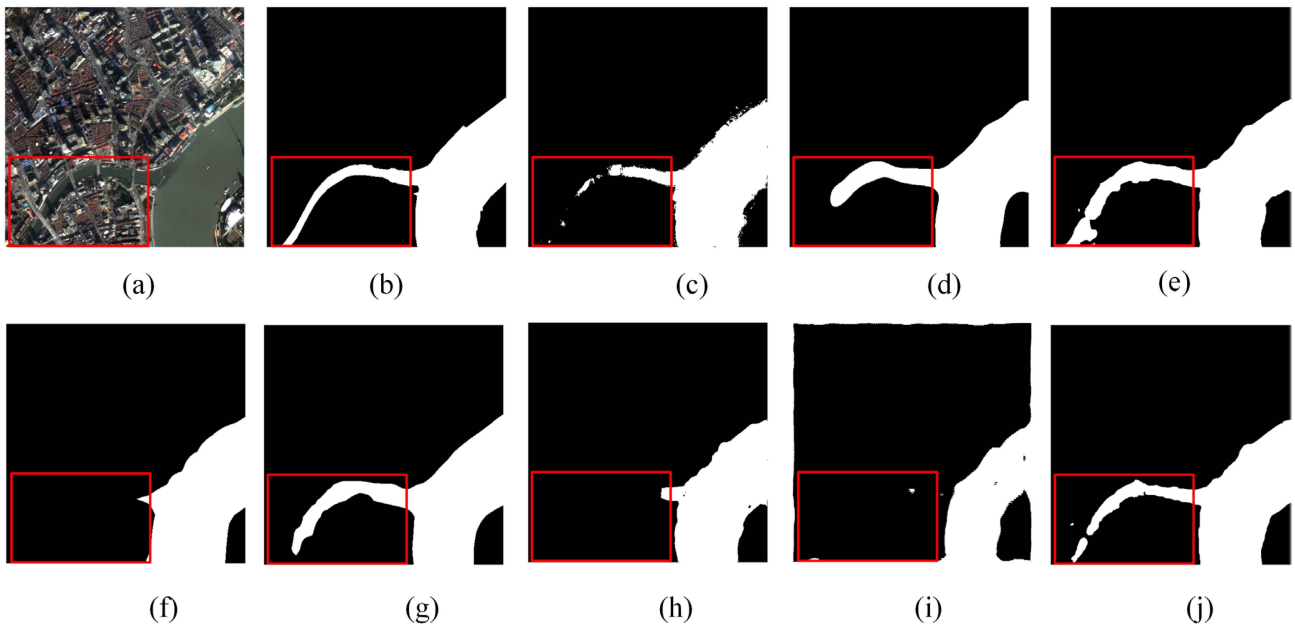

Fig. 5. Water extraction results of different methods on GID dataset, where white pixels refer to water and black pixels refer to nonwater. (a) Raw image. (b) Ground truth. (c) U-Net. (d) PSPNet. (e) DeepLabv3+. (f) CBAM. (g) DANet. (h) MECNet. (i) MSResNet. (j) AEDNet.

D. Comparative Experiments

Experiments were conducted using the GID and LoveDA datasets to compare the performance of AEDNet with several state-of-the-art models, including several general semantic segmentation models: U-Net [26], PSPNet [27], DeepLabv3+ [31], CBAM [41], and DANet [42], as well as two models for water extraction tasks: MECNet [52], [53] and MSResNet [37]. The quantitative analysis of the GID dataset is presented in Table II. The bold entities represents the best performance in a certain accuracy evaluation metric (a certain column), and the same applies to the tables in the following contents. From the results, it can be observed that CBAM and DANet exhibit relatively poor performance in terms of accuracy. This indicates that the conventional attention mechanism, which overlooks the significance of low-level features, is not effective for urban water extraction. On the other hand, DeepLabv3+ performs better among the state-of-the-art methods due to its ability to extract multiscale features and combine low-level and high-level features. MECNet demonstrates superior performance in comparison to general semantic segmentation models. However, it is important to acknowledge that MSResNet falls short in terms of accuracy when contrasted with the other models. This underscores the fact that due to the inherent complexity of urban remote sensing images, conventional water body extraction models may struggle to produce satisfactory results. However, the proposed AEDNet surpasses all the compared methods in terms of accuracy on the GID dataset. It achieves a precision of 95.52%, recall of 95.66%, and IoU of 91.55%. Moreover, AEDNet also demonstrates higher efficiency with relatively fewer parameters, making it an excellent choice for urban water extraction tasks.

