remote sensing

MDPI

Article

MSNet: Multifunctional Feature-Sharing Network for Land-Cover Segmentation

Liguo Weng [1,*], Jiahong Gao [1], Min Xia [1,2] and Haifeng Lin [2]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China
* Correspondence: 002311@nuist.edu.cn

Abstract: In recent years, the resolution of remote sensing images, especially aerial images, has become higher and higher, and the spans of time and space have become larger and larger. The phenomenon in which one class of objects can produce several kinds of spectra may lead to more errors in detection methods that are based on spectra. For different convolution methods, downsampling can provide some advanced information, which will lead to rough detail extraction; too deep of a network will greatly increase the complexity and calculation time of a model. To solve these problems, a multifunctional feature extraction model called MSNet (multifunctional feature-sharing network) is proposed, which is improved on two levels: depth feature extraction and feature fusion. Firstly, a residual shuffle reorganization branch is proposed; secondly, linear index upsampling with different levels is proposed; finally, the proposed edge feature attention module allows the recovery of detailed features. The combination of the edge feature attention module and linear index upsampling can not only provide benefits in learning detailed information, but can also ensure the accuracy of deep feature extraction. The experiments showed that MSNet achieved 81.33% MIoU on the Landover dataset.

Keywords: land-cover detection; hyperspectral; remote sensing images; feature fusion

check for updates

Citation: Weng, L.; Gao, J.; Xia, M.; Lin, H. MSNet: Multifunctional Feature-Sharing Network for Land-Cover Segmentation. *Remote Sens.* **2022**, *14*, 5209. https://doi.org/10.3390/rs14205209

Academic Editor: Xiaodong Li

Received: 25 August 2022
Accepted: 6 October 2022
Published: 18 October 2022

1. Introduction

Remote sensing image processing technology has played an important role in the study of urban and rural land conditions [1,2]. Accurate information on land cover is a key data resource for urban planning and other fields [3]. Semantic segmentation in aerial orthophoto images is very important for the detection of the real-time situations of buildings, plants, and surface water. Existing land-cover segmentation models still have some defects. Among the archaic remote sensing image segmentation practices, the valuation measure from mathematical statistics has been widely used. Methods that use this obtain the mean and variance of each category in the target area by learning the receptive field so as to obtain the classification results. Other relevant methods rely on the spectral discrimination ability [4] of a training model to obtain the spatio-temporal features of an image. However, with the progress of science and technology, the resolution of remote sensing images continues to improve, and the spectral features are becoming more and more complex; as a result, small differences in different objects of the same category will have a great impact on the segmentation results. Therefore, only using spectral features to extract targets is often not enough. Classification algorithms based on machine learning, such as width learning [5], non-deep neural networks [6], and so on, are not suitable for large amounts of data. When used for detection, target maps only undergo a small amount of linear or nonlinear transformation, but the classification of complex detailed information and high-order semantic information by the above means can be terrible. Especially for hyperspectral remote sensing maps with large feature differences and numerous space–time indices, the above classification results are not satisfactory. For this type of land detection analysis [7], aerial images are widely used. There are many ways to occupy land, and the influence of tall architecture on low architecture is complex and changeable; for vegetation, areas covered by shrubs are difficult to separate from forest areas [8]. In addition, a wide variety of trees grow in different ways and soil types. Water is divided into living water and dead water, including natural pools and man-made ponds, but ditches and riverbeds are excluded. These features make it very difficult to extract features from remote sensing images. Last but not least, the above traditional methods [3,5,9] usually require manual calculation of the statistics of the obtained parameters, which further increases the complexity of deeper feature learning. Remote sensing images are developing towards higher resolutions and larger space–time spans. The same object with different spectra and different objects with the same spectrum can make classification more difficult. To sum up, the conventional means of land detection have limited feature-mining abilities and fail to adapt to different datasets.

In recent years, deep learning has been widely used in the field of remote sensing image analysis for land cover [9–12] and other applications [13]. When it comes to deep learning, since Long et al. [14] published a full convolution neural network (FCN, 2015), many achievements of scientific research based on pixel classification have emerged. For instance, Ronneberger et al. [15] proposed a UNet that can obtain detailed contextual information. The pyramid aggregation proposed by Zhao et al. [16] can integrate the contextual information of different regions so as to enhance the mechanisms to learn the overall characteristics, similarly to PSPNet [17] and DeepLabV3+ [18]. However, the excessive amounts and complexity of calculations made by the model restrict the experimental equipment and cause a certain waste of resources [19]. Therefore, Andrew and others proposed MobileNet (lightweight network) to alleviate the computational pressure, but the full release of the efficiency of the model is still a major problem faced by researchers. For example, Zhang et al. [20] proposed a channel shuffle module in 2017, which released the potential of the model by shuffling and recombining; Yu et al. proposed a feature fusion module (FFM) [21] and attention refinement module (ARM) [22] in 2020 to balance accuracy and speed; the selective kernel proposed by Li et al. in 2019 used an attention mechanism on the convolution kernel, allowing the network to select its own suitable convolution kernel. The difficulty in this kind of research is the enhancement of the accuracy of the model on the premise of limiting the weight of the model. Fully releasing the performance of the model [23] is our research direction.

Considering the existing problems, a multifunctional feature-sharing network for land-cover segmentation is proposed in this paper. For depth feature mining, an SFR module is proposed. Inspired by the residual structure [24], we take the output result of the shuffle unit as the input of ResNet and change the numbers of input and output channels. In this way, even if the SFR blocks are stacked many times, the amounts of calculations (flops) can be strictly limited to about 1G. Another branch that we propose is a linear index upsampling branch with different levels to guide upsampling after continuous downsampling, which saves the process of learning upsampling. At the same time, the outputs after two SFR modules and three SFR modules are processed by the EFAModule and then fused with the output of the last downsampling of this branch to extract logical features and reply high-resolution detailed information to improve the learning effects of the detail features and edge information. The introduction of the EFAModule can alleviate the mutual occlusion caused by different objects with the same spectrum and can also greatly avoid the influence of the shadows of high-level objects on low-level waters. The experimental results show that this branch can ensure the accuracy of deep feature extraction, and the fusion of the two branches has better performance: The average intersection union ratio (MIoU) of the multi-functional feature-sharing network is higher than that of other networks. In general, this work has three contributions:

- One branch is a linear index upsampling branch with different levels. There is no need to learn upsampling. A trainable convolution kernel is used for convolution operations to acquire a complex feature map, which not only limits the amounts of calculations, but also ensures the integrity of high-frequency information.
- The other branch combines a shuffle unit with a skip connection. Channel rearrangement makes the extraction of information more well distributed, and the residual structure ensures the accuracy of deep semantic information extraction. This branch extracts key features and pays attention to the dependencies between contexts by using the logical relationships [25] between information within a class and between classes.
- After processing of the two SFR modules, the EFAModule is introduced to extract logical features and reply high-resolution detailed information, and good results in the learning of detailed information and edge information of a feature map were achieved.

The network structure enables the model to better integrate global information, and enhances the extraction of intra-class information and inter-class information. The MLNet model improves the average accuracy (MIoU) by 1.19–6.27%, with only 10–20% of the weight of other models, such as PSP and Deeplabv3+.

2. Land-Cover Segmentation Methodology

With the improvement of remote sensing image resolution, the amount of detailed information in remote sensing images has also greatly increased. Therefore, frameworks that are applicable to land cover have great room for progress from the perspectives of detailed information and upsampling feature fusion [26]. Fully releasing the efficiency of models based on the above is the research direction of this paper.

2.1. Network Architecture

This paper proposes a special image segmentation network. Firstly, we propose a residual shuffle reorganization branch. This branch learns the deep-level information of images in the order of the channels, pays attention to the logical relationship between intra-class information and inter-class information, and reduces misclassifications of the same object and misdetection of different objects. Secondly, we propose a linear index upsampling branch with different levels, and it does not need to learn upsampling. A trainable convolution kernel is used for the convolution operation to generate a dense feature map and fully extract the semantic information of the target feature map. Then, the EFAModule is introduced to strengthen the recognition of class information and accurately segment the edge information. The feature map processed by the SFR and EFAModule is fused with the downsampled feature map of the linear index upsampling branch with different levels, which effectively limits the amounts of calculations on the premise of ensuring the integrity of high-frequency information [27]. Finally, the output of the linear index upsampling branch with different levels is fused with the output of the SFR branch, and the final prediction map is generated [28]. The two-way fused MSNet has better performance, and its mean intersection over union (MIoU) is higher than that of other networks. Its hidden units in each graph convolutional layer are explicitly indicated in Table 1, and its overall architecture is described in the following (Figure 1):

Table 1. Hidden units in each graph convolutional layer.

