remote sensing

MDPI

Article

Multi-Scale Feature Aggregation Network for Semantic Segmentation of Land Cover

Xu Shen [1], Liguo Weng [1,*], Min Xia [1] 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: Land cover semantic segmentation is an important technique in land. It is very practical in land resource protection planning, geographical classification, surveying and mapping analysis. Deep learning shows excellent performance in picture segmentation in recent years, but there are few semantic segmentation algorithms for land cover. When dealing with land cover segmentation tasks, traditional semantic segmentation networks often have disadvantages such as low segmentation precision and weak generalization due to the loss of image detail information and the limitation of weight distribution. In order to achieve high-precision land cover segmentation, this article develops a multi-scale feature aggregation network. Traditional convolutional neural network downsampling procedure has problems of detail information loss and resolution degradation; to fix these problems, a multi-scale feature extraction spatial pyramid module is made to assemble regional context data from different areas. In order to address the issue of incomplete information of traditional convolutional neural networks at multiple sizes, a multi-scale feature fusion module is developed to fuse attributes from various layers and several sizes to boost segmentation accuracy. Finally, a multi-scale convolutional attention module is presented to enhance the segmentation's attention to the target in order to address the issue that the classic convolutional neural network has low attention capacity to the building waters in land cover segmentation. Through the contrast experiment and generalization experiment, it can be clearly demonstrated that the segmentation algorithm proposed in this paper realizes the high precision segmentation of land cover.

Keywords: land cover; remote sensing image; deep learning; semantic segmentation

Citation: Shen, X.; Weng, L.; Xia, M.; Lin, H. Multi-Scale Feature Aggregation Network for Semantic Segmentation of Land Cover. *Remote Sens.* **2022**, *14*, 6156. https://doi.org/10.3390/rs14236156

Academic Editors: Zhou Zhang, Zhengxia Zou, Bin Pan and Xia Xu

Received: 29 October 2022
Accepted: 30 November 2022
Published: 5 December 2022

1. Introduction

Semantic segmentation is a hot field in computer vision and remote sensing image processing. In the processing and application of semantic segmentation images, land cover is an important indicator. lt has important practical significance in land planning, mapping analysis [1], surveying, and land classification [2]. The analysis of land resources is receiving more and more attention. Consequently, real-time semantic land cover segmentation is a crucial step in land resource planning and management [3,4].

Currently, the primary classical semantic segmentation techniques include the threshold approach, clustering method, support vector machine technique, and others. Support vector machines were introduced by Zhang, Jiang, and Xu (2013) [5] as a way to extract shorelines by increasing geometric edge properties and reducing mistakes, although this approach is challenging for large-scale data training. Normalized difference water index was proposed by McFeeters in 1996 (NDWI) [6]. The technique creates a segmentation band from the image by combining the green and near-infrared bands; however, it is quite environment-dependent. On the basis of a digital elevation model, Du et al. (2018) [7] developed a novel method for the detection of the water body and accomplished high accuracy. The issue is that the outcomes are significantly influenced by various infrared images.

The picture segmentation problem is solved by the dynamic contour model (ACM) [8] by turning it into a minimal energy functional problem. The model can produce a smooth, continuous contour curve, but it is highly dependent on the original contour. Segmentation on small samples can be completed using the traditional methods mentioned above. However, there are drawbacks for large sample segmentation, including poor precision, laborious parameter tweaking, and inadequate generalization performance. Therefore, it is impossible to accurately extract the shapes of any lakes, rivers, or constructions.

Most recently, deep learning neural networks are extensively utilized in image processing and other industries [9–11], and they significantly increase image processing accuracy as a result of the deep learning industry's quick development. Hinton [12] created a deep learning technique that trained deep belief networks using the back propagation algorithm, considerably enhancing the model's capacity for image processing. The major function of the initial stages of the network model of deep learning is to classify images. Convolution, pooling, and sigmoid are the processes that the deep learning network uses to extract the feature information of the entire picture. The output layer then calculates the probability that each picture is a particular category, taking the highest probability as the classification outcome. However, its disadvantage is that it can only effectively classify specific objects and cannot extract image position information and edge information. Therefore, for land cover segmentation objects with complex environmental backgrounds, traditional convolutional neural networks are enabled to complete effective and high-precision segmentation. To solve the above problems, computer vision scholars have recommended a series of efficient models for semantic segmentation. Convolutional neural network (CNN) [13] significantly boosts the efficiency and precision of semantic land cover segmentation compared to conventional deep learning algorithms. For the pixel-level classification of pictures, Long, Shelhamer, and Darrel [14] suggested a Fully Convolution Network (FCN) in 2014. A feature map for the final convolutional layer is upsampled using a deconvolution layer to equalize its size with that of the source image, allowing FCN to take input images of any size. This allows predictions to be made for each pixel. Spatial data from the initial input image is also retained. The parity in the up-sampled feature map pixel classification, which considerably enhances the classification accuracy of remote sensing images, is the last step. In 2015, Badrinarayanan et al. (2017) [15] recommended SegNet to address image semantic segmentation for autonomous driving and intelligent robots. SegNet and FCN ideas are very similar, but its encoder and decoder (upsampling) use inconsistent methods. The first 13 layers of the VGG16 convolutional network are also used by SegNet's encoder, and a decoder layer is matched to each encoder layer. The soft-max classifier receives the result of the final decoder and independently creates likelihoods of a class for each pixel, but its training is time-consuming and ineffective. In 2015, Ronneberger, Fischer and Brox [16] proposed a widely used semantic segmentation model UNet. UNet is a segmentation network applied to the medical field based on FCN, and has quickly become the baseline for most medical image semantic segmentation tasks. Finally, although the medical image signal is complex but the category is simple, and the distribution of human tissue has certain regularity, so the data set is relatively simple, and the segmentation task is relatively easy. In 2017, Liang-Chieh Chen, George Papandreou [17] and others proposed to revisit the DeepLabv3 network of dilated convolutions for semantic segmentation. This is a potent tool for managing the feature response resolution produced by deep convolutional neural networks in semantic picture segmentation applications, as well as for visualizing and altering the filter field of view. It suggests atrous spatial pyramid pooling (ASPP), which is equal to capturing the image's multi-scale context semantic information in various proportions, to sample the input in parallel with atrous convolution at various sampling rates. However, with the deepening of Block, the expansion coefficient (rate) of the dilated convolution continues to increase, and finally the ability of capturing global information becomes weaker and weaker.

Although the existing segmentation algorithms perform well in real-time semantic segmentation of land cover, due to the continuous down-sampling operation of the convolu-

tional neural network, many semantic features are lost, resulting in inaccurate segmentation and blurred edges [18,19]. Additionally, the contextual semantic information of the image cannot be properly captured because it is difficult to merge the global message of low-level semantic features with the detailed message of high-level semantic features [20,21]. Because of the complex background and terrain of land cover, the traditional convolutional neural network system often has problems such as misjudgment and undetectable objects, resulting in low segmentation accuracy. In order to solve the problems and shortcomings of traditional convolutional neural network in semantic segmentation task, a network is designed in this paper by considering various elements of convolutional neural network. The network extracts the feature information of land cover images by downsampling the adapted RestNet50. The multi-scale feature extraction pyramid module aggregates the context information of different regions, thereby improving its ability to obtain global information of land cover. Then, the multi-scale feature fusion module is used to combine the low-level global information and high-level detail information to improve the segmentation ability of land cover edges and details. Land cover has complex background, unpredictable terrain and intricate edges. Finally, a multi-scale convolutional attention module is proposed to focus on the key target area of land cover image, which helps the network to locate the target of interest more accurately. It can pay more attention to the important information of land cover, dilute the interference and unimportant information, so that the segmentation effect has been significantly improved, and the high-precision segmentation could be achieved.