The water extraction results of different methods on the GID dataset are depicted in Fig. 5. It can be observed that U-Net, PSPNet, CBAM, MECNet, and MSResNet have, to varying degrees, overlooked the small river branches that are interspersed

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE LOVEDA DATASET

Model	Backbone	Params (M)	FLOPs (G)	Precision (%)	Recall (%)	IoU (%)
U-Net	ResNet-101	31.03	873.83	77.56	83.89	67.51
PSPNet	ResNet-101	72.20	189.11	**85.50**	79.36	69.95
DeepLabv3+	ResNet-101	59.23	263.76	83.41	82.96	71.21
CBAM	ResNet-101	51.86	168.08	69.37	75.58	56.67
DANet	ResNet-101	66.31	187.49	81.97	83.37	70.45
MECNet	-	30.07	743.39	75.50	90.51	69.96
MSResNet	ResNet-101	69.55	137.53	72.74	76.34	59.36
AEDNet	ResNet-101	44.77	334.02	82.00	**88.57**	**74.15**

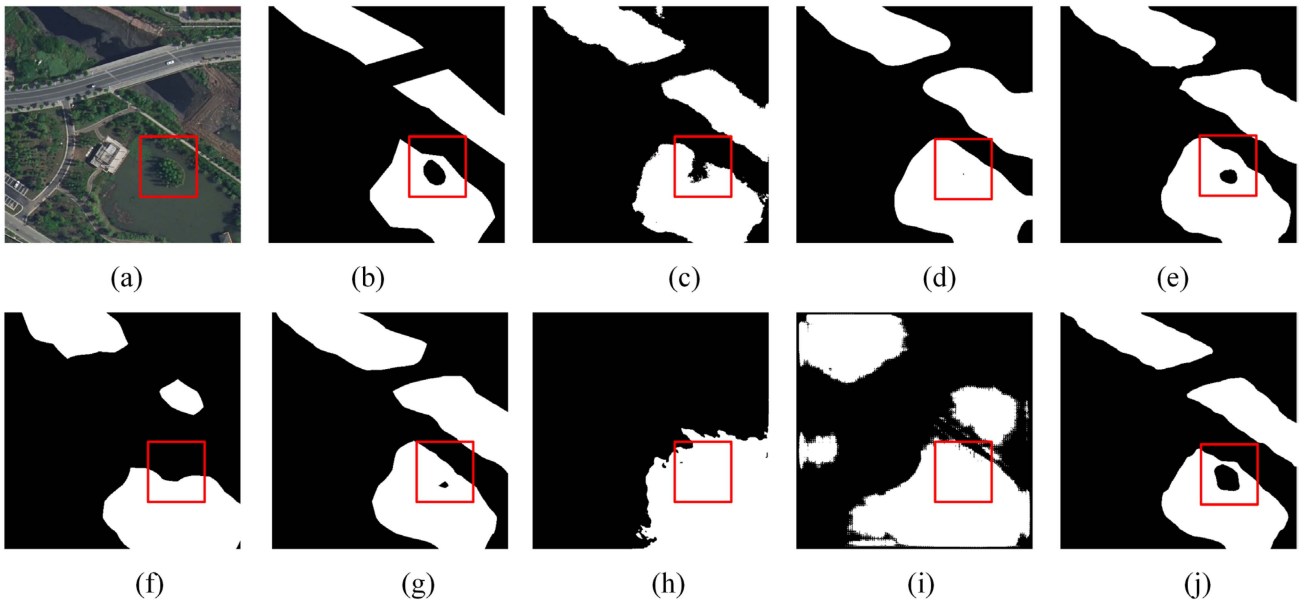

Fig. 6. Water extraction results of different methods on LoveDA dataset, where white pixels refer to water and black pixels refer to nonwater. (a) Raw image. (b) Ground truth. (c) U-Net. (d) PSPNet. (e) DeepLabv3+. (f) CBAM. (g) DANet. (h) MECNet. (i) MSResNet. (j) AEDNet.

within the urban area (highlighted in the red box) in the original image. DeepLabv3+ and DANet are able to recognize the small rivers in the example, but the extracted river boundaries appear wider than the actual boundaries due to misclassification caused by buildings and their shadows. In contrast, AEDNet extracts this section of the river more comprehensively and accurately. The extracted boundary is closer to the ground truth and shows better delineation of the water bodies. It is important to note that the "Ground Truth" of the sample data may not be completely accurate. For instance, in Fig. 5(a), several bridges on the small river are not reflected in Fig. 5(b). However, it can be observed that AEDNet is still able to recognize these unmarked nonwater bodies to some extent, as evidenced by the truncation of the small river in its extraction result. This further indicates the excellent performance of AEDNet in extracting urban water bodies.

