*remote sensing*

MDPI

*Article*

# MSResNet: Multiscale Residual Network via Self-Supervised Learning for Water-Body Detection in Remote Sensing Imagery

**Bo Dang and Yansheng Li ***

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; bodang@whu.edu.cn

* Correspondence: yansheng.li@whu.edu.cn

**Abstract:** Driven by the urgent demand for flood monitoring, water resource management and environmental protection, water-body detection in remote sensing imagery has attracted increasing research attention. Deep semantic segmentation networks (DSSNs) have gradually become the mainstream technology used for remote sensing image water-body detection, but two vital problems remain. One problem is that the traditional structure of DSSNs does not consider multiscale and multishape characteristics of water bodies. Another problem is that a large amount of unlabeled data is not fully utilized during the training process, but the unlabeled data often contain meaningful supervision information. In this paper, we propose a novel multiscale residual network (MSResNet) that uses self-supervised learning (SSL) for water-body detection. More specifically, our well-designed MSResNet distinguishes water bodies with different scales and shapes and helps retain the detailed boundaries of water bodies. In addition, the optimization of MSResNet with our SSL strategy can improve the stability and universality of the method, and the presented SSL approach can be flexibly extended to practical applications. Extensive experiments on two publicly open datasets, including the 2020 Gaofen Challenge water-body segmentation dataset and the GID dataset, demonstrate that our MSResNet can obviously outperform state-of-the-art deep learning backbones and that our SSL strategy can further improve the water-body detection performance.

**Keywords:** water-body detection; multiscale residual network (MSResNet); self-supervised learning (SSL); high-resolution remote sensing imagery



**Citation:** Dang, B.; Li, Y. MSResNet: Multiscale Residual Network via Self-Supervised Learning for Water-Body Detection in Remote Sensing Imagery. *Remote Sens.* **2021**, *13*, 3122. https://doi.org/10.3390/rs13163122

Academic Editor: Wei Yang

Received: 3 July 2021
Accepted: 4 August 2021
Published: 6 August 2021



## 1. Introduction

Along with the rapid development of remote sensing (RS) technology, the availability of RS images has been growing rapidly [1], and human beings have already entered the era of big data RS [2,3], making large-scale water-body monitoring possible [4]. To pursue timely large-scale water-body monitoring, automatic water-body detection from RS imagery has become incredibly urgent. As is well known, the water-body regions in optical RS imagery generally present multiscale characteristics, as water-body regions comprise water bodies of different types, such as thin and wide regions, and large differences exist within these water-body categories. Therefore, accurately detecting water bodies from RS images remains a challenging problem. Although existing studies have demonstrated that synthetic aperture radar (SAR) images have the potential to monitor water bodies [5–7], effectively monitoring water bodies using optical RS images is still very important because optical RS images are easier to collect than SAR images. Hence, this paper mainly focuses on exploiting water-body detections using optical RS imagery.

In the literature, the existing methods of monitoring water bodies mainly include two perspectives. The traditional methods are mainly based on handcrafted features, such as the normalized difference water index (NDWI) [8], the band analysis-based method [9], the gray-level cooccurrence matrix (GLCM) and the supported vector machine (SVM)-based method [10]. As mentioned above, most traditional water-body detection methods

strongly depend on the spectral information of RS images, pay little attention to contextual spatial information, and have poor universal applicability. Currently, the more mainstream methods are deep learning-based water-body detection methods, which attract increasing attention. Many general deep networks [11–14] in the field of computer vision are convenient to use for water-body detection, but they do not adequately consider the characteristics of the water bodies. Feng et al. [15] proposed a method combined with deep U-Net and a superpixel-based conditional random field model, but the drawback of this method is that it ignores multiscale information and is thus insufficient in complex cases. Guo et al. [16] presented a multiscale feature extractor added to the FCN-based backbone. Duan et al. [17] proposed a multiscale refinement network (MSR-Net) to take advantage of multiscale information and designed an erasing-attention module to embed features during a multiscale refinement strategy. However, none of the studies mentioned above made full use of unsupervised labels. Unsupervised labels are worth more exploration as they do not increase the additional annotation cost.

To address the aforementioned problems, this paper proposes a novel deep semantic segmentation network (DSSN) called the multiscale residual network (MSResNet) by considering the various characteristics of water bodies in optical RS imagery. Different from the most related LinkNet [12], MSResNet uses some parallel branches to obtain multiscale semantic features, thus increasing the sensitivity of different sizes and shapes of water bodies and retaining detailed information about water boundaries. The residual connection prevents the gradient from vanishing. In addition, the center part of MSResNet includes a multiscale dilated convolution (MSDC) module and a multikernel max pooling (MKMP) module. The MSDC module can effectively encode high-level semantic feature maps and gather different levels of context information by controlling the receptive field without changing the size of the feature map. To enhance scale invariance, the MKMP module we propose contains four-level outputs and upsamples the feature maps to obtain the same size features as those in the original feature map via interpolation.

Generally, to obtain a better performance in image segmentation, a large amount of labeled data is required to train the DSSN. With the rapid development of satellites with different types of sensors, many RS images can be easily obtained. However, the annotation and collection of semantic labels are time-consuming and costly and can consume a great deal of manpower and financial resources. As a branch of unsupervised learning methods, self-supervised learning (SSL) methods are proposed to enhance performance when training a deep neural network without sufficient human-annotated labels. In previous studies, many scholars, in applying SSL techniques, defined some pretext tasks (image painting [18], grayscale image colorization [19], predicting image rotation [20] and exemplar-based image prediction [21], relative depth prediction [22] and edge detection [23]) that can only be finished via unsupervised data so that deep neural networks can learn useful image representations. As a result, many SSL methods have recently demonstrated promising results [20,24,25]. For example, Chen et al. [25] presented a simple framework for the contrastive self-supervised learning of visual representations that shows that the composition of data augmentations plays a critical role in defining effective predictive tasks and that the learnable nonlinear transformation between representative and contrastive losses enhances the quality of learned representations. This method considerably outperformed previous methods for self-supervised and semi-supervised learning on ImageNet [26]. Nevertheless, SSL is not widely used in the field of RS, and we are the first to explore and apply SSL to water-body detection.

Massive experiments on the 2020 Gaofen Challenge dataset [27] and the Gaofen image dataset [28] demonstrate that our proposed MSResNet can outperform the state-of-the-art competitive baselines (e.g., FCN [11], LinkNet [12] and HR-Net [13]) and that our SSL can use prior knowledge (i.e., unsupervised invariant facts) to make full use of unlabeled data via geometric transformation learning, noise disturbance learning, image resolution learning and image context fusion learning. Our contributions are three-fold:

1. According to the special characteristics of water bodies, this paper proposes the novel MSResNet which includes two well-designed modules: MSDC and MKMP. More specifically, MSDC benefits the encoding of high-level semantic feature maps and the gathering of different levels of context information, and MKMP helps capture the scale-invariant characteristics of water bodies;
2. In the water-body detection background, this paper, for the first time, explores an SSL strategy to train the DSSN model by fully leveraging unlabeled data. More specifically, four kinds of SSL strategies (e.g., geometric transformation learning, noise disturbance learning, image resolution learning and image context fusion learning) are comprehensively explored and evaluated, providing references for further applications;
3. Our proposed water-body detection method in which MSResNet is combined with the SSL strategy demonstrates improved results compared with state-of-the-art methods operating on two publicly open datasets.

The rest of this paper is organized as follows. Related work is discussed in Section 2. Section 3 describes our proposed method, including the detailed introduction of MSResNet and the details of the SSL strategy (geometric transformation learning, noise disturbance learning, image resolution learning and image context fusion learning). In Section 4, we show the evaluation datasets and metrics and the experimental results. Some discussions are provided in Section 5. Finally, Section 6 concludes this paper.

## 2. Related Work

In the following, we review the existing water-body detection methods from two perspectives: handcrafted feature-based methods and deep learning-based methods.

### *2.1. Handcrafted Feature-Based Water-Body Detection Methods*

In the early years, traditional water-body detection methods developed many indexes for extraction; however, these indexes usually strongly rely on spectral information and pay less attention to contextual and spatial information. For instance, NDWI [8] and the modified normalized difference water index (MNDWI) [29] were developed. Depending on the information of multiband such as short-wave infrared band (SWIR) and so on, Feyisa et al. [30] presented Automated Water Extraction Index (AWEI) and enhanced the performance by distinguishing water-bodies from shadows and other dark surfaces. Fisher et al. [31] designed an accurate water index called $WI_{2015}$ for Landsat data, and comprehensively assessed seven water index approaches to certificate methods' validity. Zhao et al. [32] proposed a spectral information-based approach by forming a decision tree of rules that represented characteristic knowledge for water-body information extraction. Unmanned Aerial Systems (UAS) have become a popular tool in water research [33], and the multispectral images and the Light Detection and Ranging (LiDAR) data that UAS acquired are extensively used in water bodies detection. Jakovljević et al. [34] established an object-based water body delineation method via LiDAR data and automatic water body detection with Sentinel 2 multispectral images. Morsy et al. [35] applied a seeded region growing algorithm to discriminate the land and water automatically for multispectral LiDAR data. However, there are several satellite optical RS images for which more spectrum channels, such as infrared bands, are not available. Especially for high-resolution images, these methods lack robustness. The main reason for this is that high-resolution images contain more spatial details with the enhancement of the image resolution but no infrared band. Due to the complicated distribution of water body appearances and the similarities among water bodies, buildings and shadows, it is difficult to determine an appropriate threshold value when using this method. In addition, methods using the GLCM and SVM [10], and other machine learning methods [36] have been proposed, but the accurate extraction of water bodies from a variety of RS images remains challenging. In conclusion, traditional water-body detection methods based on handcrafted features have limited extraction accuracy and are greatly influenced by the selection of thresholds

and external conditions. The majority of the manual detection features are only suitable for specific situations, which shows that the universality and stability of the methods are poor.

