

NFANet: A Novel Method for Weakly Supervised Water Extraction from High-Resolution Remote Sensing Imagery

Ming Lu, Leyuan Fang, *Senior Member, IEEE*, Muxing Li, Bob Zhang, *Senior Member, IEEE*, Yi Zhang, *Senior Member, IEEE*, and Pedram Ghamisi, *Senior Member, IEEE*

Abstract—The use of deep learning for water extraction requires precise pixel-level labels. However, it is very difficult to label high-resolution remote sensing images at the pixel level. Therefore, we study how to utilize point labels to extract water bodies and propose a novel method called the neighbor feature aggregation network (NFANet). Compared with pixel-level labels, point labels are much easier to obtain, but they will lose much information. In this paper, we take advantage of the similarity between the adjacent pixels of a local water-body, and propose a neighbor sampler to resample remote sensing images. Then, the sampled images are sent to the network for feature aggregation. In addition, we use an improved recursive training algorithm to further improve the extraction accuracy, making the water boundary more natural. Furthermore, our method utilizes neighboring features instead of global or local features to learn more representative features. The experimental results show that the proposed NFANet method not only outperforms other studied weakly supervised approaches, but also obtains similar results as the state-of-the-art ones.

Index Terms—Convolutional neural network (CNN), deep learning, semantic segmentation, water extraction, weak supervision.

I. INTRODUCTION

WATER body extraction from high-resolution remote sensing images is an important research topic in the field of remote sensing applications. Although the traditional algorithms have made some progress in water-body extraction,

This work was supported in part by the National Natural Science Fund of China under Grant 61922029, in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2019RS2016, in part by the Key Research and Development Program of Hunan under Grant 2021SK2039, and in part by the Natural Science Fund of Hunan Province under Grant 2021JJ30003. (Corresponding author: Leyuan Fang.)

Ming Lu is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xian, 710126, China (e-mail: 1148462196@qq.com).

Leyuan Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: fangleyuan@gmail.com).

Muxing Li is with the College of Engineering & Computer Science, Australian National University, ACT 2601, Australia (e-mail: u6481081@anu.edu.au)

Bob Zhang is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: bobzhang@um.edu.mo)

Y. Zhang is with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: yzhang@scu.edu.cn).

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

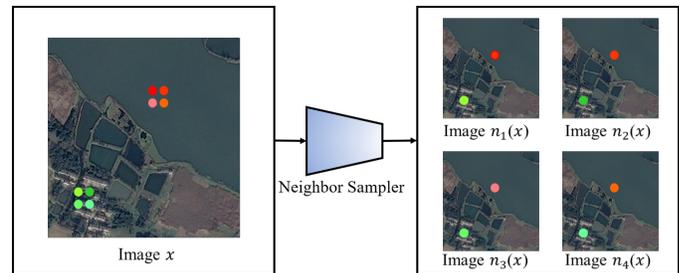


Fig. 1. Use of the neighbor sampler to sample the input image, where the left is the original image, and the right is the neighbor image group. The dots of different colors indicate the pixel mapping relationship.

there are still problems such as cumbersome manual feature extraction and insufficient extraction accuracy. At the same time, with the rapid advancement of remote sensing technology, high-resolution remote sensing images contain increasingly detailed texture information and edge structure information. With traditional remote sensing water-body extraction methods, it is difficult to make full use of the semantically rich high-resolution remote sensing image information, and also difficult to meet the growing demand of remote sensing applications.

In recent years, deep learning has become an emerging research hot spot in the field of artificial intelligence [1]–[4]. The rapid development of deep learning technology and the improvement of computer hardware performance have enabled deep learning, especially convolutional neural network (CNN) [5]–[7], to be successfully applied to many important tasks, such as image classification [8], [9], target detection [10], [11], and semantic segmentation [12], [13]. In addition, due to its excellent performance, deep learning has been widely used in the remote sensing image fields, such as remote sensing image classification [14]–[18], change detection [19], [20], and ground object extraction [21], [22]. Hong *et al.* [23] proposed a variety of fusion architectures to solve the special case of multi-modal learning and cross-modality learning that is widely used in remote sensing image classification applications. Hong *et al.* [24] established a method that combined graph convolutional network (GCN) and CNN to fuse different hyperspectral features to improve the performance of hyperspectral classification.

Deep learning is also used for water extraction, and the deep learning-based water extraction method can significantly

improve the extraction accuracy of traditional techniques. Zhang *et al.* [25] proposed a cascaded fully-convolutional network to improve the performance of water-body detection, and introduced fully-convolutional conditional random fields (CRF) to realize automatic learning of Gaussian kernels in CRF. Li *et al.* [26] proposed a dense local feature compression network. Each layer of this network was densely connected to receive all its previous feature maps. In addition, the network included a local feature compression (LFC) module to integrate spatial and spectral information. He *et al.* [27] applied UNet and attention-UNet to Landsat-8 images for glacial lakes extraction. Specifically, the self-attention module was added to the skip connection to strengthen the extraction of the network features of the glacial lake. Wang *et al.* [28] combined the multi-scale convolutional network with Google Earth Engine and proposed an offline training and online prediction method to monitor urban water. In order to improve the ability to extract the boundaries of water bodies, Miao *et al.* [29] proposed a new loss function called edges weighting loss, which calculated the Euclidean distance to obtain a higher weight for the boundary.

In general, the success of deep learning for feature extraction is highly dependent on the availability of sufficient pixel-level labels for training. However, high-resolution remote sensing images are large in both scale and data volume, which makes pixel-level labeling extremely laborious. Pixel-level annotation usually requires a lot of time and labor costs, and professional knowledge to accurately mark uncertain boundaries between different classes of interest, which hinders the development of high-resolution remote sensing image feature extraction to a certain extent. Training algorithm models using weak labels has received increasing attention in the field of computer vision [30]–[32]. Compared with fully-supervised semantic segmentation, weak-supervised learning [33]–[35] does not require pixel-level labels, and has the characteristics of fast labeling and low time cost. However, the use of weak annotations makes the supervision information seriously insufficient, and thus the key information, such as shape, texture, and edges, is usually lost, which makes it difficult to extract water from high-resolution remote sensing images with complex scenes.

In this paper, we propose a new solution to the above problems. Unlike other natural objects, water-bodies are usually liquid, whose colors and textures are very similar. Therefore, there is a high degree of similarity between neighbor pixels in water-bodies, which makes the inherent difference between the neighbor pixels of the water-body generally smaller than those of the non-water-body. Thus, we attempt to map the neighbor pixels of remote sensing images to the same location in space, and then extract neighbor features from multiple neighbor pixels. Next, we use the neighbor features to jointly decide whether the pixel at this location belongs to the water-body. Based on the above motivation, we propose the neighbor feature aggregation network (NFANet) to make full use of this property. Specifically, we utilize a sampling method called neighbor sampler to generate a set of neighbor images from high-resolution remote sensing images. The neighbor pixels of the original image are allocated separately to each neighbor

image, so that the pixel values of any two neighbor images at the same position are similar, but the pixel values are different. On the whole, neighbor image groups have similar but different characteristics (see Fig. 1). Once the neighbor images are generated, we use an end-to-end model to perform feature extraction on each image of the neighbor image groups, and aggregate the features by using a feature aggregation module. Compared with other methods that only use the local or global information from the image, neighbor feature aggregation effectively utilizes the neighbor information, and so more representative features can be learned. Finally, we used recursive training to further refine the water-body boundary, making the water-body edge smoother and more natural. Our experiments demonstrate that the proposed NFANet can utilize neighbor information and achieve good results, that are comparable to the results of the current state-of-the-art fully-supervised methods.

The main contributions of this paper are summarized as follows.

- 1) We propose a new sampling method to resample high-resolution remote sensing images and generate neighbor image groups in order to extract neighbor information, which helps to distinguish between water-bodies and land boundaries.
- 2) We develop a weakly supervised deep learning algorithm model called NFANet. Compared with other methods that only use local features or global features, the proposed approach enables using neighbor information to be used to learn more representative characteristics.
- 3) We propose a recursive training strategy, which can further refine the boundary of the water-body, making the boundary smooth and closer to the ground truth.

This paper is organized as follows. Section II introduces some related works and the main contribution of our method. Section III discusses the proposed NFANet and related modules in detail. In section IV, a comprehensive experiment is carried out on the proposed method on the water-body data-set, and a comparison to other state-of-the-art methods is given. Section V summarizes the conclusions of this paper.

II. RELATED WORKS AND CONTRIBUTIONS

A. Related Works

In the past few years, a large number of approaches have been proposed to solve the problem of water-body extraction from remote sensing images. Generally, these methods can be divided into two categories: methods based on manual features and methods based on deep learning.

