International Journal of Geo-Information

MDPI

Article

A Multi-Scale Water Extraction Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images

Hongxiang Guo [1,2], Guojin He [1,3,4,*,†], Wei Jiang [5,6,†], Ranyu Yin [1,2], Lei Yan [1,2] and Wanchun Leng [1,2]

1 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; guohx@radi.ac.cn (H.G.); yinry@radi.ac.cn (R.Y.); yanlei@aircas.ac.cn (L.Y.); lengwch@radi.ac.cn (W.L.)
2 College of Resource and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
3 Satellite Remote Sensing Technology Department, Key Laboratory of Earth Observation Hainan Province, Sanya 572029, Hainan, China
4 Satellite Remote Sensing Technology Department, Sanya Institute of Remote Sensing, Sanya 572029, Hainan, China
5 State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing,100038, China; jiangwei@iwhr.com
6 Remote Sensing Technology Application Center, Research Center of Flood and Drought Disaster Reduction of the Ministry of Water Resources, Beijing,100038, China
* Correspondence: hegj@radi.ac.cn
† These authors contributed equally to this work.

Received: 12 January 2020; Accepted: 22 March 2020; Published: 25 March 2020

Abstract: Automatic water body extraction method is important for monitoring floods, droughts, and water resources. In this study, a new semantic segmentation convolutional neural network named the multi-scale water extraction convolutional neural network (MWEN) is proposed to automatically extract water bodies from GaoFen-1 (GF-1) remote sensing images. Three convolutional neural networks for semantic segmentation (fully convolutional network (FCN), Unet, and Deeplab V3+) are employed to compare with the water bodies extraction performance of MWEN. Visual comparison and five evaluation metrics are used to evaluate the performance of these convolutional neural networks (CNNs). The results show the following. (1) The results of water body extraction in multiple scenes using the MWEN are better than those of the other comparison methods based on the indicators. (2) The MWEN method has the capability to accurately extract various types of water bodies, such as urban water bodies, open ponds, and plateau lakes. (3) By fusing features extracted at different scales, the MWEN has the capability to extract water bodies with different sizes and suppress noise, such as building shadows and highways. Therefore, MWEN is a robust water extraction algorithm for GaoFen-1 satellite images and has the potential to conduct water body mapping with multisource high-resolution satellite remote sensing data.

Keywords: convolutional neural network; water body extraction; GaoFen-1; multiple scales; deep learning

1. Introduction

Water is the basic substance for human society's production and development [1]. Surface water bodies play important roles in Earth's material and energy cycles [2,3]. Since satellite remote sensing data can capture large-scale surface information in little time and with low costs, the data have been used in water body surveys [4]. Multiple remote sensing data, including optical data [5] and

radar data [6], have been used for water body information extraction. The current water information extraction methods include the threshold method [7], machine learning [8,9], and deep learning [10,11], etc. The threshold method is a conventional method for water body extraction. The threshold method selects an appropriate threshold to distinguish water bodies and other objects in one or more bands [7]. Because the spectral characteristics of water in the near-infrared (NIR) band are significantly different from those of other objects, the NIR band is very popular in threshold segmentation [12]. To further highlight the difference between water bodies and surrounding features, water indexes have been developed [13]. However, the water index method has some problems. One is that objects with similar spectral characteristics, such as mountain shadows, cloud shadows, and highways, can be easily confused with water bodies, which makes it difficult to select thresholds. In addition, the threshold selected in large-scale water extraction may not be applicable to local areas [14]. With the development of machine learning, traditional machine learning algorithms, such as decision tree (DT) [15], support vector machine (SVM) [6], and random forest (RF) [9], have been widely used in water body extraction. These algorithms perform classification by using artificially designed features, including spectral and textural features. However, artificially designed features require considerable professional domain knowledge and artificially designed features are usually based on a specific scale of images. A standard way to extract artificially designed features from images at multiple scales is resampling the images to different scales and extract features based on the images with different scales. Thus, the process requires intensive computation with time consuming. In addition, different feature vectors are needed for different images and the feature vectors have great impacts on the final classification results. These issues make applying machine learning to water extraction challenging.

Deep learning is a popular method in image processing during the past several years [16,17]. Convolutional neural networks (CNNs) have been used in scene classification [18], semantic segmentation [19], and object detection [20,21]. The advantage of CNNs is to capture the features from raw images directly by multiple convolutional layers [22], which can avoid the complex feature processing. CNNs for semantic segmentation are capable of performing image classification at pixel level, which is important for information extraction from remote sensing images. In CNNs, the shallow convolutional layers are able to capture the pixel position information and the deep convolutional layers are used to label the pixels [22]. The fully convolutional network (FCN) is the first end-to-end CNN designed for semantic segmentation [19]. FCN extracts abstract features from the input image and labels each pixel in the feature maps extracted by the last convolutional layer. However, FCN loses information contained in low-level features extracted by shallow convolutional layers. In recent years, many models, such as Unet [23] and Deeplab V3+ [24], have been developed to improve the performance of CNNs for semantic segmentation in the field of computer vision. CNNs are gradually being applied to water information extraction with remote sensing images. In [10], CNN was firstly used for water body extraction in Landsat ETM+ images. The structure of the CNN contained only two convolutional layers and a fully connected layer. The shallow structure allows it to capture only low-level features which results in poor robustness in complex scenes. In addition, the input tile (19 $\times$ 19) is small in the CNN model. Thus, it cannot be used to extract features at large scales. With the improvement of the spatial resolution of satellite images [25], various methods based on deep learning have been proposed for water body extraction in high-resolution images. A CNN method that combines the super pixel was proposed by Chen, Y, et al. [11]. The core idea is to combine artificial designed features and CNN extraction features. However, the process reduces the fluidity of the water extraction and misses some useful information during forward propagation. In recent years, end-to-end CNNs, such as fully convolutional network (FCN) [26] and DeepWaterMap [27] have been applied to water body extraction. These end-to-end CNNs greatly improved the accuracy and efficiency of water body extraction. There are still challenges in the application of CNNs in water body extraction: (1) In the process of forward propagation, the resolution of feature maps is reduced due to the repeated max-pooling layers, which leads to the loss of detailed water body information. (2) The receptive fields of pixels are different in the feature maps extracted by the convolutional layers at

different depths, which allows these feature maps to contain feature information at different scales [22]. The combination of the features extracted at multiple scales in water body extraction still needs to be explored.

This paper aims to propose an improved convolutional neural network (CNN), named multi-scale water extraction convolutional neural network (MWEN), for water body extraction for GaoFen-1 images. For the first challenge, the encoder-decoder structure is used in the MWEN inspired by the Unet [23]. The encoder extracts the features from the input images and obtains feature maps with low resolution. The role of the decoder is to map the feature maps to the input resolution feature maps. For the second challenge, a structure, named the multi-scale feature extractor (MTFE), is proposed to capture features at multiple scales. Objects exist at various scales in remote sensing images and geological correlations may exist between adjacent objects. Features extracted by CNNs at different scales contain various information [28]. In the MTFE, four dilated convolutional layers with different dilation rates are used to learn features from images with different receptive fields.

The structure of the remainder of this article is as follows. First, GaoFen-1 high-resolution remote sensing satellite images in Beijing-Tianjin-Hebei region, Zhejiang province, and Tibet province in China are collected for the dataset and preprocessed. Then, four CNNs are employed to extract water body information. Finally, the accuracies of these algorithms are compared based on five accuracy metrics and a visual comparison.

2. Materials and Methods

2.1. Data

In this study, 24 GaoFen-1 images (17 for training and 7 for testing) located in Beijing-Tianjin-Hebei region, Zhejiang province, and Tibet province in China were collected as the experiment dataset and these images are showed in Figure 1. Four multispectral bands with a spatial resolution of 8 m and panchromatic band with a spatial resolution of 2 m are included in GaoFen-1 images. The radiation resolution of both the panchromatic band and multispectral bands is 16 bits. The spectral and textural characteristics of the water bodies in different regions are quite different, and the environments surrounding the water bodies are complex. To test the universality of these CNNs for water body extraction, environment characteristics, such as spectral, textural, season, water environment characteristics and confusing areas, such as shadows, highways, and ice are considered in the dataset. The detail information of the dataset is shown in Table 1.

Table 1. Detailed information of dataset.