2. Methodology

Convolutional neural network (CNN) architecture-based models are intensively used in computer vision due to its explosive growth. However, in remote sensing images, the land background is complex and diverse, the terrain trend is unpredictable, and the detailed information and spatial information are rich. Traditional convolutional neural networks fail to accomplish high-precision semantic land cover segmentation. In order to segment land cover images more accurately, multi-scale and context semantic information must be used efficiently. It is not only satisfied with the segmentation of buildings and waters in land cover images, but more importantly, the segmentation of their edges and small details. This study recommends ResNet [22] as the foundation network for a multi-scale feature aggregation network. The network is made up of three modules: a multi-scale feature fusion module, a multi-scale feature extraction space pyramid module, and a multi-scale convolutional attention module. In this study, the input land cover remote sensing image is used to train the model, and the current model's output image is acquired by performing a forward propagation computation. The discrepancy between the produced image and the label is calculated using the cross entropy loss function, and the chain rule is used to return the resulting error to the network. When doing a back propagation computation, the adaptive moment estimation (Adam) optimizer is used to update the learning rate and other parameters of the model. The Adam optimizer [23] utilizes an exponential decay rate with a coefficient of 0.9 to regulate the redistribution of control and an exponential decay rate with a coefficient of 0.999 to control the impact of the preceding gradient square. The variety of learning tactics are used, such as "fixed", "step", and "ploy" strategies. Current experiments show that the "ploy" strategy is better in semantic segmentation experiments [24].

2.1. Network Structure

This research introduces a unique semantic segmentation deep learning network. It performs incredibly well in real time land cover segmentation. Its frame structure is shown in Figure 1. In its composition, as the foundation network for extracting features, we employ the modified ResNet50. Then, this paper presents a multi-scale feature extraction space pyramid module. Its goal is to combine the context data from several places, enhancing the ability to obtain global information, so as to accurately and effectively identify

regions of any scale. Next, a multi-scale feature fusion module is constructed to combine various scale features, which is crucial for enhancing performance. Compared with high-level features, low-level features have more location, detail, and better resolution data, less convolution, lower semantics, and more noise. Although high-level characteristics have weaker resolution and poorer detail perception, they have stronger semantic information. The multi-scale feature fusion module is utilized to fully integrate low-level features and high-level features. Finally, a multi-scale convolutional attention module is suggested. It can efficiently lmitigate the interference of invalid targets, enhance the detection effect of concerned targets, and improve the attention to buildings and waters in land cover images and change the allocation of resources according to the importance of the target. It allocates more resources to the object of attention [25], and thereby increases the model's total segmentation accuracy, which makes the real-time semantic segmentation of land cover more precise.

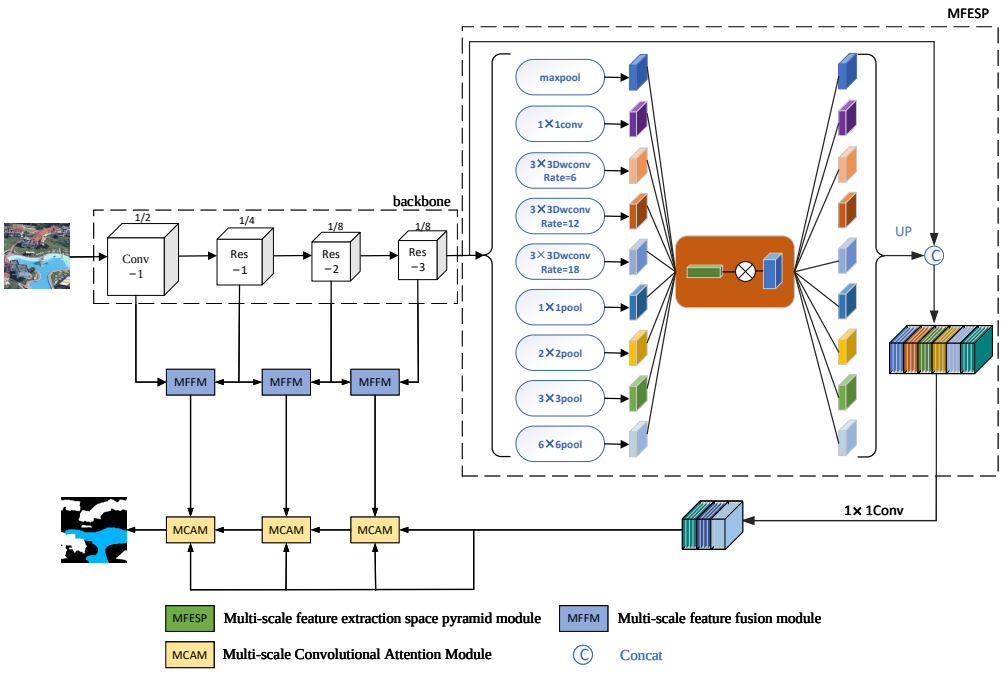

Figure 1. Multi-Scale Feature Aggregation Network framework structure.

2.2. Backbone

The extraction of semantic information and detail information is a key step in image segmentation. The selection and optimization of the backbone network directly affect the semantic segmentation. Typical backbone networks include VGGNet [26], MobileNet [27], Transformer [28], ResNet, Xception [29], and DenseNet [30]. In the semantic segmentation of land cover, the typical backbone network in the past is often undersampled too deeply, which leads to heavy loss of global information in the land cover image; as a result, effective information of small targets in the image decrease sharply. At the same time, the deep undersampling leads to the multiple increase in each parameter, which makes the model too heavy in the calculation of high volume, or there is too much pursuit of the lightweight, as it is unable to focus on the essential problems faced in the downsampling process. In the task of the real-time semantic segmentation of land cover, in order to extract high-precision semantic information and detailed information, it prevents the explosion of the gradient caused by increasing depth. Therefore, this paper uses modified ResNet50 as the backbone network of semantic segmentation of land cover. Because ResNet50 is down-sampled 32 times, too deep down-sampling can easily lead to the loss of more location and detail information. With both at once, the last layer of the ResNet50 channel is expanded to 2048 layers, which makes the calculation heavy and model complexity high. Thus, this article

uses the modified ResNet50, using only the first four layers of ResNet50, the channel to the 1024th layer, and the fourth layer stride equals one downsampling only eight times. In this way, sufficient location information and detailed information are retained, and the lightweight of the network is guaranteed. It solves the problems that the traditional typical backbone network has too deep sampling and too much pursuit of lightweight, resulting in the loss of the key information of land coverage. The expression of the residual unit in the ResNet residual block is as follows:

$$x_{m+1} = W_{m+1}\sigma(W_1 x_1) + x_1, \tag{1}$$

where x_1 is the first residual unit's input vector, $x_m + 1$ is the $m + 1$ residual unit's output vector, σ is a nonlinear function ReLU, the weight matrix is represented by W_1 and $W_m + 1$.

2.3. Multi-Scale Feature Extraction Space Pyramid Module

Capturing multi-scale context data is crucial for semantic segmentation in order to gather comprehensive data. The land cover image has a complex background and criss-crossed rivers and houses. Therefore, it is necessary to design a module to aggregate multi-scale context information to solve the problem of context information loss, which could cause blurred edge and undetectable small targets, thereby improving the accuracy of the real-time semantic segmentation of land cover. Therefore, to combine the context information from several locations and enhance the capacity to access global information, this research refers to pyramid pooling module in PSPNet [31]. The ASPP module of DeepLabv3Plus network of classical convolutional neural networks is also referenced. In order to ensure the image resolution, the dilated convolution is employed to expand the receptive field of land cover image information. Figure 2 illustrates its particular structure.

In land cover segmentation, ordinary convolution operations can only be used to process the local area, and its receptive field is limited; thus, it is difficult to capture multi-scale context information. Firstly, three 3×3 dilated convolutions are used, and their expansion coefficients are 6, 12 and 18. The dilated convolution shows the extent of the convolutional expansion by the expansion factor. It improves the network's ability to acquire multi-scale context while expanding the network's receptive field without altering the convolution kernel's form, or without downsampling. The two-dimensional dilated convolution receptive field growth formula can be expressed as [32]

$$R_x = (4 \times x - 1)^2. \tag{2}$$

Then, 1×1, 2×2, 3×3, 6×6 adaptive AvgPool is used. Global information is gained by pooling instead of convolution. This method reduces the quantity of convolution layers on the premise that the backbone is deep enough. The pooling of different sizes divides the feature mapping into different sub-regions and forms a set of different positions for the land cover image. Pyramid pooling module integrates the features of land cover images in four different scales, fully aggregating the context information and global information for different regions. Then, the CBAM attention module is also added [33]. It can change the way of resource allocation according to the importance of the target, so that the resources are more inclined to the construction and water objects in the land cover image, which solves the problem of edge blurring and small object omission because the traditional neural network does not pay enough attention to important objects. The traditional attention mechanism in the deep learning network is more concerned with the analysis of the channel domain, which is limiting the relationship between feature map channels. CBAM begins with the two scopes of channel and space. The sequential attention structure from channel to space is realized by introducing two analytical dimensions of spatial attention and channel attention. Spatial attention might direct the neural network to focus more on the pixel regions that are important for segmentation and ignores the unimportant parts. The feature map channel's distribution connection is processed using channel attention, and the attention across the two dimensions strengthens the influence of the attention mechanism.