1) *Methods based on manual features:* The key problem of water-body extraction is how to determine the difference between the water-body and other natural features. For this reason, methods based on manual features usually hope to find one or more distinctive features that can distinguish water bodies and non-water bodies. Gao *et al.* [36] proposed the normalized differential water index method (NDWI), which uses a specific band of remote sensing images for normalized difference processing. This method can highlight the water

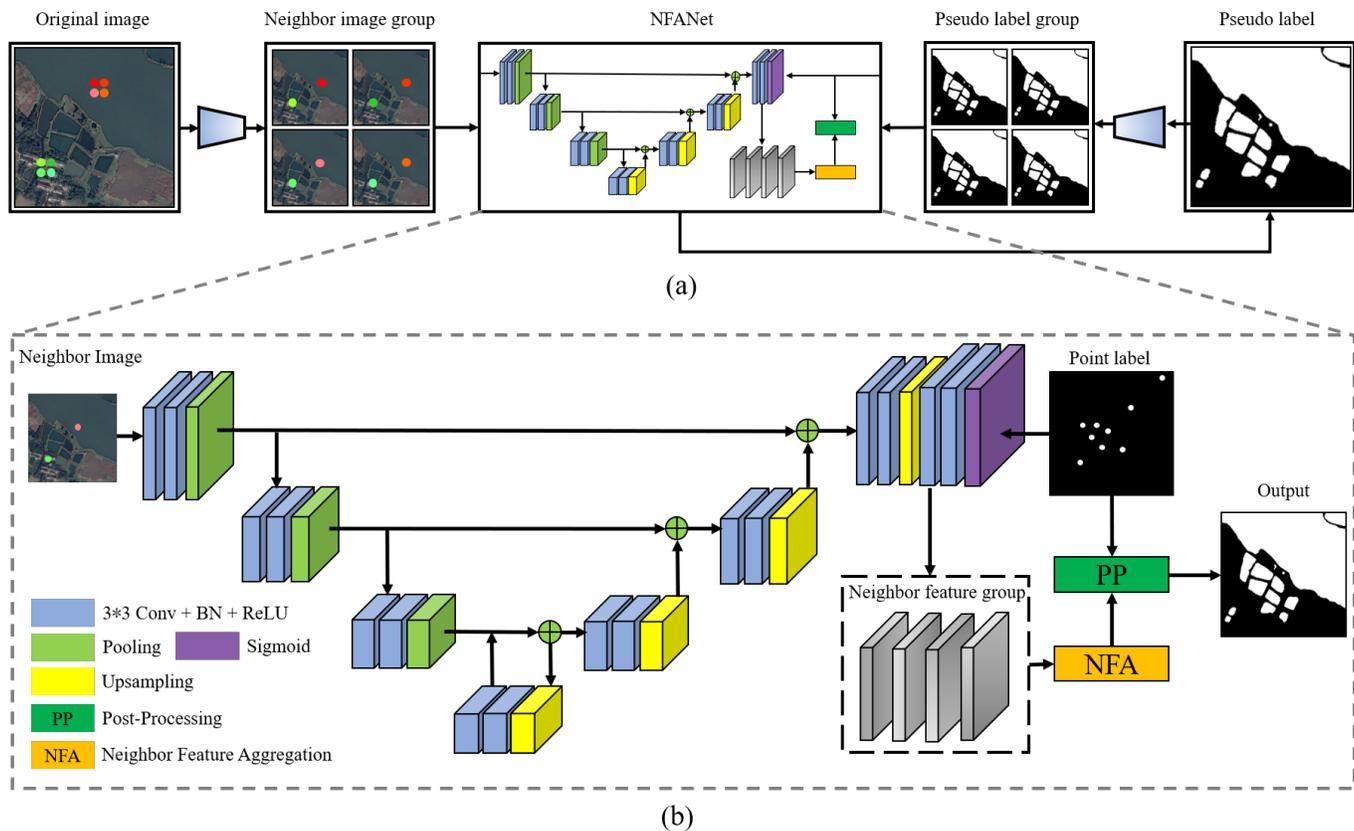


Fig. 2. The proposed weakly supervised water-body extraction framework; (a) represents the recursive training process, and (b) is the proposed NFANet.

information in the image and weaken the influence of non-water factors. Trias-Sanz *et al.* [37] used a hierarchical segmentation method combining color space transformation and texture features to segment the lake. Barton *et al.* [38] used the brightness temperature extracted by AVHRR channel 4 to identify the water-body and monitor flooding.

2) *Methods based on deep learning:* Although there has been some progress in the task of water-body extraction, manual features still have problems, such as cumbersome design and limited extraction accuracy. Recently, with the powerful generalization ability of deep learning, related models have been widely used in the field of water extraction. In order to balance the background semantics and edge details, Zhang *et al.* [39] proposed a network based on R-CNN called Mask-R-CNN, which combines semantic segmentation and target detection at a decision-level, and improved the segmentation accuracy of water, pools, roads, and buildings. Li *et al.* [40] argued that a single-scale model may lead to inaccurate extraction of water bodies. They combined DeepLab v3+ with multi-scale joint prediction and used fully connected CRF to optimize the water-body boundary. Wang *et al.* [41] used a model accuracy variation degree algorithm and a sample quality analysis algorithm to reduce the impact of incorrect water labeling on the model, thereby improving prediction accuracy. Cui *et al.* [42] proposed a multi-scale feature extraction module and an adaptive feature fusion module and applied them to sea-land segmentation. Chu *et al.* [43] constructed Res-

UNet to handle complex sea-land segmentation scenes, and then used fully connected CRF and morphological operations to further refine the water-body boundary. Dong *et al.* [44] introduced the spatial position and feature space of adjacent pixels into the loss function, and proposed a sub-neighbor system constraint. Feng *et al.* [45] combined the super-pixels-based CRF and prediction results of UNet to enhance water extraction.

B. Weakly supervised learning

Fully-supervised learning can extract the target object in a remote sensing image very accurately, but usually requires the use of pixel-level labels, which consumes a lot of manpower and financial resources, and needs a certain degree of professional knowledge for labeling. Some researchers have tried to use traditional methods combined with deep learning to solve weak supervision problems. Fu *et al.* [46] combined super-pixels and a local map to obtain rough pseudo-labels to train a water extraction model. Chen *et al.* [47] combined super-pixel pooling with multi-scale feature fusion to detect buildings. Other researchers attempted to obtain better results by using the extraction capabilities of the neural network itself. Wang *et al.* [48] learned from the principle of class activation maps (CAM) [49] and extracts feature maps from UNet [50] for hard-threshold processing to obtain segmentation predictions. These methods have achieved promising results in the field of

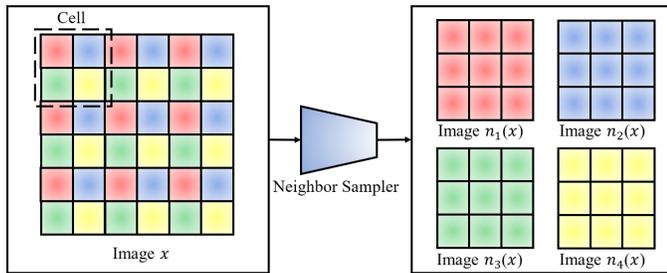


Fig. 3. Proposed neighbor sampler. When k is set to 2, a cell contains neighbor pixels of 2×2 size and is reallocated to a neighbor image group. Best viewed in color.

weak-supervised learning, but do not consider the characteristics of the image itself.

C. The novelty of the proposed method

The proposed method belongs to the category of weakly supervised learning methods. Due to the similarity of neighbor pixels in local water bodies, the potential features between neighbor pixels in local water-bodies are more alike than those in non-water-bodies. Based on this characteristic, a neighbor sampler is proposed to resample remote sensing images and use a feature aggregation module to obtain rough results. The proposed method uses the features of neighbor pixels instead of only the global features or local features of the image, and so more features can be learned, which is more representative. Compared with methods such as Unet-based CAM (U-CAM) [48], our approach exhibits better performance and is comparable to the visual results of the fully-supervised method. In addition, it is inspired by the fact that when applying the resulting model over the training set, the outputs of the network can significantly better capture the shape of objects than that of just pseudo-labels [51]. We have observed through experiments that when the training set is input into the network again, the obtained network output will become smoother than the coarse-grained pseudo-label, which improves the accuracy of the prediction result to a certain extent.

III. PROPOSED METHOD

Fig. 2 illustrates the proposed weakly supervised water extraction framework. Fig. 2a shows the entire recursive training process, which is described in subsection C below. The acquisition of pseudo-labels is shown in Fig. 2b. We input neighbor images into the network and use point labels for supervision to obtain neighbor features. The feature aggregation module is then used for feature aggregation. Finally, post-processing is performed to obtain pseudo-labels. We describe the details of each of the above steps in the following sections.

A. Neighbor Sampler

First, we introduce a neighbor sampler to obtain a neighbor images group $(n_1(x), n_2(x), \dots, n_L(x))$ from a single optical remote sensing image x . L represents the number of neighbor images. Fig. 3 shows the schematic diagram of generating a group of neighbor images using the neighbor

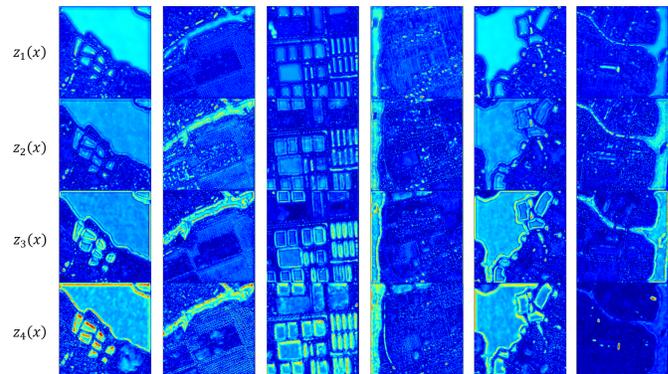


Fig. 4. Visualization results of neighbor feature groups after CMax pooling. Columns represent the same neighbor feature group.

sampler. Let us assume that the width, height, and channel of the input image x are W, H , and C , respectively. The implementation of the neighbor sampler $N = (n_1, n_2, \dots, n_L)$ is described as follows:

- 1) The image x is divided into $\frac{W}{K} \times \frac{H}{K}$ cells, where the size of each cell is $K \times K \times C$. We experimentally set K to 2 and, therefore, $L = K \times K = 4$.
- 2) For the i -th row and j -th column of the cell, the pixels in the adjacent positions of each cell are selected in the order from top to bottom and from left to right, which are regarded as the (i, j) -th elements of $N = (n_1(x), n_2(x), \dots, n_L(x))$. When K is set to 2, the pixels at the upper left, upper right, lower left, and lower right adjacent positions are selected, respectively.
- 3) For all $\frac{W}{K} \times \frac{H}{K}$ cells being divided in step 1), step 2) will be repeated until all the cells are resampled, and a neighbor sampler $N = (n_1, n_2, \dots, n_L)$ is generated. Given an optical remote sensing image x , a neighbor images group $(n_1(x), n_2(x), \dots, n_L(x))$ is generated, where the size of each neighbor image is $\frac{W}{K} \times \frac{H}{K} \times C$.

In this way, the neighbor image dataset can be generated from the original dataset. Neighbor images are similar but not identical, because for any two neighbor images, the (i, j) -th pixel comes from the neighboring location of the original remote sensing image.