Images	Location	Acquisition Times	Water Types	Major Confusing Objects
a1-a8	Tibet province	July, 2014 and August, 2016	Plateau lake, Plateau river, Saline lake	Cloud shadows, Saline land
b1-b7	Beijing-Tianjin-Hebei region	January, September and October, 2019	Agricultural water, town water, city water	Building shadows, sports field, highways.
c1-c9	Zhejiang province	April, 2017 and October, 2019	Agricultural water, town water, woodland water, city water	Mountain shadows, wetland, roads

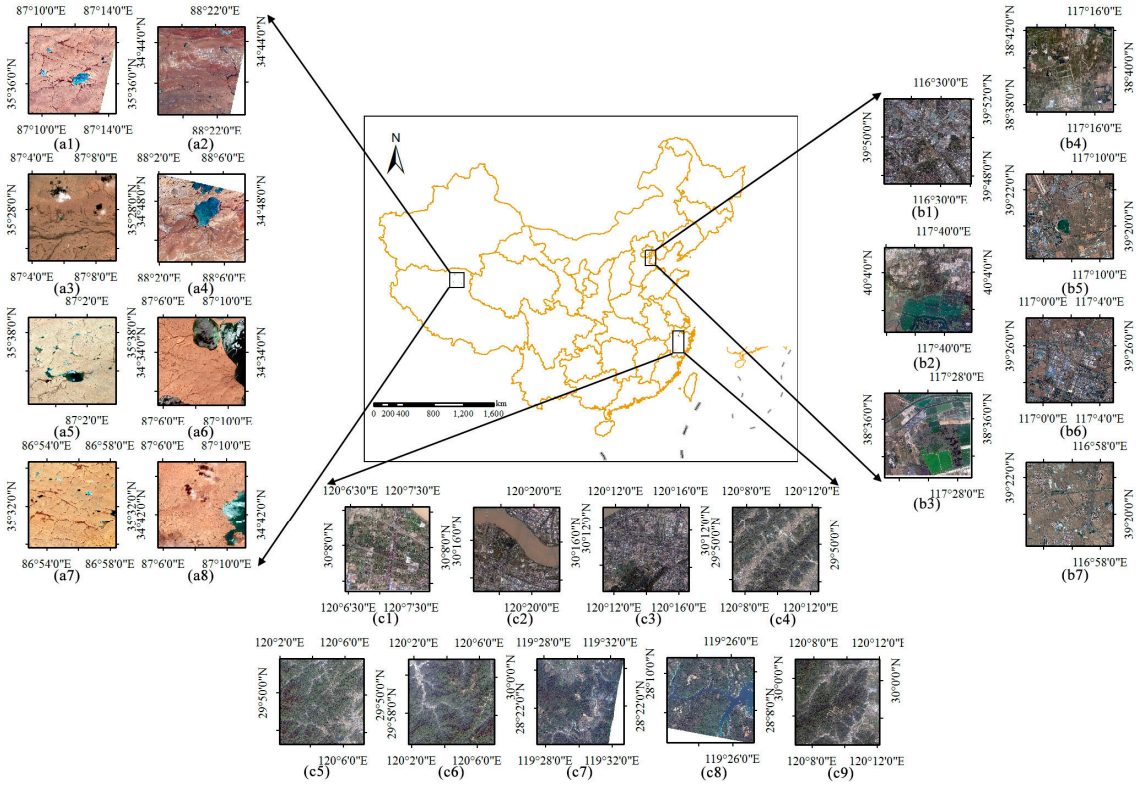

Figure 1. The GaoFen-1 (GF-1) dataset (a1, a3, a5, a6, a7, a8, b1, b2, b5, b6, b7, c1, c2, c3, c4, c5, and c6 are used for training images. a2, a4, b1, b3, b4, c7, and c8 are used for test images.).

2.2. *Methods*

The methods can be divided into four parts: image preprocessing, sample generation, water information extraction, and accuracy assessment. In the image preprocessing part, the Rational Polynomial Coefficient (RPC) model is used to geometrically correct these images [29]. Then, the multispectral and panchromatic images fusion was conducted using PANSHARP method [30]. The image preprocessing part was conducted based on the PCI Geo Imaging Accelerator software. The geometric errors of the images after preprocessing were within 1 pixel. In the second part, the water bodies in the fused images are labeled. These images and labels are clipped to 512 × 512 pixels and divided into a training dataset and a validation dataset. In the third step, MWEN (multi-scale water extraction convolutional neural network), MWEN "without MTFE", FCN, Unet, and Deeplab V3+ are employed to extract the water bodies. Finally, the accuracy comparison for different methods are conducted using visual comparison and quantitative evaluation metrics. The flowchart is shown in Figure 2.

2.2.1. Sample Generation

The labels in the dataset are from the fusion images and cover all water types mentioned in Section 2.1. The labels consist of water areas and background areas. All the labels in the dataset are binary images, where 1 represents water body and 0 represents background. All of the images were labeled via visual interpretation. These images were divided into training images and test images (17 for training and 7 for test). Both the training images and test images contain all water types mentioned in Table 1. These training images and training labels were clipped to samples with 512 × 512 pixels. A training sample library containing 13,509 samples from training images was obtained. The samples in the training sample library contains all water pixels in training images. Some areas without surface water bodies are also contained in these samples. The training sample library was divided into two parts. Ninety percent of the training samples were used as the training

dataset and the remaining small part was used for the validation dataset. The role of the validation dataset is to reflect the generalization ability of the model parameters and indicate whether the model is overfitting during training process. Both the validation dataset and training dataset were from the training images, which reduced the generalized representation of the validation dataset. To get a more generalized training model, the samples from the images other than the training image are needed for the validation dataset. In this study, a random part of each image in the test images was selected and clipped to 512 × 512 pixels to enrich the validation dataset. The final validation dataset consisted of 1651 samples from test images and 1350 samples from the training images.

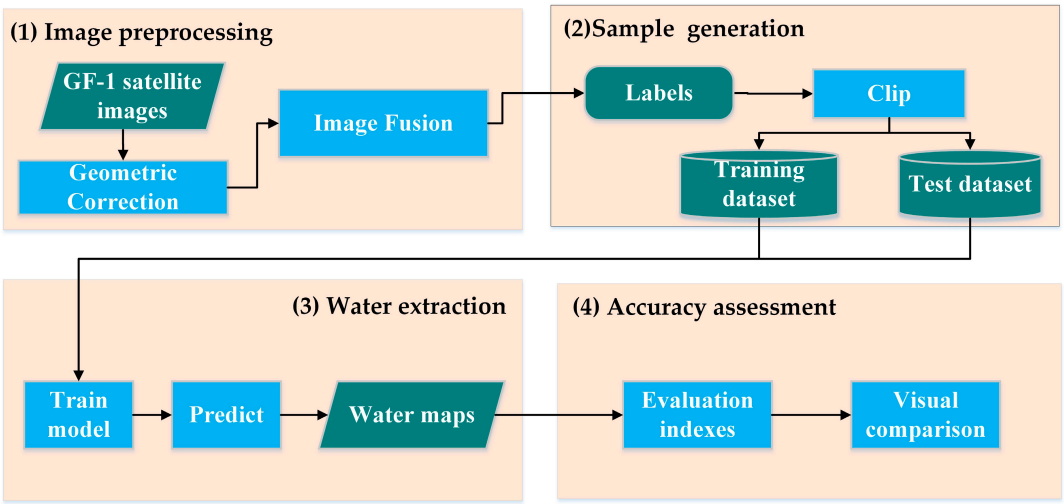

Figure 2. Flowchart of this study.

2.2.2. Multi-Scale Feature Extractor

Dilated convolution was originally used for the wavelet transform [31] and has been used in convolutional neural networks for semantic segmentation [32]. The convolution kernel with holes (or gaps) is used in the dilated convolution. The number of gaps inserted in the kernel depends on the dilation rate r. The dilation rate is prerequisite when a convolution kernel is defined. The dilated convolution with filter dilation rates of 0, 1, and 2 are shown in Figure 3. The kernel with a dilation rate of 0 is the same as the standard convolution kernel. The convolution kernels with different dilation rates have different receptive fields. The combination of dilated convolutions with different dilation rate kernels can capture the features at different scales.