It can improve the representation of the model's, effectively reduce the interference of invalid targets such as trees, cars and seats in the land cover image, raise the effectiveness of segmentation of the concerned target, and overall improve the accuracy of real-time semantic segmentation. Finally, maxpool, 1 × 1 Conv, three 3 × 3 Atrous Convolution and four AdaptiveAvgPool are combined. Then, they focus on important features through the CBAM module to suppress unnecessary features. Finally, the multi-scale feature extraction space pyramid module is combined.

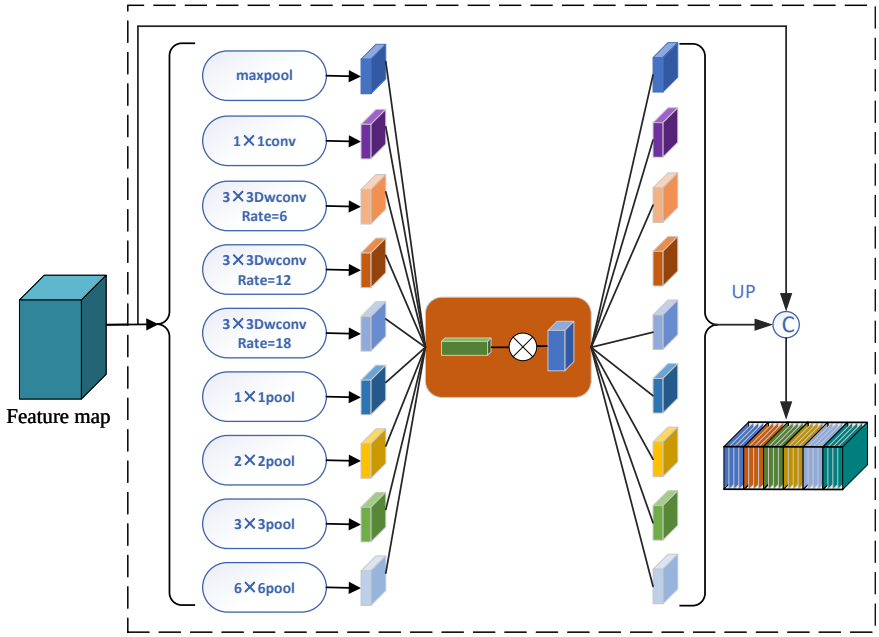

Figure 2. Multi-Scale Feature Extraction Space Pyramid Module

2.4. Multi-Scale Feature Fusion Module

In real-time land cover semantic segmentation, the complex background and interlaced housing construction make small buildings and rivers very difficult to identify [34]. Higher resolution, additional location and detail information can be found in low-level features. Stronger semantic information is present in high-level features; however, these features have limited resolution and poor detail perception. In order to address the issue of incomplete data at various scales of land cover photographs, we fused multi-scale information. Therefore, the blending of characteristics on various scales improves the segmentation performance by combining the segmentation outcomes of various layers. In addition, the existing network generally uses maximum pooling or average pooling to process channels, which will lose the spatial position information of the object. In lightweight networks, model capacity is strictly limited. The application of attention is very lagging, mainly because the computational overhead of most attention mechanisms is unaffordable for lightweight.

As a result, this study develops a multi-scale feature fusion module. Figure 3 illustrates its particular structure. It uses X Avgpool and Y Avgpool [35] for average pooling in two dimensions. The module incorporates channel attention with location data. In contrast to channel attention only, 2D global pooling transforms the input into an individual feature vector. So as to analyze channel attention, the module employs encoding of two 1D features techniques that aggregate features in several directions. This allows for the acquisition of dependencies over a long distance in a single spatial direction and the retention of precise position data along another. This is very important because the target position detection is often inaccurate or the target position cannot be directly detected in the land cover segmentation. The produced land cover feature maps are then independently programmed

to create a pair of position-sensitive and direction-aware feature maps. They can be applied in conjunction with the input feature map to make the target of interest more accurately represented. This is to encode exact location information and draw attention to the breadth and height of the land cover image. Prior to obtaining the feature maps in the width and height directions, the input feature map is first separated into two directions for global average pooling. Each channel has its own unique horizontal and vertical encoding for a particular input X using the pooling kernel's spatial ranges (H, 1) or (1, W). As a result, the channel c output at height a can be written as:

$$z_c^a(a) = \frac{1}{W} \sum_{0 \le n \le W} x_c(a, n), \tag{3}$$

where channel c output at width b can be expressed as follows:

$$z_c^b(b) = \frac{1}{H} \sum_{0 \le m \le H} x_c(m, b), \tag{4}$$

where a is the height, b is the width, c is the cth channel, n is the cth channel's nth position pixel with height a, and m is the pixel at the mth position of the cth channel, with width b. Shifting in two directions also enables the attention module to save the exact position information in one spatial direction and long-range dependency along another. Afterward, two feature maps are produced via cascading. Using a shared 1×1 convolution transformation, the spatial data in the horizontal and vertical axes are combined to create a feature map in the middle of $f \in R^{C/r \times (H+W)}$. The following is the expression for f:

$$f = \delta\Big(F\Big(\Big[z^a, z^b\Big]\Big)\Big), \tag{5}$$

where δ is a nonlinear activation function [36], a represents height, and b represents width. f is then split into two distinct tensors $f^a \in R^{C/r \times H}$ and $f^b \in R^{C/r \times W}$ along the spatial dimension. Two 1×1 Conv F_a and F_b are used to transform the feature maps f^a and f^b to the same number of channels as the input X. The outcome is as shown below.

$$g^a = \delta(F_a(f^a)), \tag{6}$$

$$g^b = \delta\Big(F_b\Big(f^b\Big)\Big). \tag{7}$$

Then, it extends g^a and g^b. As the attention weight, its final output expression is as follows:

$$g_c(n, m) = x_c(n, m) \times g_c^a(n) \times g_c^b(m). \tag{8}$$

Finally, the low-level features are changed by 3×3 Conv, batch normalization and ReLU, and multiplied by the above high-level feature extraction module. The advanced feature is then changed in the number of channels by 1×1 Conv, and the output results after changing the channel are added to the above multiplied output results. These elements come together to form a multi-scale feature fusion module. From the point of view of effectiveness, the multi-scale feature fusion module enhances the pixel information and spatial information at the background edge of the land cover image, and the dependence between the capture channels can also well model the location information and long-range dependence. It weakens or eliminates interference information such as debris, forest shade, road and so on, and combines the high-level features and low-level features. It can better aggregate different scale features and solve the problem that the contour and size of objects are different. It solves the problem of pain points such as serious interference of objects or loss of multi-scale information in land cover segmentation tasks, thereby improving the accuracy of real-time semantic segmentation of land cover.