Branch 1	Branch 2	Middle Branch
3 × 3 conv	shuffleNet ×2 +PReLU, 16	1 × 3 conv, 3 × 1 conv
maxpool	shuffleNet ×2 +PReLU, 28	3 × 1 conv, 1 × 3 conv
groups conv + point conv	shuffleNet ×2 +PReLU, 40	EFAModule
deconv ×2	shuffleNet ×2 +PReLU, 56	
softmax		

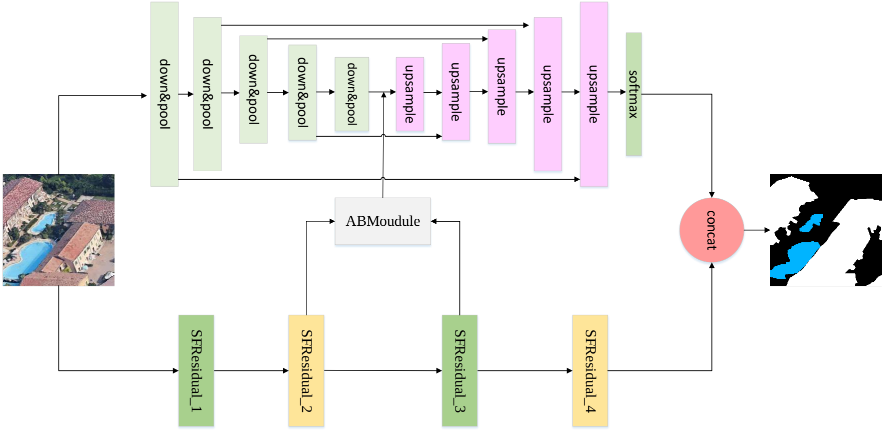

Figure 1. Diagram of the structure of MSNet.

2.2. SFResidual

Inspired by shufflenet [29], we took the output of the shuffle unit as the input of the residual structure to form the SFResidual module, as shown in Figure 2. The shuffle unit's shuffling can cover the global information and make the extraction of information more uniform [30]; the residual structure uses the classic "skip connection", which can efficiently complete the recognition task with a large number of classifications, so the introduction of the residual structure can make up for the deficiencies of a lightweight network. The fusion of the two can greatly improve the spectral recognition ability, alleviate the problem of image misclassification, and greatly improve the accuracy of segmentation. The structure is shown in Figure 3.

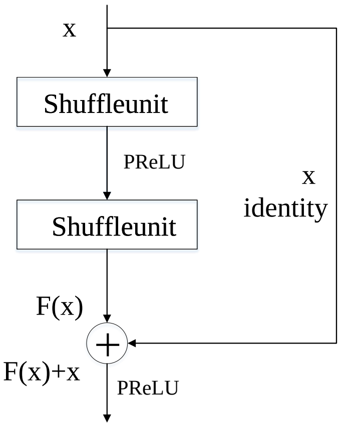

Figure 2. Structure of the SFResidual module.

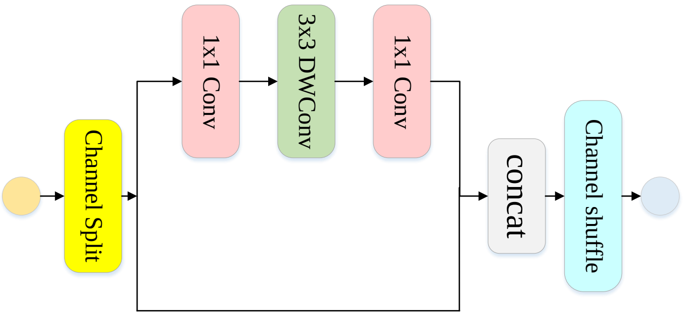

Figure 3. Structure of the shuffle unit.

2.3. LIU Branch

Note that with more convolutional layers, more corresponding features will be extracted, but a very deep network will cause gradient disappearance and gradient expulsion. VGG-16 uses convolutions to simulate fully connected layers, which can effectively alleviate this problem, so we propose a linear index upsampling branch with different levels (LIU) to optimize VGG-16 [31,32] and to achieve a better improvement.

2.4. EFAModule

Generally speaking, dilation convolution is used to prevent the loss of spatial hierarchical information. A convolution with a dilation rate of 2 and a convolution kernel of 3 actually becomes a 7 × 7 convolution, which will also produce grid effects when increasing the receptive field. In order to reduce the influence of similar problems, an EFA unit is proposed based on the lightweight structure of BiSeNet [33]. As shown in Figure 4, we adopt a two-branch model composed of strip convolution. One branch is used to obtain local information, and the other branch introduces dilation parameters to obtain edge semantic information. The comprehensive extraction of multiscale information enhances the ability to learn the module's edge information.

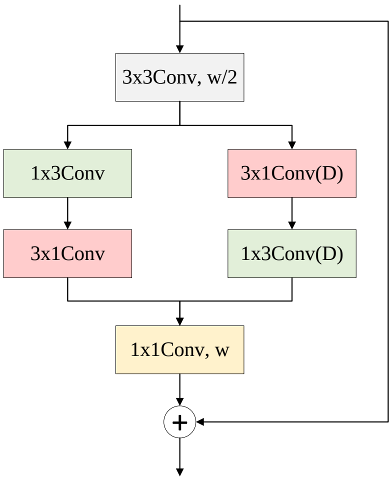

Figure 4. Structure of the EFA unit.

The output graph (after ABunit processing) is pyramid pooled to obtain a characteristic graph with the number of C1 channels and converted into the dimensions of (H · W, C1) and (C1, H · W), as shown in Figure 5; then, they are cross-multiplied. The resulting graph is processed with the softmax function, and the two processing results after the softmax are cross-multiplied to obtain a large characteristic graph with dimensions of (H · W, H · w). The characteristic image is cross-multiplied, and it is then reshaped with the image before softmax (H · W, C1); finally, concatenation is performed on the channel dimension. After the fusion, the module's logic information extraction ability is significantly enhanced, and the accuracy of edge information and detailed information recognition is improved [34]. For example, buildings covered by tree shadows are no longer misclassified as background, and the segmentation of water edges is no longer affected by coasts and ships. The detailed structure of the edge feature attention module is shown in Figure 5.

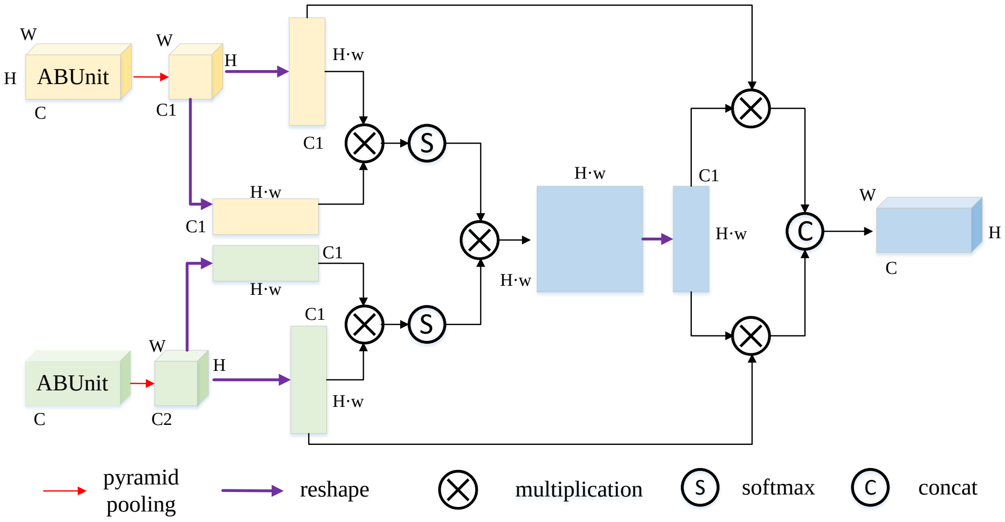

Figure 5. Edge feature attention module.

In summary, the linear index upsampling branch with different levels not only limits the amount of calculation, but also ensures the integrity of high-frequency information. The SFResidual module extracts key features and pays attention to the logical relationship between information within a class and between classes so that it can more fully focus on the dependencies between contexts. The edge feature attention module can provide high-resolution detailed information, and it has achieved good results in the learning of the detailed information and edge information of a feature map.