### *2.2. Deep Learning-Based Water Body Detection Methods*

In recent years, due to the large amount of labeled or unlabeled data available, deep learning has become increasingly popular in the computer vision field and has been successfully used for image interpretation. Long et al. [11] proposed the first end-to-end deep learning network for image segmentation, which is called the fully connected network (FCN). Ronneberger O. et al. [14] introduced the encoder-decoder architecture called U-Net, which uses skip connections to extract multiscale information. As a representative encoder-decoder architecture, DeepLab V3+ [37] applied depthwise separable convolution to atrous spatial pyramid pooling and achieved great results. Recently, HR-Net [13] achieved superior results when used on many official datasets. The reason why HR-Net shows the top performance in many fields is that it builds a multilevel representation from a high-resolution representation and applies it to other frameworks. As a popular branch of semantic segmentation, instant semantic segmentation is developing rapidly, DDRNet [38] which includes a novel contextual information extractor achieved state-of-the-art performance on road scenes recently. In the field of RS, image scene classification [39], retrieval [40], segmentation [41] and detection [42,43] have gained great breakthroughs. However, among a variety of classes, water bodies have their own complicated features including detailed information and imagery that is always mixed shadows. As reported in [44], generative adversarial networks (GANs) and U-Net have been proposed to perform efficient segmentations of auto-annotated water-body and land data. In the case of limited training samples, Li et al. [45] used FCN model in water-body detection from very high spatial resolution (VHR) optical images. As shown in [46], a modified structure of the Mask R-CNN that integrates bottom-up and top-down processes for water-body recognition was proposed. Yu et al. [47] presented a DSSN framework for water-body segmentation by considering both its spectral and spatial information. To decrease the error and omissive classification of water boundary pixels, Miao et al. [48] proposed the restricted receptive field deconvolution network (RRF DeconvNet) and a loss function called edge weighting loss (EWLoss) to train the segmentation network to be more sensitive to boundaries. Duan et al. [17] proposed MSR-Net to take advantage of multiscale information and designed an erasing-attention module to effectively embed features during a multiscale refinement strategy. Chen et al. [49] successfully utilized the global spatial-spectral convolution (GSSC) module and the surface water-body boundary refinement (SWBBR) module to enhance the surface features and boundaries of water bodies, respectively. To deal with the characteristics of VHR RS imagery, a method combined with deep U-Net and a superpixel-based conditional random field model was proposed by Feng et al. [15] It was proven that this method could enhance the consistency of connected areas. Li et al. [50] applied the classical DeepLab V3+ and the fully connected conditional random field (CRF) as a postprocessing optimization approach. Zhang et al. [51] proposed MECNet that includes a multi-feature extraction and combination module, and designed the loss function according to different layers in the decoding stage. To enhance the efficiency of water-body variation detection, Wu et al. [52] applied a bag of visual words to capture neighboring feature information to classify pixel-labels and then locate irregular changes, additionally, they utilized distributed processing to accelerate. Although there are many scholars researching the extraction of water bodies from RS imagery, obtaining the different appearances, sizes, shapes and scales of water-body objects such as seas, lakes, small ponds, and narrow streams at the same time remains a challenge. Moreover, although some weakly supervised learning methods have been proposed [53], the existing networks used for water-body detection do not make full use of homologous or heterologous unlabeled images in addition to supervised learning.

## 3. Methodology

In this section, we first describe the structure of MSResNet and explain its characteristics. Then, we provide a detailed introduction of the proposed SSL strategy.

### *3.1. Multiscale Residual Network (MSResNet) for Water-Body Detection*

Due to the different sizes and shapes of water bodies in optical RS images, the key to improving the performance of water-body detections is to utilize multiscale structural information and merge it effectively. Based on this idea, we propose MSResNet to solve this problem, as shown in Figure 1. This network is inspired by LinkNet [12] and similarly designs the structure of the encoder-decoder. However, during the process of multiple downsampling operations in the encoder, some spatial information that is significant for images can be lost. Traditionally, U-Net [14] designs skip connections to strengthen the feature maps and combines low-level detail information from the encoder and high-level semantic information from the decoder. The proposed MSResNet utilizes the residual blocks to replace the normal skip connection, which not only retains more high-resolution detailed information in high-level feature maps by combining low-level features with high-level features but also solves the problems of the exploding gradient and vanishing gradient during the process of training the relatively deep network. The proposed MSResNet uses a pretrained ResNet [54] that retains the feature extraction blocks without the average pooling layer and the fully connected layer as the feature extraction network and obtains a satisfactory performance (in our experiments, we just use ResNet-34 as the feature extractor in consideration of the complexity of the network).

In the central part of MSResNet, there is an MSDC module and an MKMP module.

#### 3.1.1. The Multiscale Dilated Convolution (MSDC) Module

In past semantic segmentation tasks, a deep convolutional layer is confirmed to be able to extract the feature representations from optical RS images. However, due to the limitation of the receptive field, massive context information is ignored along with the significant prior global scenery. To make better use of the multiscale context features and overcome the limitation, we propose the MSDC module as shown in Figure 1b.

We use the different levels of global average pooling (GAP) [55] to obtain the multiscale global context information and reduce the dimension of features by using a $1 \times 1$ convolution layer, which is helpful in distinguishing water from nonwater. To fuse and maintain the multiscale information, we upsample the low-dimension feature maps, obtain the same size features as those in the output feature map, and finally concatenate them as a new feature map. This is different than the D-linkNet [56], which performs well in road extraction tasks, and to maintain the high spatial resolution and expand the receptive field of the information, we modify the filters' fields of view by changing the dilation rate $r$ and making $r$ = 1,2,4,8,16, then adopting six parallel branches to obtain the different receptive fields. In the final steps of each branch, we use one $1 \times 1$ convolution to reduce the dimension of the channels, reduce the number of parameters in MSResNet and accelerate the network convergence. Thus, each layer in the module can learn the characteristics of "sparse" and "not sparse", which represent multiscale traits.

#### 3.1.2. The Multikernel Max Pooling (MKMP) Module

Inspired by spatial pyramid pooling (SPP) [57], which was first applied to natural image classifications and object detections, we proposed the MKMP module for semantic segmentation, as shown in Figure 1a.

Different from the fixed-length representation and fully connected layers of SPP [57], and the invariable kernel size and stride of multiple kernel CNN [58], the proposed MKMP module contains contextual information for four different sizes of receptive fields:

$2 \times 2$, $3 \times 3$, $5 \times 5$ and $6 \times 6$. Without the massive increase of the number of neurons, the multikernel approach considers that multiscale water bodies (such as rivers, ponds, and seas) are problematic for distinguishing them through a single pooling kernel. At the

same time, it also makes use of more context information via expanding receptive fields. Additionally, its four-level output includes feature maps of diverse sizes. Moreover, we utilize a $1 \times 1$ convolution to diminish the dimensional weight after each level of max pooling, reducing the dimension of the feature map to $\frac{1}{n}$ of the original dimension, where n indicates the number of channels in the original feature map. Then, we upsample the low-dimensional feature map to gain the equivalent size through bilinear interpolation. Finally, we fuse the original feature map and four-level outputs to gain a total feature map. Through the MKMP module, local features and global features are fused at the feature map level, enriching the expression ability of the final feature map and improving the accuracy of the water-body detection.

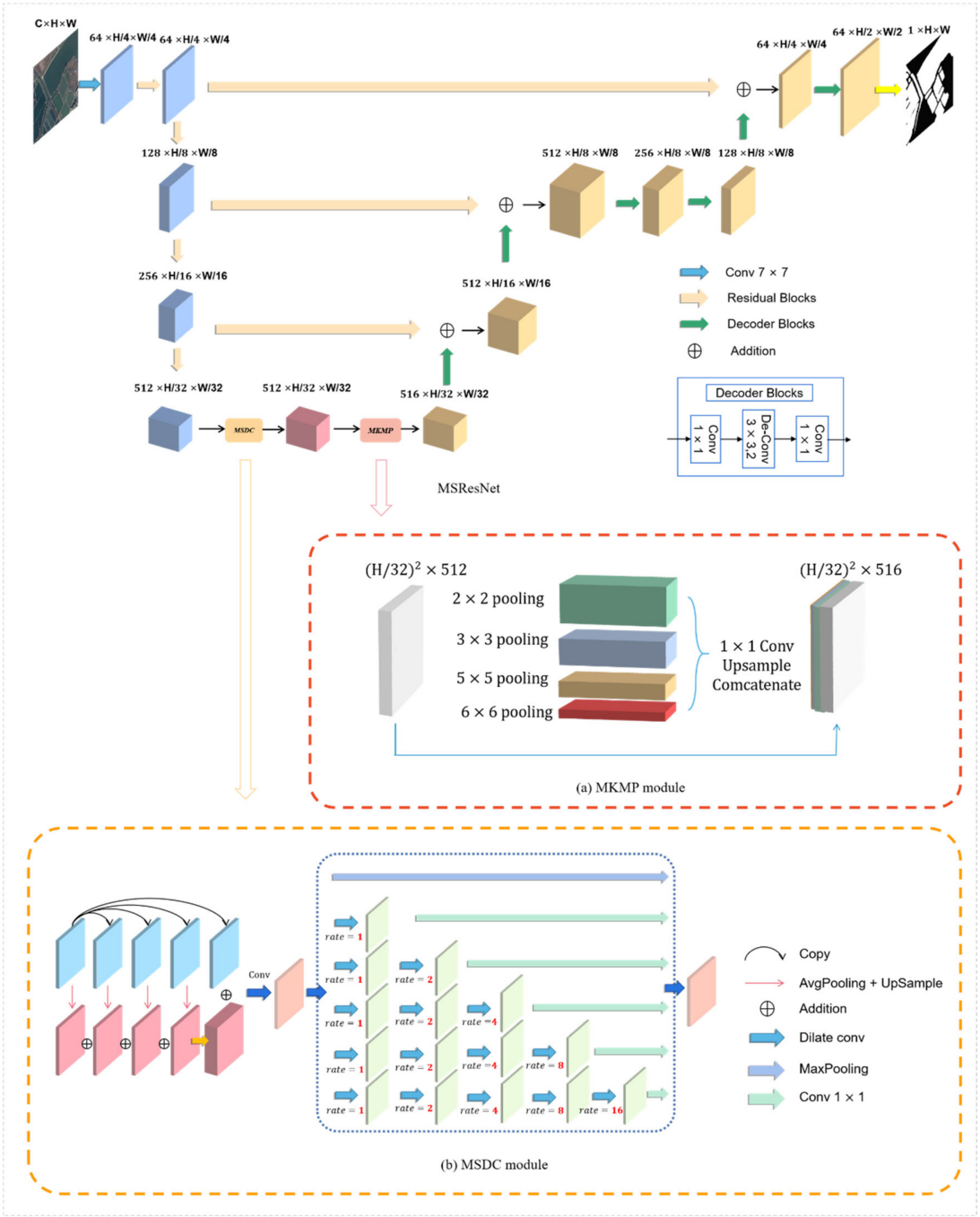



**Figure 1.** The structure of the proposed MSResNet including the (**b**) MSDC module and (**a**) MKMP module.

### *3.2. Self-Supervised Learning (SSL) Strategy for Optimizing MSResNet*

In general, applying a large amount of manually labeled data to train a DSSN and obtain a decent performance is common, but this method is extremely time-consuming and expensive. Additionally, SSL could directly learn the crucial characteristics of data from unlabeled data which is simple to obtain in RS field; thus, according to general knowledge of water bodies and image fusion strategy, we propose a novel SSL strategy for optimizing MSResNet that can make better use of the information extracted from the unlabeled validation and test datasets, and its SSL loss function can constrain the training process of MSResNet. In detail, during the training processing, we propose a comprehensive loss function by mixing the supervised learning loss and SSL loss. The following formula describes the total loss:

$$L = L_{CE}(X_{train}, Y_{train}) + \alpha \cdot L_{SSL}\left(f(X_{all}), T^{-1}(f(T(X_{all})))\right) \quad (1)$$

where $L_{CE}(\cdot)$ represents cross-entropy loss in supervised learning, and $X_{train}$ and $Y_{train}$ represent the images and labels of the training datasets, respectively. $L_{SSL}(\cdot)$ is the generalized self-supervised loss function, including geometric transformation learning loss, noise disturbance learning loss, image resolution learning loss and image context fusion learning loss in this paper. $X_{all}$ represents all images in the training, validation, and test datasets. $f(\cdot)$ and $T(\cdot)$ are the prediction function of the DSSN and the generalized transformation function, respectively. $T^{-1}(\cdot)$ is the inverse transformation output, and $\alpha$ is the weight of the SSL loss.

Considering the prior knowledge of water bodies in optical RS imagery, we propose geometric transformation learning, noise disturbance learning, image resolution learning and image context fusion learning as our SSL strategy, as shown in Figure 2.

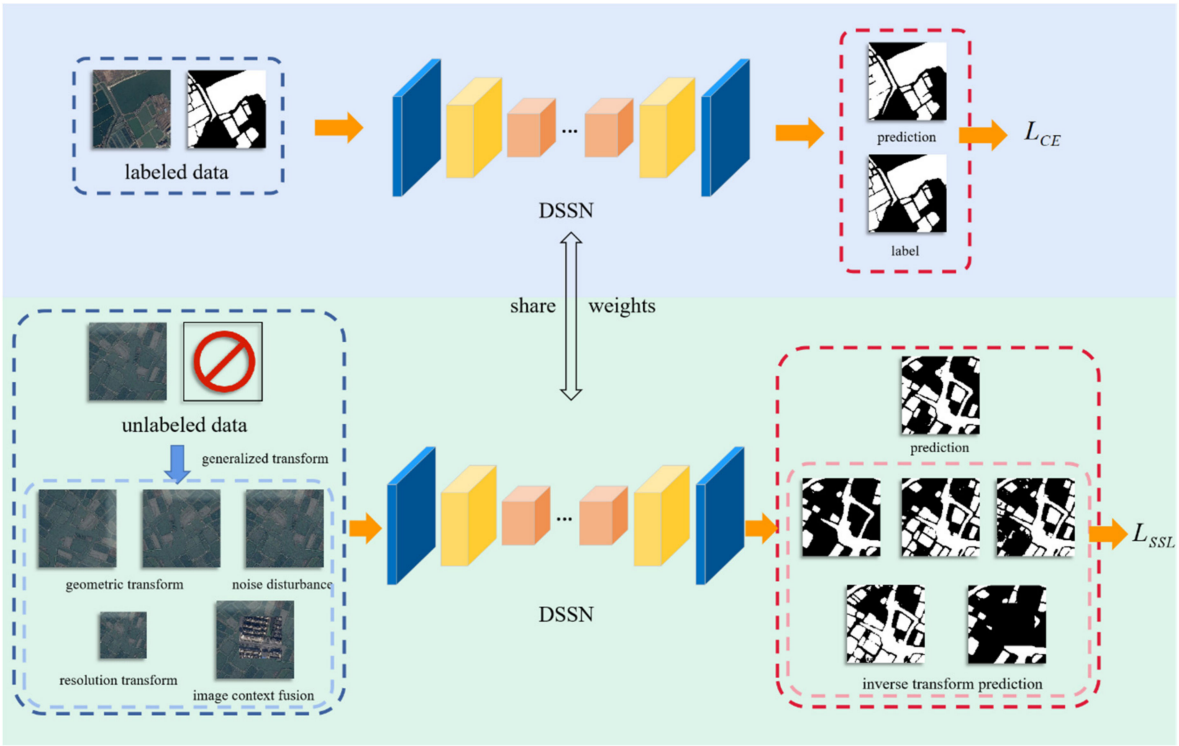



**Figure 2.** The workflow of the proposed SSL strategy, including geometric transformation learning, noise disturbance learning, image resolution learning and image context fusion learning. During the process of training, the total training loss contains supervised cross-entropy loss and SSL loss.

#### 3.2.1. Geometric Transformation Learning

There is no doubt that the classification of each pixel in optical RS imagery is unlikely to change via rotation or flipping. According to this prior knowledge, during the supervised learning of our proposed MSResNet, we rotate the unlabeled images by 90, 180, and 270 degrees to explore the semantic information contained in the imagery and share the network parameters with supervised learning. Similarly, horizontal and vertical flips are also applied to the images, as this can effectively help MSResNet learn the spatial structure information and improve the robustness of the network. These transformations are shown in Figure 3.

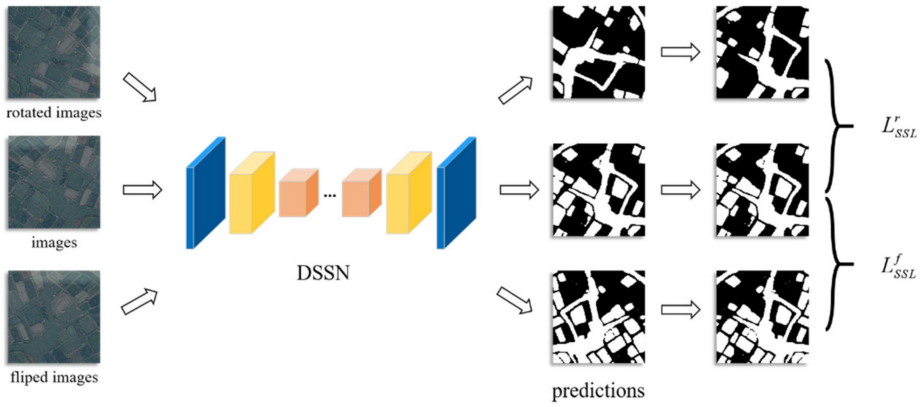



**Figure 3.** The structure of geometric transformation learning.

Simply put, we design the geometric transformations learning loss that is described as Equations (2) and (3) and add it to the total loss:

$$L^r_{SSL} = L^d_{MSE}\left(f(X_{all}), R_d{}^{-1}(f(R_d(X_{all})))\right) \tag{2}$$

$$L^f_{SSL} = L_{MSE}\left(f(X_{all}), F^{-1}(f(F(X_{all})))\right) \tag{3}$$

where $L^r_{SSL}$ and $L^f_{SSL}$ represent the rotation and flip transformation learning loss, respectively, $d$ is the rotation degree, and $R_d(\cdot)$ and $R_d^{-1}(\cdot)$ represent the rotation and counterrotation with $d$ degrees, respectively. $F(\cdot)$ and $F^{-1}(\cdot)$ represent the horizontal and vertical flips and their inverse flips. $L_{MSE}(\cdot)$ represents the mean squared error loss, which encourages the pixel-level consistency of the output under different rotation transforms or flip transforms.

It should be noted that geometric transformation learning contains two transformation approaches: different rotation transforms and horizontal and vertical flip transforms.

#### 3.2.2. Noise Disturbance Learning

Noise disturbance always has an important influence on segmentation performance, so many clear images are widely applied. However, some semantic information can also be obtained from noisy images. We design the strategy of noise disturbance learning based on random noise disturbance that never changes the result of water-body extraction and define the noise disturbance learning loss, which is described as follows:

$$L^n_{SSL} = L_{MSE}(f(X_{all}), f(X_{all} + G(X_{all}))) \tag{4}$$

where $L^n_{SSL}$ represents the noise disturbance learning loss and $G(X_{all})$ represents adding Gaussian noise to the original images and can be defined by $G(X_{all}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_{all}-\mu)^2}{2\sigma^2}\right)$.

3.2.3. Image Resolution Learning

As an important factor in determining the segmentation results, the image spatial resolution is determined by the type of satellite. In general, the higher the image spatial resolution is, the better the water-body detection performance can be. However, a large number of water-body images with low spatial resolution is easily obtained; thus, we apply the upsampling and downsampling functions to change the spatial resolutions of the datasets and help MSResNet learn more spatially detailed information. The image resolution learning loss can be defined as Equation (5):

$$L_{SSL}^{resolution} = L_{MSE}\left(f(X_{all}), S^{-1}(f(S(X_{all})))\right) \tag{5}$$

where $S(\cdot)$ and $S^{-1}(\cdot)$ represent upsampling and downsampling, respectively, which apply nearest-neighbor interpolation. The original image's height and width are $H_{in}$ and $W_{in}$, respectively, and the shape of the output image is $H_{out}$, $W_{out}$, which is sampled with scale factor $\beta$ and is described as follows.

$$H_{\text{out}} = \lfloor H_{\text{in}} \times \beta \rfloor \tag{6}$$

$$W_{\text{out}} = \lfloor W_{\text{in}} \times \beta \rfloor \tag{7}$$

3.2.4. Image Context Fusion Learning

By randomly combining optical RS imagery with water bodies, the shapes and scales of water bodies can be expanded widely, which has the advantage of helping the network learn more contextual information. We design the strategy of image context fusion learning inspired by the classical data augmentation CutMix [59], but it is different in that ground truth is not cut and pasted and we define the consistency difference between two conditions: the prediction of the water-body probability of the CutMix-ed images and the processed water-body confidence by CutMix, as depicted in Figure 4.