B. Feature aggregation module and post-processing

We input the neighbor images group into an end-to-end network to extract features, and obtain the corresponding neighbor feature group $(f_1(X), f_2(X), \dots, f_L(X)) \in \mathbb{R}^{H \times W \times C \times L}$, where $f_l(X) \in \mathbb{R}^{H \times W \times C}$ represents the feature maps extracted from the l -th image in the neighbor images group. We use the encoder-decoder structure as the feature extraction network. Specifically, the feature maps are extracted from the penultimate convolutional layer. The network structure is shown in Fig. 2b. It is worth noting that the network is replaceable (in the experimental part, a variety of network structures are used for feature extraction). CMax pooling is adopted to reduce the number of channels of each neighbor feature to one. CMax pooling is defined mathematically in detail as follows: Given a three-dimensional

feature maps tensor group $F = (f_1(x), f_2(x), \dots, f_L(x)) \in \mathbb{R}^{H \times W \times C \times L}$, the operation of CMax pooling is as follows:

$$z_{i,j,l}(x) = \max(f_{i,j,1,l}(x), f_{i,j,2,l}(x), \dots, f_{i,j,3,l}(x)), \quad (1)$$

$$i = 1, 2, \dots, H, j = 1, 2, \dots, w, l = 1, 2, \dots, L.$$

As a result, the feature maps group $Z = (z_1(x), z_2(x), \dots, z_L(x)) \in \mathbb{R}^{H \times W \times L}$ is obtained. Fig. 4 shows the visualization results of the feature map group Z . It can be seen that different neighbor images of the same remote sensing image have a certain degree of attention to the water-body. Some features pay more attention to the boundary between a water-body and non-water-body, and some features pay more attention to texture areas.

Next, the OTSU algorithm is used to binarize each feature in Z to obtain the result $O = (o_1(x), o_2(x), \dots, o_L(x)) \in \mathbb{R}^{H \times W \times L}$. The formula is as follows:

$$o_l = \text{Otsu}(z_l), l = 1, 2, \dots, L. \quad (2)$$

Finally, we vote for all binarized neighbor features of the neighbor feature group to obtain the aggregated result V . The specific formula is calculated as:

$$V_{i,j} = \begin{cases} 1, & \sum_{l=1}^L o_{i,j,l} \geq \frac{2}{L} \\ 0, & \sum_{l=1}^L o_{i,j,l} < \frac{2}{L} \end{cases}, \quad (3)$$

To summarize, the mathematical definition of the feature aggregation module is detailed as follows:

$$V = \text{Vote}(\text{Otsu}(\text{CMax}(F))), \quad (4)$$

where $F \in \mathbb{R}^{H \times W \times C \times L}$ represents the neighbor features group and $V \in \mathbb{R}^{H \times W}$ is the output. Next, the aggregated result V is input into the post-processing module. Fig. 5 shows the feature aggregation module and post-processing module. The specific operations include filling small holes in the closed area by using area filling and removing noise by using morphological operations. Then we apply a point label constraint to the processed results. The point label constraint is expressed as follows:

$$q = \begin{cases} 1, & p \subset Q \\ 0, & p \not\subset Q \end{cases}. \quad (5)$$

where $Q \in \mathbb{R}^{H \times W}$ represents any independent water-body area in V , q is any pixel in Q , and p represents the point labels. If the area in the result contains point labels, the entire area is retained, otherwise it is not retained; The generated results are used as pseudo-labels and input into recursive training as supervision information.

C. Recursive training

There exist several recursive training approaches in the field of deep learning [52], [53]. The work in [54] uses a semantic segmentation and target detection branch for cyclic guidance to improve the quality of pseudo-labels. In [55], data noise and model noise are used for recursive training to improve the generalization ability of the model. Unlike the above methods, we embed the neighbor sampler into the recursive training so that the network can learn neighbor features (the flowchart is

shown in Fig. 2a). The specific steps of our method are as follows:

- Input the images into the neighbor sampler to generate neighbor image groups, and input these neighbor images with point labels into the model for training.
- Input the neighbor images into the model again, and use the feature aggregation module to generate pseudo-labels as a replacement for point labels.
- Input pseudo-labels into the neighbor sampler to generate pseudo-labels groups, and input the neighbor images groups with pseudo-labels groups into the model for training.
- Input the neighbor images groups into the model again for prediction, and generate results groups (average the group results to obtain a new pseudo-label) as the new pseudo-labels to replace the previous set of labels.
- Go to step c.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets and Evaluation

To verify the effectiveness of the proposed method, we applied the proposed method to high-resolution visible spectrum images for water extraction. This water-body dataset comes from the Gaofen Challenge [56], which contains RGB pan-sharpened images with a resolution of 0.5 m and does not contain infrared bands or digital elevation models. All images are taken from Wuhan and Suzhou, China, mainly in rural areas supplemented by urban areas. The positive labels in the dataset include rivers, reservoirs, rice fields, ditches, ponds, and lakes, while all other non-water pixels are considered negative. The data-set was cropped into 1000 images with the size of 492×492 without any overlap. The original data-set only contained pixel-level labels, so we re-annotated the data-set. Specifically, the data-set was annotated by two professionals and checked by an expert in the field of remote sensing image processing. The rule was that each independent water-body was randomly labeled with a point label of size 5×5 , which simulated the labeling behavior of the labelers in a real scene. We utilized the pixel-level labels of the original data-set as ground truth. The data sets are shown in Fig. 6. In the experiment, the weak supervision models adopt the point labels as the initial supervision information, while the full supervision models use pixel-level labels. Because the remote sensing image segmentation/classification evaluation index of overall accuracy or kappa coefficient cannot effectively describe the real structure of image segmentation geometry, we choose the fgIoU (foreground IoU), bgIoU (background IoU), mIoU (mean IoU), fgDice (foreground Dice), bgDice (background Dice), and mDice (mean Dice) to comprehensively evaluate the results. For each model, we performed five independent runs to calculate these evaluation indicators and standard deviations.

B. Parameter Settings

In our experiment, we use multiple mainstream models as feature extraction networks and embed the neighbor sampler and feature aggregation module, including the FCN [57],

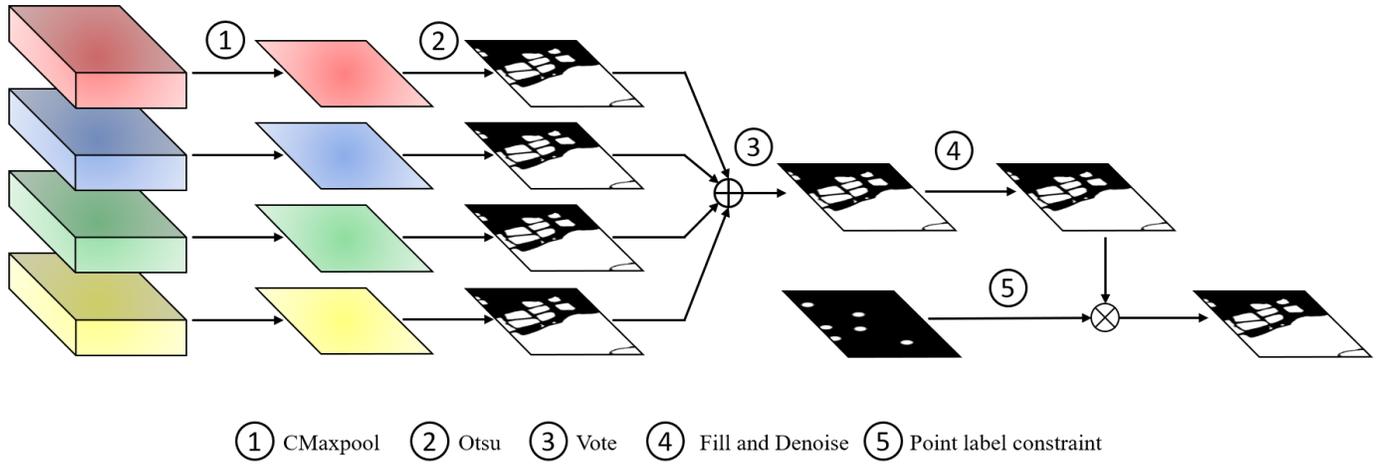


Fig. 5. Feature aggregation module and post-processing.

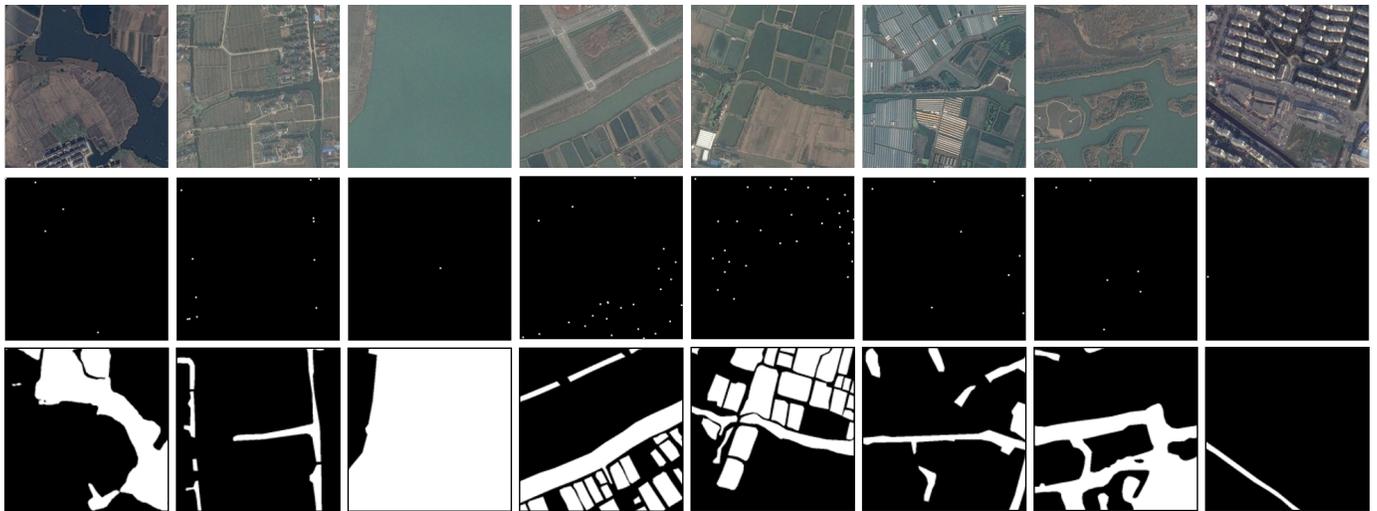


Fig. 6. Water-body extraction data-set. The first line represents the original images, the second line represents the point labels, and the third line represents ground truth.