Quantitative analysis on the LoveDA dataset is provided in Table III. Mirroring the results observed in the GID dataset, CBAM, and MSResNet display comparatively lower performance in terms of accuracy. It is noteworthy that MECNet does not surpass the performance of typical semantic segmentation models. Conversely, AEDNet excels by achieving the highest recall and IoU scores. Nevertheless, the precision of AEDNet falls slightly short, which can be attributed to the emphasis of the proposed loss function on optimizing the IoU metric, rather than precision.

The water extraction results of different methods on the LoveDA dataset are illustrated in Fig. 6. The challenging aspect of segmenting this scene lies in a small "island" located within a large body of water. The densely planted trees on the island exhibit spectral features that are similar to those of water bodies in true color images (red, green, and blue). In addition, the shadows cast by the trees can further confuse the segmentation algorithm when distinguishing between water and nonwater bodies. As observed from the example, U-Net, PSPNet, CBAM, and DANet do not perform well in accurately segmenting the water bodies. DeepLabv3+ demonstrates better performance but still

TABLE IV
QUANTITATIVE EVALUATION RESULTS OF ABLATION EXPERIMENTS ON THE GID DATASET

Model	Params (M)	FLOPs (G)	Precision (%)	Recall (%)	IoU (%)
Backbone	47.11	41.96	94.06	91.02	86.07
Backbone + AC	47.11	56.65	94.91	93.40	88.95
Backbone + AED	44.18	61.81	95.81	92.37	88.78
Backbone + JLF	47.11	41.96	93.20	93.99	87.95
Backbone + AC + AED	44.77	334.02	**96.47**	92.07	89.06
Backbone + AC + JLF	47.11	56.65	93.76	94.25	88.70
Backbone + AED + JLF	44.18	61.81	94.35	**95.94**	90.73
Backbone + AC + AED + JLF	44.77	334.02	95.52	95.66	**91.55**

misclassifies numerous nonwater pixels as water bodies. Remarkably, in this particular scene, both MECNet and MSResNet displayed notably inadequate performance, resulting in the generation of blurry water boundaries. The LoveDA dataset has higher spatial resolution than GID, and this observation underscores the limited capability of these two models in extracting complex water bodies from high-resolution remote sensing images. Only AEDNet successfully extracts the boundary between the small island and the surrounding water body, showcasing its accurate segmentation capabilities in challenging scenes.

E. Ablation Experiments

The effects of each improvement in AEDNet were investigated through relevant ablation experiments. ResNet-101 was used as the backbone, and the three improvements: AC, AED and JLF were incrementally integrated to the model. These networks were then tested on the GID dataset, and the results are summarized in Table IV. When JLF is not incorporated, the conventional cross-entropy loss function is employed. It is evident that "Backbone + AC + AED + JLF" confers clear advantages in terms of IoU and demonstrates relatively high precision and recall. Each enhancement, when compared to the baseline "Backbone," leads to a substantial improvement in all three accuracy evaluation metrics. "Backbone + AC + AED" achieves the highest precision, reinforcing the idea on the side that our proposed JLF, as opposed to conventional cross-entropy loss, places a greater emphasis on IoU, making it more compelling for semantic segmentation tasks. On the other hand, "Backbone + AED + JLF" exhibits suboptimal performance in precision and IoU, underscoring the significance of employing AC to enhance accuracy. "Backbone + AC + JLF" performs least effectively among networks that incorporate two enhancements, with accuracy even lower than some networks utilizing only one single improvement. This emphasizes that haphazardly stacking enhancements may not necessarily yield the desired results. Conversely, "Backbone + AC + AED + JLF" demonstrates a significant improvement compared to "Backbone + AED," highlighting the synergistic advantages of integrating AED with AC and JLF.

Fig. 7 illustrates the results of water extraction in ablation experiments conducted on the GID dataset. Fig. 7(b) serves as a reference for accurately delineating larger water bodies, although it may occasionally overlook the boundaries between water bodies and small rivers in this scene. Upon comparing the segmentation results, it becomes evident that all the networks outperform the backbone. Networks lacking the AED component tend to miss numerous water pixels, resulting in indistinct boundaries between the extracted water bodies and urban areas. In the red box at the top of this scene, Fig. 7(e), (g), and (j) demonstrates superior extraction of narrow water features, with Fig. 7(j) displaying the best performance. When focusing on the red box at the top of this scene, Fig. 7(j) exhibits enhanced performance in delineating water body boundaries compared to Fig. 7(e), (g), and (i), underscoring the beneficial role of AC and JLF in enhancing segmentation details. These visual results provide further validation of the effectiveness of the proposed enhancements in AEDNet, highlighting the pivotal role of AED, AC, and JLF in achieving superior performance in urban water extraction.