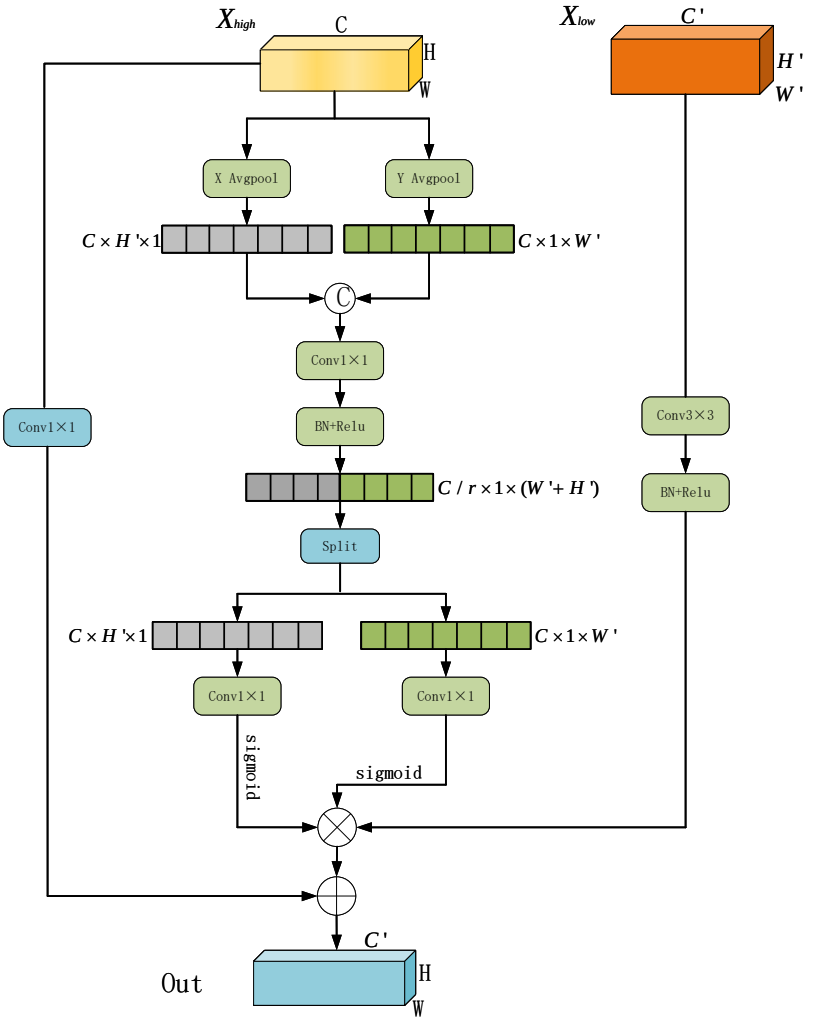

Figure 3. Multi-scale feature fusion module structure diagram.

2.5. Multi-Scale Convolutional Attention Module

In semantic segmentation of land cover, complex background and staggered house rivers make the segmentation difficult. Therefore, in the semantic segmentation, it is particularly crucial to concentrate on key targets, effectively decrease the interference of invalid targets, and enhance the module's representation capability [37]. In order to address the issue of low attention ability of traditional convolutional neural network to waters and buildings, a multi-scale convolution attention module is proposed; it could reduce the interference of invalid targets such as tree shadows. Furthermore, it makes the edge segmentation of buildings and waters is more accurate, and the information of small rivers is more detailed. Simultaneously, the 3D attention module is used to effectively alleviate or solve the shortcomings of channel attention and spatial attention in the segmentation. Its specific structure is shown in Figure 4. Finally, this method improves the accuracy of land cover semantic segmentation. Its specific structure is shown in Figure 5.

A crucial component of the multi-scale convolution attention module is the SimAM module [38]. The spatial attention or channel modules are distinct from the SimAM module. Without using any other parameters, the module produces 3D attention weights for the feature map. The 1D channel attention module treats different channels differently and treats all locations equally [39]. The 2D spatial attention module treats different locations differently and equally for all channels [40]. Compared with the existing channel and spatial

attention modules, it provides three-dimensional attention weights for feature mapping in the feature layer. In the face of the complex and diverse background of land cover, it is often difficult for ordinary attention modules to combine such complex variables, but the SimAM module maps the three-dimensional attention weight, which solves the pain point that the attention module does not perform well in the complex backgrounds. Its particular structure is displayed in Figure 4.

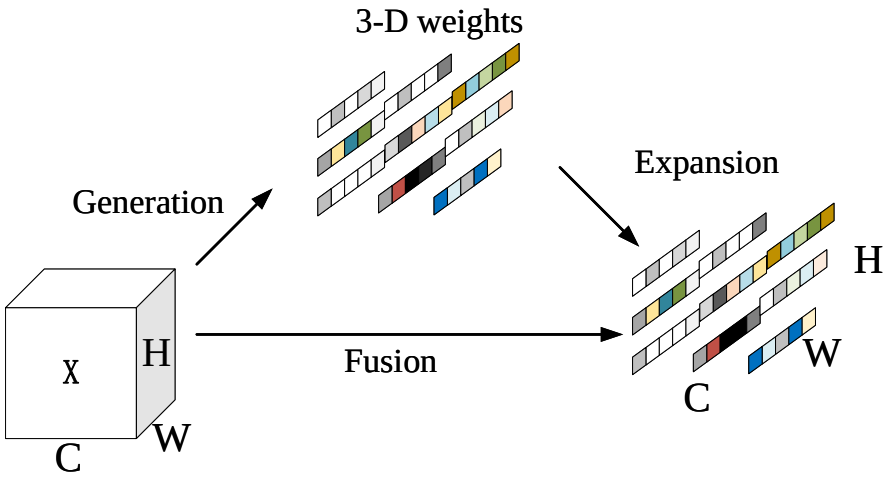

Figure 4. SimAM module structure diagram.

In the multi-scale convolution attention module design, the input value is first combined with the feature map of the multi-scale feature extraction space pyramid module and the 1×1 conv reduction channel to increase the input feature map, with more detailed information and semantic information. Then, through 1×1 Conv to reduce the quantity of channels, to avoid too many channels and information is too redundant, computing waste. The output feature maps are then averaged and maximum pooled to produce two $1 \times 1 \times C$ feature maps. After that, a neural network with two layers is supplied with the data (MLP) [41]. ReLU serves as the activation function, and C/r is the amount of neurons in the first layer, where r is the reduction rate. The second layer of neural networks has C neurons, and the two levels are shared. The ultimate channel attention feature map is then produced from the MLP output's features using a sigmoid activation technique. Following that, an element-wise multiplication operation is carried out on the output and input feature maps. Finally, the feature maps before the Maxpool and the Avgpool are subjected to the SimAM module, and the two outputs are added to obtain the output of the first layer. The second layer of this module's module uses the feature map output from the first layer as its feature map input. First, two $H \times W \times 1$ feature maps are produced using AdaptiveMaxPool and AdaptiveAvgPool. Then, one concats the two feature maps based on the channel. The dimension is then decreased to 1 channel with a 7×7 convolution operation, and the feature map's size is $H \times W \times 1$. Then, through the sigmoid, the output and the module input feature map are multiplied. In order to create the final generated map, the output feature map of the first layer is handled by the SimAM module, the two outputs are combined, and channel splicing is conducted with the input feature map. The multi-scale convolution attention module can change the allocation of resources according to the importance of the target, so that the neural network focuses more on the pixel area that has a significant impact on the segmentation and ignores the irrelevant area. Specifically, the resources are more inclined to houses and rivers. Taking into account both 2D attention and 3D attention, the problem that the ordinary attention module cannot take into account both the land cover image channel and space is solved, and effectiveness of the entire land cover's semantic segmentation is enhanced.

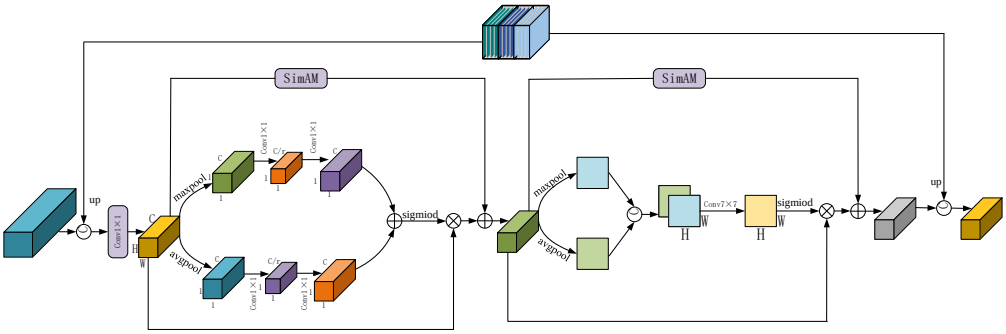

Figure 5. Multi-scale convolution attention module.