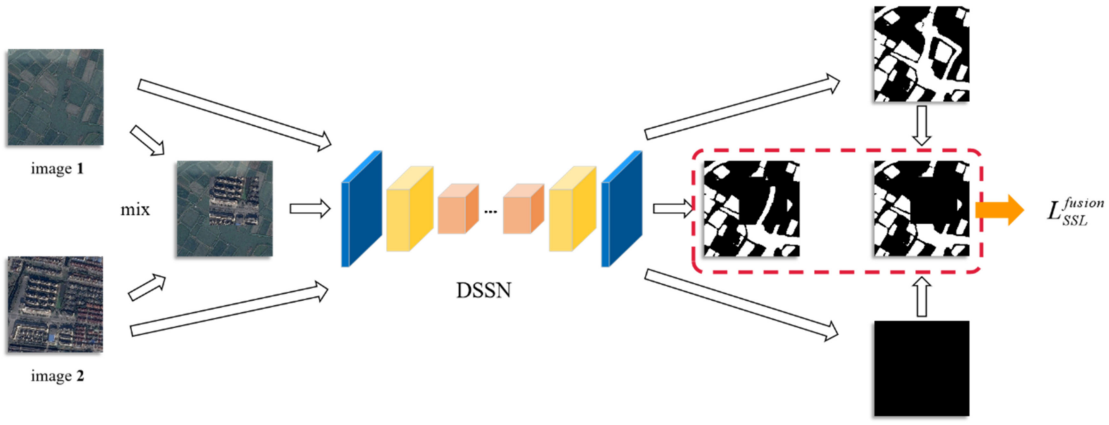



**Figure 4.** The structure of image context fusion learning.

The image context fusion learning loss is defined as follows:

$$L_{SSL}^{fusion} = L_{MSE}\left(f\left(mix\left({X_1}^{all}, {X_2}^{all}\right)\right) - mix\left(f\left({X_1}^{all}\right), f\left({X_2}^{all}\right)\right)\right) \tag{8}$$

where $X_1^{all}$ and $X_2^{all}$ are two randomly selected images among the training, validation, and test images. $mix(\cdot)$ can be described as follows:

$$mix(x_1, x_2) = M \otimes x_1 + (1 - M) \otimes x_2 \tag{9}$$

where $M \in \{0,1\}^{W \times H}$ denotes a binary mask indicating where to drop out and fill in from the two images, 1 represents a binary mask filled with ones and $\otimes$ is elementwise multiplication.

## 4. Experimental Results

In this section, the datasets' descriptions are introduced first. Afterward, the evaluation metrics and details of the experimental settings are shown. Finally, the experimental results and related analyses are given.

### *4.1. Evaluation Datasets*

To verify our methodology, we conducted experiments on the 2020 Gaofen Challenge water-body segmentation dataset [27] and the Gaofen image dataset (GID) [28], and some examples and their labels are shown in Figures 5 and 6, respectively.

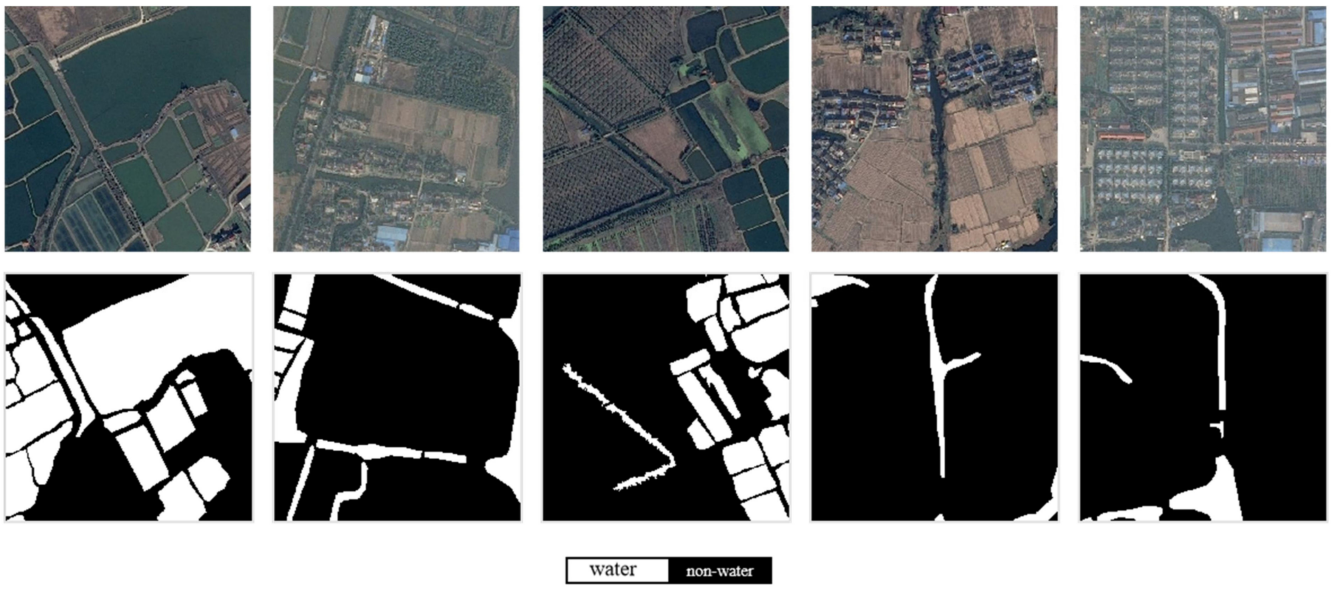



**Figure 5.** Raw images and ground truth masks from the 2020 Gaofen Challenge water-body segmentation dataset.


**Figure 6.** Raw images and ground truth masks from the GID dataset.

The 2020 Gaofen Challenge water-body segmentation dataset contains 1000 RGB images with the size of 492 × 492 containing multiscale and multishape water bodies. The images are pansharpened and have a 0.5 m spatial resolution with no infrared bands. The positive annotations include lakes, ponds, rivers, paddies and so on, while the other pixels are treated as negative. These images and their labels are randomly divided into three sets. In total, 60% are used for training, 20% are used for validation, and 20% are used for testing.

The original GID contains 150 large-scale annotated Gaofen-2 satellite images with the advantages of large coverage, wide distribution and high spatial resolution. To meet the needs and requirements for our experiments, we only select the high intraclass difference images and crop them to a size of 256 × 256, and the original labels are processed. The positive annotations include lakes, ponds, rivers and sea, while pixels of other classes are treated as negative. Finally, we obtain 11700 GF-2 satellite images with NIR-RGB bands. Similarly, these cropped images are randomly divided into a training set, validation set, and test set with 7020, 2340 and 2340 images and proportions of 60%, 20%, and 20%, respectively.

### *4.2. Evaluation Metrics*

In this paper, we calculate the overall accuracy (OA), the mean intersection over union (MIoU), and the frequency weighted intersection over union (FWIoU) to evaluate the performance of the water-body detection [60]. These evaluation indicators can be defined as Equations (10)–(12):

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$MIoU = \frac{1}{n+1}\sum_{i=0}^{n}\frac{TP}{FN + FP + TP} \tag{11}$$

$$FWIoU = \frac{1}{n+1}\sum_{i=0}^{n}\left(\frac{TP_i}{TP_i + TN_i + FN_i}\cdot\frac{TP_i + FN_i}{TP_i + FP_i + TN_i + FN_i}\right) \tag{12}$$

where $TP$, $TN$, $FP$, and $FN$ are the number of true positives, true negatives, false positives, and false negatives, respectively, and $n$ is the number of classes.

### *4.3. Experiment Settings*

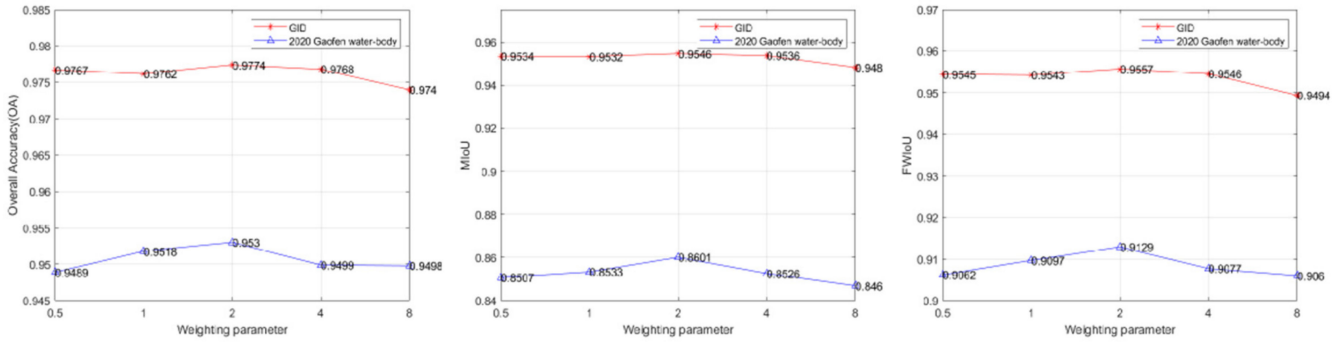
In our experiments, our MSResNet adopts residual blocks 1–4 from ResNet-34 [54], which is initialized with the weights trained on ImageNet [26] and fine-tuned with our datasets for feature extraction; the decoder uses convolution and transposed convolution and a residual block skip connection to replace the normal skip connection. Four SSL strategies we propose are applied in MSResNet. In the experiments on the SSL strategy, we only use images without labels in the training, validation, and test datasets.

During the training of the DSSN, for a fair comparison between our and other methods, all methods are trained with the Adam optimizer [61], and the learning rate is initially set to 0.0001 which gradually decays during the training process. We set the batch-size to 4 or 8 and maximum epoch to 100 and determine the most appropriate epoch according to the performance on validation dataset. Cross-entropy is applied as the supervised loss function, and different self-supervised losses are adopted in the related methods. Extensive experiments, such as sensitivity analyses of the weighting parameters, can help in choosing the most appropriate weighting parameter and examining the effectiveness of our proposed SSL strategy. Moreover, to verify the performance of our proposed method, we compare our proposed MSResNet (without or with different SSL sub-strategies) with LinkNet [12], HR-Net [13], FCN [11], DeepLab V3+ [37], and MECNet [51]. LinkNet is a classical network that uses the structure of an encoder-decoder, and DeepLab V3+ and HR-Net achieve state-of-the-art semantic segmentation performance on natural images. MECNet [51] is a novel network for water-body segmentation and obtains superior performance. All experiments are conducted on the PyTorch framework with a single NVIDIA GeForce RTX 3090 GPU.