3. Land-Cover Segmentation Experiment

3.1. Dataset

3.1.1. Land-Cover Dataset

The way in which we verified the model for land-cover segmentation proposed in this paper was by using a dataset that we made. The dataset came from Google Earth. Google Earth is a virtual Earth software developed by Google. It presents satellite photos, aerial photos, and a GIS in the form of three-dimensional models. The authors first obtained 1000 large images with a resolution of 1500 × 800 px on Google Earth on 20 March 2021; these were cut into 23,915 small images with a resolution of 224 × 224 px. These large images had a large space span and a variety of shooting angles. They were roughly divided into the following categories: private villas in wealthy areas of North America and Europe, villages and forests in Western European countries, and China's coastal rivers. In summary, the dataset covered a wide area, including many environments with complex terrain, and it was suitable for investigating the true detection capabilities of the model. As shown in Figure 6, these images were manually labeled as 3 types of objects: buildings (white, RGB [255, 255, 255]), water (blue, RGB [0, 180, 255]), and background (black, RGB [0, 0, 0]). The work of making the labels corresponding to the original images was carried out with Adobe Photoshop 2020.

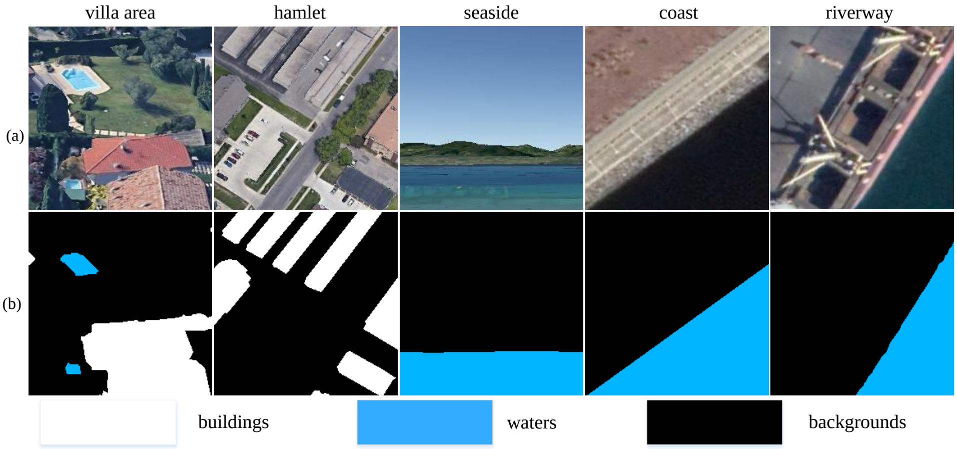

Figure 6. Land segmentation dataset: (**a**) a remote sensing image; (**b**) an artificial label.

The semantic segmentation of this dataset presented great difficulties. In addition, there were some more difficult problems [35]. Buildings are stationary objects, but the differences in their height are large. The projection of the shadows of high buildings will affect the edge contour segmentation of low buildings, and the same is true for tall trees; remote sensing images with projections that are similar in appearance, indistinguishability is likely. As shown in the figures below, some vehicles were similar to buildings. Although they are small in size, large areas of stationary vehicles are easily misclassified as buildings; a close-to-horizontal viewing angle can cause trees to hide the water, which would make a training set more difficult to learn; the tops of some buildings are similar in color to vegetation, and can easily be misclassified as the background; the same water area (private swimming pool) has two colors of blue and green, making it more difficult to segment water objects. In summary, this dataset is relatively difficult to learn [36], and it is also difficult to use it to perform accurate land-cover segmentation and perfect target classification, as shown in Figure 7.

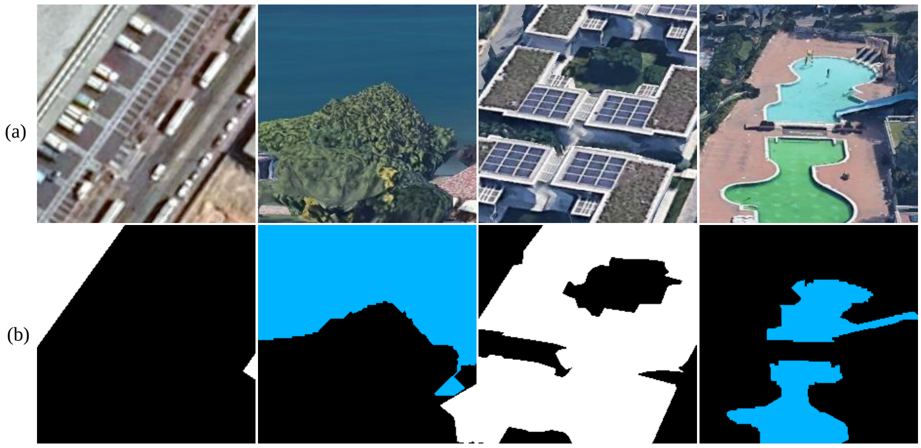

Figure 7. Land segmentation dataset. (**a**) Original image; (**b**) artificial label.

To facilitate the experiments, all pictures were cut in a certain order—from left to right and from top to bottom, and there was no area overlap during the segmentation; images with only one category were excluded to obtain a final dataset of more than 12,000 images (224 × 224 px). The dataset is randomly divided into a training set and a test set at a ratio of 7:3.

3.1.2. Public Dataset

This dataset consisted of 310 aerial images in the Boston area, each with 1500 × 1500 pixels, and it contained hyperspectral, multispectral, and SAR-type images (the reader can search for the 'Massachusetts Roads Dataset' on the official website to find it easily). The complete dataset contained 34,000 hectares. We cut the dataset into 10,620 small images (256 px), which were divided into a training set and verification set according to the ratio of 7:3. There were only the buildings (white, RGB [0, 0, 0]) and the background (black, RGB [255, 255, 255]) in it, as shown in Figure 8.

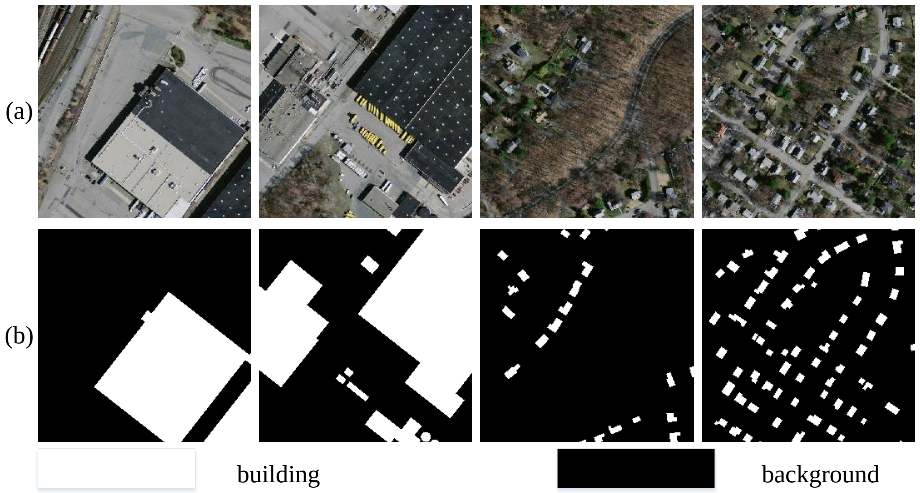

Figure 8. Public dataset. (**a**) Original image; (**b**) artificial label.

3.1.3. Four-Class Public Dataset

For the sake of verifying the effect of the proposed network in the segmentation task, the LandCover dataset [37] was used, as shown in Figure 9. The dataset was composed of images chosen from remote sensing photos covering 216.27 square kilometers of land in Poland. It contained four types of labels: buildings (red, RGB [128, 0, 0]), woodlands (green, RGB [0, 128, 0]), water (gray, RGB [128, 0, 0]), and background (black, RGB [0, 0, 0]), as shown in Figure 9. The dataset had 33 pictures with a resolution of 25 cm (about 9000 × 9500 px) and 8 pictures with a resolution of 50 cm (about 4200 × 4700 px). It was definitely not easy to accurately segment this dataset. In addition to the difficulties mentioned above, there was still a problem with how we could accurately define these four types of objects. "Building" refers to a regular solid object with a certain height that will not move; "vegetation" refers to tall trees, flower beds, green belts, etc., but does not include pure grassland; "water" includes rivers and streams, but does not include waterless ponds. The projections of these objects blocked each other, which could easily cause false detection. In Figure 8, the objects encircled by the yellow ellipse are single trees, which are easy to misclassify as forests. The low shrubs marked by the yellow rectangle are also easily misclassified as forests; buildings marked with blue rectangles are easily misclassified as background; those marked with pink ellipses are greenhouses, which can easily be mistaken for buildings, but they should be classified as background. To sum up, it is not easy to perfectly classify land cover in this dataset.