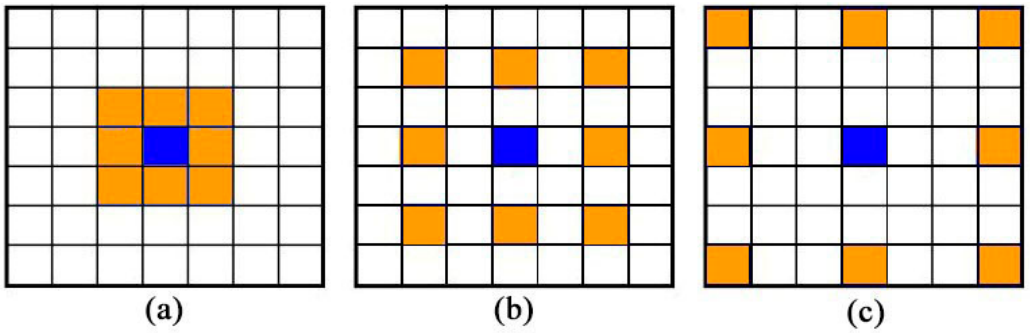

Figure 3. Dilated convolution kernels with different rates. (**a**), (**b**), (**c**) are dilated convolution kernels with dilation rate 0, 1, 2, respectively.

In remote sensing images, the sizes of water bodies are diverse and there are many confusing objects in high-resolution images, such as building shadows, mountain shadows, and sports fields, whose spectral characteristics are similar to those of water body. The combination of features extracted

at multiple scales is important in dealing with these issues. In this study, a structure, named multi-scale feature extractor (MTFE) is proposed. Dilated convolutions with various rates are used in the MTFE to extract the features at multiple scales. The structure of the MTFE is given in Figure 5. An example of feature extraction at multiple scales by dilated convolution with different rates is shown in Figure 4. As we can see in Figure 4b, the standard convolution (dilated convolution with a rate of 0) can only get the information of the surrounding 9 pixels, all of which lie in building shadows. It is difficult to identify the pixel at the center of the convolution kernel because shadows and water bodies have similar spectral characteristics. In the dilated convolutions with rates of 2, 4, and 8, the features are extracted at different scales and the information of the buildings and woods is captured. The combination of extracted features at these different scales is important for the distinction of building shadows.

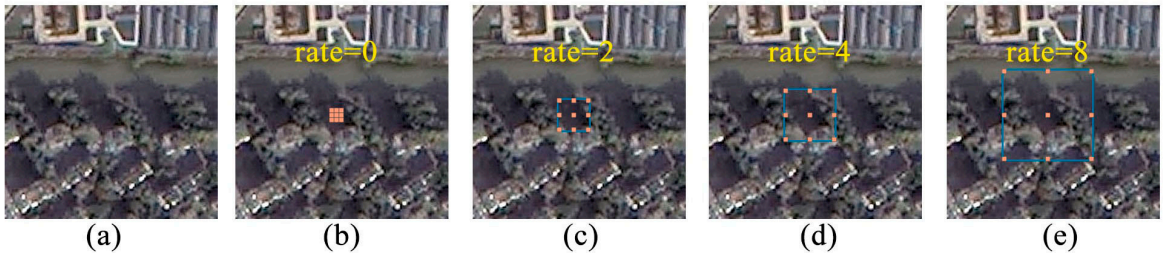

Figure 4. Examples of dilated convolution with different rates. (**a**) is the sample image; (**b**–**e**) are examples of dilated convolution with dilation rate 0, 2, 4, 8, respectively.

2.2.3. Convolutional Neural Networks (CNNs) for Water Extraction

A multi-scale water extraction convolutional neural network (MWEN) for surface water information extraction is proposed. The structure of the MWEN is shown in Figure 5. The MWEN can be divided into three parts: encoder, multi-scale feature extractor (MTFE), and decoder. In the first part, the input data are encoded by the encoder and feature maps with an output stride of 16 are obtained. In the multi-scale feature extractor (MTFE) part, the feature maps from the encoder are fed to four dilated convolutions with different rates. These dilated convolutions with different rates can learn features at different scales. Then, the feature maps generated by these dilated convolutions are concatenated and integrated by three convolutional layers. In the decoding part, the feature maps are decoded by the decoder to obtain the water segmented images.

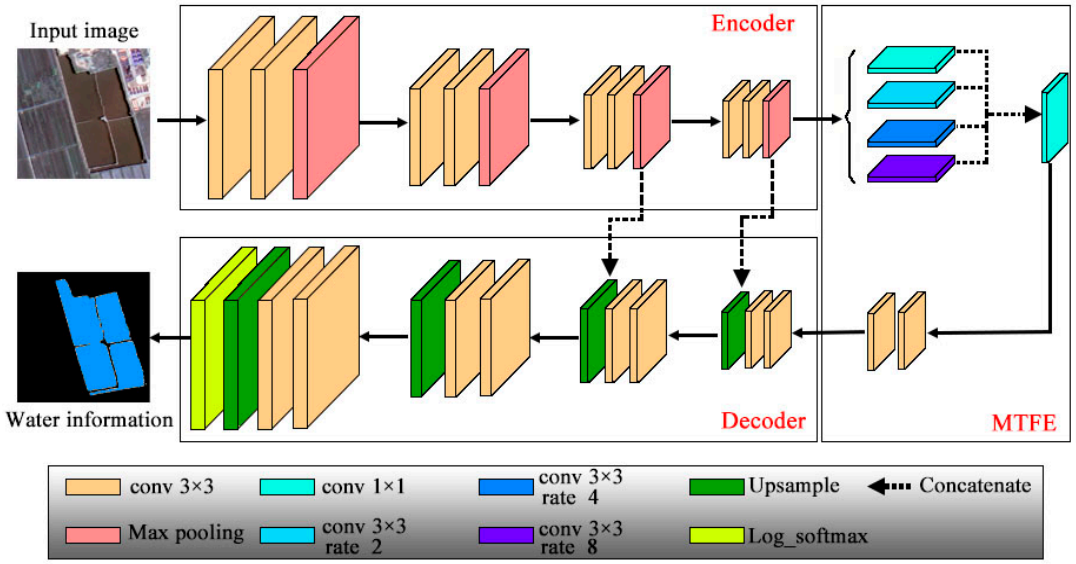

Figure 5. The structure of the multi-scale water extraction convolutional neural network (MWEN).

To examine the importance of MTFE to the segmentation results, both of the MWEN structure "with MTFE" and "without MTFE" were trained for water body extraction. The other three kinds of convolutional neural networks (CNNs) used for semantic segmentation, the FCN [33], Unet [23], and DeepLab V3+ [24], were also selected in this study for comparison. The water body extraction process using CNNs contains three steps: data augmentation, forward propagation, and model training.

- Data augmentation: Date augmentation is performed before training. In this step, the input samples are randomly processed in three ways, including flipping, zooming, and panning. All samples in the training dataset are randomly processed before every training epoch, and the number of training samples for every training epoch does not change. The data augmentation results for the three samples are shown in Figure 6.

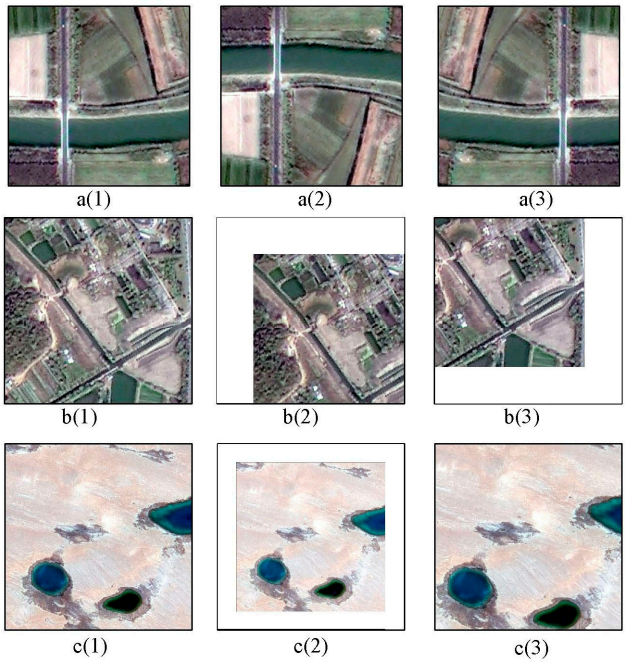

Figure 6. Data augmentation of the three samples. a(2) and a(3) are the results of flipping a(1), b(2) and b(3) are the results of panning b(1), and c(2) and c(3) are the results of zooming c(1).