IV. DISCUSSION

In this section, the superiority of AEDNet compared to traditional methods in urban water extraction tasks, the values of hyperparameters in the proposed JLF, and the defects and mitigation measures of AEDNet will be discussed.

A. Comparison With NDWI-Based Methods

Comparative experiments have unequivocally demonstrated that AEDNet outperforms other state-of-the-art semantic segmentation models in terms of water extraction accuracy. In this section, a detailed comparison will be conducted between AEDNet and traditional NDWI-based water extraction methods. The NDWI equation is as follows:

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR}}{\mathrm{Green} + \mathrm{NIR}} \tag{18}$$

where *Green* refers to the green band and *NIR* refers to the near-infrared band. Given that the LoveDA dataset lacks the near-infrared band necessary for NDWI calculations, this comparative experiment was restricted solely to the validation set of the GID dataset, as indicated in Table I.

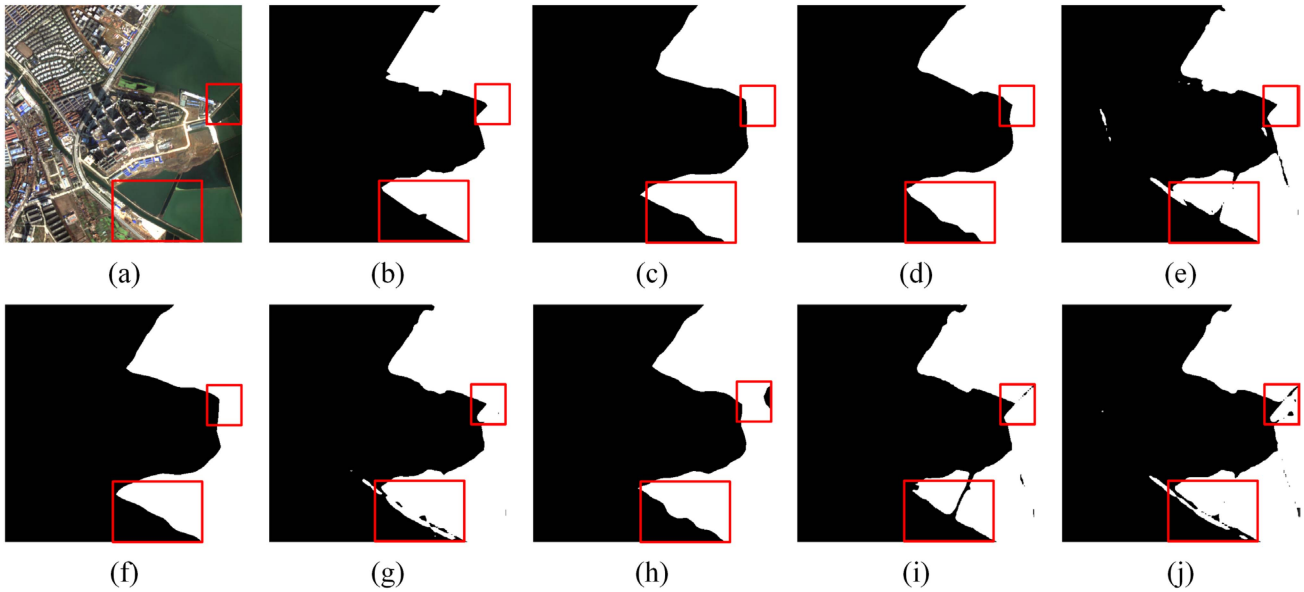

Fig. 7. Water extraction results of ablation experiments on GID dataset, where white pixels refer to water and black pixels refer to nonwater. (a) Raw image. (b) Ground truth. (c) Backbone. (d) Backbone with AC. (e) Backbone with AED. (f) Backbone with JLF. (g) Backbone with AC and AED. (h) Backbone with AC and JLF. (i) Backbone with AED and JLF. (j) Backbone with AC, AED, and JLF.