TABLE I
COMPARISON OF WATER EXTRACTION RESULTS (%) AND FULL SUPERVISION APPROACHES

Method	Sup.	Validation						Test					
		bgIoU	fgIoU	mIoU	bgDice	fgDice	mDice	bgIoU	fgIoU	mIoU	bgDice	fgDice	mDice
FCN	full	91.2±0.4	66.9±1.0	79.1±0.5	95.4±0.2	80.2±0.7	87.8±0.4	91.0±0.6	66.4±0.5	78.7±0.4	95.3±0.3	79.8±0.3	87.6±0.2
FCN(Ours)	weak	89.6±0.2	61.5±0.6	75.5±0.4	94.5±0.1	76.1±0.5	85.3±0.3	89.3±0.4	60.7±1.0	75.0±0.7	94.4±0.2	75.5±0.7	85.0±0.5
UNet	full	92.7±0.2	73.3±0.4	83.0±0.3	96.2±0.1	84.6±0.3	90.4±0.2	92.6±0.1	73.0±0.3	82.8±0.2	96.2±0.1	84.4±0.2	90.3±0.1
UNet(Ours)	weak	91.4±0.4	67.8±1.3	79.6±0.8	95.5±0.2	80.8±0.9	88.2±0.5	91.1±0.3	67.0±0.7	79.1±0.5	95.3±0.2	80.3±0.5	87.8±0.3
Res-UNet	full	92.7±0.2	73.3±0.6	83.0±0.4	96.2±0.1	84.6±0.4	90.4±0.2	92.6±0.1	72.9±0.4	82.7±0.2	96.1±0.1	84.3±0.2	90.2±0.2
Res-UNet(Ours)	weak	91.2±0.3	67.4±0.7	79.3±0.5	95.4±0.2	80.5±0.5	88.0±0.3	91.0±0.2	67.0±0.4	78.9±0.4	95.3±0.1	80.1±0.4	87.7±0.2
NestedUNet	full	93.0±0.4	74.2±0.4	83.6±0.3	96.4±0.2	85.2±0.2	90.8±0.2	92.8±0.3	73.3±0.6	83.1±0.2	96.3±0.2	84.6±0.4	90.4±0.2
NestedUNet(Ours)	weak	91.7±0.2	69.0±0.6	80.4±0.4	95.7±0.1	81.7±0.4	88.7±0.3	91.6±0.2	68.8±0.6	80.2±0.4	95.6±0.1	81.5±0.4	88.6±0.3
D-LinkNet	full	92.3±0.2	72.3±0.7	82.3±0.5	96.0±0.1	83.9±0.5	90.0±0.3	92.0±0.2	71.1±0.3	81.6±0.2	95.9±0.1	83.1±0.2	89.5±0.2
D-LinkNet(Ours)	weak	91.1±0.2	67.1±0.4	79.1±0.3	95.3±0.1	80.3±0.3	87.8±0.2	90.8±0.2	66.1±0.6	78.5±0.4	95.2±0.1	79.6±0.4	87.4±0.3
DeepLab V3+	full	93.3±0.2	75.3±0.4	84.3±0.3	96.5±0.1	85.9±0.2	91.2±0.2	92.9±0.3	74.3±0.7	83.6±0.5	96.3±0.1	85.3±0.4	90.8±0.3
DeepLab V3+(Ours)	weak	91.5±0.1	68.5±0.3	80.0±0.2	95.6±0.1	81.3±0.2	88.4±0.1	91.2±0.2	67.6±0.6	79.4±0.4	95.4±0.1	80.6±0.5	88.0±0.3
MFDeepLab V3+	full	93.4±0.2	75.6±0.4	84.5±0.3	96.6±0.1	86.1±0.3	91.3±0.2	93.0±0.2	74.5±0.5	83.7±0.3	96.4±0.1	85.4±0.3	90.9±0.2
MFDeepLab V3+(Ours)	weak	91.6±0.1	68.7±0.3	80.1±0.2	95.6±0.0	81.4±0.2	88.5±0.1	91.2±0.2	67.8±0.6	79.5±0.4	95.4±0.1	80.8±0.4	88.1±0.3
SANet	full	93.0±0.1	74.2±0.2	83.6±0.2	96.4±0.1	85.2±0.1	90.8±0.1	92.7±0.1	73.3±0.2	83.0±0.2	96.2±0.1	84.6±0.2	90.4±0.1
SANet(Ours)	weak	91.4±0.2	68.1±0.3	79.8±0.3	95.5±0.1	81.0±0.2	88.3±0.2	91.1±0.2	67.2±0.6	79.2±0.4	95.3±0.1	80.4±0.4	87.9±0.3
SNS-CNN	full	93.2±0.2	74.6±0.2	83.9±0.2	96.3±0.5	85.4±0.1	91.0±0.1	92.9±0.2	73.8±0.4	83.3±0.2	96.3±0.1	84.9±0.3	90.6±0.1
SNS-CNN(Ours)	weak	91.6±0.1	68.6±0.4	80.1±0.3	95.6±0.1	81.4±0.3	88.5±0.2	91.4±0.3	68.2±0.7	79.7±0.5	95.5±0.2	81.1±0.5	88.3±0.3
DeepUNet	full	92.9±0.1	73.6±0.3	83.2±0.2	96.3±0.1	84.8±0.2	90.5±0.1	92.6±0.1	72.9±0.3	82.8±0.2	96.2±0.1	84.3±0.2	90.2±0.1
DeepUNet(Ours)	weak	91.3±0.2	68.0±0.4	79.7±0.3	95.4±0.1	81.0±0.3	88.2±0.2	91.1±0.1	67.6±0.4	79.3±0.3	95.3±0.1	80.7±0.3	88.0±0.2
MRRResNet	full	94.0±0.1	76.3±0.5	85.2±0.3	96.9±0.1	86.6±0.3	91.7±0.2	93.3±0.3	74.8±0.4	84.0±0.4	96.5±0.2	85.6±0.3	91.1±0.2
MRRResNet(Ours)	weak	91.6±0.1	69.0±0.3	80.3±0.2	95.6±0.1	81.6±0.2	88.6±0.1	91.5±0.2	68.3±0.5	79.9±0.3	95.5±0.1	81.1±0.3	88.3±0.2

UNet [50], ResUNet [43], NestedUNet [58], D-LinkNet [59], DeepLab V3+ [60], MFDeepLab V3+ [40], SANet [42], SNS-CNN [44], DeepUNet [45], and MRResNet [61]. The specific settings are as follows. In the FCN model, we choose to use FCN-8s. Specifically, the last three layers of feature maps are gradually up-sampled and merged, and finally the output was obtained. The UNet model uses a four-layers encoder-decoder structure. The encoder consists of two consecutive 3×3 Conv-BN-ReLU layers and a maxpooling layer. The decoder consists of two 3×3 Conv-BN-ReLU layers and an upsampling layer. We connect two consecutive 3×3 Conv-BN-ReLU modules into a residual structure for ResUNet; other structures are the same as UNet. The depth of NestedUNet is set to four layers, and dense connections are added to each layer. DLinkNet uses pre-trained ResNet as the backbone. We set the size of the kernel for different dilated convolutions in Dblock to 3×3 , and the dilated rate is 6, 12, 18, and 24. For the ASPP in Deeplab V3+, we set up a point convolution, a global average pooling layer, and three dilated convolutions, with the dilated rate set to 6, 12, and 18. In MFDeepLab V3+, the weighting coefficients of three different scale features are set to 0.4, 0.3 and 0.3. In SANet, we use an adaptive multi-scale feature learning module (AML) to replace the convolution module of UNet. The SNA-CNN model uses a four-layer encoder-decoder structure. DeepUNet uses a six-layer encoder-decoder structure; for fair comparison, superpixel-based segmentation and CRF are not used. In MRResNet, the dilated rates of the Multiscale Dilated Convolution (MSDC) Module are set to 1, 2, 4, 8 and 16. The Multikernel Max Pooling (MKMP) Module contains contextual information for four different sizes of receptive fields: 2×2 , 3×3 , 5×5 , and 6×6 . When the value of K is 2, an additional decoder is added to make the output size consistent with the label. In other cases, the output is upsampled to the same size as the label.

In the proposed NFANet, only random horizontal flips, random up and down flips, and random rotations of 90 degrees are used for data augmentation. We set the batch size to 4. In recursive training, cross entropy plus dice loss is used as the loss function. We set the learning rate to 0.0001 and use the early stop strategy to prevent over fitting. The specific early stopping strategy is to halve the learning rate if the training loss does not decrease every 3 epochs, and terminate the training if the training loss does not decrease for 6 epochs; and the total epoch upper limit is set to 100 times. The optimizer adopts the Adam optimizer with weight decay, and the weight decay is set to 0.001.