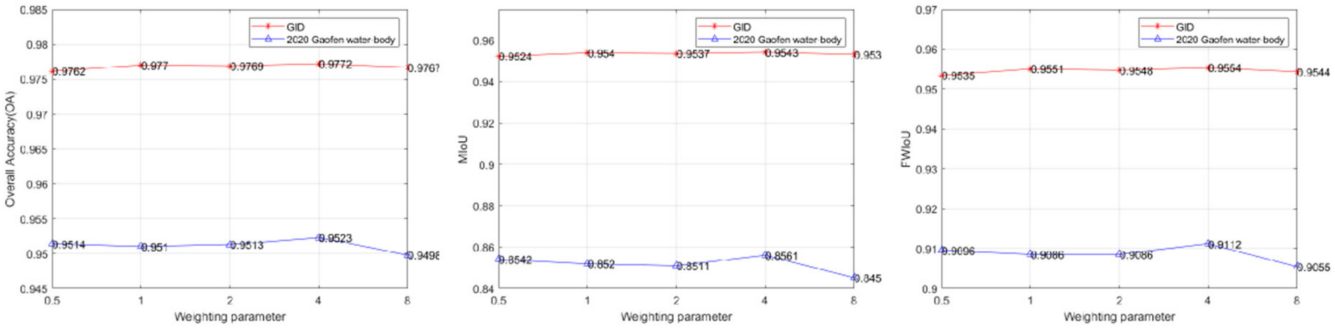
### *4.4. Sensitivity Analysis of the Weighting Parameter*

The weighting Parameter $\alpha$ in Equation (1) is an important hyperparameter that determines the influence degree of the SSL strategy in the total training process. In general, the meaning of loss during the training process indicates the difference between the predicted value and target value. The smaller the loss value is, the closer the predicted detection result is to the label, and the more convergent the DSSN is. In our experiments, the total loss mainly includes supervised cross-entropy loss and self-supervised learning loss, so the weighting parameter indicates the proportion of SSL loss and influences the time of network convergence. To discover the best weighting parameter in the two datasets, we experiment on the performance of our methods in the validation datasets of the 2020 Gaofen Challenge water-body segmentation dataset and GID dataset when $\alpha$ is equal to 0.5, 1, 2, 4, and 8.

Figure 7 shows how the results of the five SSL strategies we propose change with the adjustment of the weight parameters, including OA, MIoU, and FWIoU. In the test of the GID dataset, our MSResNet via noise disturbance learning performs best when the weight coefficient is 0.5. For image resolution learning and image context fusion learning, a weight coefficient of 1 may be the most suitable. Rotation transformation learning and flip transformation learning achieve the best scores when the weight coefficients are 2 and 4, respectively. Similarly, after the sensitivity analysis on the 2020 Gaofen Challenge water-body segmentation dataset, it is concluded that (1) the choice of weight parameter has a greater impact on the detection accuracy; (2) three utilized methods, such as rotation transformation learning, obtain the best performance when the weight parameter is 2; (3) the remaining two methods are more suitable for training with a weight parameter of 4 or 8.

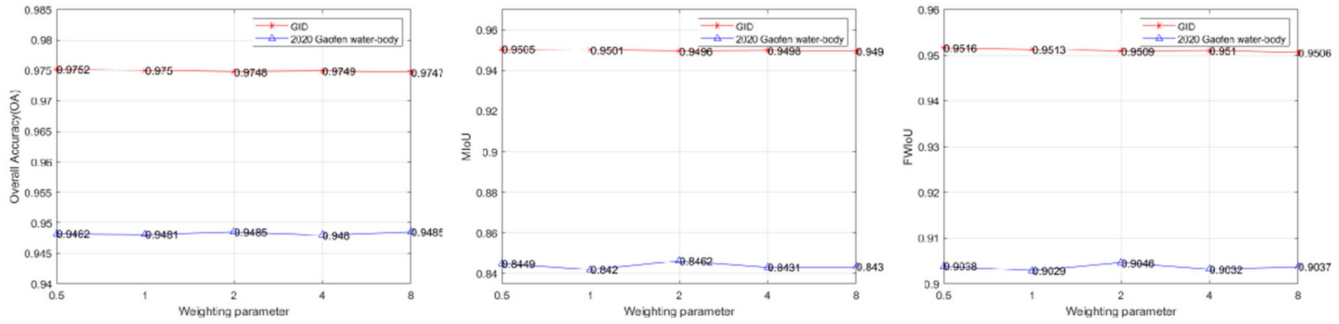



(a) Sensitivity analysis results of rotation transformation learning

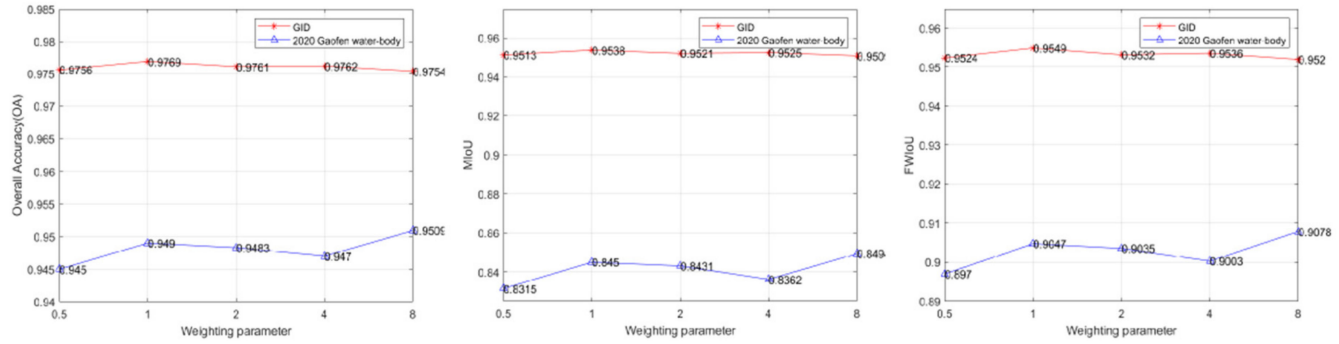



(b) Sensitivity analysis results of flip transformation learning

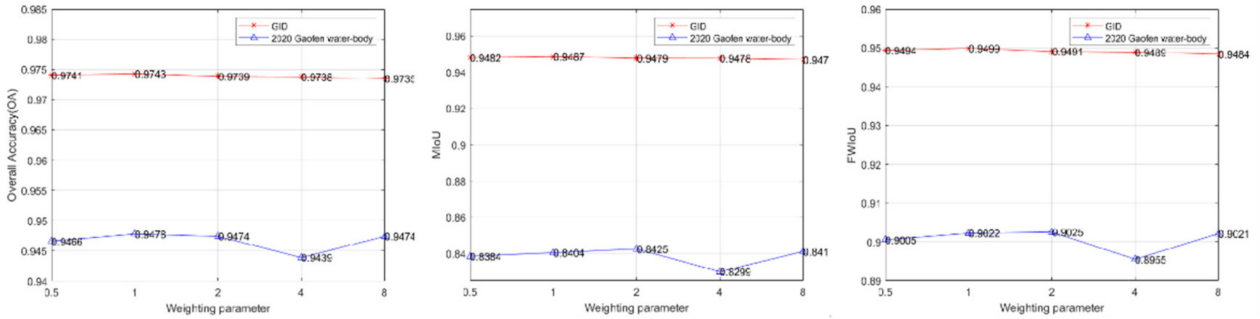
**Figure 7.** *Cont.*




(c) Sensitivity analysis results of noise disturbance learning




(d) Sensitivity analysis results of image resolution learning




(e) Sensitivity analysis results of image context fusion learning

**Figure 7.** The sensitivity analysis results of five different SSL strategies we proposed, including (**a**) sensitivity analysis results of rotation transformation learning, (**b**) sensitivity analysis results of flip transformation learning, (**c**) sensitivity analysis results of noise disturbance learning, (**d**) sensitivity analysis results of image resolution learning, and (**e**) sensitivity analysis results of image context fusion learning.

### *4.5. Ablation Study*

This part verifies the effectiveness of our five proposed SSL strategies and selects the SSL strategy with the best performance on the abovementioned datasets. As shown in Tables 1 and 2, the abovementioned five SSL strategies are proven to play important roles in improving the performance of our MSResNet approach and learning the prior knowledge of optical RS imagery via the SSL strategies we proposed has an important role in enhancing the detection performance. Geometric transformation learning, including rotation transformation learning and flip transformation learning, can achieve especially great performance on the GID dataset and the 2020 Gaofen Challenge water-body segmentation dataset, respectively.

**Table 1.** Quantification of the effectiveness of our MSResNet via five SSL strategies on the 2020 Gaofen Challenge water-body segmentation dataset (%).

| Network Architecture | SSL | OA | MIoU | FWIoU |
|---|---|---|---|---|
| Our MSResNet | - | 94.93 | 85.34 | 90.57 |
| Our MSResNet | Rotation Transformation Learning | 94.94 | 85.69 | 90.68 |
| **Our MSResNet** | **Flip Transformation Learning** | **95.10** | **85.82** | **90.88** |
| Our MSResNet | Noise Disturbance Learning | 94.87 | 85.12 | 90.45 |
| Our MSResNet | Image Resolution Learning | 94.89 | 85.62 | 90.64 |
| Our MSResNet | Image Context Fusion Learning | 95.07 | 85.58 | 90.77 |

**Table 2.** Quantification of the effectiveness of our MSResNet via five SSL strategies on the GID dataset (%).

| Network Architecture | SSL | OA | MIoU | FWIoU |
|---|---|---|---|---|
| Our MSResNet | - | 96.96 | 93.94 | 94.12 |
| **Our MSResNet** | **Rotation Transformation Learning** | **97.47** | **94.94** | **95.08** |
| Our MSResNet | Flip Transformation Learning | 97.39 | 94.76 | 94.91 |
| Our MSResNet | Noise Disturbance Learning | 97.08 | 94.17 | 94.34 |
| Our MSResNet | Image Resolution Learning | 97.36 | 94.71 | 94.86 |
| Our MSResNet | Image Context Fusion Learning | 96.93 | 93.87 | 94.05 |

### *4.6. Comparison with the State-of-the-Art Methods*

To evaluate the performance and measure the time consuming of our proposed MSResNet via SSL strategies, we conduct experiments on the 2020 Gaofen Challenge water-body segmentation dataset and GID dataset. We compare our proposed methods with several outstanding methods, including LinkNet [12], HR-Net [13], FCN [11], DeepLab V3+ [37] and MECNet [51]. For a fair comparison, all compared methods adopt the same device, maximum epoch, learning rate and Adam optimizer [61]. What is worth mentioning is that popular water indexes, such as AWEI [30] and $WI_{2015}$ [31], are not considered because of the spectrum bands limitation of our datasets.