2.6. Visual Contract of Different Modules

The model's focus on various areas of the image is depicted by the heat map. Following the red zone in importance in the image are the orange and green regions, and the blue region is typically the area that does not require attention. However, the heat map is not equivalent to the final segmentation result. Specifically, on the land cover dataset, housing construction and water area are the most concerned areas, which are shown in red. Areas of secondary interest include the building handover and water edge detail, which are depicted in yellow and green, respectively. Additionally, the model must take into account the fact that complex background details such as trees and traffic would affect the segmentation [42]. The particular image is displayed in Figure 6.

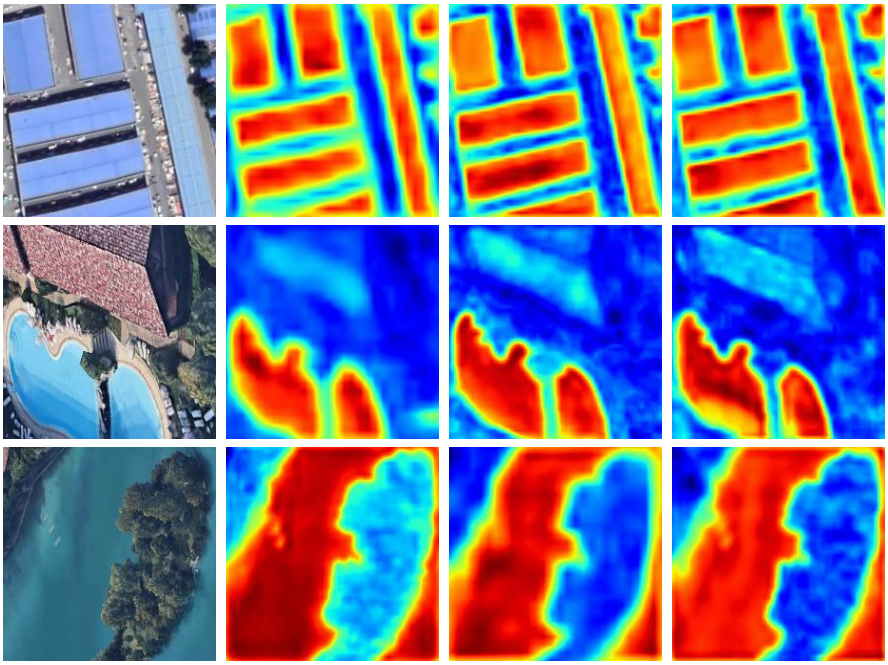

Figure 6. Visual contracts for different modules. (**a**) Test image; (**b**) method without MFFM and MCAM; (**c**) method without MCAM; (**d**) method with both MFFM and MCAM.

Figure 6b,c show the visual effects of buildings or waters without or with MFFM. By comparison, we find that models with MFFM can identify targets more clearly and focus more on the building and the waters. The segmentation results without MFFM are more blurred, and the attention to housing construction and the water area is not enough, as in the other modules. Heat maps better demonstrate this effect.

3. Experiment

3.1. Dataset Introductions

3.1.1. Land Cover Dataset

We developed a land cover dataset to train the model and assess the precision of its segmentation so as to validate the model's performance. The datasets in this study was created using high quality Landsat8 satellite and Google Earth (GE) remote sensing photographs [43]. Landsat8 satellite was successfully launched by Atlas-V rocket on 11 February 2013. Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) are carried by Landsat8. A 15-m panchromatic band with an imaging width of 185 × 185 km is a band among nine with a 30 m spatial resolution that make up the OLI land imager. The land cover dataset used in this paper selects a wide range of space and different angles for shooting. Therefore, the background environment of the dataset is complex and the model performance requirements are high. We selected 10,000 Google Earth images of land cover, which are 1600 × 900 high-resolution images intercepted on Google Earth (GE), including European rural parks and American private homes and coastal lake residential areas in Asia. Next, the captured high-resolution image is sliced into 224 × 224 images and processed using data enhancement in image processing [44]. The sliced image is then horizontally flipped (50%), vertically flipped (50%), and randomly rotated (-10° to 10°). This method not only increases the size of the land cover data set but also improves the way the data set interferes with the model-training process and the model's capacity for generalization. These land cover images are manually marked as three different object types, namely buildings, waters and backgrounds. The final partial land cover dataset is shown in Figure 7. Then, the segmented and labeled images are randomly split into a training set and a verification set in a 6:4 ratio.C

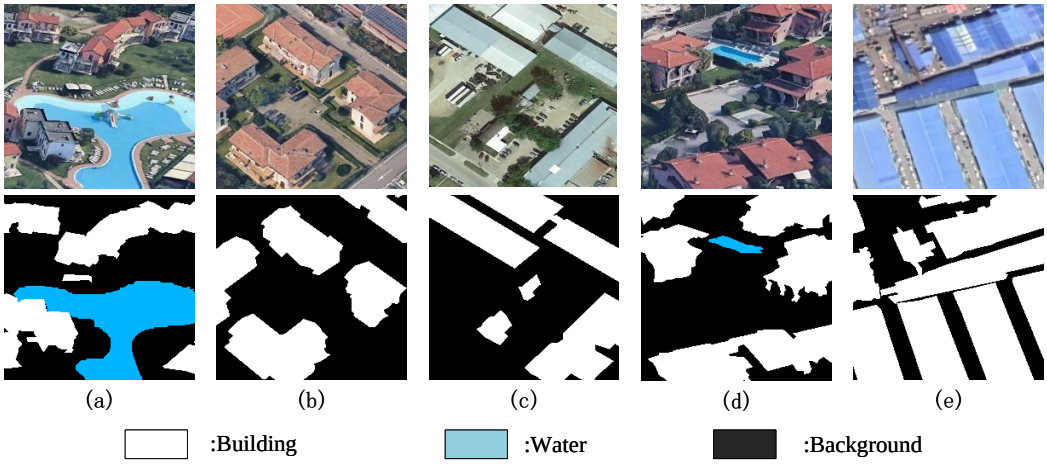

Figure 7. Homemade land cover dataset part of the picture display. (**a**) Villa area; (**b**) Urban residential area; (**c**) Rural area; (**d**) Urban residential area; (**e**) Factory area.

The land cover dataset used in this study possesses the following features: (1) In the image, containers, big vehicles, and comparable buildings and some roofs of houses are similar to water and background, which further highlights the representation ability of the model. This is as depicted in Figure 7c–e. (2) As a result of the various intercepted angles, the data set image shows a particular positional deviation, not only a single pitch angle, as shown in Figure 7a,b,d. (3) Land cover background is complex and diverse, including farmland, factories, roads and houses, which highlights the model's capability to segment. (4) Buildings in residential neighborhoods have a fairly wide range in height. The semantic segmentation of low-rise buildings will be hampered by the shadow cast by high-rise structures; this could further test the model's capacity to prevent interference. (5) In order to gather precise segmentation data, it is necessary to segment all the ground observations manually, but due to various factors, land cover data sets are difficult to achieve perfectly.

3.1.2. Waters' Segmentation Dataset

The water segmentation dataset serves as the test dataset in this study to verify the model's capacity for generalization. This dataset was manually segmented from high-resolution remote sensing photos taken by the Landsat8 satellite and from Google Earth (GE). In order to make the data more real and diverse, we selected a broad spectrum of distribution, and as to the choice of water, we selected rivers of different colors, shapes and widths. In addition, we selected a complex and diverse background, including urban, hilly, woodland, farmland and other areas, which can fully verify the model's ability to deal with different scenarios. The average size of Landsat8 satellite images is 10,000 × 10,000 pixels. We intercepted multiple images with a resolution of 4800 × 2742 from Google Earth (GE), and then cut them into 256 × 256 images. Finally, through data enhancement operations such as translation, rotation, and folding, 9580 images were finally obtained. To overcome the problem of overfitting, we used HoldOut cross-validation to randomly split and label the water dataset into training and validation sets at a ratio of nearly 7:3, 7185 were used as training datasets and 2395 as validation datasets. Some of the water segmentation data sets were exhibited in Figure 8.