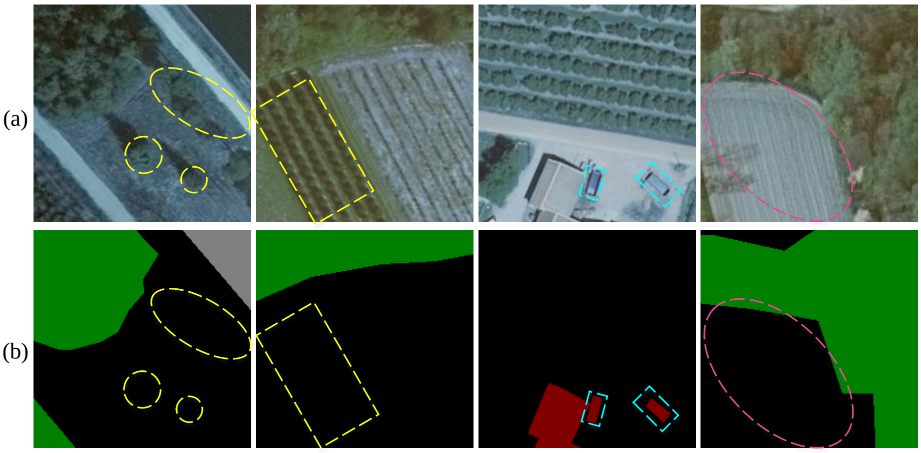

Figure 9. An example from the LandCover dataset. (**a**) Original image; (**b**) artificial label. The parts shown by the dotted line belong to the background class, though they are easily misclassified. The objects encircled by the solid line are the target classes, though they are easily misrecognized as background.

We processed datasets as follows: All images were cut in a certain order without overlapping and omission to create images with 224 × 224 px. Pure-color pictures were removed, and the remaining images were randomly divided into the training set, validation set, and test set according to the ratio of 7:3.

3.2. Evaluation Index

In this experiment, we selected three evaluation indicators: the pixel accuracy (pixel accuracy, *PA*), mean pixel accuracy (*MPA*), and mean Intersection over union (*MIoU*). They are calculated as follows:

$$PA\frac{p_{ii}}{\sum\limits_{i=0}^{k}\sum\limits_{j=0}^{k}p_{ij}}, \tag{1}$$

$$MPA=\frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum\limits_{j=0}^{k}P_{ij}}, \tag{2}$$

$$MIoU=\frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum\limits_{j=0}^{k}P_{ij}+\sum\limits_{j=0}^{k}P_{ji}-P_{ii}}, \tag{3}$$

where k is the number of categories, p_{ii} is a pixel, and its correct mark is i, but its prediction is j if the correct sign is i. When $i \neq j$, p_{ii} is a true positive, p_{ij} is a false negative, p_{ji} is a false positive, and p_{jj} is a true negative. True positive: for a real example, the model prediction provides a positive example, and it actually is a positive example. False negative: For a false counterexample, the model predicts that it is a counterexample, but it is actually a positive example. False positive: for a false positive example, the model predicts that it is a positive example, but it is actually a negative example. True Negative: for a true counterexample, the model predicts that it is a counterexample, but it is actually a counterexample. Pixel-like precision indicates the ratio of the intersection of the real index and the predicted index of each class in the three classifications [38,39]; the average pixel accuracy is the ratio of the intersection of the real value and the predicted value to the real value [40]; the average intersection and union ratio is an important index for measuring the effect of land segmentation [41]. It refers to the ratio of the intersection and union of the real value and the predicted value, and then the average value is taken [42]. This index can reflect the quality of a network and the advantages and disadvantages of a model well.

3.3. Supplementary Experimental Procedures

This experiment was based on the public platform PyTorch. In this work, we used the ’poly’ learning rate strategy [43] and the Adam optimizer. We believed that Adam optimizer was the most suitable for this dataset and network—if the data were dense, the SGD optimizer was adopted. Although it takes a long time and is easily trapped at saddle points, it can quickly reach the maximum value. On the contrary, the Adam optimizer can converge quickly, and the rising curve is relatively stable. The experiments showed that the Adam optimizer can make the most of the model in terms of the density of the land segmentation dataset. Too high of a learning rate will lead to too a large span, and it is easy to miss the best advantage; too low of a learning rate will lead to too slow of a convergence speed. The training effect was the best when the learning rate = 0.001 in this experiment, so the learning rate was set to 0.001. When the power was lower than 0.9, the rising speed of the first 100 epochs was too slow, and when the power was higher than 0.9, the last 150 epochs were completely saturated. So, the basic learning rate was set to 0.001, the power was set to 0.9, and the upper iteration limit was set to 300. The momentum and weight decay rates were set to 0.9 and 0.0001. Considering the actual situation of GPU memory in this experiment, the batch size of the training was set to 4. All experiments were carried out on a Windows 10 system with an Intel (R) corei5 10400F/10500 CPU, 2.90 GHZ, 16 G memory, and NVIDIA GeForce RTX 3070s (8GB) graphics card. This experiment used Python version 3.8 with cuda10.1. We used the cross-entropy loss function [14] to calculate the loss of the neural network. Shannon proposed that the probability of the occurrence of information decreases with the increase in the amount of information, and vice versa. If the probability of an event is $P(x)$, the amount of information is expressed as:

$$I(x) = -log(P(x)). \tag{4}$$

Information entropy is used to express the expectation of the amount of information:

$$H(x) = -\sum_{i=1}^{n} P(x_i)\log(P(x_i)). \tag{5}$$

If there are two separate probability distributions and $P(x)$ and $Q(x)$ can describe the same random variable, we can use the relative entropy (in this paper, the predicted value and the loss value of the label) to quantify the difference between the two probability distributions:

$$D_{KL} = \sum_{i=1}^{n} p(x_i)\log(\frac{q(x_i)}{p(x_i)}), \tag{6}$$

$$loss = -D_{KL}=\sum_{i=1}^{n} p(x_i)\log(p(x_i)) - \sum_{i=1}^{n} p(x_i)\log(q(x_i)), \tag{7}$$

where x_i is the sample, p and q are two independent probability distributions of random variables, and n is the number of samples. The gradient descent algorithm was used in the training process. By comparing tags and predictions, the parameters were updated through backpropagation. The optimal parameters of the training model were all saved. For the problem of land segmentation and cover, the effect of the cross-entropy loss function was better than that of the mean square error loss function [44], so the cross-entropy loss function [45] was used in this experiment.

3.4. Analysis of the Results

The optimal values are bold. As shown in Table 2, the main module used the SHResidual module as the backbone network, and the training parameters of all models were set to the same values. According to the information in the table, the EFAModule was improved by 1.03% because of the module's ability to recover detailed information and capture boundary information. The branch composed of the DCModule and EFAModule improved by 4.03%. The GDC_branch, which was connected with the DCModule and EFAModule, was able to greatly improve the segmentation effect. This combination paid attention to contextual information, detail features, and boundary information at the same time, which improved the accuracy of the final MSNet by 5.71% (learning rate = 0.001, power = 0.9, weight decay rate = 0.0001, batch = 4, epoch = 300).

Table 2. Ablation experiment.

Method	PA (%)	MPA (%)	MIoU (%)	Flops (G)	Parameters (M)
SFR	87.69	87.96	75.62	0.96	0.039
EFA	79.56	77.96	70.62	0.05	0.0001
LIU	88.75	88.51	79.59	134.27	73.34
SFR + EFA	88.48	87.27	76.65	1.01	0.039
SFR + LIU	88.09	88.56	78.13	20	17.5
SFR + LIU + EFA	**89.84**	**89.62**	**81.33**	**21.86**	**18.7**

The maps that included the EFAModule obviously avoided many misclassifications and achieved a better edge segmentation effect, which was beneficial in that it was possible to extract logical features and output high-resolution detailed information. In comparison with the SFRModule alone, the combination did not misdetect the underground with respect to the buildings. Obviously, the combination of SFR, LIU, and the EFA basically allowed all of the misclassifications and edge blur to be avoided, which largely improved the segmentation effect. A diagram of the effects in the ablation experiment is shown in Figure 10.