TABLE V
ACCURACY EVALUATION RESULTS OF NDWI METHODS AND AEDNET ON THE GID DATASET

Method	Precision (%)	Recall (%)	IoU (%)
Threshold = 0.2	61.66	89.92	56.83
Threshold = 0.3	68.94	77.78	60.19
Threshold = 0.4	74.80	68.87	58.77
Otsu	51.74	92.89	49.87
AEDNet	**95.52**	**95.66**	**91.55**

After calculating the NDWI image for each sample, two distinct methods were employed for binary segmentation. The first method involved manually setting a threshold (0.2, 0.3, 0.4) and applying this threshold uniformly to all NDWI images. The second method utilized the Otsu algorithm to automatically compute the threshold for each NDWI image and perform segmentation. The implementation of the Otsu algorithm involves the calculation of interclass variance, with the equation as follows:

$$\delta^2 = p_w \times (M_w - M)^2 + p_{nw} \times (M_{nw} - M)^2 \tag{19}$$

where δ is the interclass variance of water and nonwater, p_w and p_{nw} are the proportions of the water and nonwater classes, M_w and M_{nw} are the mean values of water and nonwater classes, M is the mean value of the whole image. The Otsu algorithm iterates through all possible thresholds and finds the one that maximizes the interclass variance.

Comparison between the NDWI-based methods and AEDNet (utilizing ResNet-101) is presented in Table V. It is important to note that NDWI-based methods and deep learning methods differ significantly in terms of efficiency, making it impractical to calculate directly comparable indicators in this experiment. The results clearly indicate that AEDNet achieves significantly higher water extraction accuracy when compared to NDWI-based methods. Among the NDWI-based methods, the accuracy of Otsu methods is not ideal, while the accuracy of the manual threshold method is highly sensitive to different threshold settings.

Fig. 8 showcases the water extraction results of NDWI-based methods and AEDNet on the GID dataset, featuring six samples, including those previously presented in Figs. 5 and 7. For the NDWI-based methods, segmentation results with a manually set threshold of 0.3 (yielding the highest IoU) and the Otsu threshold segmentation results were displayed. A clear observation is that the water extraction results of AEDNet exhibit a higher degree of consistency with the "Ground Truth" when compared to NDWI-based methods. Both manual and Otsu threshold segmentation methods tend to misclassify building roofs and shadows as water bodies, a challenge that AEDNet minimizes. However, it is worth noting that AEDNet may sacrifice some details at the boundaries between water bodies and nonwater bodies, such as narrower roads and bridges spanning water bodies, whereas NDWI-based methods excel in preserving these fine-grained distinctions.

B. Experiment About the Hyperparameter in Loss

The hyperparameter γ in (12) significantly affects the performance of the model, so its optimal value needs to be discussed. Different γ were set and the performance of AEDNet is tested on GID dataset. The IoU curves in the validation stage are shown in Fig. 9. As γ increases, the IoU tends to increase until $\gamma = 0.9$, illustrating the obvious effect of Lovász hinge on improving the prediction accuracy of the model. Then, the IoU decrease when $\gamma = 1$, indicating that the model with the JLF has better