Then, the data are normalized. The fused GF-1 data have a radiation resolution of 16 bits, with DN values ranging from 0 to 65535. To improve the accuracy and training efficiency of convolutional neural networks (CNNs), the input images are normalized. The normalization converses each input image into a feature map with a mean of 0 and a variance of 1. The formulas are as follows:

$$\mu = \frac{1}{w \times h \times c} \sum_{i=1}^{w} \sum_{j=1}^{h} \sum_{z=1}^{c} DN_{i,j,z} \tag{1}$$

$$\sigma^2 = \frac{1}{w \times h \times c} \sum_{m=1}^{w} \sum_{n=1}^{h} \sum_{z=1}^{c} (DN_{m,n,z} - \mu)^2 \tag{2}$$

$$\overline{DN_{m,n,z}} = \frac{DN_{m,n,z} - \mu}{\sqrt{\sigma^2}} \tag{3}$$

where μ is the average of the input image array, and w, h, and c are the width, height, and the number of channels of the input image, respectively. $DN_{m,n,z}$ is the DN value of the pixel in row n, column m, and channel z. σ^2 is the variance of the input image. $\overline{DN_{m,n,z}}$ is the DN value of the pixel in row n, column m, and channel z after normalization.

- Forward propagation: The normalized sample is fed into the CNN and a feature map is obtained after forward propagation. The output of the CNN is a feature map with a size of 512 × 512 × channels (where the channels are the number of classes). In this study, the number of channels is 2 (water bodies and backgrounds). Then, the feature map is activated by an activation function. The log softmax function is used as the activation function and the argmax function [34] is used to get the final water maps in this study. The formula of the activation function for each pixel in the feature maps is as follows:

$$P_{(m)} = \log\left(\frac{e^m}{\sum_{n=1}^{c} e^n}\right) \quad (4)$$

where $P_{(m)}$ is the data value of the pixel in channel m. c is the number of classes (2 in this study to reflect the water and background).

- Model training: The cross-entropy loss function [35] and the back propagation algorithm [36] are used when training the CNNs. The mean cross-entropy and the sparse categorical accuracy [37] are calculated between the labels and the predicted maps by the CNN forward propagation. To minimize the cross entropy, the Adam optimizer [38] is applied to identify the weights and biases in the back-propagation process. In this study, the weights of the CNNs model are trained on training dataset and weights with the highest parse categorical accuracies on the validation dataset are selected as the training results.

2.2.4. Accuracy Assessment

The performances of these convolutional neural networks (CNNs) are thoroughly evaluated via visual comparison and five evaluation metrics. The visual comparisons contain the comparison between MWEN "with MTFE" and "without MTFE" and the comparison between MWEN, FCN, Unet, and Deeplab V3+ on regions with different types of surface water bodies and confusing objects. Regarding the evaluation metrics, five evaluation metrics are used to evaluate the accuracy in this study, including the Overall Accuracy (OA) [30], the True Water Rate (TWR), the False Water Rate(FWR), the Water Intersection over Union (WIoU) [30], and the Mean Intersection over Union (MIoU) [39]. The definitions and formulas of these indicators are listed in Table 2.

Table 2. Five evaluation metrics for the accuracy assessment.

Evaluation Index	Definition	Formula
OA	The ratio of the correctly classified number of pixels and the total number of pixels	$OA = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$
TWR	The ratio of the number of properly classified water pixels and the number of labeled water pixels	$TWR = \frac{TP}{TP+FP} \times 100\%$
FWR	The ratio of the number of misclassified water pixels and the number of labeled water pixels	$FWR = \frac{FP}{FP+TP} \times 100\%$
WIoU	The ratio of the intersection and the union of the ground truth water and the predicted water area.	$WIoU = \frac{TP}{FN+TP+FP}$
MIoU	The average IoU for all classes (water and background)	$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{TP}{FN+TP+FP}$

where TP, TN, FN, and FP represent the numbers of pixels of true water, true background, false background, and false water, respectively.

3. Results

3.1. Model Training

The training processes were conducted using Python3.6, Keras, and TensorFlow on a NVIDIA Titan GPU with cuDNN 10.0 acceleration. The categorical accuracies on the training dataset and validation dataset are calculated at the end of each training epoch. The weights with the highest categorical accuracies are used for water extraction in next steps. The highest validation accuracies of these models are shown in Table 3. The training accuracy and validation accuracy curves are shown in Figure 7. The training and validation accuracy curves of these models grow slowly after the 15th epoch and some even show downward trends after the 25th epoch. There is a large gap between the training accuracy curve and the validation accuracy curve of the Deeplab V3+. The Deeplab V3+ appeared to overfit when it is directly used in water body extraction from remote sensing images. The efficiency of training models is affected by many factors. The efficiency of the CNNs are simply compared via the number of trainable parameters and training time in this study. The efficiency comparison of these CNNs are shown in Table 4. The FCN has the most parameters but less training time. The Deeplab V3+ has the longest train time due to its complex and deep model structure. The MWEN and Unet have fewer parameters and less training time.

Table 3. The highest validation accuracy of CNN models in training process.

CNN	MWEN	MWEN without MTFE	FCN	Unet	Deeplab V3+
Highest validation accuracy	0.987	0.981	0.978	0.983	0.957

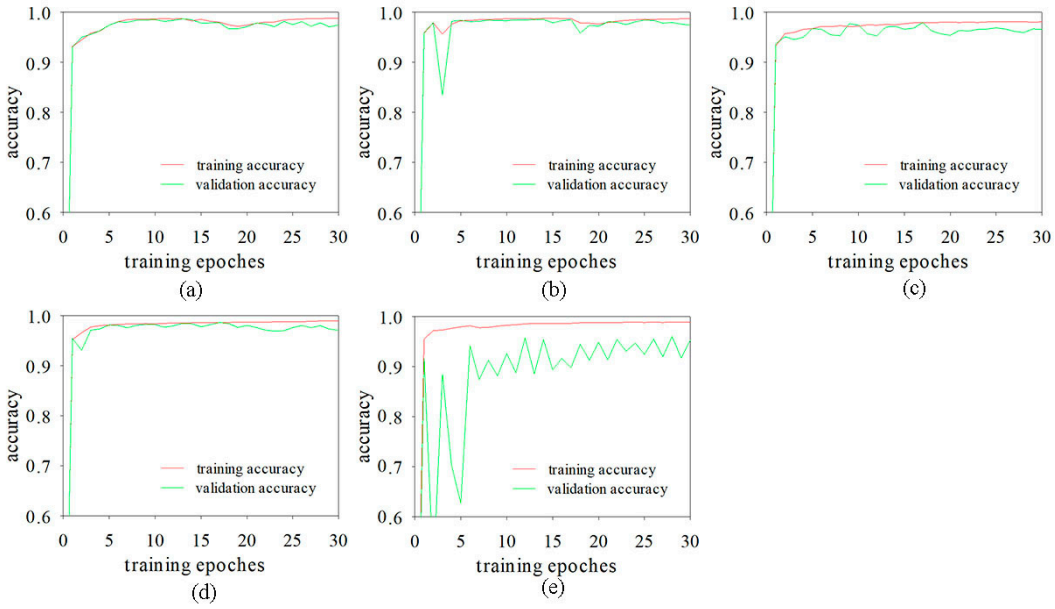

Figure 7. The training and validation accuracy curves of the convolutional neural networks (CNN) models. (**a**), (**b**), (**c**), (**d**), and (**e**) represent training and validation accuracy curves of the MWEN, MWEN "without multi-scale feature extractor (MTFE)", fully convolutional network (FCN), Unet, and Deeplab V3+, respectively.

Table 4. The efficiency comparison of the various methods.

CNN	Number of Trainable Parameters (Million)	Training Time (s/epoch)
MWEN	3.72	1343
MWEN without MTFE	1.57	1161
FCN	5.71	1345
Unet	3.11	1366
Deeplab V3+	4.11	2161

3.2. Water Extraction Results on the Test Dataset

The results of the water body extraction using these CNNs on the test images are shown in Figure 8. As can be seen from the figure, the water body prediction results of these CNNs are different. For Regions a and g, more confusing objects are contained in these two regions than the others, which makes the CNNs more prone to make mistakes. The roads and the building shadows are misclassified using Unet and Deeplab V3+ in these two regions. For Regions e and f, there are some detailed water bodies that are missed by the FCN and MWEN "without MTFE". Although performances of these CNNs are similar in Regions b, c, and d across these images, there are still differences in details. Some details are derived from these results and shown in Section 3.3. Figure 8 shows that MWEN has the capability to capture detailed water and suppresses noise better than the others.