C. Comparison with Fully-Supervised Approaches

We compared our method with several other fully supervised semantic segmentation approaches. In general, datasets are randomly divided into three sets, where 60% of the images are used for training, 20% of the images are used for validation, and 20% of the images are used for testing. The following 11 methods are considered for comparison:

- 1) Fine-tuned FCN. The feature extractor is a pre-trained VGG-Net, and we set the output dimension of the last layer to the dimension of the category.
- 2) UNet. We use a four-layer encoder-decoder structure, and the other settings are basically the same as the weakly supervised parameter settings. UNet usually performs well in two-category semantic segmentation.
- 3) Res-UNet. Based on an improvement of the UNet structure, we composed the convolutional layer in the encoder or decoder into residual connections.
- 4) NestedUNet. Based on an improvement of the UNet structure, it contains more long and short connections, and it has been proved that the feature extraction ability is more powerful.
- 5) D-LinkNet. This is the model that won first place in the 2018 CVPR road extraction competition. It was good at two-category semantic segmentation; the Dblock in the middle part continuously uses multiple dilated convolutions for dense connection.
- 6) DeepLab V3+. The ASPP module extracts deep semantic information from different receptive fields, and uses global average pooling to obtain global features and more semantic details.
- 7) MFDeepLab V3+. The decoder uses multi-scale features weighted fusion to further refine the details of the water area, and the weight coefficients of different features need to be set as hyperparameters.
- 8) SANet. The model combines multi-feature extraction and adaptive feature fusion to form an AML.
- 9) SNS-CNN. This model embeds the dilated convolution into the up-sampling operation to keep the decoding details.
- 10) DeepUNet. The model uses a deep UNet network for water extraction, and uses superpixel-based segmentation and CRF to further refine the results.
- 11) MRResNet. The model uses MSDC Module and MKMP Module, which helps preserve the boundaries of the water-bodies.

In Table I, we report the water extraction performance of the proposed approach and compare it with the fully-supervised approaches. The best scores are shown in bold in the table. Fig. 7 also illustrates the visual performance of all approaches. Experiments demonstrate that our method achieves the best score together with the NestedUNet-based model. The visual performance shows that the prediction results obtained by our method are very close to the ground truth. The bgIoU of our method reaches 91.6%, the fgIoU reaches 68.8%, the mIoU reaches 80.2%, the bgDice reaches 95.6%, the fgDice reaches 81.5%, and the mDice reaches 88.6%. Compared with the best fully-supervised model MRResNet, the mIoU of our method is only reduced by 3.8 percentage points, and mDice is only reduced by 2.5 percentage points. However, the labeling cost of our method is much less than that of the fully-supervised method. From Fig. 7, we can observe the following results. The edges of the weakly-supervised prediction results are not very smooth compared to the fully-supervised prediction results, and there are still some defects in some local small water-bodies. We observe that these incomplete small water-bodies are more difficult to identify, and are usually mistaken for grass or bare soil. This is likely to be the reason for the

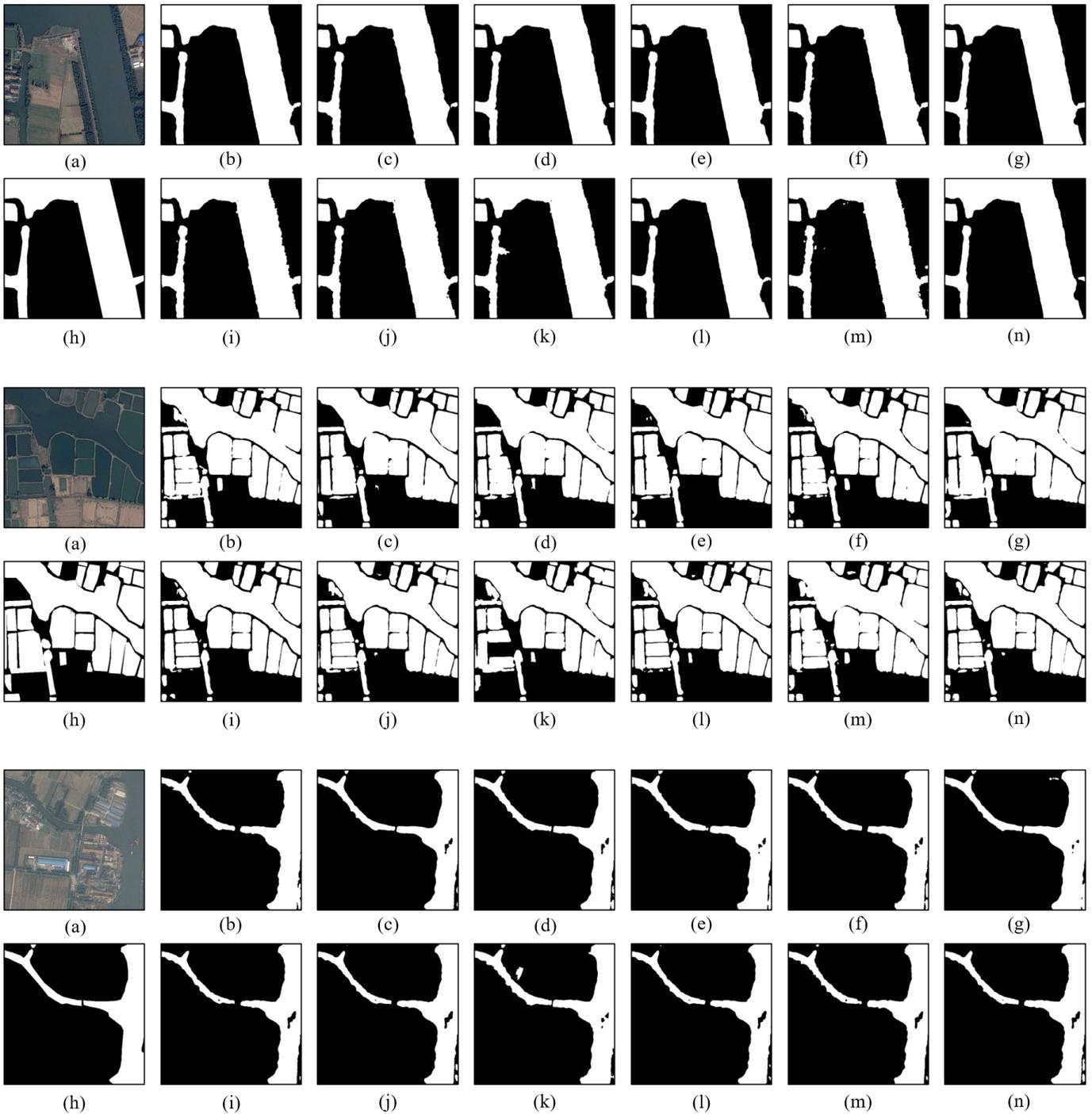


Fig. 7. Prediction results of the fully-supervised methods and our methods. The visualization results of some new models are shown below. (a) and (h) represent the original image and ground truth. (b)-(g) represent the prediction results of NestedUNet, MFDeepLab V3+, SANet, SNS-CNN, DeepUNet and MRResNet, respectively. (i)-(n) represent the prediction results of our method based on NestedUNet, MFDeepLab V3+, SANet, SNS-CNN, DeepUNet and MRResNet, respectively.

TABLE II
COMPARISON OF WATER EXTRACTION RESULTS (%) AND WEAK SUPERVISION

Method	Validation						Test					
	bgIoU	fgIoU	mIoU	bgDice	fgDice	mDice	bgIoU	fgIoU	mIoU	bgDice	fgDice	mDice
LocMap	87.6±0.4	54.5±1.0	71.0±0.7	93.4±0.2	70.6±0.8	82.0±0.5	87.1±0.4	53.1±0.8	70.1±0.6	93.1±0.2	69.4±0.7	81.2±0.5
SPMF-Net	85.4±0.4	48.5±0.6	66.9±0.5	92.1±0.2	65.3±0.6	78.7±0.4	85.0±0.3	47.7±0.5	66.4±0.4	91.9±0.2	64.6±0.5	78.2±0.3
U-CAM	86.8±0.4	52.1±1.0	69.5±0.7	92.9±0.2	68.5±0.9	80.7±0.6	86.3±0.4	51.0±0.8	68.7±0.6	92.7±0.2	67.5±0.7	80.1±0.5
Baseline	41.1±5.9	14.2±5.8	27.7±5.5	58.1±5.8	24.5±8.8	41.3±6.9	41.4±5.9	14.1±5.8	27.7±5.6	58.3±5.8	24.4±8.8	41.4±7.0
Ours	91.4±0.4	67.8±1.3	79.6±0.8	95.5±0.2	80.8±0.9	88.2±0.5	91.1±0.3	67.0±0.7	79.1±0.5	95.3±0.2	80.3±0.5	87.8±0.3