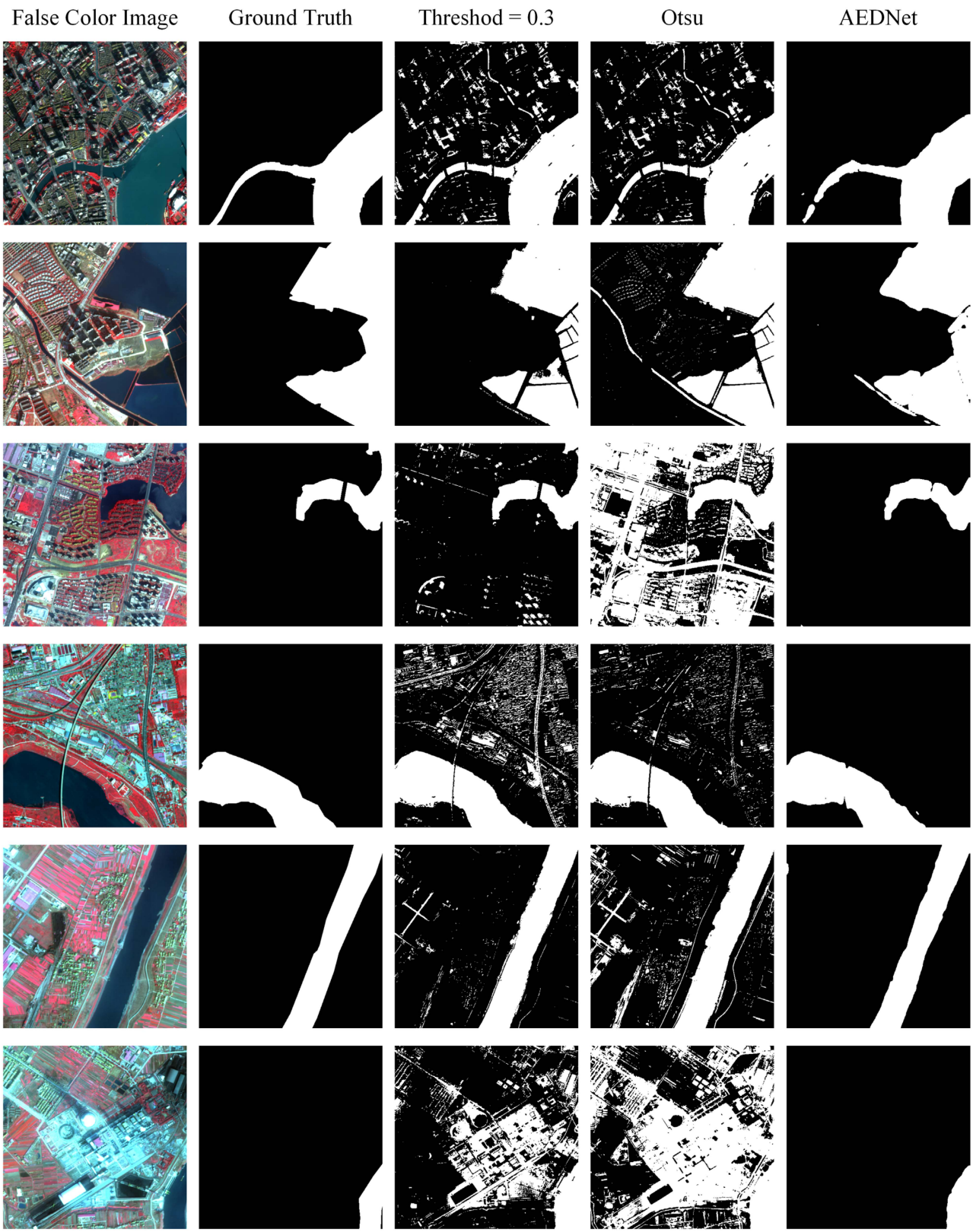

Fig. 8. Comparison of NDWI-based methods and AEDNet on GID dataset, where white pixels refer to water and black pixels refer to nonwater. The false color image refers to the image displayed in the order of NIR–Red–Green bands.

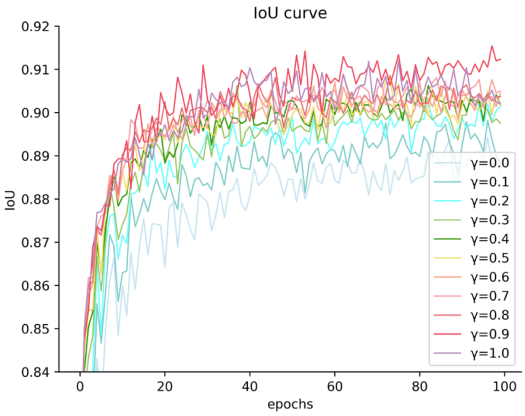

Fig. 9. Comparison of IoU curves of AEDNet with different γ in the validation stage on GID dataset.

performance compared to using Lovász hinge alone. Based on these experiments, it is found that the model performs best when $\gamma = 0.9$, and this optimal value have been used in all other experiments using the JLF previously.

C. Deficiencies and Future Steps

In this article, an AEDNet was proposed for urban water extraction from high-resolution remote sensing images, and we have proved the effectiveness of this model through a series of experiments. However, AEDNet still has some deficiencies. First, as shown in Fig. 8, compared with traditional methods, AEDNet has fewer misclassifications, but it cannot obtain as clear boundaries as threshold segmentation results. Fuzzy boundaries represent a common challenge in the realm of semantic segmentation. Contemporary research endeavors are increasingly dedicated to addressing this issue, encompassing approaches such as introducing supplementary losses related to boundary information [52], [53] and segregating the features of boundaries and those within the regions separately [54]. In the future, our exploration will focus on incorporating these akin concepts into AEDNet to enhance the precision of urban water boundary delineation.

Second, the training process of AEDNet demands substantial computing resources, and GPUs with limited memory often face challenges in efficiently completing the training. For this reason, this method is difficult to use in large areas (such as the global scale), thus limiting its application value. Presently, numerous scholars have introduced light-weight models designed specifically for remote sensing image semantic segmentation [55], [56]. Our future efforts will be directed toward amalgamating the merits of these lightweight models with AEDNet, with the aim of enhancing its computational efficiency and practical applicability.

V. Conclusion

In this article, an AEDNet was proposed for urban water extraction from high-resolution remote sensing images. The proposed ResNet-101 with AC can extract features at different levels and scales as a backbone. The AED structure with dual attention modules can effectively capture global feature dependencies in both the spatial and channel domains. The proposed JLF in combination with Lovász hinge and weighted cross entropy can further improve the model performance on urban water extraction.