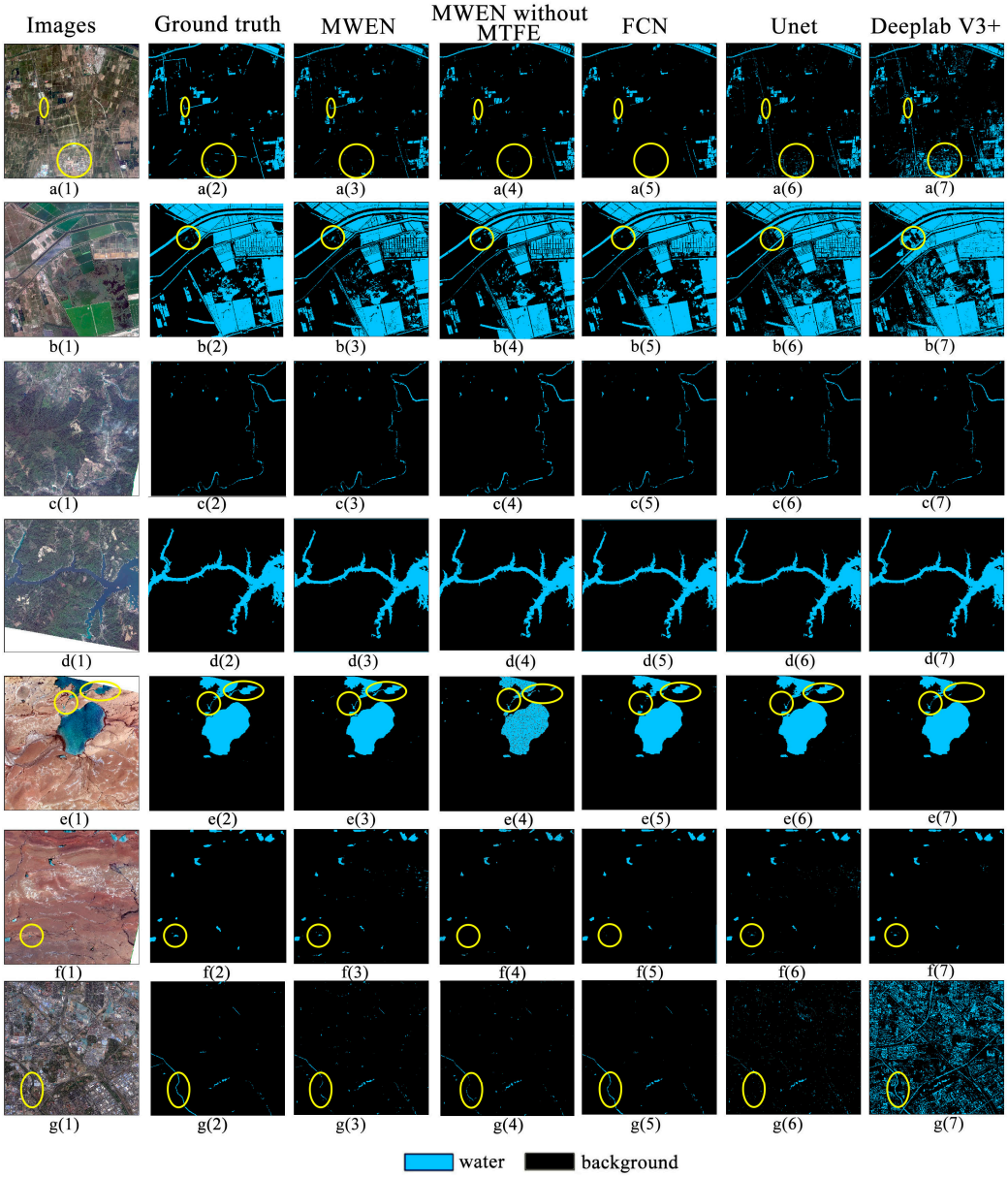

Figure 8. The results classified by the four CNNs on the test dataset. (a1–g1) are the original images, (a2-g2), (a3-g3), (a4–g4), (a5–g5), (a6–g6), (a7–g7) are the water body information extracted by artificial interpretation, MWEN, MWEN "without MTFE", FCN, Unet, Deeplab V3+, respectively. The areas in yellow circles are the areas water bodies greatly differ. Blue parts of the pictures represent the extracted water bodies and black parts of the pictures represent the background.

3.3. Accuracy Analysis

To analyze the universality of the MWEN method, different water types are analyzed. The accuracy comparisons via the evaluation metrics are shown in Section 3.3.1, the comparisons between MWEN "with MTFE" and "without MTFE" are shown in Section 3.3.2, and the accuracy comparisons via the visual comparison between MWEN, FCN, Unet, and Deeplab V3+ are shown in Sections 3.3.3 and 3.3.4.

3.3.1. Accuracy Comparisons via the Evaluation Metrics

To quantitatively analyze the water body extraction accuracy, the metrics mentioned in 2.2.3 were calculated based on the water maps predicted by the CNNs and the ground truth. Results are summarized in Table 5. As can be seen from the table, the MWEN outperforms the others in the OA, FWR, WIoU, and MIoU [30]. Deeplab V3+ is one of the best CNNs for semantic segmentation. In this study, Deeplab V3+ performs poorly in the OA, FWR, WIoU, and MIoU, but it performs the best in the TWR. Deeplab V3+ may be suitable for datasets with complex scenes, but it appears to be overfitting when training for water extraction.

Table 5. Water body extraction accuracies of the various methods.

CNN	OA (%)	TWR (%)	FWR (%)	WIoU	MIoU
MWEN	98.62	92.34	0.61	0.880	0.932
MWEN without MTFE	98.35	91.58	0.86	0.863	0.916
FCN	98.52	91.40	0.62	0.870	0.927
Unet	98.18	92.82	1.16	0.849	0.914
Deeplab V3+	91.82	96.92	8.81	0.566	0.737

3.3.2. Performance Comparison for MWEN and MWEN "Without Multi-Scale Feature Extractor (MTFE)"

Feature maps extracted by CNN at different scales contain various information. In this study, the multi-scale feature extractor (MTFE) is proposed to capture the features at multiple scales. In order to examine the importance of features extracted by MTFE for water extraction, results containing ponds and rivers with different sizes, and building shadows are derived from the result water maps mentioned in Section 3.2. The comparisons between the MWEN "with MTFE" and "without MTFE" are shown in Figure 9.

For the pools with different sizes in Figure 9a, both of the MWEN "with MTFE" and "without MTFE" can identify larger ponds, but the latter has obvious disadvantages for addressing the smaller pool information in Figure 9(a4). Moreover, tiny rivers cannot be identified by the MWEN "without MTFE" in Figure 9(b4,c4). Regarding confusing objects, the highway and some building shadows are mixed by the MWEN "without MTFE" in Figure 9(d4,e4). This may result from the relevance information between objects, such as the relationship between buildings and shadows, being ignored by MWEN "without MTFE". The relevance information may be contained in the features extracted by the convolution kernel with a large expansion rate. Figure 9 shows that MTFE plays an important role in extracting water bodies with various sizes and suppressing noise.

3.3.3. Performance Comparison for Different Water Types

Different surface water bodies, including open ponds, plateau rivers and lakes, city waters and agricultural water bodies, are taken from the results to assess the universality of the MWEN algorithm. The performances of the MWEN are compared with those of the FCN, Unet, and Deeplab V3+ based on the visual inspection. The performance comparison is shown in Figure 10.

For the open pools in Figure 10a, the comparison shows that all four CNNs are able to extract the large open pools. The smaller open pools are missed when using the FCN in Figure 10(a4). The results for agricultural waters show that detailed boundary information is missing by the FCN and Deeplab

V3+ in Figure 10(b4,c4,c6). Rough boundaries and mixing between water and wetlands appear when using the Unet in Figure 10(c5). Regarding plateau rivers and lakes, it can clearly be seen that the parts of rivers and lakes are missing by the FCN and Deeplab V3+ in Figure 10(d4,d6,e4,e6). The results for small puddle and tiny rivers in city demonstrate that the small puddle and tiny rivers are missed by the FCN and Unet in Figure 10(f4,g4,g5). Affected by urban buildings and other objects, the results extracted by the Unet and Deeplab V3+ contain more noises in Figure 10(f5,f6,g6).