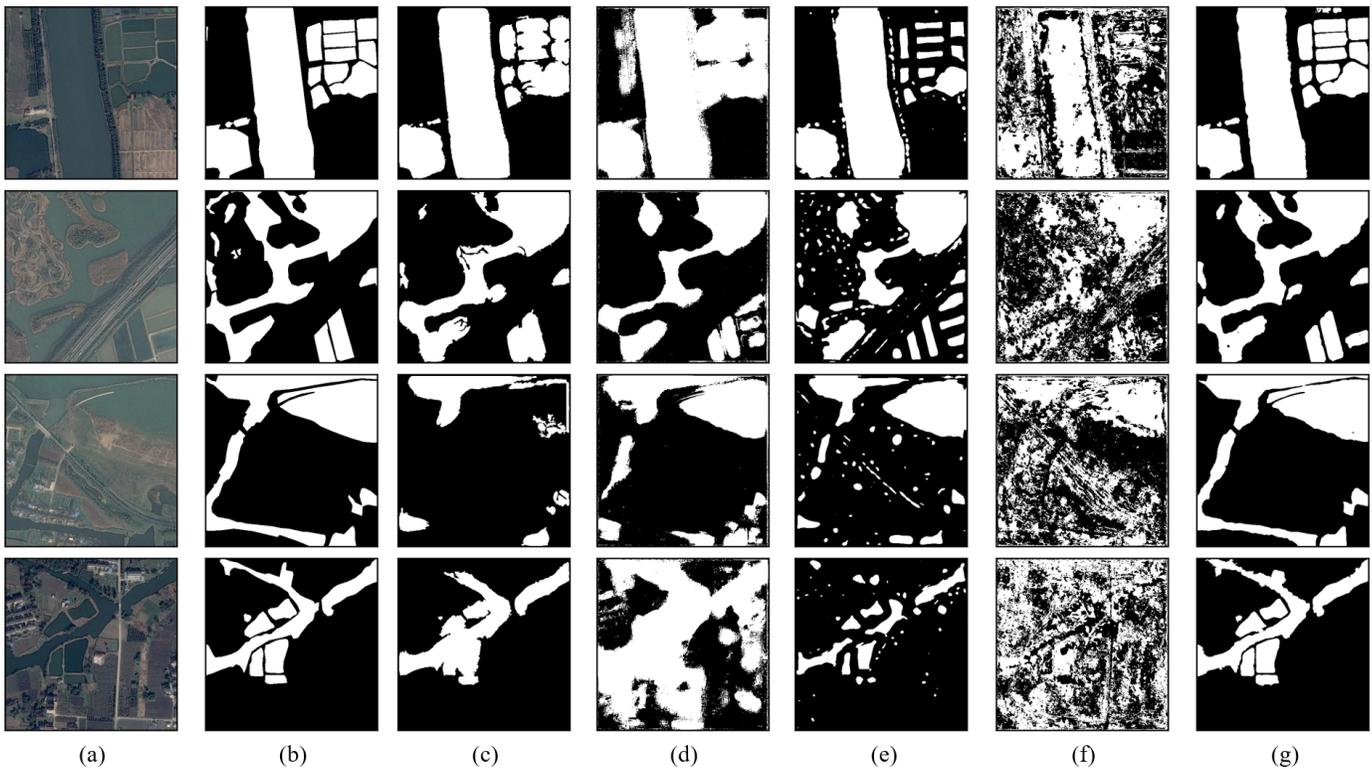


Fig. 8. Prediction results of the other weakly supervised methods and our method. (a) and (b) represent the original image and ground truth. (c) represents Loc-Maps method. (d) represents SPMF-Net. (e) represents U-CAM. (f) represents baseline. (g) represents our method.

gap between our results and the results of the fully-supervised method. For the impact of the amount of training data, we conduct a more detailed ablation experiment (see Table IV and Fig. 12).

D. Comparison with Weakly Supervised Approaches

We compared our method with several other weakly supervised remote sensing approaches. Data sets are randomly divided into three sets, where 60% of the images are used for training, 20% of the images are used for validation, and 20% of the images are used for testing. The other weak supervision methods used for comparison are as follows.

- 1) In [46], the author uses image-level labels supervision to generate localization maps, combined with traditional algorithms based on graph algorithms to obtain pseudo-labels, and then uses pseudo-label supervision to train the network. For fair comparison, we use point labels for supervision, and then CMax pooling to generate localization maps.
- 2) In [47], the model uses image-level labels for supervision. We remove the global average pooling layer and the fully connected layer, and used point labels for supervision.
- 3) In [48], the author combines UNet and CAM to form a U-CAM. Specifically, U-CAM is obtained by the weightes sum of the filter dimensions output by the last convolutional layer, where the weight comes from the fully connected layer. Then threshold processing

is performed on U-CAM to obtain hard segmentation prediction.

- 4) We train a UNet that only uses point labels for supervision. The loss function only calculates positive labels, and negative labels do not participate in the loss calculation. We use it as the baseline of point supervision.

The experimental results of the comparison of our method with other weakly supervised remote sensing approaches are shown in Table II. Fig. 8 shows the prediction results of other weakly supervised methods in comparison to the proposed method. To be fair, all methods are based on UNet. The bgIoU of our method reaches 91.1%, the fgIoU reaches 67.0%, the mIoU reaches 79.1%, the bgDice reaches 95.3%, the fgDice reaches 80.3%, and the mDice reaches 87.8%. Compared with the best weakly supervised method, LocMap, the mIoU of our method improves 9.0 percentage points, and mDice improves 6.6 percentage points. Although other weak supervision methods can predict the local area of the water-body, there are errors in the detection of the water-body boundary, while our method is relatively more accurate. The compared weak supervision methods cannot detect small objects appropriately, while this issue is largely solved to by the proposed method. If only point labels are used as positive labels to train the network, it is difficult for the network to predict water-bodies and non-water-bodies.

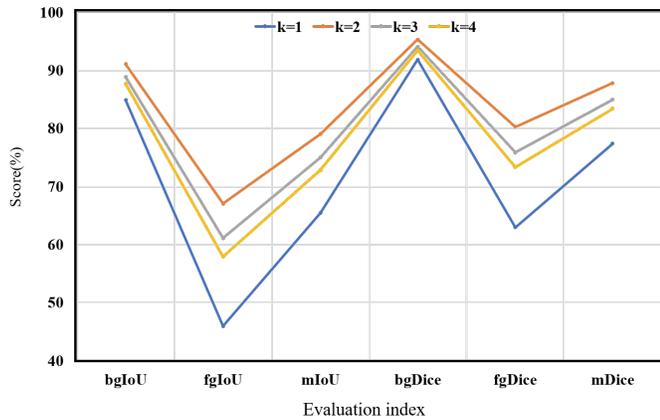


Fig. 9. Effectiveness of neighbor sampling.

E. Ablation Experiments

The method proposed in this paper mainly includes two modules and one strategy, i.e., the neighbor sampler module, the feature aggregation module, and the recursive training strategy. In order to prove the effectiveness of each module or strategy, ablation experiments were performed on the dataset.

1) *Effectiveness of neighbor sampling*: In the ablation experiment, the other settings are unchanged, and only the value of K is changed. We set the neighbor sampling parameter K of our proposed network from 1 to 4 and only use cross entropy and dice loss to train the model. For different K values of NFANet, in order to avoid interference from other modules, we only select UNet as the feature extraction network for comparative experiments. In particular, when the value of K is set to 1, the neighbor image group degenerates into the input image. As shown in Fig. 9, with the gradual increase of K , mIoU first increases and then decreases. As the neighbor sampling parameter K increases, the number of adjacent pixels that need to be considered increase geometrically, resulting in information redundancy, and the size of each reconstructed neighbor image is gradually reduced. Therefore, we set K equal to 2, because the neighbor features require less computation and achieves better performance.

2) *Effectiveness of feature aggregation*: K is set to 2, and the pseudo-label is obtained separately for each image in the neighbor image group (Eqs. (1) and (2)). As shown in Table III, when the features of the i -th neighbor image are f_i , $[f_1 : f_2 : f_3 : f_4]$, then feature aggregation is used. “RT” means that a recursive training strategy is used. Compared with the result of the best single neighbor image, the neighbor feature aggregation module improves mIoU and mDice by 9.29 percentage points and 7.02 percentage points, respectively. Compared with the worst single neighbor image result, the neighbor feature aggregation module improves mIoU and mDice by 14.83 percentage points and 11.90 percentage points, respectively. Due to the different contributions of neighbor pixels to prediction results, the scores of the pseudo-labels generated by a single neighbor image are different. The visualization results after using CMax pooling are shown

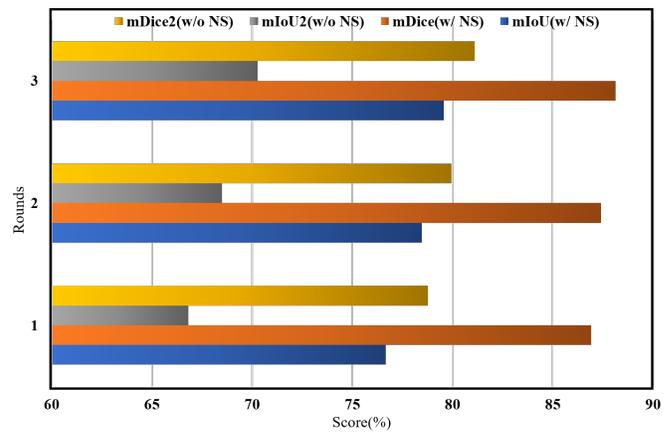


Fig. 10. Effectiveness of recursive training.

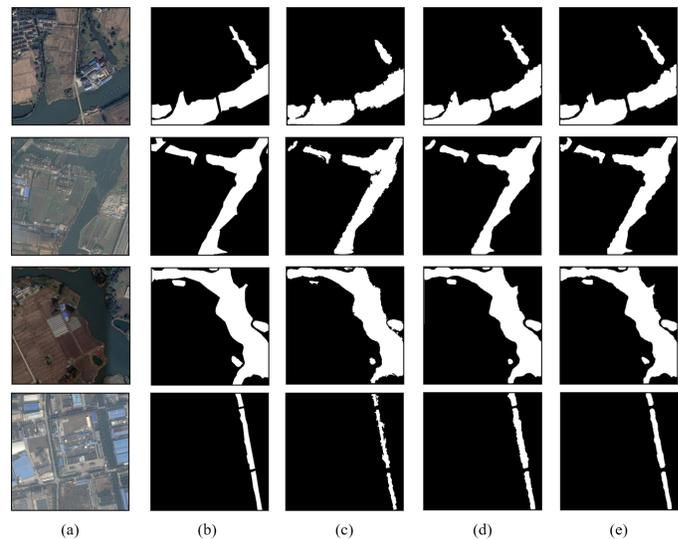


Fig. 11. Predicted results of recursive training. (a) and (b) represent the original image and ground truth. (c)-(e) represents the results of the first, second and third rounds of recursive training respectively.

in Fig. 4. The feature aggregation module uses the features extracted from neighbor pixels to make joint judgments, and the information of different neighbor pixels complement each other, so it generates more high-quality pseudo-labels and ultimately improves the score of the water-bodies prediction result.