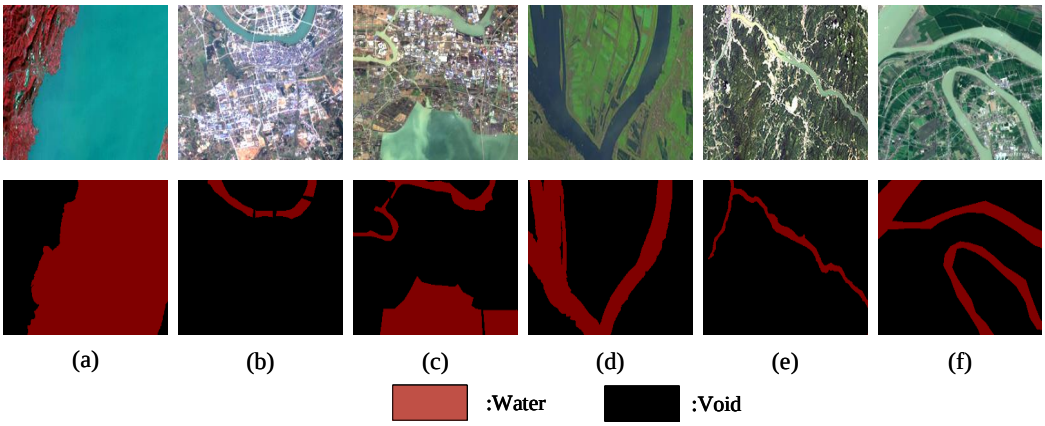

Figure 8. Waters' segmentation dataset part of the picture display. (**a**) Sea; (**b**) Urban residential area; (**c**) Coastal area; (**d**) Rivers; (**e**) Highland area; (**f**) Terraced area.

3.2. Implementation Details

All the experimental tasks in this paper are carried out on the Intel Core i5-11400 CPU and NVIDIA RTX3070 GPU, Windows10 and pytorch deep learning framework. In this experiment, the initial learning rate of the model network is set to 0.001, the batch size is 12, and the iteration period is 300. In this paper, precision rate (P), recall rate (R), harmonic mean (F_1), pixel accuracy (PA), average pixel accuracy (MPA) and average intersection ratio ($MIOU$) are used as evaluation indexes. The calculation formulas are as follows:

$$P = \frac{(TP)}{(TP) + (FP)}, \tag{9}$$

$$R = \frac{(TP)}{(TP) + (FN)}, \tag{10}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}, \tag{11}$$

$$PA = \frac{\sum_{n=0}^{t} p_{n,n}}{\sum_{n=0}^{t} \sum_{m=0}^{t} p_{n,m}}, \tag{12}$$

$$MPA = \frac{1}{t+1} \sum_{n=0}^{t} \frac{p_{n,n}}{\sum_{m=0}^{t} p_{n,m}}, \tag{13}$$

$$MIOU = \frac{1}{t+1} \sum_{n=0}^{t} \frac{p_{n,n}}{\sum_{m=0}^{t} p_{n,m} + \sum_{m=0}^{t} p_{m,n} - p_{n,n}}, \tag{14}$$

where TP is the sample that was really positive and was anticipated to be positive, FN was expected to be a negative sample but is really positive, t denotes the class of object segmentation (excluding background), $p_{n,n}$ denotes the true quantity, $p_{m,n}$ denotes the number of pixels in category n that are actually expected to be in category m.

3.3. Ablation Experiment

To test the efficacy of various modules, several modules are introduced to the backbone network in the ablation experiment. In the ablation experiment, this paper uses the modified ResNet50 as the backbone network for feature extraction. In this experiment, mean intersection-over-union (MIOU) is used to evaluate the performance of different modules. The optimum land cover semantic segmentation is achieved by combining all modules, as indicated in Table 1.

Table 1. Performance comparison of different modules in the model.

Method	Mean IOU (%)
Baseline	85.22
Baseline+MFESP	86.92
Baseline+MFESP+MFFM	87.46
Baseline+MFESP+MFFM+MCAM	88.05

The ablation of spatial pyramid modules for multi-scale feature extraction proved effective. The pyramid module captures multi-scale context information by using maxpool, 1×1 Conv, three 3×3 Atrous Convolution and 4 AdaptiveAvgPool. At last, the CBAM is used to enhance the attention to important targets and reduce the interference of invalid targets, so as to increase the module's capacity for representation. As displayed in Table 1, the multi-scale feature extraction pyramid module increases the model MIOU from 85.22% to 86.92%.

Ablation for multi-scale feature fusion modules proved effective. Low-level features have more noise and less convolution, but they have greater resolution and more location and detail information. Stronger semantic information is present in high-level features; however, these features have limited resolution and poor detail perception. As a result, by combining features from several scales, the multi-scale feature fusion module enhances the module's performance. As displayed in Table 1, The model MIOU is further enhanced by 0.54% thanks to the multi-scale feature fusion module proposed in this research.

The ablation for multi-scale convolutional attention modules is proven effective. Change the allocation of resources according to the importance of the goal of attention, so that resources are more inclined to the goal of attention. As displayed in Table 1, the multi-scale convolution attention module improves the model MIOU to 88.05%.

3.4. Generalization Experiment

To further confirm the generalizability of the proposed model, it is evaluated on the water segmentation dataset to verify. We selected PA, MPA and MIOU as the indicators to evaluate the generalization ability of the model. The comparison results are displayed in Table 2. The model proposed in this study also outperforms other models on the water segmentation dataset, which verifies that our model not only performs well in the semantic segmentation task of land cover, but has strong performance in multi-class task segmentation.

Figure 9 demonstrates the segmentation performance of different models on water data sets. From the figure, we can observe that traditional semantic segmentation networks such as FCN8sAtOnce perform poorly in water segmentation. They can only segment approximate contour and shape; they miss a lot of detail information such as edges and boundaries. SegNet has improved the segmentation effect of river waters, as shown in the fourth group of Figure 9, but the improvement could only be observed occasionally. Finally, ShuffleNetV2 and UNet have improved the segmentation of water segmentation data set obviously, but their edge detection is not clear enough, and their detection of small tributaries is poor. In order to enhance the capacity to gather global information, the model put forward in this research suggests a multi-scale feature extraction pyramid module that can aggregate information from various scale contexts. This module combines the characteristics of various scales, which is the key to improve its performance. A multi-scale convolution attention module is also proposed, it improves the representation ability of the model and effectively reduces the interference of invalid targets, thus performing well in water segmentation tasks.

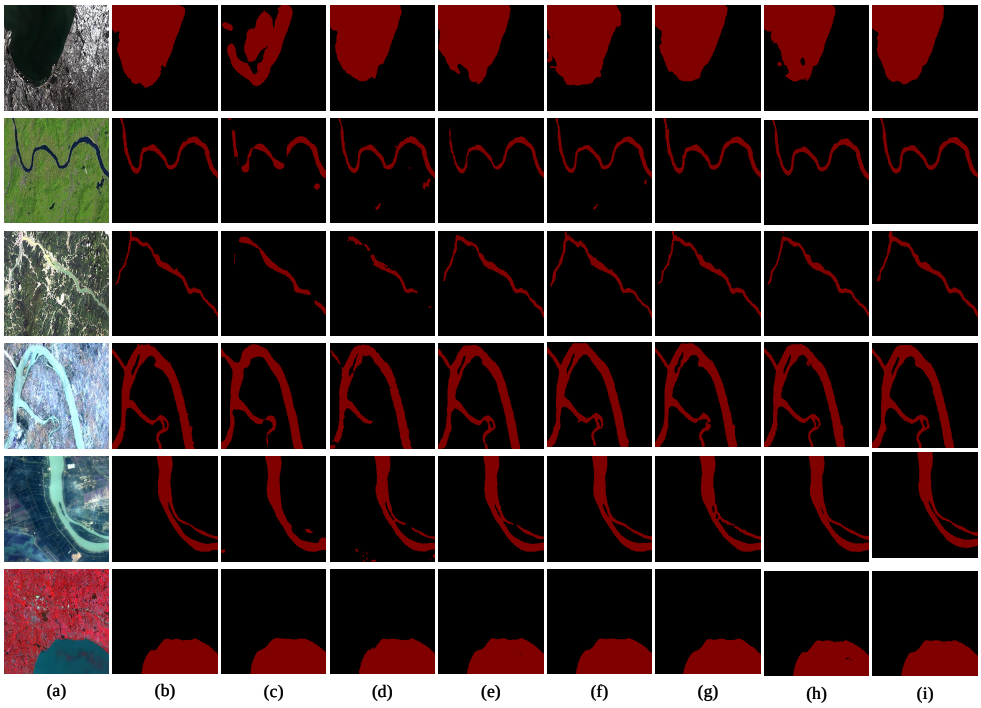

Figure 9. Comparison of different models on on the water segmentation datasets. (**a**) Real image; (**b**) label; (**c**) FCN8s; (**d**) SegNet; (**e**) DFNet; (**f**) DeepLabV3+; (**g**) ShuffleNetV2; (**h**) UNet; (**i**) our segmention results.