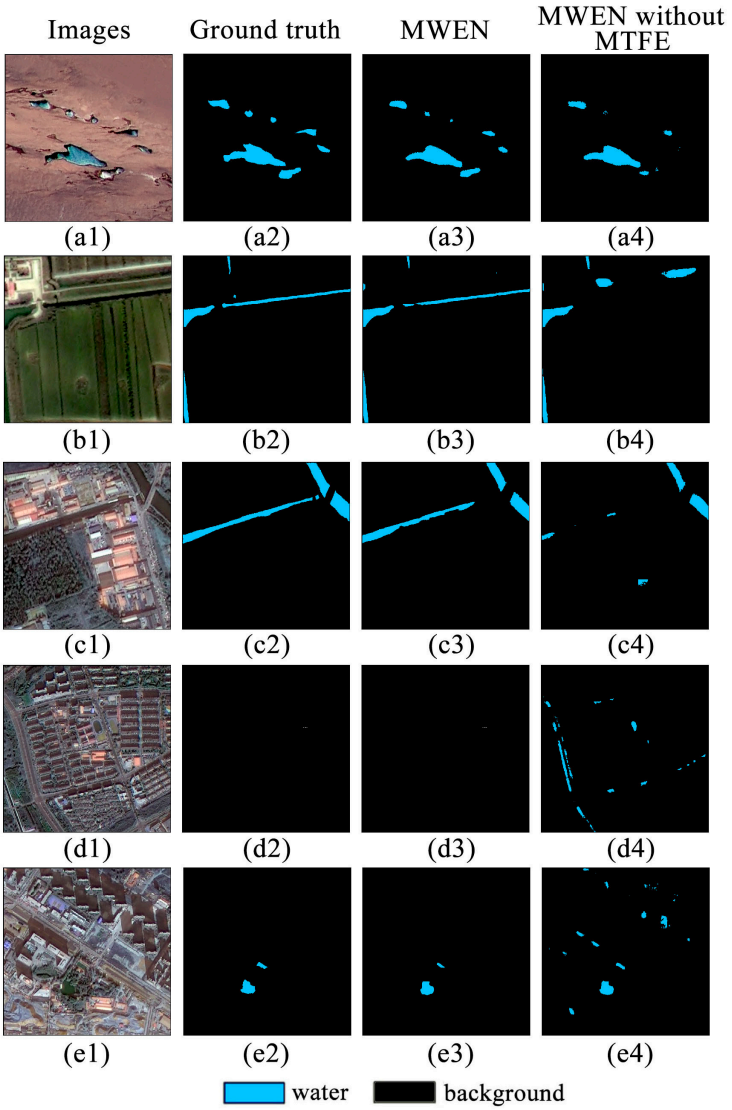

Figure 9. Results comparison between MWEN "with MTFE" and "without MTFE". (a1–e1) are the images, (a2–e2) are the ground truth, (a3–e3) are the water maps extracted by the MWEN "with MTFE", (a4–e4) are the water maps extracted by the MWEN "without MTFE".

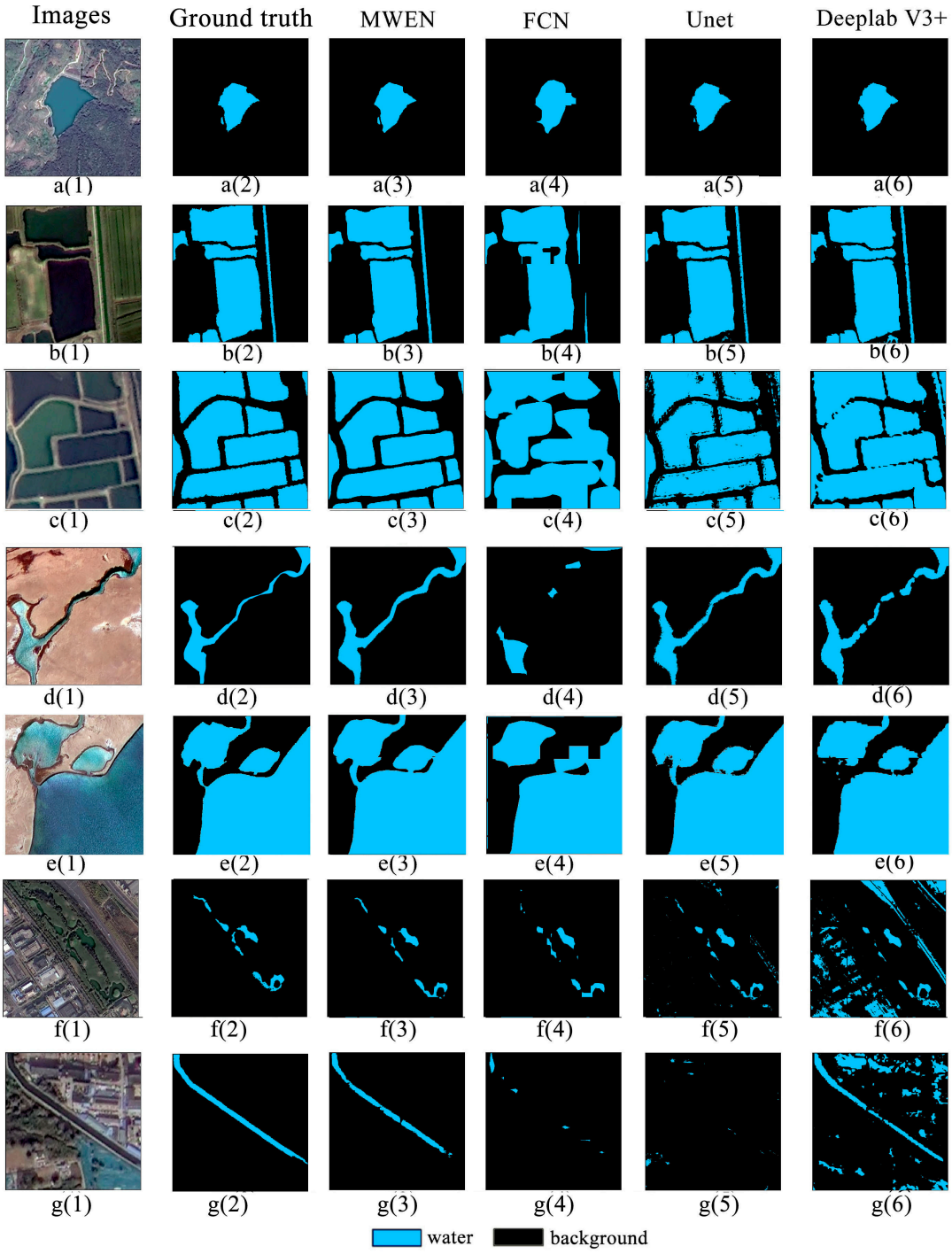

Figure 10. Typical surface water classification results. a(1) is the original image with open pools, and a(2–6) represent the water body information extracted from a(1) by artificial interpretation, MWEN, FCN, Unet, DeepLab V3+, respectively. Additionally, b, c, d, e, f, g give the experimental results of water body extraction from different images with agricultural water, plateau river, plateau lakes, small water bodies, and tiny rivers, respectively. Blue parts of the pictures are the extracted water bodies and black parts of the pictures are the backgrounds.

From Figure 10, it can be seen that MWEN performs better than the other algorithms. The FCN loses much detailed information for surface water body, which leads to blurred boundaries and the absence of small water bodies. Unet and Deeplab V3+ can better extract detail information of the water body compared with FCN but may be confused with objects with spectral characteristics to similar

water. Figure 10 shows that the MWEN has the ability to extract different types of water bodies and the universal performance is better than other.

3.3.4. Performance Comparison for Confusing Areas

In high-resolution remote sensing images, some objects have spectral features or texture features similar to those of water bodies. It is a challenge to distinguish water bodies from these objects. To examine the reliability of these CNNs in distinguishing water bodies from confusing areas, the water body extraction results for confusing areas, such as building shadows, sports fields, and highways, are shown in Figure 11.