3) *Effectiveness of recursive training*: We set recursion times to 3 and compare the evaluation indicators of each recursive training. As shown in Fig. 10, “w/o NS” means neighbor sampling was not used. As can be seen from the figure, whether using neighbor sampling or not, mIoU and mDice are improved. Compared with the preliminary results, mIoU and mDice increased by 2.88 percentage points and 1.22 percentage points, respectively, after using recursive training. Fig. 11 shows the visualization results of three rounds of recursion. It can be seen from the figure that there are many burrs on the water boundary in the first round, and there are still some errors in the local small water-bodies. After recursive training, the ability of the network to capture

TABLE III
EFFECTIVENESS OF FEATURE AGGREGATION(%)

	NF ¹	RT ²	bgIoU	fgIoU	mIoU	bgDice	fgDice	mDice
f_1			84.08	39.78	61.93	91.35	56.92	74.13
f_1		✓	87.45	42.01	64.73	93.30	59.17	76.23
f_2			84.31	49.33	66.82	91.49	66.07	78.78
f_2		✓	88.98	51.57	70.27	94.17	68.04	81.11
f_3			83.55	48.37	65.96	91.04	65.21	78.12
f_3		✓	87.52	51.48	69.50	93.35	67.97	80.66
f_4			82.21	45.75	63.98	90.24	62.78	76.51
f_4		✓	87.27	48.09	67.68	93.20	64.95	79.07
$[f_1 : f_2 : f_3 : f_4]$			89.63	63.73	76.68	94.53	77.85	86.19
$[f_1 : f_2 : f_3 : f_4]$		✓	91.40	67.73	79.56	95.51	80.76	88.13

¹ “NF” Means Neighbor Features.

² “RT” Means Using Recursive Training.

TABLE IV
SEGMENTATION RESULTS(%) USING DIFFERENT PROPORTIONS OF THE TRAINING SET

Method	Train=100		Train=200		Train=300		Train=400		Train=500		Train=600	
	mIoU	mDice	mIoU	mDice								
UNet	80.5	88.7	81.0	89.1	82.2	89.9	82.6	90.1	82.7	90.2	83.1	90.5
Ours	66.9	78.9	69.1	80.8	72.8	83.4	74.8	84.7	76.6	86.1	79.5	88.1

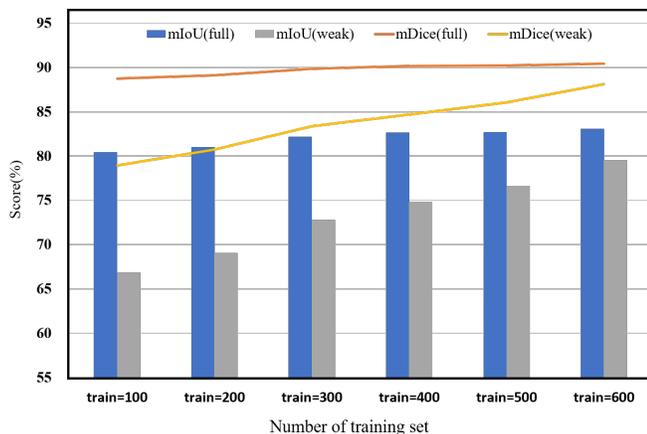


Fig. 12. Impact of data split.

the water contour is improved, and the water boundary is gradually smoothed. The boundaries between water-bodies are also clearer, even the small boundaries that are mislabeled in ground truth are correctly distinguished.

4) *Impact of data split*: This section explores the impact of the data split. We compare the fully-supervised method with our method (Table IV). When the amount of data in the training set is increased from 100 to 600, the fully-supervised mIoU score increases by 2.6 percentage points, and the mIoU score of our method increases by 12.6 percentage points. The following conclusions can be drawn from the Fig. 12. As the amount of data gradually increases, both the fully-supervised method and our method are improved. When the amount of data is small, the point-level label seriously lacks the supervision information that it can provide, the neighbor features are not enough to distinguish water-bodies from non-water-

bodies, and the generated pseudo-labels have much wrong information. The supervised information provided by pixel-level labels is more comprehensive, which enables the neural network to learn some key features. When the amount of data increases, neighbor features play an increasingly important role, which enables the neural network to better distinguish water-bodies from remote sensing images, and the quality of pseudo-labels is significantly improved.

F. Time Consumption and Model Parameters

The hardware configurations for the experiments in this paper consisted of Intel Core i7-9700k 3.60 GHz CPU, GeForce RTX 2080Ti GPU, and 16GB RAM. The results of the GPU inference time, shown in Table V, are the average GPU inference time of the data-set. After using recursive training to improve the quality of pseudo-labels, we input the pseudo-labels and original images into the models that are consistent with the fully-supervised methods for training and inference. Therefore, the GPU inference time of the proposed method is the same as that of the fully-supervised methods. It can be observed from the table that the inference time of D-LinkNet is the shortest. This is because D-LinkNet compresses the feature channel in the decoder to reduce the computational cost. MFDeepLab V3+ takes the longest time, due to its use of multi-scale feature weighted fusion, which requires multiple DeepLab V3+ models for prediction.

The above methods are divided into two categories: UNet-based models and ResNet-based models, and so the model parameters are different. In the UNet-based models, the parameter of SANet is the smallest at 13 MB, and the parameter of SCS-CNN is the greatest at 62.9 MB. In the ResNet-based model, the parameter of FCN is the smallest at 74.6 MB, and

TABLE V
GPU INFERENCE TIME AND MODEL PARAMETERS

Method	FCN	UNet	Res-UNet	NestedUNet	D-LinkNet	DeepLab V3+	MFDeepLab V3+	SANet	SNS-CNN	DeepUNet	MRResNet
Times(MS)	19.0	16.2	18.9	35.7	12.9	26.5	75.7	34.2	25.0	15.5	28.4
Param.(MB)	74.6	31.5	33.2	36.7	124.5	161.7	485.1	13	62.9	39.2	278.7

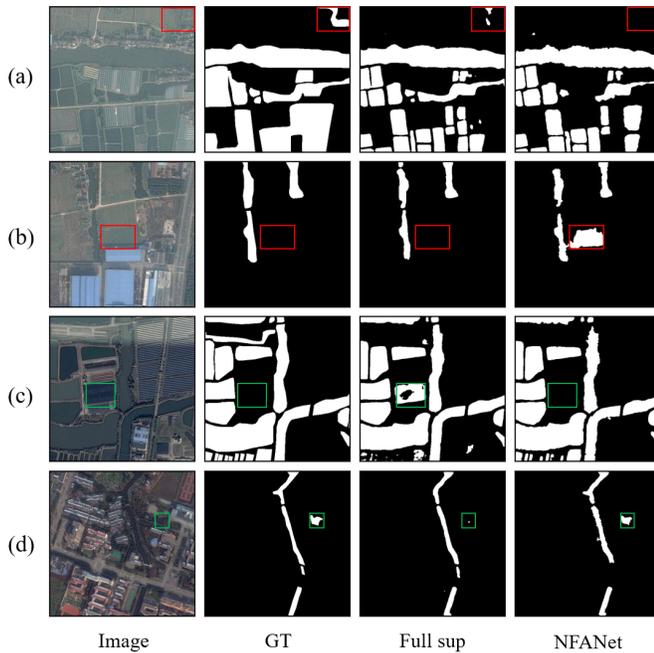


Fig. 13. Impact of data split.

the parameter of MFDeepLab V3+ is the greatest at 485.1 MB.

G. Hard case

Although our method yields good results, there are still some errors. As shown in Figure 13, the entire water body area in the red box in Fig. 13a has missed predictions. We think that the border between the cultivated land on its left and the water body is fuzzy and difficult to distinguish. The cultivated land in the red box in Fig. 13b is wrongly predicted to become a water body. It is worth noting that in some images that are difficult to distinguish water bodies, the proposed method is better than full supervision. As shown in the green box of Fig. 13c, the fully supervised model incorrectly predicted the greenhouse as a water body. A small piece of water in Fig. 13d is not predicted, and our method can find it well.

V. CONCLUSION

In this paper, we proposed a network entitled NFANet. Unlike traditional convolutional neural networks that only use global or local features for discrimination, NFANet uses neighbor features, which allows more representative features to be learned. We propose a method called neighbor sampling to obtain neighboring images. By using the feature aggregation module, we fuse these neighbor features to obtain pseudo-labels, and improve the label quality by recursive training. We tested it on water data sets and compared it with advanced

fully supervised and weakly supervised methods. By using only point labels, the proposed method obtains comparable results with that of full supervision.

As a possible future work, we will conduct research on weakly supervised or semi-supervised methods of self-correction. Moreover, we will also consider the characteristics of other different ground objects and design a dedicated module for feature extraction to further improve the ground object extraction performance of CNN.

REFERENCES

- [1] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2063–2079, 2018.
- [2] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, 2021.
- [3] X. Lu, Y. Zhong, Z. Zheng, Y. Liu, J. Zhao, A. Ma, and J. Yang, "Multi-scale and multi-task deep learning framework for automatic road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9362–9377, 2019.
- [4] R. Tombe and S. Viriri, "Adaptive deep co-occurrence feature learning based on classifier-fusion for remote sensing scene classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 155–164, 2021.
- [5] H. Guo, J. Liu, J. Yang, Z. Xiao, and Z. Wu, "Deep collaborative attention network for hyperspectral image classification by combining 2-d cnn and 3-d cnn," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4789–4802, 2020.
- [6] T. Shao, Y. Yang, Y. Weng, Q. Hou, and K. Zhou, "H-cnn: Spatial hashing based cnn for 3d shape analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 7, pp. 2403–2416, 2020.
- [7] H. Lee, S. Eum, and H. Kwon, "Me r-cnn: Multi-expert r-cnn for object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 1030–1044, 2020.
- [8] P. Tang, Q. Liang, X. Yan, S. Xiang, and D. Zhang, "Gp-cnn-dtel: Global-part cnn model with data-transformed ensemble learning for skin lesion classification," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2870–2882, 2020.
- [9] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2019.
- [10] F. Fang, L. Li, H. Zhu, and J.-H. Lim, "Combining faster r-cnn and model-driven clustering for elongated object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 2052–2065, 2020.
- [11] Y. Zhao, Z. Yuan, and B. Chen, "Training cascade compact cnn with region-iou for accurate pedestrian detection," *IEEE trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3777–3787, 2020.
- [12] J. C. Nascimento and G. Carneiro, "One shot segmentation: Unifying rigid detection and non-rigid segmentation using elastic regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 3054–3070, 2020.
- [13] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "Connnet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2518–2529, 2019.
- [14] L. Ru, B. Du, and C. Wu, "Multi-temporal scene classification and scene change detection with correlation based fusion," *Trans. Image Process.*, vol. 30, pp. 1382–1394, 2021.
- [15] D. Guo, Y. Xia, and X. Luo, "Self-supervised gans with similarity loss for remote sensing image scene classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2508–2521, 2021.