Table 2. Comparison of evaluation indexes of different models on the water segmentation dataset.

Method	PA (%)	MPA (%)	MIOU (%)
GhostNet	96.80	96.32	91.61
FCN8sAtOnce	97.45	97.01	93.28
SegNet	97.48	97.04	93.35
DANet(ResNet50)	98.03	97.72	94.76
DFNet	98.11	97.82	94.96
DeepLabv3+(ResNet101)	98.08	97.72	94.89
CCNet(ResNet50)	98.10	97.83	94.93
ShuffleNetV2	98.17	98.02	95.10
UNet	98.23	98.13	95.26
Ours	98.81	98.86	96.06

4. Discussion

4.1. Research Results

The experiment in this paper only performs three-category segmentation at most, and its complexity and segmentation difficulty are far lower than multi-category segmentation. It distinguishes fewer categories and is not suitable for using too many convolution kernels and very deep convolution layers, otherwise it will lead to information redundancy or loss of important semantic information. Therefore, this article only uses the first four layers of the modified ResNet50 to avoid this problem. Compared with other classical networks, first of all, the classical convolutional neural network will inevitably lose the context information of different regions during the downsampling process, resulting in the loss of global information. Our network uses the multi-scale feature extraction spatial pyramid module (MFESP) to aggregate the context information of multiple regions, thereby improving the control ability of global information. Secondly, the segmentation of large buildings and rivers requires shallow features with a global receptive field; small houses and tributaries at the bottom of the high-resolution prediction is better, because some details such as the edge of the enlarged prediction results are better. Traditional neural networks often take care of one thing and lose another, so our proposed multi-scale feature fusion module (MFFM) better balances different degrees of features. Finally, focusing on the segmentation object is the key to high-precision segmentation. Traditional neural networks are often overwhelmed in the face of complex and changeable land cover scenarios. The proposed multi-scale convolutional attention module (MCAM) pays more attention to buildings and waters in land cover. In summary, this work has four contributions:

1. A pyramid module for multi-scale feature extraction is proposed. The context semantic information of different regions is aggregated by multiple dilated convolutions and global pooling to get global information acquisition capabilities. Global information can effectively produce high-quality results in land cover analysis.
2. Module for multi-scale feature fusion is proposed. Performance can be increased by combining data from multiple scales. It completely makes use of both the extra location and detail information provided by low-level features and the extra semantic data provided by high-level features.
3. A multi-scale convolution attention module is proposed. It can improve the representation ability of the model, effectively reduce the interference of invalid targets, improve the segmentation of concerned targets, and enhance the model's overall segmentation precision.
4. Three models are proposed and combined to build a high-precision semantic segmentation network to attain highly accurate land cover segmentation.

4.2. Comparative Experiments with Other Classical Networks

We compare our proposed semantic segmentation model of land cover with other classical semantic segmentation networks in the comparison experiment with other classical networks to completely test the performance of our model. We select R, F_1, PA, MPA and MIOU as evaluation indexes to fully demonstrate the superiority of our model. The details exhibited in Table 3.

As shown in Table 3, in the experimental environment with the same learning rate, batch size, iteration period and other factors, the comparison results of different models clearly demonstrate that our proposed land cover semantic segmentation model is superior to current segmentation algorithms in all indicators. In all networks, the performance of DDRNet model [45] is the worst. With the continuous improvement of the classical model, its performance evaluation indicators are also growing, but our proposed model still performs best in the semantic segmentation task of land cover.

According to Table 3, the proposed model can achieve high-precision real-time semantic segmentation of land cover datasets. To demonstrate the effectiveness of our approach, Figure 10 shows some comparison experimental results of the semantic segmentation of land cover. The segmentation results of our network are contrasted with the segmentation

results of the classical models, where white represents the building, blue represents the water, and black represents the default background.

Table 3. Comparison of evaluation indexes of different models.

Method	R (%)	F_1 (%)	PA (%)	MPA (%)	MIOU (%)
DDRNet	95.67	94.66	91.11	90.93	82.26
BiseNetv2 [46]	96.21	95.62	91.07	91.69	82.45
FCN8s	93.39	94.87	91.04	90.68	82.61
SegNet	94.65	95.76	91.02	91.17	82.85
HRNet(HRNetv2-W48)	96.41	95.40	92.28	92.06	84.40
DFN(ResNet101)	96.72	96.31	92.18	92.24	84.74
OCRNet(ResNet101)	96.94	95.82	92.60	92.74	85.03
ENet	97.22	97.43	92.55	92.86	86.04
DANet(ResNet50)	96.40	96.55	92.89	92.76	86.16
DeepLabv3+(ResNet101)	95.85	96.83	92.68	92.44	86.19
UNet	96.98	96.27	93.13	92.86	86.32
PSPNet(ResNet101)	97.12	97.22	92.93	92.93	86.61
CCNet(ResNet50) [47]	96.39	96.66	93.18	93.19	86.66
ACFNet(ResNet50) [48]	96.86	97.19	93.30	93.69	86.69
Ours	97.25	97.50	93.82	94.17	88.05

In order to more intuitively show the representation ability of the model in the semantic segmentation task of land cover, As shown in Figure 10, it can be observed that our network can effectively mitigate background interference on the segmentation of buildings and waters, as well as the interference of different perspectives on segmentation. Specifically, as shown in the third line of Figure 10, when container area is used as the segmentation object, because the multi-scale feature extraction space pyramid module (MFESP) is used in our network, it can aggregate multi-scale context information and eliminate the issue of lost detail information and resolution reduction in the sampling process of land cover semantic segmentation, so it can still clearly segment the details and corners of the container area edge. As shown in the first line of Figure 10, because of the use of multi-scale feature fusion module (MFFM), the semantic information and detail information of different scales are fused, so the edge detail segmentation is excellent and the object distinction is clear. As display in the second, fourth and seventh lines of Figure 10, even the trees block the house and the interception angle is different, more attention was paid to the building and the water area, and better segment the water area and the building is achieved because of the multi-scale convolutional attention module (MCAM).

In addition, we add the prediction maps generated by the test results in multiple complex environments such as desertification areas, urban residential areas, reservoirs, and container areas to Figure 10. As shown in column d of Figure 10, the FCN8s network often has partial segmentation targets that cannot be detected, and our network uses the multi-scale convolutional attention module (MCAM) module to increase the focus on land cover segmentation targets. All the target objects to be segmented in the land cover image are detected. As shown in the e column of Figure 10, the SegNet network has problems such as blurred edge and misjudgment, whereas our network aggregates the context features of different regions and pays more attention to land cover segmentation targets. Therefore, these problems are perfectly solved, as shown in the i column. As shown in column g of Figure 10, the UNet network clearly solves the problem of FCN8s, but it does not perform well on building edges and small tributaries. Our network uses multi-scale feature extraction space pyramid module (MFESP) and multi-scale feature fusion module (MFFM) modules to aggregate global information and performs well on segmentation of edge details and small tributaries. We can clearly find that our model can achieve efficient and reliable segmentation of land cover images in multiple complex environments.