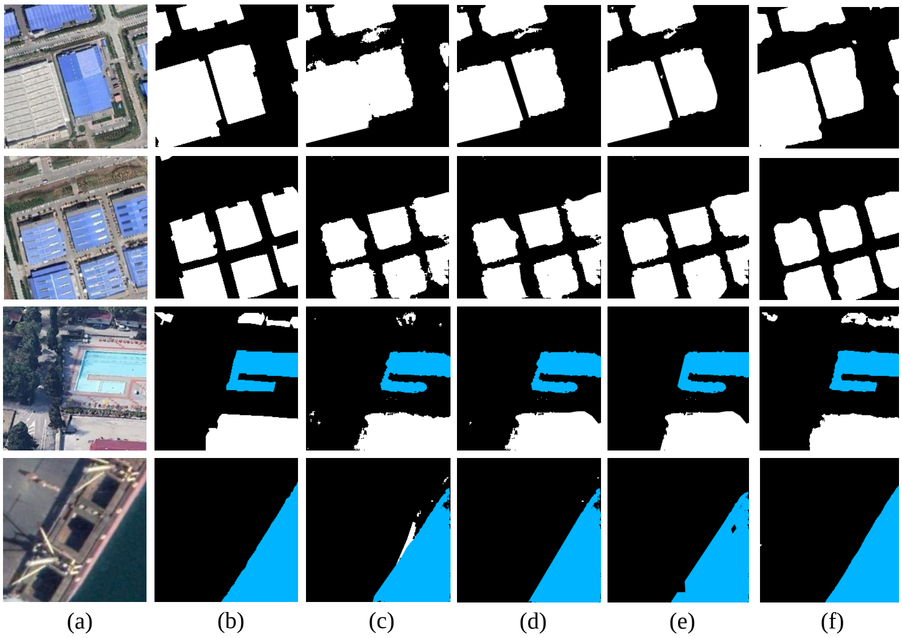

Figure 10. Diagram of the effects in the ablation experiment.(**a**) Real image; (**b**) label; (**c**) SFR; (**d**) SFR + LIU; (**e**) SFR + EFA; (**f**) MSNet.

To compare the performance of each model, the models were tested under the same conditions. Figure 11 shows a chart comparing the effects of MSNet and the other models. In the first and second sets of images, under the influence of vehicles, lawns, and other objects, networks such as FCN and SegNet showed different degrees of misdetection. SegNet mistakenly recognized the background as the buildings. The problem of edge blurring in images assessed with ExtremeC3Net was obvious, but our model achieved accurate segmentation. In the third set of images, the low buildings above the swimming pool were easily recognized as the background, though networks such as SegNet and PSPNet had obvious omissions in their recognition of buildings. In the fourth and fifth groups of images, although there was interference from the boat and the cement on the land by the sea, our model still achieved a more accurate segmentation of the water, whereas the other models misclassified the boat and the cement on the land as buildings, especially SegNet and ExtremeC3Net. In fact, the boat and the cement on the land should have been classified as background. In the sixth set of images, the blue buildings were easy to recognize as water, as they were by ExtremeC3Net, but our model still avoided such mistakes. This was caused by the synchronous learning of intra-class information and inter-class information by the SFRModule; by combining it with the EFAModule, the high-frequency detailed information was restored and the accuracy of the edge segmentation was ensured. Finally, the fusion with LIU caused our model to achieve a great effect. As shown in Figure 11, the actual segmentation effect of MSNet was better than that of the other networks (learning rate = 0.001, power = 0.9, weight decay rate = 0.0001, batch = 4, epoch = 300) And the heat map of this data set is shown as Figure 12.

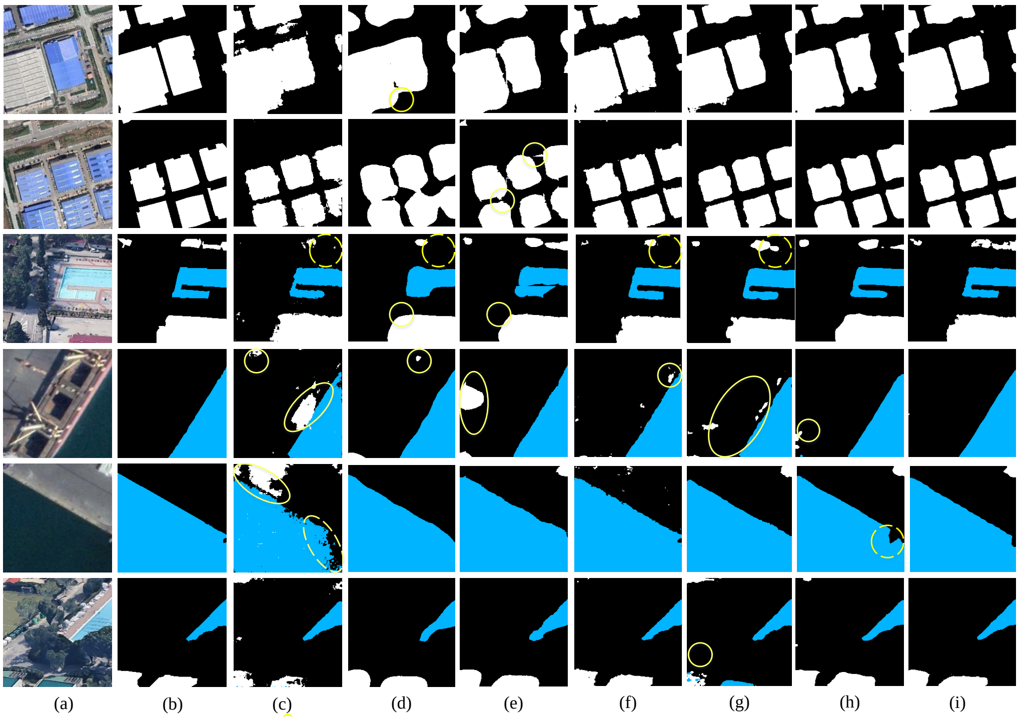

Figure 11. Comparison of the actual segmentation results. (**a**) Real image; (**b**) label; (**c**) SegNet; (**d**) FCN8s; (**e**) FCN32s, (**f**) PSPNet; (**g**) ExtremeC3Net; (**h**) DABNet; (**i**) MSNet.

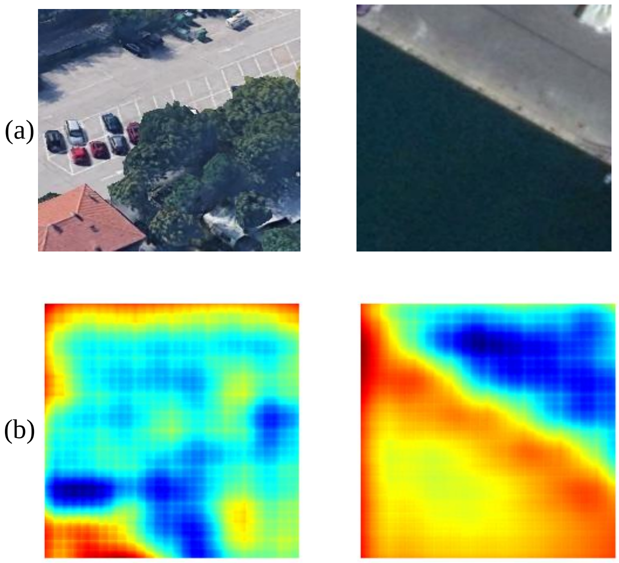

Figure 12. Feature space analysis of MSNet: (**a**) real Image; (**b**) thermodynamic diagram

Table 3 shows the evaluation metrics of each model, and the PA represents the pixel accuracy of each category in the three categories. In comparison with FCN-32s, SegNet, DABNet, UNet [33], EspNet [46], ShuffleNetv1 [20], and the other models, MSNet was able to achieve the best results, which were 1.19% higher than the second best index.

Table 3. Results of the land-cover test set.

Method	PA (%)	MPA (%)	MIoU (%)	Parameters (M)	Flops (G)
UNet	86.89	85.66	75.18	17.27	40
SegNet [47]	87.95	88.35	75.95	29.44	40.07
FCN8s	89.43	88.51	79.59	134.27	73.35
FCN32s [48]	89.75	88.35	79.37	134.29	73.34
PSPNet [16]	89.29	89.75	80.14	48.94	44.3
DABNet [49]	89.55	89.90	79.89	0.75	2.82
EspNetV2 [50]	89.46	89.59	79.95	1.24	0.66
OcrNet [51]	89.97	89.51	80.09	70.35	40.4
ExtremeC3Net [52]	88.69	88.05	78.04	**0.04**	**1.27**
MSNet	**89.84**	**89.62**	**81.33**	18.7	21.86

The MIoU curve of the model is shown in Figure 13 below. In the first 50 generations of ExtremeC3Net [52], the growth rate was very fast—better than that of MSNet (MFNet) after 100 generations—but the MSNet curve could be steadily maintained above other models. The same was true for the training loss curve (Figure 14). In the first 50 generations, it was significantly higher than that of DABNet, and it was stable at the bottom of all curves after 100 generations. From the point of view of the convergence speed and long-term effect, MSNet had great superiority.