#### 4.6.1. Results for the 2020 Gaofen Challenge Water-Body Segmentation Dataset

The overall accuracy (OA), mean intersection over union (MIoU) and frequency-weighted intersection over union (FWIoU) of the segmentation on the 2020 Gaofen Challenge water-body segmentation dataset are shown in Table 3. We observe that our proposed MSResNet via the SSL strategy achieves higher scores for OA, MIoU, and FWIoU than LinkNet, HR-Net, FCN, DeepLab V3+, and MECNet. In particular, the OA/MIoU/FWIoU increased by 1.03%/2.98%/1.85% for our proposed MSResNet compared with HR-Net, and the OA/MIoU/FWIoU increased by 1.45%/3.48%/2.4% compared with DeepLab V3+, which certifies the effectiveness of our proposed MSResNet.

**Table 3.** Performances for the 2020 Gaofen Challenge water-body segmentation dataset (%).

| Method | OA | MIoU | FWIoU |
|---|---|---|---|
| LinkNet [12] | 93.85 | 82.68 | 88.76 |
| HR-Net [13] | 93.90 | 82.36 | 88.72 |
| FCN [11] | 92.44 | 78.99 | 86.36 |
| DeepLab V3+ [37] | 93.48 | 81.86 | 88.17 |
| MECNet [51] | 94.66 | 84.96 | 90.21 |
| Our MSResNet | 94.93 | 85.34 | 90.57 |
| Our MSResNet via Rotation Transformation Learning | 94.94 | 85.69 | 90.68 |
| **Our MSResNet via Flip Transformation Learning** | **95.10** | **85.82** | **90.88** |

We also observe that our proposed MSResNet via flip transformation learning achieves the best performance among all methods, which improves the OA/MIoU/FWIoU by 0.44%/0.86%/0.67% compared with MECNet and gains an improvement of 0.17%/0.48%/0.31% for OA/MIoU/FWIoU.

Moreover, the visualization of water-body detection results via our methods and other methods are shown in Figure 8; the figure shows the visual results of the LinkNet, HR-Net, FCN, DeepLab V3+, MECNet, our MSResNet, our MSResNet via rotation transformation learning, and our MSResNet via flip transformation learning from (c) to (j). Through these predictions, we can find that our proposed method has a great impact on the accuracy of the water-body detection.

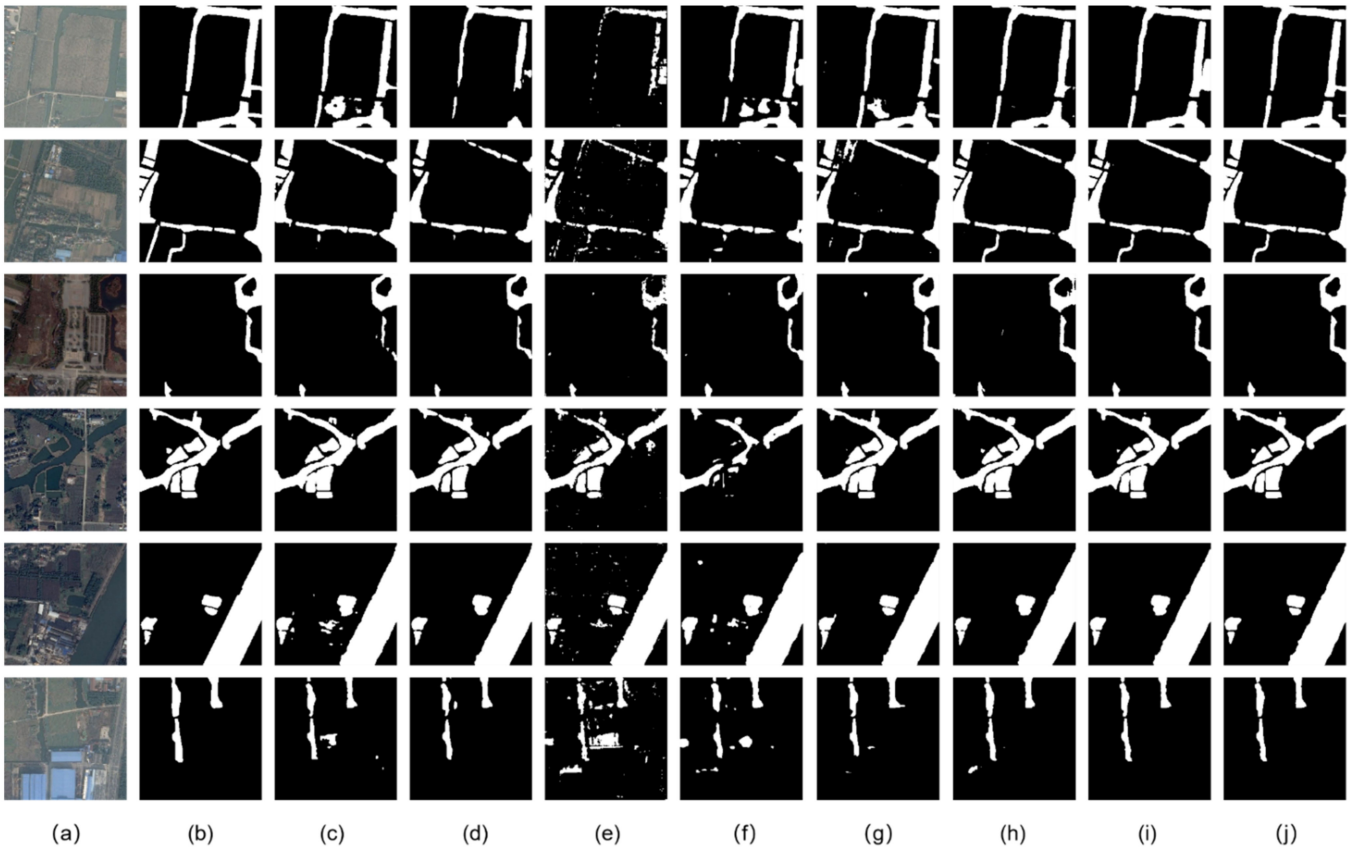



**Figure 8.** Visible water-body detection of the 2020 Gaofem Challenge water-body segmentation dataset. (**a**,**b**) Raw images and ground truth, respectively. (**c**,**d**) The results of LinkNet and the results of HR-Net, respectively. The results of FCN and the results of DeepLab V3+ are shown in (**e**,**f**), respectively. (**g**) The results of MECNet, and (**h**) the results of our MSResNet. The results of our proposed MSResNet via rotation transformation learning and flip transformation learning are displayed in (**i**,**j**), respectively.

Additionally, the running times for detecting water bodies from one single image via our method and other methods on the 2020 Gaofen Challenge water-body segmentation dataset are measured, as shown in Table 4. The time consumption of our MSResNet is relatively short and ranks 4th among all aforementioned methods. Due to the same number of parameters, our proposed MSResNet via SSL strategies do not spend extra time compared with our MSResNet.

**Table 4.** The running times for detecting water bodies from one single image by different methods on the 2020 Gaofen Challenge water-body segmentation dataset.

| Method | Time of Single Image (ms) |
|---|---|
| LinkNet [12] | 89 |
| HR-Net [13] | 117 |
| **FCN** [11] | **86** |
| DeepLab V3+ [37] | 104 |
| MECNet [51] | 119 |
| Our MSResNet | 108 |
| Our MSResNet via Rotation Transformation Learning | 106 |
| Our MSResNet via Flip Transformation Learning | 108 |

#### 4.6.2. Results for the GID Dataset

Table 5 compares the proposed approach with five published network results from the GID dataset. Our proposed MSResNet via rotation transformation learning achieves the best performance of OA/MIoU/FWIoU. In detail, our proposed MSResNet via rotation transformation learning achieves 94.94% on MIoU, while the highest MIoU among other published networks is 93.13%. Compared to MSResNet, our proposed MSResNet via rotation transformation learning obtains an improvement of 1.00% for MIoU and an improvement of 0.96% for FWIoU. It can be observed that the use of our MSResNet via SSL for water-body detection has more image information to increase the detection precision.

**Table 5.** Performances for the GID dataset (%).

| Method | OA | MIoU | FWIoU |
|---|---|---|---|
| LinkNet [12] | 96.17 | 92.49 | 92.64 |
| HR-Net [13] | 96.02 | 92.15 | 92.37 |
| FCN [11] | 94.26 | 88.89 | 89.19 |
| DeepLab V3+ [37] | 96.54 | 93.13 | 93.33 |
| MECNet [51] | 96.44 | 92.94 | 93.15 |
| Our MSResNet | 96.96 | 93.94 | 94.12 |
| **Our MSResNet via Rotation Transformation Learning** | **97.47** | **94.94** | **95.08** |
| Our MSResNet via Flip Transformation Learning | 97.39 | 94.76 | 94.91 |

Figure 9 displays some samples of the water-body detection results and their ground truth for the GID dataset. Figure 9 shows that the results from (c) to (j) are the water-body detection results of LinkNet, HR-Net, FCN, DeepLab V3+, MECNet, our MSResNet, our MSResNet via rotation transformation learning, and our MSResNet via flip transformation learning. The performance for both the 2020 Gaofen Challenge water-body segmentation and GID datasets shows the effectiveness and robustness of our approach.

In addition, Table 6 presents the running times for detecting water bodies from one single image by eight approaches on the GID dataset. The time consumption of our proposed method is relatively short and ranks 3rd among all aforementioned methods.