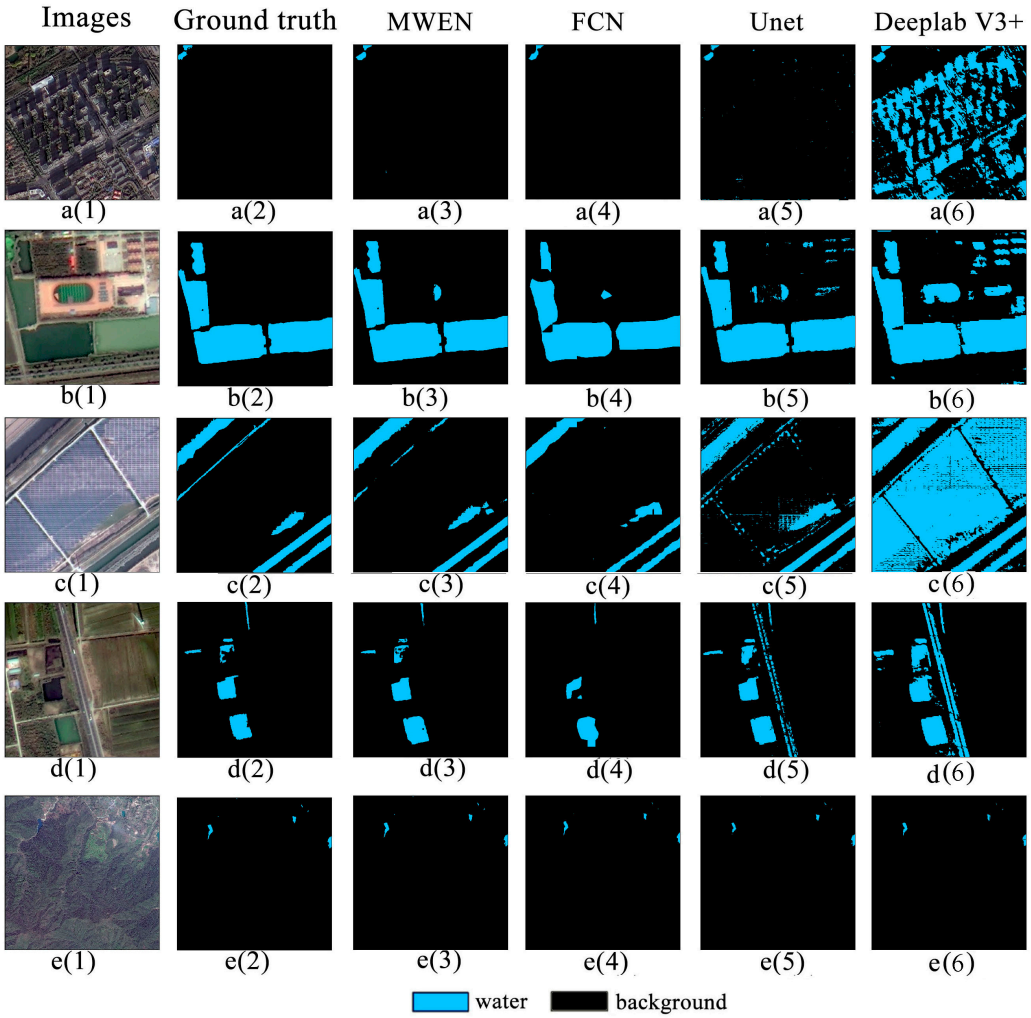

Figure 11. The results of the four methods for confusing regions. a(1) is the original image with building shadows, and a(2–6) represent the water body information extracted from a(1) by artificial interpretation, MWEN, FCN, Unet, DeepLab V3+, respectively. Additionally, b, c, d, e give the experimental results of water body extraction from different images with playgrounds, shade net, highways and mountain shadows, respectively. Blue parts of the pictures are the water bodies and black parts of the pictures are the backgrounds.

For the building shadows shown in Figure 11a, the MWEN, FCN, and Unet can better suppress noise, while Deeplab V3+ does not remove the building shadows, which may be caused by overfitting during training. Figure 11b demonstrates that all of these CNNs cannot clearly remove the noises from the sports field, but the MWEN and FCN perform better than the others. For the areas in Figure 11c,d, the Unet and Deeplab V3+ obviously mix the surface water body and other objects. For the mountain

shadow area in Figure 11e, all four CNNs can clearly remove the noise. The performance comparison in confusing areas shows that the noises from the sports field, shade net and highway still exist in the results based on Unet and Deeplab V3+. The MWEN and FCN achieve better performances in suppressing the noise than the others.

4. Discussion

With the improvement in the temporal and spatial resolution of remote sensing data [25], many meaningful works have been conducted on water body information extraction with high-resolution remote sensing data [40,41]. Deep learning has been a hot topic in recent years [42], and it shows great promise in water body extraction with high-resolution remote sensing data. In this study, a new CNN named MWEN is proposed for water body extraction for GaoFen-1 images. The extraction accuracy of water bodies on the test dataset is evaluated by five evaluation metrics and visual comparison. The results show that MWEN has the ability to extract water bodies with different sizes and can accurately capture the boundaries of water bodies. In addition, MWEN can suppress noise better than Unet and Deeplab V3+.

The different performance in water body extraction may relate to the structures of these CNNs. FCN has been applied to water body extraction in previous research [26]. The FCN based methods extract features by several convolutional layers from the image and then perform water body segmentation based only on the low-resolution feature maps extracted by the last convolutional layer. The water maps are mapped to the original image resolution by upsampling. However, the upsampling process is not sensitive to the details in the image, which leads to small water bodies to be ignored and the boundaries of water bodies are smoothed. The Unet combines the structure of the encoder and decoder, and features at multiple scales are fused through skip connection between the encoder and decoder [23]. This is good for extracting the accurate boundaries of water bodies and capturing detailed information in the image. However, the Unet fuses too many low-level features extracted by the shallow convolutional layers. These low-level feature maps may be related to mistakes for noises that have similar spectral characteristics with water bodies. Deeplab V3+ is one of the state-of-the-art CNNs in the field of computer vision [24]. Deeplab V3+ uses ASPP pyramids to extract features at multiple scales and uses a decoder to restore the resolution of the feature maps. The Deeplab V3+ does not perform well in this study, which may be related to its complex structure. It may be suitable for pixel-level segmentation in complex scenes. It is prone to overfit in water body extraction. Motivated by the Unet [23] and Deeplab V3+ [24], the MWEN is proposed in this study. In the MWEN, the MEFT structure is proposed for capturing features at multiple scales and the encoder-decoder structure is used to restore the resolution. Compared with Deeplab V3+, the MWEN contains fewer convolutional layers and fewer trainable parameters, which effectively suppresses overfitting. The structure of MWEN makes it perform better in water body extraction for high-resolution images. Although MWEN obtains good accuracy on the test images, there are factors that affect the classification accuracy.

One is that new challenges appear in high-resolution image water extraction compared to mid-resolution images. The noise in water extraction based on medium resolution images, such as mountain shadows [42], can be easily distinguished in high-resolution images. Small water bodies may be difficult to extract in medium-resolution images, but they can be easily identified in high-resolution images. However, building shadows, highways, dark lawns, and dark roofs may result in new errors. In this study, the MWEN performs better in suppressing noise compared to the Unet and Deeplab V3+, but it does not completely remove the noise, such as noise from sports fields. In addition, very detailed water information is contained in high-resolution images, which brings new challenges for more accurate water body extraction.

The other is the dataset. The CNN with trained weights can perform well on images similar to the samples in the sample library. Its applicability to images that are quite different from the samples in the sample library needs further study. A dataset based on high-resolution remote sensing images containing multiple types of water bodies and easily confused areas, such as shadows, is needed.

Although the dataset proposed in this article contains common water bodies and easily confused areas, which can meet some data requirements in certain areas, the sample library needs to be enriched in the future.

5. Conclusions

Convolutional neural networks have been shown to have strong image classification and semantic segmentation abilities for remote sensing images. A new convolutional neural network named the MWEN for water body extraction for GF-1 high-resolution satellite images is proposed in this study. Three CNNs that conduct semantic segmentation in computer vision field are employed for comparison. The performances of the water body extraction results are evaluated based on five evaluation metrics and visual comparisons. The conclusions are as following:

(1) The performance of the MWEN is better than that of the FCN, Unet, and DeepLab V3+ when extracting surface water according to the visual comparison. The quantitative metrics show that results of the MWEN on the OA, TWR, FWR, WIoU, and MIoU are better than those of the others.