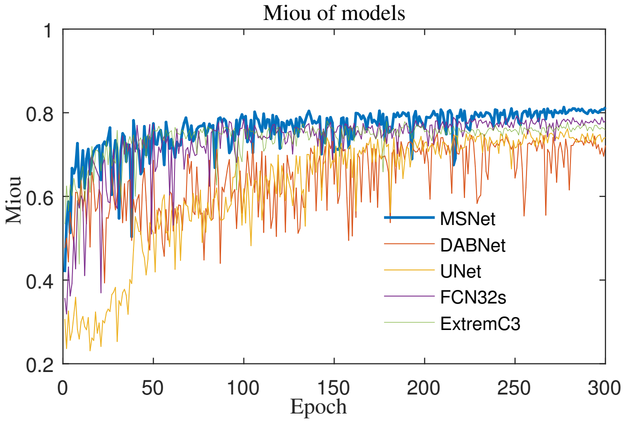

Figure 13. The MIoU of the models.

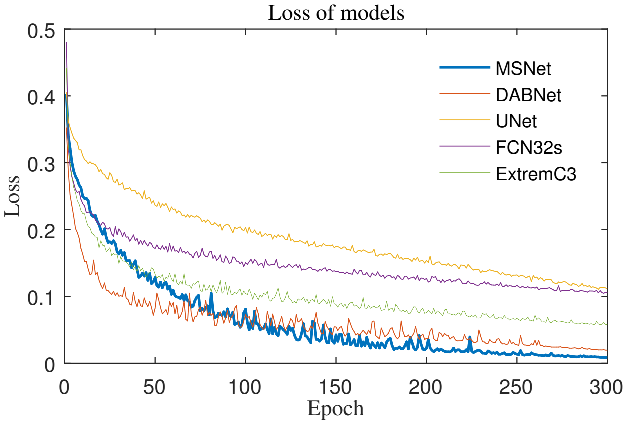

Figure 14. The loss of the models.

We provide a "feature space analysis" to show the segmentation of MSNet. In the Figure 12, red represents the segmentation object and blue represents the background. Through the graphic analysis of the thermal diagram, it could be seen that MSNet's segmentation of the buildings at the lower-left corner was more accurate in the first image, but the shadows at the upper-left corner were wrongly detected as buildings. For the second image, MSNet was able to accurately detect the overall scope of the water area, but the center of the water area had a false detection.

For the sake of verifying the generalization ability of the model, further experiments were carried out on the public land-cover dataset. The dataset consisted of 310 aerial images in the Boston area, each with 1500 × 1500 pixels and an area of 225 hectares. The entire dataset covered about 34,000 hectares. It was cut into 10,620 small images (256 × 256), which were divided into a training set and verification set at a ratio of 7:3 [45]; this dataset had only the building (white, RGB [0, 0, 0]) and the background (black, RGB [255, 255, 255]) types. Without data enhancement, the settings of the various hyperparameters, except for the batch size of 3, were the same as those in the previous experiment. For the first set of images, it was obvious that MSNet's edge segmentation effect was much better than those of FCN32s and DeepLabV3Plus. For the second, third, and fourth sets of images, the abilities of PSPNet and DABNet to segment small and dense buildings were relatively poor, but our model had accurate recognition of those buildings, including their edge information and detailed information. In terms of indicators, it was 1.01% better than the second best model and 9.44% better than the lowest model. The experimental results are shown in Table 4 and Figure 15 (learning rate = 0.001, power = 0.9, weight decay rates = 0.0001, batch = 4, epoch = 300).

Table 4. Generalization experiment on the public land-cover dataset.

Method	MIoU(%)	Parameters (M)	Flops (G)
UNet	69.5	17.27	40
SegNet	72.17	29.44	40.07
PSPNet	77.08	48.94	44.3
ENet [53]	70.52	0.35	**0.45**
DeeplabV3+Net	75.28	91.77	64.42
DABNet	75.14	0.752	1.27
Pan [54]	74.48	23.65	1.27
BiseNetV2 [55]	74.63	3.62	3.2
MSNet	**79.90**	**18.7**	**21.01**

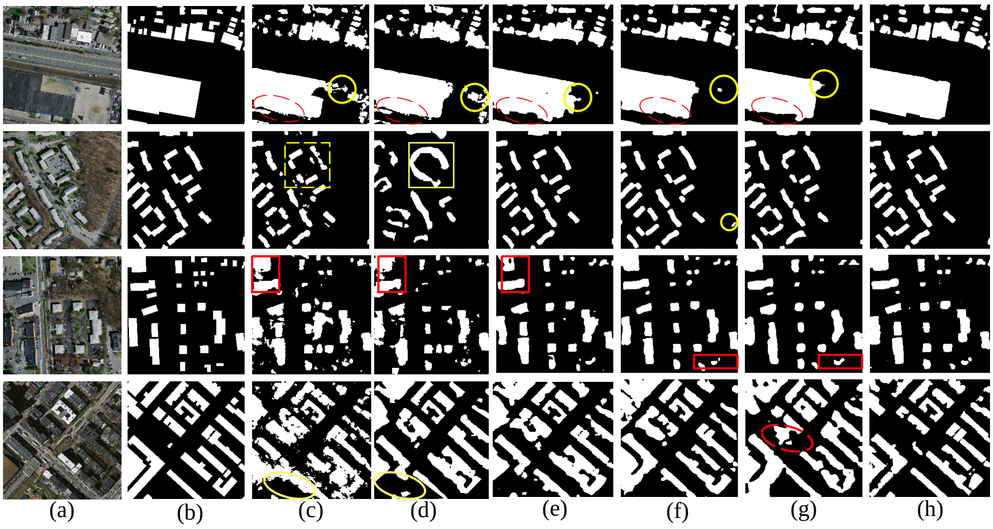

Figure 15. Diagram comparing the effects of the MIoU values of different models. (**a**) Real image; (**b**) label; (**c**) SegNet; (**d**) FCN32s; (**e**) DeepLabV3+; (**f**) PSPNet; (**g**) DABNet; (**h**) MSNet.

To verify the performance of the model with other categories of datasets, further experiments were carried out on a public water-cover dataset (Figure 16). The data in this paper came from high-resolution remote sensing images selected from Google Earth, and the number of images in the data was 26,200. In order to make the data more authentic, we used a wide range of distributions, and in terms of river selection, we chose rivers with different widths and colors and small and rugged rivers. On the other hand, we selected complex environments surrounding the rivers, including hills, forests, urban areas, farmlands, and other areas, which could fully test the generalization performance of the model. Some of the images of the river that were collected are shown in Figure 15. The average size of the Google Earth images was 4800 × 2742 pixels, and this was cut to 224 × 224 for model training. The training set and test set contained 20,960 and 5240 images, respectively. This dataset had only the building (red, RGB[128, 0, 0]) and the background (black, RGB [255, 255, 255]) types. Without data enhancement, the settings of the various hyperparameters, except for the batch size of 3, were the same as those in the previous experiment. For the first set of images, it was obvious that MSNet's edge segmentation effect was much better than those of SegNet and DeepLabV3Plus. There were also obvious fractures in SegNet's effect for the first and second sets of images. For the third and fourth groups of maps, SegNet and DABNet mistakenly detected the grassland and buildings as water areas, and DeeplabV3+ mistakenly detected an intersection of rivers as the background. The edge detection accuracy of PSPNet was relatively low, and it also mistakenly detected a water area as the background. However, MSNet could not only distinguish water areas from grasslands, buildings, and other backgrounds, but could also accurately extract edge information. In terms of indicators, it is 2.82% better than the second better model and 10.4% higher than the worst model. The experimental results are shown in Table 5 and Figure 15 (learning rate = 0.001, power = 0.9, weight decay rates = 0.0001, batch = 4, epoch = 300).