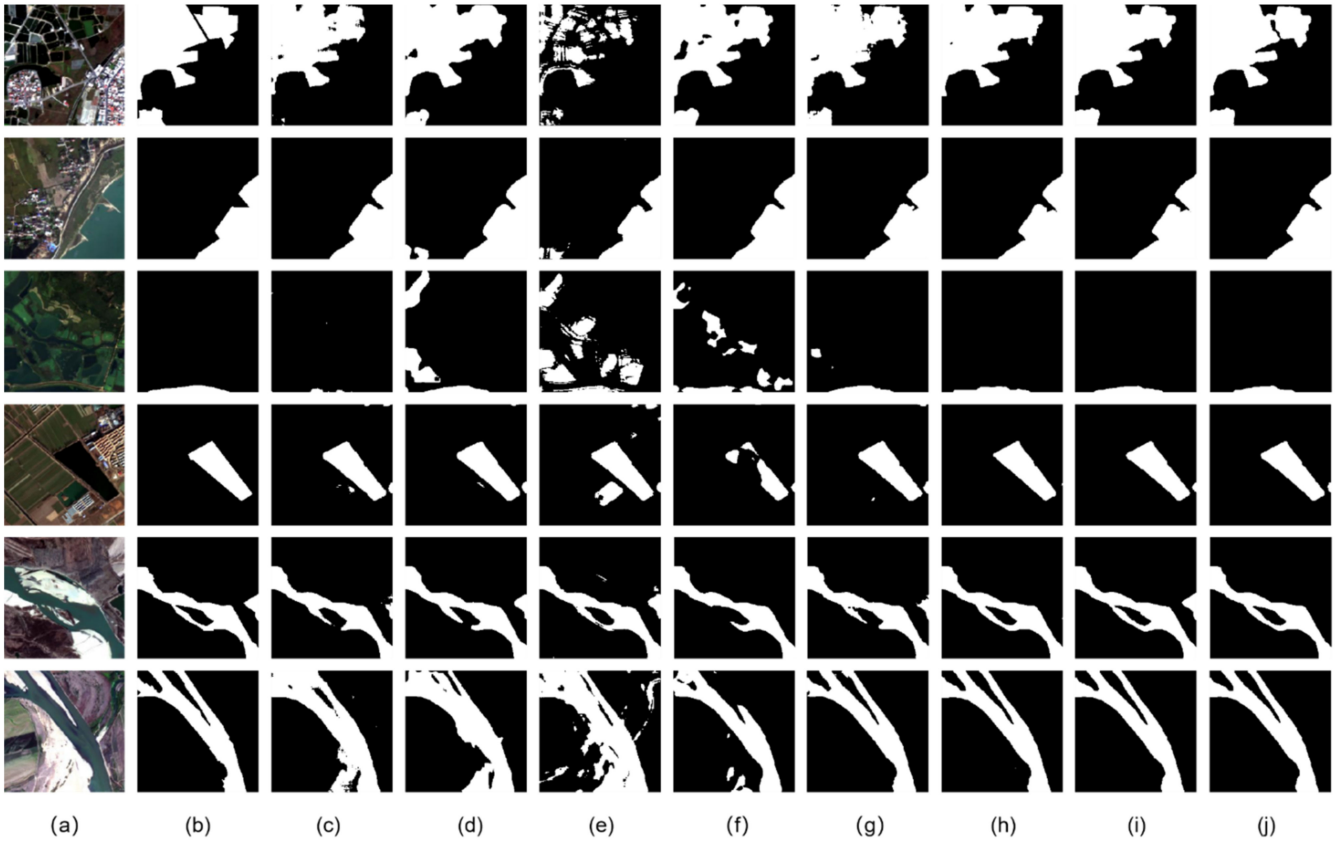



**Figure 9.** Visible water-body detection of the GID dataset. (**a**,**b**) are raw images and ground truth, respectively. (**c**,**d**) are the results of LinkNet and the results of HR-Net, respectively. The results of FCN and the results of DeepLab V3+ are shown in (**e**,**f**), respectively. (**g**) The results of MECNet, and (**h**) the results of our MSResNet. The results of our proposed MSResNet via rotation transformation learning and flip transformation learning are displayed in (**i**,**j**), respectively.

**Table 6.** The times for detecting water bodies from one single image by different methods on the GID dataset.

| Method | Time of Single Image (ms) |
|---|---|
| LinkNet [12] | 11 |
| HR-Net [13] | 29 |
| **FCN [11]** | **7** |
| DeepLab V3+ [37] | 24 |
| MECNet [51] | 23 |
| Our MSResNet | 22 |
| Our MSResNet via Rotation Transformation Learning | 21 |
| Our MSResNet via Flip Transformation Learning | 21 |

## 5. Discussion

The weighting parameter was viewed as a hyperparameter, and we analyzed the detection accuracies of the different kinds of SSL strategies we proposed for validation. We selected the most appropriate weighting parameters and regarded their performance as the results of the corresponding approaches. In the ablation study, we compared our proposed MSResNet and five SSL strategies, exploring the degree of influence on the results. The experimental results compared with the state-of-the-art methods show that our method can achieve the best performance. On the 2020 Gaofen challenge water-body segmentation dataset, our MSResNet via flip transformation learning achieves the best performance, and our OA (95.10%), MIoU (85.82%), and FWIoU (90.88%) are greater than those of the other approaches. In addition, on the GID dataset, our MSResNet via rotation transformation

learning achieves the best results, and our OA (97.47%), MIoU (94.94%), and FWIoU (95.08%) are also greater than those of other approaches. The sensitivity of the weighting parameter has a significant influence on the SSL strategy. We argue that the improvement of MSResNet may result from multiscale and multilayer characteristics, which is beneficial for the water-body detection. Tables 1 and 2 demonstrate the improvements of each SSL strategy. (1) The majority of SSL strategies we mention are suitable for two water-body segmentation datasets, revealing the universal adaptability and robustness of our method. (2) Although we do not utilize the labels of the test or validation datasets, we successfully extract the helpful imagery information.

However, we observe in Table 1 that the improvement of our approaches is not vast. Our best method on the 2020 Gaofen Challenge water-body segmentation dataset merely gains an improvement of 0.17%/0.48%/0.31% for OA/MIoU/FWIoU, respectively. The main reason may be that the number of test images in the 2020 Gaofen Challenge water-body segmentation dataset is not enough. Figure 10 shows the visual results of water-body detection. It is not difficult to see that our methods can pay more attention to the spatial structure information and have common adaptability of different scales of water bodies. However, it is still a challenging task for our proposed method to more easily identify categories such as water bodies, farms and barren areas. As shown in the red circles in Figure 10, the results of our MSResNet via flip transformation learning are better than other methods, but all methods cannot adapt to all conditions.

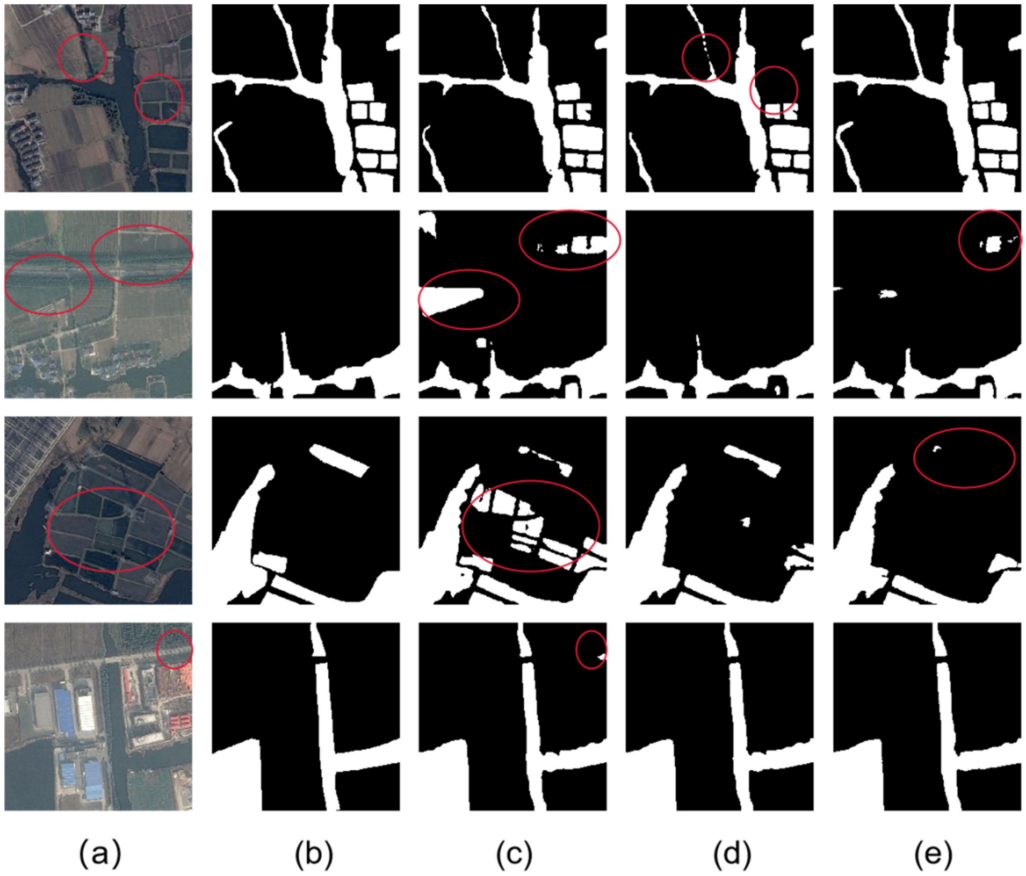



**Figure 10.** The results of water-body detection. (**a**) Raw image, (**b**) ground truth, (**c**) results of our MSResNet via rotation transformation learning and (**d**) results of our MSResNet via flip transformation learning, (**e**) results of our MSResNet via rotation transformation learning, flip transformation learning, noise disturbance learning and image resolution learning.

## 6. Conclusions

Due to their excellent feature-extraction abilities, DSSNs have been widely used in water-body detections of RS images and have achieved great success. However, a general DSSN is insufficient to take the full characteristics of water bodies into account. Moreover, a large amount of unlabeled data is ignored during the training process. Accurate water-body detection remains a challenging task. Thus, in this paper, according to the characteristics of water-body detections, we proposed MSResNet via SSL for water-body detections. To apply multiscale structural information and merge it effectively, we designed an MSDC module and an MKMP module that maintain a high spatial resolution and expand the receptive field. In addition, novel SSL strategies, including geometric transformation learning, noise disturbance learning, image resolution learning and image context fusion learning, were able to flexibly utilize the image information of test and validation datasets and consider the prior knowledge about image contexts by adding the self-supervised learning loss to the total loss during the process of training.