(2) The comparison between MWEN "with MTFE" and "without MTFE" demonstrates that the combination of features extracted at multiple scales is important to water extraction. The MTFE is helpful for dealing with confusing areas and water bodies with different sizes.

(3) Compared with the FCN and Unet, the results of the MWEN show that it can accurately extract water bodies in different scenes, such as the details of city water and plateau lakes. In addition, the MWEN has the ability to suppress noises, such as mountain shadows, highways, vegetation shadows, and dark lawns.

With the further enrichment of dataset, the MWEN has the application potential in large scale surface water mapping with high resolution satellite images, which can provide data support for surface water resource survey.

Author Contributions: Guojin He, Hongxiang Guo conceived of and designed the experiments. Hongxiang Guo. Wei Jiang and Ranyu Yin performed the experiments. Hongxiang Guo Wanchun Leng and Lei Yan made the dataset. Hongxiang Guo and Wei Jiang analyzed data. Hongxiang Guo wrote the whole paper, and all authors edited the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China (61731022), the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (Grant No.2019QZKK0307) and the National Key Research and Development Program of China (2016YFA0600302).

Acknowledgments: The authors thank the anonymous reviewers and the editors for their valuable comments to improve our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Oki, T.; Kanae, S. Global hydrological cycles and world water resources. *Science* **2006**, *313*, 1068–1072. [CrossRef]
2. Van Oost, K.; Quine, T.A.; Govers, G.; De Gryze, S.; Six, J.; Harden, J.W.; Ritchie, J.C.; McCarty, G.W.; Heckrath, G.; Kosmas, C.; et al. The impact of agricultural soil erosion on the global carbon cycle. *Science* **2007**, *318*, 626–629. [CrossRef]
3. Wei, J.; Guojin, H.; Zhiguo, P.; Hongxiang, G.; Tengfei, L.; Yuan, N. Surface water map of china for 2015 (swmc-2015) derived from landsat 8 satellite imagery. *Remote Sens. Lett.* **2020**, *11*, 265–273.
4. Ji, L.Y.; Gong, P.; Wang, J.; Shi, J.C.; Zhu, Z.L. Construction of the 500-m resolution daily global surface water change database (2001-2016). *Water Resour. Res.* **2018**, *54*, 10270–10292. [CrossRef]
5. Fang, Y.; Ceola, S.; Paik, K.; McGrath, G.; Rao, P.S.C.; Montanari, A.; Jawitz, J.W. Globally universal fractal pattern of human settlements in river networks. *Earths Future* **2018**, *6*, 1134–1145. [CrossRef]
6. Lv, W.; Yu, Q.; Yu, W. Water extraction in sar images using glcm and support vector machine. In Proceedings of the 2010 IEEE 10th International Conference on Signal Processing Proceedings (Icsp2010), Beijing, China, 24–28 October 2010; pp. 740–743.

7. Xiao, Y.; Zhao, W.; Zhu, L. A study on information extraction of water body using band1 and band7 of tm imagery. *Sci. Surv. Mapp.* **2010**, *35*, 226–227.
8. Song, X.F.; Duan, Z.; Jiang, X.G. Comparison of artificial neural networks and support vector machine classifiers for land cover classification in northern china using a spot-5 hrg image. *Int. J. Remote Sens.* **2012**, *33*, 3301–3320. [CrossRef]
9. Ko, B.C.; Kim, H.H.; Nam, J.Y. Classification of potential water bodies using landsat 8 oli and a combination of two boosted random forest classifiers. *Sensors* **2015**, *15*, 13763–13777. [CrossRef] [PubMed]
10. Yu, L.; Wang, Z.; Tian, S.; Ye, F.; Ding, J.; Kong, J. Convolutional neural networks for water body extraction from landsat imagery. *Int. J. Comput. Intell. and Appl.* **2017**, *16*. [CrossRef]
11. Chen, Y.; Fan, R.S.; Yang, X.C.; Wang, J.X.; Latif, A. Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning. *Water* **2018**, *10*, 585. [CrossRef]
12. Frazier, P.S.; Page, K.J. Water body detection and delineation with landsat tm data. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 1461–1467.
13. Gao, B.C. Ndwi—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
14. Zhou, Y.A.; Luo, J.C.; Shen, Z.F.; Hu, X.D.; Yang, H.P. Multiscale water body extraction in urban environments from satellite images. *IEEE J. Sel. Topics Appl. Earth.Observ. Remote Sens.* **2014**, *7*, 4301–4312. [CrossRef]
15. Acharya, T.D.; Lee, D.H.; Yang, I.T.; Lee, J.K. Identification of water bodies in a landsat 8 oli image using a j48 decision tree. *Sensors* **2016**, *16*, 1075. [CrossRef]
16. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *arXiv* **2019**, arXiv:1909.00133. [CrossRef]
17. Lu, D.; Weng, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870. [CrossRef]
18. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *39*, 640–651.
20. He, K.M.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask r-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision (Iccv), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
21. Pan, H.D.; Chen, G.F.; Jiang, J. Adaptively dense feature pyramid network for object detection. *Ieee Access* **2019**, *7*, 81132–81144. [CrossRef]
22. Wu, Z.; Gao, Y.; Li, L.; Xue, J.; Li, Y. Semantic segmentation of high-resolution remote sensing images using fully convolutional network with adaptive threshold. *Connect. Sci.* **2019**, *31*, 169–184. [CrossRef]
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
24. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
25. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.M.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.M.; Anders, K.; Gloaguen, R.; et al. Multisource and multitemporal data fusion in remote sensing a comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [CrossRef]
26. Li, L.W.; Yan, Z.; Shen, Q.; Cheng, G.; Gao, L.R.; Zhang, B. Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks. *Remote Sens.* **2019**, *11*, 1162. [CrossRef]
27. Isikdogan, F.; Bovik, A.C.; Passalacqua, P. Surface water mapping by deep learning. *IEEE J. Sel. Topics Appl. Earth.Observ. Remote Sens.* **2017**, *10*. [CrossRef]
28. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
29. Long, T.F.; Jiao, W.L.; He, G.J. Nested regression based optimal selection (nrbos) of rational polynomial coefficients. *Photogramm. Eng. Remote Sens.* **2014**, *80*, 261–269.
30. Peng, Y.; Zhang, Z.M.; He, G.J.; Wei, M.Y. An improved grabcut method based on a visual attention model for rare-earth ore mining area recognition with high-resolution remote sensing images. *Remote Sens.* **2019**, *11*, 987. [CrossRef]

31. Holschneider, M.; Kronland-Martinet, R.; Morlet, J.; Tchamitchian, P. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 286–297.
32. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
33. Li, Y.; Qi, H.Z.; Dai, J.; Ji, X.Y.; Wei, Y.C. Fully convolutional instance-aware semantic segmentation. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (Cvpr 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
34. Gould, S.; Fernando, B.; Cherian, A.; Anderson, P.; Cruz, R.S.; Guo, E. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv* **2016**, arXiv:1607.05447.
35. De Boer, P.-T.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A tutorial on the cross-entropy method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]
36. Leung, H.; Haykin, S. The complex backpropagation algorithm. *IEEE Trans. Signal Process.* **1991**, *39*, 2101–2104. [CrossRef]
37. Von Davier, M. Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a monte carlo study. *Methods Psychol. Res. Online* **1997**, *2*, 29–48.
38. Bello, I.; Zoph, B.; Vasudevan, V.; Le, Q.V. Neural optimizer search with reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 459–468.
39. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. *arXiv* **2019**, arXiv:1907.05740.
40. Miao, Z.; Fu, K.; Sun, H.; Sun, X.; Yan, M. Automatic water-body segmentation from high-resolution satellite images via deep networks. *IEEE Geosci. Remote Sens. Lett.* **2018**. [CrossRef]
41. Yao, F.F.; Wang, C.; Dong, D.; Luo, J.C.; Shen, Z.F.; Yang, K.H. High-resolution mapping of urban surface water using zy-3 multi-spectral imagery. *Remote Sens.* **2015**, *7*, 12336–12355. [CrossRef]
42. Jiang, W.; He, G.; Long, T.; Ni, Y.; Liu, H.; Peng, Y.; Lv, K.; Wang, G. Multilayer perceptron neural network for surface water extraction in landsat 8 oli satellite images. *Remote Sens.* **2018**, *10*, 755. [CrossRef]