- [16] C. Li, Y. Zhuang, W. Liu, S. Dong, H. Du, H. Chen, and B. Zhao, "Effective multiscale residual network with high-order feature representation for optical remote sensing scene classification," *IEEE Geosci. Remote Sens.*, pp. 1–5, 2021.
- [17] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, 2021.
- [18] H. Gao, Z. Chen, and C. Li, "Sandwich convolutional neural network for hyperspectral image classification using spectral feature enhancement," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 3006–3015, 2021.
- [19] X. Luo, X. Li, Y. Wu, W. Hou, M. Wang, Y. Jin, and W. Xu, "Research on change detection method of high-resolution remote sensing images based on subpixel convolution," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 1447–1457, 2021.
- [20] Y. Gao, F. Gao, J. Dong, and H.-C. Li, "Sar image change detection based on multiscale capsule network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 484–488, 2021.
- [21] Y. Zhang, Q. Zhu, Y. Zhong, Q. Guan, L. Zhang, and D. Li, "A modified d-linknet with transfer learning for road extraction from high-resolution remote sensing," in *IGARSS 2020 – 2020 IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1817–1820.
- [22] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using cnn-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, 2020.
- [23] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, 2021.
- [24] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, 2021.
- [25] J. Zhang, M. Xing, G.-C. Sun, J. Chen, M. Li, Y. Hu, and Z. Bao, "Water body detection in high-resolution sar images with cascaded fully-convolutional network and variable focal loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 316–332, 2021.
- [26] M. Li, P. Wu, B. Wang, H. Park, H. Yang, and Y. Wu, "A deep learning method of water body extraction from high resolution remote sensing images with multisensors," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 3120–3132, 2021.
- [27] Y. He, S. Yao, W. Yang, H. Yan, L. Zhang, Z. Wen, Y. Zhang, and T. Liu, "An extraction method for glacial lakes based on landsat-8 imagery using an improved u-net network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 6544–6558, 2021.
- [28] 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. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 769–782, 2020.
- [29] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, "Automatic water-body segmentation from high-resolution satellite images via deep networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 602–606, 2018.
- [30] X. Wang, M. Liu, D. S. Raychaudhuri, S. Paul, Y. Wang, and A. K. Roy-Chowdhury, "Learning person re-identification models from videos with weak supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 3017–3028, 2021.
- [31] Z. Jia, X. Huang, E. I.-C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE Trans. Med. Imaging.*, vol. 36, no. 11, pp. 2376–2388, 2017.
- [32] L. Yang, J. Han, D. Zhang, N. Liu, and D. Zhang, "Segmentation in weakly labeled videos via a semantic ranking and optical warping network," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4025–4037, 2018.
- [33] A. Huang and T. Zhao, "Weak supervision learning for object co-segmentation," *IEEE Trans. Big Data.*, pp. 1–1, 2020.
- [34] H. Zhang, Y. Zeng, H. Lu, L. Zhang, J. Li, and J. Qi, "Learning to detect salient object with multi-source weak supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2021.
- [35] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Proposal-based instance segmentation with point supervision," in *2020 Proc. IEEE Int. Conf. Image Processing.*, 2020, pp. 2126–2130.
- [36] B. cai Gao, "Ndw—a normalized difference water index for remote sensing of vegetation liquid water from space," in *Imaging Spectrometry*, 1995, pp. 257–266.
- [37] R. Trias-Sanz, "Texture orientation and period estimator for discriminating between forests, orchards, vineyards, and tilled fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2755–2760, 2006.
- [38] I. J. Barton and J. M. Bathols, "Monitoring floods with avhrr," *Remote Sens. Environ.*, vol. 30, no. 1, pp. 89–94, 1989.
- [39] Y. Zhang and M. Chi, "Mask-r-fcn: A deep fusion network for semantic segmentation," *IEEE Access*, vol. 8, pp. 155 753–155 765, 2020.
- [40] Z. Li, R. Wang, W. Zhang, F. Hu, and L. Meng, "Multiscale features supported deeplabv3+ optimization scheme for accurate water semantic segmentation," *IEEE Access*, vol. 7, pp. 155 787–155 804, 2019.
- [41] R. Wang, Y. Meng, W. Zhang, Z. Li, F. Hu, and L. Meng, "Remote sensing semantic segregation for water information extraction: Optimization of samples via training error performance," *IEEE Access*, vol. 7, pp. 13 383–13 395, 2019.
- [42] B. Cui, W. Jing, L. Huang, Z. Li, and Y. Lu, "Sanet: A sea-land segmentation network via adaptive multiscale feature learning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 116–126, 2021.
- [43] Z. Chu, T. Tian, R. Feng, and L. Wang, "Sea-land segmentation with res-unet and fully connected crf," in *IGARSS 2019 - 2019 IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3840–3843.
- [44] S. Dong, L. Pang, Y. Zhuang, W. Liu, Z. Yang, and T. Long, "Optical remote sensing water-land segmentation representation based on proposed sns-cnn network," in *IGARSS 2019 - 2019 IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3895–3898.
- [45] 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, 2019.
- [46] K. Fu, W. Lu, W. Diao, M. Yan, H. Sun, Y. Zhang, and X. Sun, "Wsf-net: Weakly supervised feature-fusion network for binary segmentation in remote sensing image," *Remote Sens.*, vol. 10, no. 12, 2018.
- [47] J. Chen, F. He, Y. Zhang, G. Sun, and M. Deng, "Spmf-net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion," *Remote Sens.*, vol. 12, no. 6, p. 1049, 2020.
- [48] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sens.*, vol. 12, no. 2, 2020.
- [49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.
- [50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, vol. 9351, 2015.
- [51] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *CVPR*, 2017, pp. 1665–1674.
- [52] Q. Liang, Y. Nan, G. Coppola, K. Zou, W. Sun, D. Zhang, Y. Wang, and G. Yu, "Weakly supervised biomedical image segmentation by reiterative learning," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1205–1214, 2019.
- [53] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *CVPR*, June 2018.
- [54] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, "Cyclic guidance for weakly supervised joint detection and segmentation," in *CVPR*, June 2019.
- [55] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR*, June 2020.
- [56] X. Sun, A. Shi, H. Huang, and H. Mayer, "Bas4net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5398–5413, 2020.
- [57] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2015.
- [58] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *DLIA*, 2018.
- [59] L. Zhou, C. Zhang, and W. Ming, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *CVPRW*, 2018.
- [60] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *ECCV*, pp. 801–818, 2018.
- [61] 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.



Ming Lu was born in Hubei, China, in 1996. He received the B.S. degree from the Hubei University of Technology, Wuhan, China, in 2019. He is currently pursuing the masters degree with the College of Electrical and Information Engineering, Hunan University, Changsha. His research interests focus on weakly supervised learning and remote sensing images semantic segmentation.



Leyuan Fang received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015. From September 2011 to September 2012, he was a Visiting Ph.D. Student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council. From August 2016 to September 2017, he was a Post-Doctoral Researcher with the Department of Biomedical Engineering, Duke University, Durham, NC, USA. He is a Professor with the College of

Electrical and Information Engineering, Hunan University, and an Adjunct Researcher with the Peng Cheng Laboratory, Shenzhen, China. His research interests include sparse representation and multiresolution analysis in remote sensing and medical image processing. Dr. Fang was a recipient of one 2nd-Grade National Award at the Nature and Science Progress of China in 2019. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and Neurocomputing.



Muxing Li was born in Hunan, China, in 1999. He received a bachelor of advanced computing (Honours) degree from The Australian National University, Canberra, Australia. He is currently preparing to pursue a masters of philosophy degree with The University of Sydney, Australia. His research interests focus on computer vision.



Bob Zhang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2011. He was with the Center for Pattern Recognition and Machine Intelligence and later was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently an Associate Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His current research interests

include medical biometrics, pattern recognition, and image processing.



Yi Zhang received his B.S., M.S., and Ph.D. degrees from the College of Computer Science, Sichuan University, Chengdu, China, in 2005, 2008, and 2012, respectively. From 2014 to 2015, he was with the Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA, as a Postdoctoral Researcher. He is currently a Full Professor with the College of Computer Science, Sichuan University, and is the director of the deep imaging group (DIG). His research interests include medical imaging, compressive sensing, and deep learning. He

authored more than 70 papers in the field of image processing. These papers were published in several leading journals, including IEEE Transactions on Medical Imaging, IEEE Transactions on Computational Imaging, Medical Image Analysis, European Radiology, Optics Express, etc., and reported by the Institute of Physics (IOP) and during the Lindau Nobel Laureate Meeting. He received major funding from the National Key R&D Program of China, the National Natural Science Foundation of China, and the Science and Technology Support Project of Sichuan Province, China. He is a Guest Editor of the International Journal of Biomedical Imaging, Sensing and Imaging, and an Associate Editor of IEEE Access. He is a Senior Member of IEEE.



Pedram Ghamisi (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2015. He works as the Head of the Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany, and as the Visiting Professor with the Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria. He is also the CTO and a Co-Founder of VasoGnosis Inc., Milwaukee, WI, USA. His research interests include interdisciplinary research on machine

(deep) learning, image and signal processing, and multisensor data fusion. Dr. Ghamisi was a recipient of the Best Researcher Award for M.Sc. students at the K. N. Toosi University of Technology in the academic year of 2010–2011, the IEEE Mikio Takagi Prize for winning the Student Paper Competition at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2013, the Talented International Researcher by Iran's National Elites Foundation in 2016, the First Prize of the Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF) of the IEEE-GRSS in 2017, the Best Reviewer Prize of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL) in 2017, the IEEE Geoscience and Remote Sensing Society 2020 Highest Impact Paper Award, the Alexander von Humboldt Fellowship from the Technical University of Munich, and the High Potential Program Award from HZDR. He is also the Co-Chair of the IEEE Image Analysis and Data Fusion Committee. For detailed info, please see <http://pedram-ghamisi.com/>.