Through a series of rigorous comparative experiments and meticulous ablation studies on the GID and LoveDA datasets, the efficacy of the three proposed enhancements has been substantiated. The results underscore the superior accuracy and efficiency of AEDNet in urban water extraction tasks when compared to analogous methods. Furthermore, our research has established that AEDNet achieves greater accuracy than traditional NDWI-based methods in urban scenarios.

In our forthcoming research, we will concentrate on enhancing the clarity of water boundary extraction results generated by this model. Simultaneously, our efforts will be directed toward optimizing the model's computational efficiency by making it more lightweight.

Acknowledgment

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] 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.

[2] F. Chen, X. Chen, T. Van de Voorde, D. Roberts, H. Jiang, and W. Xu, "Open water detection in urban environments using high spatial resolution remote sensing imagery," *Remote Sens. Environ.*, vol. 242, 2020, Art. no. 111706, doi: 10.1016/j.rse.2020.111706.

[3] X. Yang, Q. Qin, P. Grussenmeyer, and M. Koehl, "Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery," *Remote Sens. Environ.*, vol. 219, pp. 259–270, 2018, doi: 10.1016/j.rse.2018.09.016.

[4] L. Yue, B. Li, S. Zhu, Q. Yuan, and H. Shen, "A fully automatic and high-accuracy surface water mapping framework on Google Earth Engine using Landsat time-series," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 210–233, Dec. 2023, doi: 10.1080/17538947.2023.2166606.

[5] S. Longfei, L. I. Zhengxuan, G. Fei, and Y. Min, "A review of remote sensing image water extraction," *Remote Sens. Natural Resour.*, vol. 33, no. 1, pp. 9–11, 2021.

[6] 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.

[7] H. Xu, "A study on information extraction of water body with the modified normalized difference water index (MNDWI)," *J. Remote Sens.*, vol. 9, no. 5, 2005, Art. no. 595.

[8] 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, doi: 10.1016/j.rse.2013.08.029.

[9] 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.

[10] 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, 2022, Art. no. 128202, doi: 10.1016/j.jhydrol.2022.128202.

[11] Z. Li, X. Zhang, and P. Xiao, "Spectral index-driven FCN model training for water extraction from multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 192, pp. 344–360, 2022, doi: 10.1016/j.isprsjprs.2022.08.019.