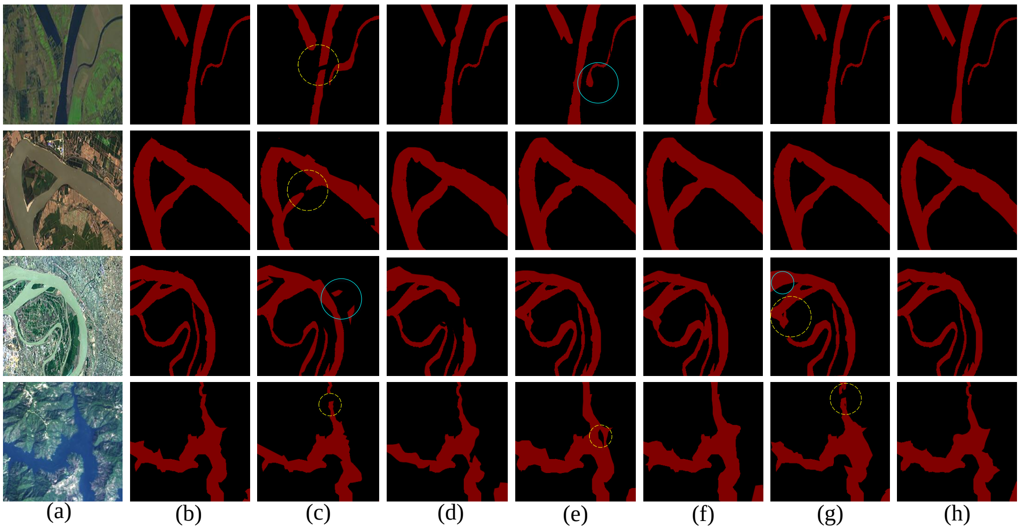

Figure 16. Diagram comparing the effects of the MIoU values of different models. (**a**) Real image; (**b**) label; (**c**) SegNet; (**d**) FCN32s; (**e**) DeepLabV3+; (**f**) PSPNet; (**g**) DABNet; (**h**) MSNet.

Table 5. Generalization experiment on a public water-cover dataset.

Method	MIoU (%)	Parameters (M)	Flops (G)
UNet	89.5	17.27	40
SegNet	95.17	29.44	40.07
PSPNet	97.08	48.94	44.3
ENet [53]	70.52	0.35	**0.45**
DeeplabV3+Net	97.28	91.77	64.42
DABNet	97.14	0.752	1.27
Pan [54]	97.48	23.65	1.27
BiseNetV2 [55]	97.93	3.62	3.2
MSNet	**98.94**	**18.7**	**21.01**

In order to include all objects in the scene, the four-class public dataset introduced above Figure 15 was selected for a generalization experiment. As shown in Figure 17, for the first set of images, UNet directly missed all buildings in the lower-left corner. FCN32s and DeepLabV3plus missed the detection to varying degrees. DABNet mistakenly detected buildings as plants. MSNet did not have these problems, as it benefited from SFR's synchronous learning of intra-class information and inter-class information. For the second set of pictures, SegNet mistakenly classified the plants in the lower-left corner as buildings, and UNet, ExtremeC3Net, and the other networks mistakenly classified the background in the middle of the figure as plants. MSNet did not have this error. For the third and fourth sets of pictures, UNet confused plants with water, and the edge detection of the buildings was relatively fuzzy. The learning of the edge information of DeepLabV3plus and ExtremeC3 was not ideal, and the error was large. However, MSNet basically avoided the problems of edge blur and false recognition, which showed the superiority of the two-way fusion model. So, it can be seen intuitively that MSNet's MIoU was higher than that of Unet by 14.3% and higher than that of FCN32s by 1.14%, as shown in Table 6.

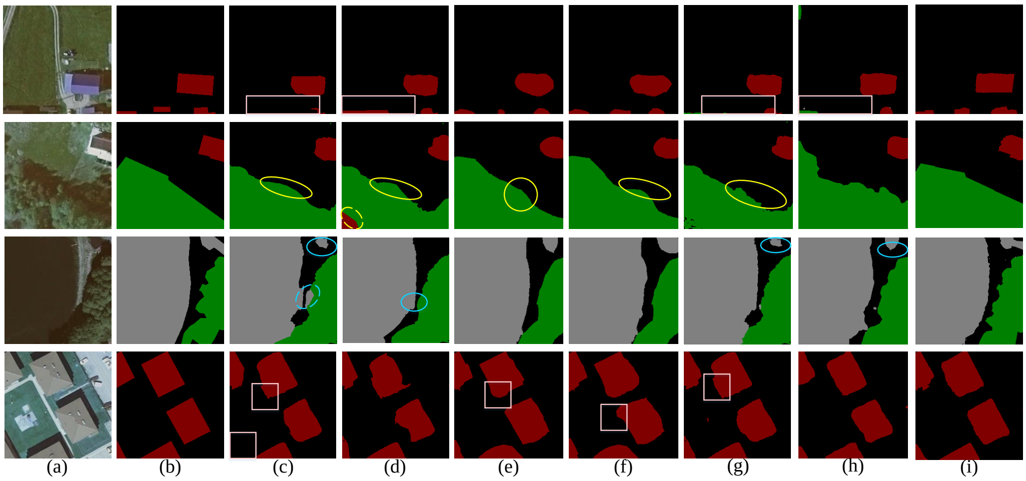

Figure 17. Comparison of the actual segmentation results. (**a**) Image; (**b**) label; (**c**) UNet; (**d**) SegNet; (**e**) FCN32s; (**f**) DeeplabV3Plus; (**g**) ExtremeC3; (**h**) DABNet; (**i**) MSNet. The part marked with a solid line represents a missing detection or a mistaken detection of the background as a real object, and the part marked with a dotted line indicates that different categories of real objects were confused.

Table 6. Results of the four-class public dataset.

Method	MIoU (%)	PA	MPA
UNet	64.49	73.31	79.91
SegNet	68.67	85.09	87.14
FCN32s	77.65	89.54	89.32
DeeplabV3+Net	76.86	87.85	87.07
DABNet	73.33	86.46	87.39
ExtremeC3Net	71.75	85.99	85.11
MSNet	**78.79**	**90.21**	**88.64**

The results show that the average intersection ratio and the other indicators of MSNet were higher than those of the other models. Therefore, the generalization and effectiveness of MSNet were proven. MSNet combined a shuffle unit with a skip connection, the channel rearrangement caused the extraction of information to be more well distributed, and the residual structure ensured the accuracy of the extraction of deep semantic information, which paid attention to the logical relationship between information within a class and information between classes. The combination of this branch and the LIU greatly improved the accuracy of segmentation, allowed information to be extracted from a deeper level, and caused better results to be achieved. A comparison of the actual segmentation results is shown in Figure 15, and the details of the indices are shown in Table 6. In this experiment, the hyperparameters were set as follows: learning rate = 0.0001, power = 0.9, weight decay rate = 0.0001, batch = 4, epoch = 300.

4. Conclusions

In this work, in order to optimize the effect of land division, a new three-way parallel feature fusion network called MSNet was proposed, and it focused on the enhancement of contextual information and compatible intra-class information and inter-class information to improve the model. The proposed LIU focused on contextual features and strengthened the learning of detailed information; a branch composed of the SFRModule and EFAModule was able to take into account the identification of intra-class information and inter-class information, filter redundant information, extract key features, and focus on the learning of boundary information. The two-way feature-sharing network was proven to have a good segmentation effect. However, the segmentation effect of the network is not ideal when faced with a large number of categories and more complex datasets. When buildings are captured from different angles, it cannot guarantee that the contours of the predicted map perfectly match those of the original image. It can only ensure the accuracy of the location. The same is true for water areas. To increase the accuracy, many studies' results have shown that adding an optimized transformer structure can significantly improve the segmentation accuracy of a model, so the next direction for research is to think about how the transformer structure can be optimized so that it can have a better effect on fusion with a convolutional neural network. In addition, the network still needs to achieve a faster computing speed with fewer parameters.

Author Contributions: Conceptualization, L.W. and M.X.; methodology, J.G. and M.X.; software, J.G.; validation, L.W. and M.X.; formal analysis, J.G. and M.X.; investigation, J.G., L.W., H.L. and M.X.; resources, M.X.; data curation, M.X.; writing—original draft preparation, J.G. and L.W.; writing—review and editing, M.X.and H.L; visualization, L.W.; supervision, L.W.; project administration, M.X.; funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of PR China under Grant 42075130.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request (002311@nuist.edu.cn).