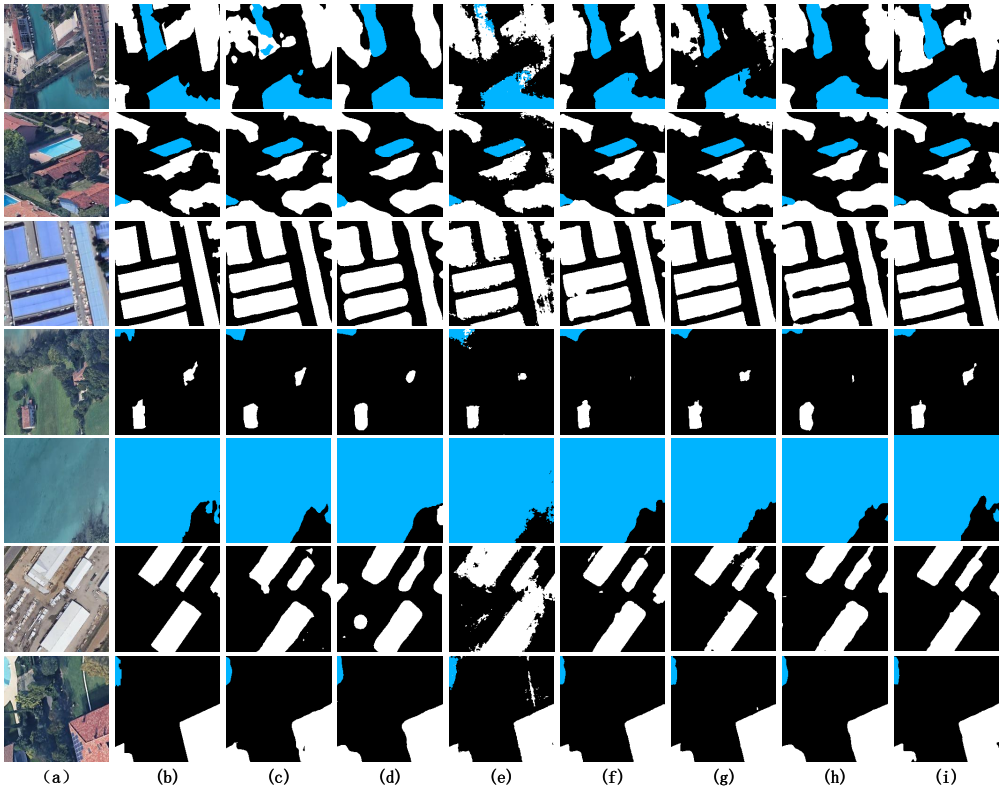

Figure 10. Comparison of different models on land cover dataset. (**a**) Real image; (**b**) label; (**c**) BiseNetv2; (**d**) FCN8s; (**e**) SegNet; (**f**) OCRNet; (**g**) UNet; (**h**) PSPNet; (**i**) our segmentation results.

4.3. Limitations and Future Research Directions

However, the land cover segmentation method in this paper still has some shortcomings, and it needs to be further optimized and perfected in later work. (1) The multi-scale feature extraction spatial pyramid module integrates too many scale features, resulting in redundant feature information. The number of channels in the feature map reaches 3328, and the model is too heavy. (2) Too dependent on the attention module, which is very easy to lead to overfitting, the model's ability to remember the data set is relatively enhanced, so only large-scale data set pre-training achieves better results, increasing the computational cost. (3) Aggregating multi-scale features and fusing different scale features make the Flops and Parameters of the model relatively high, which is compute resource consuming. We will conduct further research based on the limitations in the model to achieve further efficient and high-precision segmentation of land cover segmentation tasks.

5. Summary

A key component of processing of remote sensing images with a high level of resolution is the land cover semantic segmentation, which is a crucial landmark in remote sensing images. It is very practical in land resource protection planning, geographical classification, surveying and mapping analysis. This research proposes a multi-scale feature aggregation network for real-time semantic land cover segmentation in the area of deep learning picture segmentation. In order to address the issues with classic convolutional neural networks' poor generalization capacity and low accuracy in the context of semantic land cover segmentation, the network draws for convolutional neural networks' benefits in deep learning, and uses the modified ResNet50 as the backbone network to extract image feature information. To combine the context data from many regions, a multi-scale feature extraction space pyramid module is proposed. This module can efficiently and reliably categorize regions of any scale, enhancing the potential to gather global information. It

is suggested to use a multi-scale feature fusion module. The low-level feature is more detailed and has a better resolution, but its semantics are weaker and its noise is greater because there are fewer convolutions. Although high-level characteristics have weaker resolution and poorer detail perception, they have stronger semantic information. The fusion of high-level and low-level features helps to extract more information from the image. Finally, a multi-scale convolution attention module is proposed, which can pay more attention to the building and the river, and reduce the interference of complex backgrounds such as trees and roads. Compared with the traditional convolutional neural network model, the model in this paper greatly improves the accuracy of real-time semantic segmentation of land cover, and captures details such as small tributaries and house edge contours quickly. The experimental results demonstrate that the average intersection over union (MIOU) of this method on the land cover dataset reaches 88.05%. The generalization ability of the network is also very strong; it could reach 96.06% on the water segmentation dataset.

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

Funding: This work is supported in part by the National Natural Science Foundation of PR China (42075130).

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request.

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

References

1. 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]
2. Li, M.; Zang, S.; Zhang, B.; Li, S.; Wu, C. A review of remote sensing image classification techniques: The role of spatio-contextual information. *Eur. J. Remote Sens.* **2014**, *47*, 389–411. [CrossRef]
3. 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**, *43* , 1–23. [CrossRef]
4. 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**, *43*, 5874–5894. [CrossRef]
5. Zhang, H.; Jiang, Q.; Xu, J. Coastline extraction using support vector machine from remote sensing image. *J. Multim.* **2013**, *8*, 175–182.
6. 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]
7. Du, Y.; Feng, G.; Li, Z.; Peng, X.; Ren, Z.; Zhu, J. A method for surface water body detection and DEM generation with multigeometry TanDEM-X data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *12*, 151–161. [CrossRef]
8. Leng, Y.; Liu, Z.; Zhang, H.; Wang, Y.; Li, N. Improved ACM algorithm for Poyang lake monitoring. *J. Electron. Inf. Technol.* **2017**, *39*, 1064–1070.
9. 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. [CrossRef]
10. 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]
11. 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]
12. Mohamed, A.; Dahl, G.E.; Hinton, G. Acoustic modeling using deep belief networks. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *20*, 14–22. [CrossRef]
13. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
14. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]

16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
17. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
18. 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]
19. 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]
20. 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]
21. Lei, S.; Min, X.; Liguo, W.; Haifeng, L.; Ming, Q.; Binyu, C. Axial Cross Attention Meets CNN: Bi-Branch Fusion Network for Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022** . [CrossRef]
22. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
24. Cheriyadat, A.M. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 439–451. [CrossRef]
25. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2017.
29. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
30. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
31. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
32. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
34. 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]
35. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
36. Dubey, A.K.; Jain, V. Comparative study of convolution neural network's relu and leaky-relu activation functions. In *Applications of Computing, Automation and Wireless Systems in Electrical Engineering*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 873–880.
37. 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**, *43*, 5940–5960. [CrossRef]
38. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11863–11874.
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
40. Ye, Q.; Yuan, S.; Kim, T.K. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 346–361.
41. Karlik, B.; Olgac, A.V. Performance analysis of various activation functions in generalized MLP architectures of neural networks. *Int. J. Artif. Intell. Expert.* **2011**, *1*, 111–122.
42. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
43. Huang, H.; Chen, Y.; Clinton, N.; Wang, J.; Wang, X.; Liu, C.; Gong, P.; Yang, J.; Bai, Y.; Zheng, Y.; et al. Mapping major land cover dynamics in Beijing using all Landsat images in Google Earth Engine. *Remote Sens. Environ.* **2017**, *202*, 166–176. [CrossRef]
44. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 113–123.

45. 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.
46. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
47. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
48. Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; Ding, E. Acfnet: Attentional class feature network for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6798–6807.