[12] T. Bangira, S. M. Alfieri, M. Menenti, and A. van Niekerk, "Comparing thresholding with machine learning classifiers for mapping complex water," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1351, doi: 10.3390/rs11111351.
[13] D. Xu, D. Zhang, D. Shi, and Z. Luan, "Automatic extraction of open water using imagery of landsat series," *Water*, vol. 12, no. 7, 2020, Art. no. 1928, doi: 10.3390/w12071928.
[14] X. Li, X. Lyu, Y. Tong, S. Li, and D. Liu, "An object-based river extraction method via optimized transductive support vector machine for multi-spectral remote-sensing images," *IEEE Access*, vol. 7, pp. 46165–46175, 2019.
[15] S. Guan, X. Wang, L. Hua, and L. Li, "Quantitative ultrasonic testing for near-surface defects of large ring forgings using feature extraction and GA-SVM," *Appl. Acoust.*, vol. 173, 2021, Art. no. 107714.
[16] T. D. Acharya, D. H. Lee, I. T. Yang, and J. K. Lee, "Identification of water bodies in a landsat 8 OLI image using a J48 decision tree," *Sensors*, vol. 16, no. 7, pp. 1075, 2016.
[17] J. Yang, X. Wang, J. Wang, C. Ye, and J. Xiong, "Water extraction of hyperspectral imagery based on a fast and effective decision tree water index," *J. Appl. Remote Sens.*, vol. 15, no. 4, 2021, Art. no. 42605.
[18] B. C. Ko, H. H. Kim, and J. Y. Nam, "Classification of potential water bodies using Landsat 8 OLI and a combination of two boosted random forest classifiers," *Sensors*, vol. 15, no. 6, pp. 13763–13777, 2015.
[19] T. D. Acharya, A. Subedi, and D. H. Lee, "Evaluation of machine learning algorithms for surface water extraction in a Landsat 8 scene of Nepal," *Sensors*, vol. 19, no. 12, 2019, Art. no. 2769.
[20] K. Li, J. Wang, and J. Yao, "Effectiveness of machine learning methods for water segmentation with ROI as the label: A case study of the Tuul river in Mongolia," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102497, doi: 10.1016/j.jag.2021.102497.
[21] H. Guo, G. He, W. Jiang, R. Yin, L. Yan, and W. Leng, "A multi-scale water extraction convolutional neural network (MWEN) method for GaoFen-1 remote sensing images," *ISPRS Int. J. Geoinf.*, vol. 9, no. 4, 2020, Art. no. 189.
[22] M. Jiang, X. Zhang, Y. Sun, W. Feng, Q. Gan, and Y. Ruan, "AFSNet: Attention-guided full-scale feature aggregation network for high-resolution remote sensing image change detection," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 1882–1900, Dec. 2022, doi: 10.1080/15481603.2022.2142626.
[23] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5412012, doi: 10.1109/TGRS.2022.3207551.
[24] R. Hang, G. Li, M. Xue, C. Dong, and J. Wei, "Identifying oceanic eddy with an edge-enhanced multiscale convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9198–9207, Oct. 2022, doi: 10.1109/JSTARS.2022.3215696.
[25] 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.
[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
[27] 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.
[28] L. - C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[29] 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.
[30] L. - C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–14.
[31] 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.
[32] Y. Wang, Z. Li, C. Zeng, G. - S. Xia, and H. Shen, "An urban water extraction method combining deep learning and Google Earth engine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 769–782, Feb. 2020.
[33] W. Fang et al., "Recognizing global reservoirs from Landsat 8 images: A deep learning approach," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3168–3177, Sep. 2019.
[34] R. Nagaraj and L. S. Kumar, "Multi scale feature extraction network with machine learning algorithms for water body extraction from remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 17, pp. 6349–6387, Sep. 2022, doi: 10.1080/01431161.2022.2136505.
[35] B. Liu, S. Du, L. Bai, S. Ouyang, H. Wang, and X. Zhang, "Water extraction from optical high-resolution remote sensing imagery: A multi-scale feature extraction network with contrastive learning," *GiSci. Remote Sens.*, vol. 60, no. 1, Dec. 2023, Art. no. 2166396, doi: 10.1080/15481603.2023.2166396.
[36] Z. Zhang, M. Lu, S. Ji, H. Yu, and C. Nie, "Rich CNN features for water-body segmentation from very high resolution aerial and satellite imagery," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 31205, doi: 10.3390/rs13101912.
[37] B. Dang and Y. Li, "MSResNet: Multiscale residual network via self-supervised learning for water-body detection in remote sensing imagery," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3122, doi: 10.3390/rs13163122.
[38] X. Li et al., "A synergistical attention model for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5400916, doi: 10.1109/TGRS.2023.3243954.
[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[41] S. Woo, J. Park, J. - Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[42] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[43] J. Wang et al., "FWENet: A deep convolutional neural network for flood water body extraction based on SAR images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 345–361, Dec. 2022, doi: 10.1080/17538947.2021.1995513.
[44] X. Zhang, J. Li, and Z. Hua, "MRSE-Net: Multiscale residuals and se-attention network for water body segmentation from satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5049–5064, Jun. 2022, doi: 10.1109/JSTARS.2022.3185245.
[45] Y. Yu et al., "WaterHRNet: A multibranch hierarchical attentive network for water body extraction with remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, 2022, Art. no. 103103, doi: 10.1016/j.jag.2022.103103.
[46] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multiscale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5609519, doi: 10.1109/TGRS.2023.3276703.
[47] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4413–4421.
[48] 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.
[49] X. - Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.
[50] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021, pp. 1–12.
[51] H. Zhang et al., "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.
[52] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1623–1632.
[53] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1448–1457.
[54] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.
[55] S. Liu, J. Cheng, L. Liang, H. Bai, and W. Dang, "Light-weight semantic segmentation network for UAV remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8287–8296, Aug. 2021, doi: 10.1109/JSTARS.2021.3104382.
[56] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5614812, doi: 10.1109/TGRS.2021.3131331.

Yanjiao Song received the B.S. degree in geographic information science from Hohai University, Nanjing, China, in 2022. He is currently working toward the master's degree in remote sensing science and technology with School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include Intelligent interpretation of remote sensing images, deep learning, and mapping of surface water.

Xiaoping Rui received the Ph.D. degree in cartography and geographic information system from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2004.

He is currently a Professor with the School of Earth Sciences and Engineering, Hohai University, Nanjing, China. His research interests include machine learning, remote sensing image classification, and geo-data mining.

Junjie Li received the B.S. degree in geographic information science from Central China Normal University, Wuhan, China, in 2018. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with School of Remote Sensing and Information Engineering, Wuhan University, Wuhan.

His research interests include deep learning and remote sensing image processing.