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

References

1. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. [CrossRef]
2. Song, L.; Xia, M.; Jin, J.; Qian, M.; Zhang, Y. SUACDNet: Attentional change detection network based on siamese U-shaped structure. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102597. [CrossRef]
3. Wulder, M.A.; Masek, J.G.; Cohen, W.B.; Loveland, T.R.; Woodcock, C.E. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sens. Environ.* **2012**, *122*, 2–10. [CrossRef]
4. Bak, C.; Kocak, A.; Erdem, E.; Erdem, A. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Trans. Multimed.* **2017**, *20*, 1688–1698. [CrossRef]
5. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
6. Xia, M.; Qu, Y.; Lin, H. PADANet: Parallel asymmetric double attention network for clouds and its shadow detection. *J. Appl. Remote Sens.* **2021**, *15*, 046512. [CrossRef]
7. Potapov, P.V.; Turubanova, S.; Tyukavina, A.; Krylov, A.; McCarty, J.; Radeloff, V.; Hansen, M. Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive. *Remote Sens. Environ.* **2015**, *159*, 28–43. [CrossRef]
8. Geller, G.N.; Halpin, P.N.; Helmuth, B.; Hestir, E.L.; Skidmore, A.; Abrams, M.J.; Aguirre, N.; Blair, M.; Botha, E.; Colloff, M.; et al. Remote sensing for biodiversity. In *The GEO Handbook on Biodiversity Observation Networks*; Springer: Cham, Switzerland, 2017; pp. 187–210.
9. Gao, J.; Weng, L.; Xia, M.; Lin, H. MLNet: Multichannel feature fusion lozenge network for land segmentation. *J. Appl. Remote Sens.* **2022**, *16*, 016513. [CrossRef]
10. Qu, Y.; Xia, M.; Zhang, Y. Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. *Comput. Geosci.* **2021**, *157*, 104940. [CrossRef]
11. Lu, C.; Xia, M.; Qian, M.; Chen, B. Dual-Branch Network for Cloud and Cloud Shadow Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5410012. doi: 10.1109/TGRS.2022.3175613. [CrossRef]
12. Chen, B.; Xia, M.; Qian, M.; Huang, J. MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images. *Int. J. Remote. Sens.* **2022**. doi: 10.1080/01431161.2022.2073795. [CrossRef]
13. Wang, Z.; Xia, M.; Lu, M.; Pan, L.; Liu, J. Parameter Identification in Power Transmission Systems Based on Graph Convolution Network. *IEEE Trans. Power Deliv.* **2022**, *37*, 3155–3163. [CrossRef]
14. Shokat, S.; Riaz, R.; Rizvi, S.S.; Abbasi, A.M.; Abbasi, A.A.; Kwon, S.J. Deep learning scheme for character prediction with position-free touch screen-based Braille input method. *Hum.-Centric Comput. Inf. Sci.* **2020**, *10*, 41. [CrossRef]
15. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.

16. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–25 July 2017; pp. 2881–2890.
17. Zhou, J.; Hao, M.; Zhang, D.; Zou, P.; Zhang, W. Fusion PSPnet image segmentation based method for multi-focus image fusion. *IEEE Photonics J.* **2019**, *11*, 6501412. [CrossRef]
18. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–25 July 2017; pp. 1251–1258.
19. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote. Sens.* **2022**, doi: 10.1080/01431161.2021.2014077. [CrossRef]
20. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–21 June 2018; pp. 6848–6856.
21. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature fusion attention network for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7 February 2020; Volume 34, pp. 11908–11915.
22. Hao, S.; Zhou, Y.; Zhang, Y.; Guo, Y. Contextual attention refinement network for real-time semantic segmentation. *IEEE Access* **2020**, *8*, 55230–55240. [CrossRef]
23. O Oh, J.; Chang, H.J.; Choi, S.I. Self-Attention With Convolution and Deconvolution for Efficient Eye Gaze Estimation From a Full Face Image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 4992–5000.
24. Xia, M.; Wang, Z.; Lu, M.; Pan, L. MFAGCN: A new framework for identifying power grid branch parameters. *Electr. Power Syst. Res.* **2022**, *207*, 107855. [CrossRef]
25. Khatami, R.; Mountrakis, G.; Stehman, S.V. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sens. Environ.* **2016**, *177*, 89–100. [CrossRef]
26. Huang, J.; Weng, L.; Chen, B.; Xia, M. DFFAN: Dual function feature aggregation network for semantic segmentation of land cover. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 125. [CrossRef]
27. Zhao, J.; Du, B.; Sun, L.; Zhuang, F.; Lv, W.; Xiong, H. Multiple relational attention network for multi-task learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1123–1131.
28. Cheng, X.; Zhang, Y.; Chen, Y.; Wu, Y.; Yue, Y. Pest identification via deep residual learning in complex background. *Comput. Electron. Agric.* **2017**, *141*, 351–356. [CrossRef]
29. Ren, S.; Sun, J.; He, K.; Zhang, X. Deep residual learning for image recognition. In Proceedings of the CVPR, Vegas, NV, USA, 27–30 June 2016; Volume 2, p. 4.
30. Liu, J.; He, J.; Qiao, Y.; Ren, J.S.; Li, H. Learning to predict context-adaptive convolution for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Springer: Berlin/Heidelberg, Germany, 2020; pp. 769–786.
31. Sengupta, A.; Ye, Y.; Wang, R.; Liu, C.; Roy, K. Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **2019**, *13*, 95. [CrossRef] [PubMed]
32. Pang, K.; Weng, L.; Zhang, Y.; Liu, J.; Lin, H.; Xia, M. SGBNet: An Ultra Light-weight Network for Real-time Semantic Segmentation of Land Cover. *Int. J. Remote. Sens.* **2022**, doi: 10.1080/01431161.2021.2022805. [CrossRef]
33. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 13 September 2018; pp. 325–341.
34. Li, X.; Li, X.; Zhang, L.; Cheng, G.; Shi, J.; Lin, Z.; Tan, S.; Tong, Y. Improving Semantic Segmentation via Decoupled Body and Edge Supervision Supplementary. In Proceedings of the ECCV, Glasgow, UK, 23–28 August 2020, Volume 17, pp. 435–452.
35. Mehta, S.; Paunwala, C.; Vaidya, B. CNN based traffic sign classification using adam optimizer. In Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE, Madurai, India, 15–17 May 2019; pp. 1293–1298.
36. Hu, H.; Ji, D.; Gan, W.; Bai, S.; Wu, W.; Yan, J. Class-wise dynamic graph convolution for semantic segmentation. In Proceedings of the European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–17.
37. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Dziedzic, T.; Zambrzycka, A. LandCover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19–25 June 2021; pp. 1102–1110.
38. Marugg, J.D.; Gonzalez, C.F.; Kunka, B.S.; Ledeboer, A.M.; Pucci, M.J.; Toonen, M.Y.; Walker, S.A.; Zoetmulder, L.C.; Vandenbergh, P.A. Cloning, expression, and nucleotide sequence of genes involved in production of pediocin PA-1, and bacteriocin from Pediococcus acidilactici PAC1.0. *Appl. Environ. Microbiol.* **1992**, *58*, 2360–2367. [CrossRef] [PubMed]
39. Xia, M.; Zhang, X.; Weng, L.; Xu, Y.; et al. Multi-stage feature constraints learning for age estimation. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 2417–2428. [CrossRef]
40. Li, S.; Zhao, X.; Zhou, G. Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network. *Comput.-Aided Civ. Infrastruct. Eng.* **2019**, *34*, 616–634. [CrossRef]
41. Seifi, S.; Tuytelaars, T. Attend and segment: Attention guided active semantic segmentation. In Proceedings of the European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 305–321.

42. Chen, Y.; Li, Y.; Wang, J.; Chen, W.; Zhang, X. Remote sensing image ship detection under complex sea conditions based on deep semantic segmentation. *Remote Sens.* **2020**, *12*, 625. [CrossRef]
43. Bock, S.; Goppold, J.; Weiß, M. An improvement of the convergence proof of the ADAM-Optimizer. *arXiv* **2018**, arXiv:1804.10587.
44. Ho, Y.; Wookey, S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* **2019**, *8*, 4806–4813. [CrossRef]
45. Gordon-Rodriguez, E.; Loaiza-Ganem, G.; Pleiss, G.; Cunningham, J.P. Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. *arXiv* **2020**, arXiv:2011.05231.
46. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 13 September 2018; pp. 552–568.
47. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
48. Zhang, S.; Wu, G.; Costeira, J.P.; Moura, J.M. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 22–25 July 2017; pp. 3667–3676.
49. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
50. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9190–9200.
51. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv* **2019**, arXiv:1909.11065.
52. Park, H.; Sjösund, L.L.; Yoo, Y.; Bang, J.; Kwak, N. Extremec3net: Extreme lightweight portrait segmentation networks using advanced c3-modules. *arXiv* **2019**, arXiv:1908.03093.
53. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
54. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A deep network architecture for pan-sharpening. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 22–25 July 2017; pp. 5449–5457.
55. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *arXiv* **2020**, arXiv:2004.02147.