To verify our method, the 2020 Gaofen Challenge water-body segmentation dataset (images with RGB bands) and the GID dataset (images with NIR-RGB bands) were used to analyze our method and compare it with existing approaches. The results confirmed that our method achieved the best performance and was greatly improved compared to other methods. To be more specific, our proposed approach achieves the state-of-the-art performance on the 2020 Gaofen Challenge water-body segmentation dataset, with OA, MIoU, and FWIoU of 95.10%, 85.82%, and 90.88%. Likewise, the performance of our method is superior to other five methods on the GID dataset, which means OA, MIoU and FWIoU achieve 97.47%, 94.94%, and 95.08%, respectively. In future work, we may explore more SSL strategies and a combination of different SSL strategies and try to utilize heterogeneous unlabeled images for SSL.

**Author Contributions:** Conceptualization, Y.L.; methodology, Y.L., B.D.; validation, B.D.; writing—original draft preparation, B.D.; writing—reviewing and editing, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the National Natural Science Foundation of China under grant 41971284; the State Key Program of the National Natural Science Foundation of China under grants 42030102 and 92038301; the Foundation for Innovative Research Groups of the Natural Science Foundation of Hubei Province under grant 2020CFA003.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The source code will be made publicly available at https://github.com/Jack-bo1220/MSResNet-via-SSL (accessed on 4 August 2021). The experiments were conducted on publicly open datasets. The download sites of the publicly open datasets can be found in the corresponding published papers; we do not repeat these sites here.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Li, Y.; Ma, J.; Zhang, Y. Image retrieval from remote sensing big data: A survey. *Inf. Fusion* **2021**, *67*, 94–115. [CrossRef]
2. Chi, M.; Plaza, A.; Benediktsson, J.A.; Sun, Z.; Shen, J.; Zhu, Y. Big data for remote sensing: Challenges and opportunities. *Proc. IEEE* **2016**, *104*, 2207–2219. [CrossRef]
3. Ma, Y.; Wu, H.; Wang, L.; Huang, B.; Ranjan, R.; Zomaya, A.; Jie, W. Remote sensing big data computing: Challenges and opportunities. *Future Gener. Comput. Syst.* **2015**, *51*, 47–60. [CrossRef]
4. Huang, C.; Chen, Y.; Zhang, S.; Wu, J. Detecting, extracting, and monitoring surface water from space using optical sensors: A review. *Rev. Geophys.* **2018**, *56*, 333–360. [CrossRef]
5. Chen, L.; Zhang, P.; Xing, J.; Li, Z.; Xing, X.; Yuan, Z. A multi-scale deep neural network for water detection from SAR images in the mountainous areas. *Remote Sens.* **2020**, *12*, 3205. [CrossRef]
6. Zhang, J.; Xing, M.; Sun, G.-C.; Chen, J.; Li, M.; Hu, Y.; Bao, Z. Water body detection in high-resolution SAR images with cascaded fully-convolutional network and variable focal loss. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 316–332. [CrossRef]

7. Balajee, J.; Durai, M.A.S. Detection of water availability in SAR images using deep learning architecture. *Int. J. Syst. Assur. Eng. Manag.* **2021**, 1–10. [CrossRef]
8. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
9. Frazier, P.S.; Page, K.J. Water body detection and delineation with Landsat TM data. Photogrammetric engineering and remote sensing. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 1461–1468.
10. Lv, W.; Yu, Q.; Yu, W. Water extraction in SAR images using GLCM and support vector machine. In Proceedings of the IEEE 10th International Conference on Signal Processing, Beijing, China, 24–28 October 2010.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
13. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
14. 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, 9 October 2015; pp. 234–241.
15. Feng, W.; Sui, H.; Huang, W.; Xu, C.; An, K. 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.* **2018**, *16*, 618–622. [CrossRef]
16. Guo, H.; He, G.; Jiang, W.; Yin, R.; Yan, L.; Leng, W. A Multi-Scale Water Extraction Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 189. [CrossRef]
17. Duan, L.; Hu, X. Multiscale Refinement Network for Water-Body Segmentation in High-Resolution Satellite Imagery. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 686–690. [CrossRef]
18. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
19. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 649–666.
20. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv* **2018**, arXiv:1803.07728.
21. Dosovitskiy, A.; Springenberg, J.T.; Riedmiller, M.; Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 766–774. [CrossRef]
22. Jiang, H.; Larsson, G.; Shakhnarovich, M.M.G.; Learned-Miller, E. Self-supervised relative depth learning for urban scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 19–35.
23. Li, Y.; Paluri, M.; Rehg, J.M.; Dollár, P. Unsupervised learning of edges. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1619–1627.
24. Jing, L.; Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *1*. [CrossRef]
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Shangri-La, China, 11–13 September 2020; pp. 1597–1607.
26. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
27. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS4Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5398–5413. [CrossRef]
28. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Learning transferable deep models for land-use classification with high-resolution remote sensing images. *arXiv* **2018**, arXiv:1807.05713.
29. Han-Qiu, X. A study on information extraction of water body with the modified normalized difference water index (MNDWI). *J. Remote Sens.* **2005**, *5*, 589–595.
30. Feyisa, G.L.; Meilby, H.; Fensholt, R.; Proud, S. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sens. Environ.* **2014**, *140*, 23–35. [CrossRef]
31. Fisher, A.; Flood, N.; Danaher, T. Comparing Landsat water index methods for automated water classification in eastern Australia. *Remote Sens. Environ.* **2016**, *175*, 167–182. [CrossRef]
32. Zhao, X.; Wang, P.; Chen, C.; Jiang, T.; Yu, Z.; Guo, B. Waterbody information extraction from remote-sensing images after disasters based on spectral information and characteristic knowledge. *Int. J. Remote Sens.* **2017**, *38*, 1404–1422. [CrossRef]
33. Vélez-Nicolás, M.; García-López, S.; Barbero, L.; Ruiz-Ortiz, V.; Sánchez-Bellón, Á. Applications of unmanned aerial systems (UASs) in hydrology: A review. *Remote Sens.* **2021**, *13*, 1359. [CrossRef]
34. Jakovljević, G.; Govedarica, M. Water Body Extraction and Flood Risk Assessment Using Lidar and Open Data. In *Climate Change Management*; Springer Science and Business Media LLC: Cham, Switzerland, 2019; pp. 93–111.

35. Morsy, S.; Shaker, A.; El-Rabbany, A. Using Multispectral airborne lidar data for land/water discrimination: A case study at Lake Ontario, Canada. *Appl. Sci.* **2018**, *8*, 349. [CrossRef]
36. Nandi, I.; Srivastava, P.K.; Shah, K. Floodplain mapping through support vector machine and optical/infrared images from Landsat 8 OLI/TIRS sensors: Case study from Varanasi. *Water Resour. Manag.* **2017**, *31*, 1157–1171. [CrossRef]
37. 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.
38. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
39. Li, Y.; Zhang, Y.; Zhu, Z. Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification. *IEEE Trans. Cybern.* **2021**, *51*, 1756–1768. [CrossRef] [PubMed]
40. Tong, X.-Y.; Xia, G.-S.; Hu, F.; Zhong, Y.; Datcu, M.; Zhang, L. Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation. *IEEE Trans. Big Data* **2020**, *6*, 507–521. [CrossRef]
41. Li, Y.; Shi, T.; Zhang, Y.; Chen, W.; Wang, Z.; Li, H. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 20–33. [CrossRef]
42. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. Cfc-net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images. *arXiv* **2021**, arXiv:2101.06849.
43. Li, Y.; Chen, W.; Zhang, Y.; Tao, C.; Xiao, R.; Tan, Y. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. *Remote Sens. Environ.* **2020**, *250*, 112045. [CrossRef]
44. Pai, M.M.M.; Mehrotra, V.; Verma, U.; Pai, R.M. Improved semantic segmentation of water bodies and land in SAR images using generative adversarial networks. *Int. J. Semant. Comput.* **2020**, *14*, 55–69. [CrossRef]
45. Li, L.; Yan, Z.; Shen, Q.; Cheng, G.; Gao, L.; Zhang, B. Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks. *Remote Sens.* **2019**, *11*, 1162. [CrossRef]
46. Song, S.; Liu, J.; Liu, Y.; Feng, G.; Han, H.; Yao, Y.; Du, M. Intelligent object recognition of urban water bodies based on deep learning for multi-source and multi-temporal high spatial resolution remote sensing imagery. *Sensors* **2020**, *20*, 397. [CrossRef] [PubMed]
47. 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. Appl.* **2017**, *16*, 1750001. [CrossRef]
48. 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**, *15*, 602–606. [CrossRef]
49. Chen, Y.; Tang, L.; Kan, Z.; Bilal, M.; Li, Q. A novel water body extraction neural network (WBE-NN) for optical high-resolution multispectral imagery. *J. Hydrol.* **2020**, *588*, 125092. [CrossRef]
50. Li, Z.; Wang, R.; Zhang, W.; Hu, F.; Meng, L. Multiscale features supported Deeplabv3+ optimization scheme for accurate water semantic segmentation. *IEEE Access* **2019**, *7*, 155787–155804. [CrossRef]
51. Zhang, Z.; Lu, M.; Ji, S.; Yu, H.; Nie, C. Rich CNN Features for water-body segmentation from very high resolution aerial and satellite imagery. *Remote Sens.* **2021**, *13*, 1912. [CrossRef]
52. Wu, Y.; Han, P.; Zheng, Z. Instant water body variation detection via analysis on remote sensing imagery. *J. Real Time Image Process.* **2021**, 1–14. [CrossRef]
53. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly Supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 1–26 July 2016; pp. 770–778. [CrossRef]
55. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
56. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
58. Shi, H.; Wang, H.; Jin, Y.; Zhao, L.; Liu, C. Automated heartbeat classification based on convolutional neural network with multiple kernel sizes. In Proceedings of the 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 4–9 April 2019; pp. 311–315.
59. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 6023–6032.
60. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
61. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.