


Article

Research on Ground Object Classification Method of High Resolution Remote-Sensing Images Based on Improved DeeplabV3+

Junjie Fu ¹, Xiaomei Yi ^{1,2,*}, Guoying Wang ^{1,2}, Lufeng Mo ^{1,2}, Peng Wu ^{1,2}  and Kasanda Ernest Kapula ¹¹ College of Mathematics and Computer Science, Zhejiang A and F University, Hangzhou 311300, China² Key Laboratory of Forestry Intelligent Monitoring and Information Technology of Zhejiang Province, Hangzhou 311300, China

* Correspondence: yxm@zafu.edu.cn

Abstract: Ground-object classification using remote-sensing images of high resolution is widely used in land planning, ecological monitoring, and resource protection. Traditional image segmentation technology has poor effect on complex scenes in high-resolution remote-sensing images. In the field of deep learning, some deep neural networks are being applied to high-resolution remote-sensing image segmentation. The DeeplabV3+ network is a deep neural network based on encoder-decoder architecture, which is commonly used to segment images with high precision. However, the segmentation accuracy of high-resolution remote-sensing images is poor, the number of network parameters is large, and the cost of training network is high. Therefore, this paper improves the DeeplabV3+ network. Firstly, MobileNetV2 network was used as the backbone feature-extraction network, and an attention-mechanism module was added after the feature-extraction module and the ASPP module to introduce focal loss balance. Our design has the following advantages: it enhances the ability of network to extract image features; it reduces network training costs; and it achieves better semantic segmentation accuracy. Experiments on high-resolution remote-sensing image datasets show that the mIoU of the proposed method on WHDLD datasets is 64.76%, 4.24% higher than traditional DeeplabV3+ network mIoU, and the mIoU on CCF BDCI datasets is 64.58%. This is 5.35% higher than traditional DeeplabV3+ network mIoU and outperforms traditional DeeplabV3+, U-NET, PSP-NET and MACU-net networks.

Keywords: high-resolution remote-sensing images; semantic segmentation; object classification

Citation: Fu, J.; Yi, X.; Wang, G.; Mo, L.; Wu, P.; Kapula, K.E. Research on Ground Object Classification Method of High Resolution Remote-Sensing Images Based on Improved DeeplabV3+. *Sensors* **2022**, *22*, 7477. <https://doi.org/10.3390/s22197477>

Academic Editor: Marco Diani

Received: 31 August 2022

Accepted: 28 September 2022

Published: 2 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In 1999, China successfully launched the first land resources remote-sensing satellite “Ziyuan-1”, with the launch of “Gao fen”, “Tian hui”, and “Zi yuan” series satellites, China’s remote-sensing satellite-mapping technology has made considerable progress since then. As remote-sensing technology advances, more and more useful information is extracted from remote-sensing images [1]. In addition to their high spatial resolution, high timeliness and large amount of information, high-resolution remote-sensing images provide a wealth of information [2,3]. High-resolution remote-sensing image feature classification plays an important role in surface-spatial information extraction, land resource management and natural resource protection. A semantic segmentation technology is used to classify the surface of high-resolution remote-sensing images. Each pixel of the image will be tagged with a category [4], and marked with the same colour on the targets of the category.

Traditional image segmentation techniques include threshold-based segmentation [5], clustering segmentation [6] and edge segmentation [7]. These methods need to manually select appropriate feature parameters, which have good results for segmenting simple scenes. However, in complex scenes, the segmentation effects of various targets vary greatly, and there are also many limitations.

The rapid development of deep learning has enabled semantic segmentation to achieve pixel-level classification of images [8], so deep-learning networks are widely used in image segmentation, medicine, engineering technology and other fields. Zhao et al. [9] added a pyramid pooling module to the FCN [10] model and developed the PSP-Net network model to aggregate context information from different regions, thereby enhancing global information accessibility. Yang et al. [11] proposed skeleton monitoring based on MSB-FCN and showed that considerable progress has been made in applying semantic segmentation to medical images. Yu et al. [12] proposed a different crack detection method based on image by using deep convolutional neural network and Enhanced Chicken Swarm Algorithm (ECSA), and conducted comparative experiments with other crack detection methods on a variety of concrete crack samples. Fan et al. [13] proposed a microcrack detection method based on ResNet to detect microcracks on polycrystalline solar cells, which effectively improved the accuracy of detecting solar-cell defects. In 2021, on the basis of the U-Net [14] model, Li et al. [15] used asymmetric convolution to enhance the feature representation and feature extraction capability of the convolution layer, and effectively improved the problem of insufficient feature utilization of the U-Net model. The proposed MACU-Net network was tested on WHDL and GID datasets, proving that the performance of the proposed network was better than that of the U-Net network. Chen et al. [16] used hole convolution on the basis of FCN to avoid the loss of pooled information, and then added the CRF (Conditional Random Fields) module at the output of the feature extraction network to propose the Deeplab network. Chen et al. [17] proposed the DeeplabV2 network. The network architecture introduced the ASPP (Atrous Spatial Pyramid Pooling) module, which improved the receptive field of the network without increasing the number of parameters, thus improving the network performance. The DeeplabV2 network takes conditional random fields and hole space pyramid pooling as the core. Later, Chen et al. [18] proposed the DeeplabV3 network. The network removes conditional random fields, uses hole-convolution cores for many times, and optimizes the ASPP module, so that the segmentation effect is better than the DeeplabV2 network. Since the features of the DeeplabV3 network do not contain too many shallow features, Chen et al. [19] proposed the DeeplabV3+ network. The DeeplabV3+ network is considered a new milestone in semantic segmentation technology. It proposes a classical coding and decoding structure based on the DeeplabV3+ network, which fuses shallow features with deep features, so as to improve the segmentation accuracy. Although the DeeplabV3+ network has shown leading advantages in multiple public datasets in the field of semantic segmentation, the segmentation of remote-sensing images is still not fine enough.

In view of the fact that the DeeplabV3+ network needs to face the complex features of images in the process of image segmentation from remote sensing, this paper proposes to take the MobilenetV2 network as the feature extraction network, add an attention mechanism module in the network structure, and introduce focal loss to balance the loss. The segmentation performance of traditional DeeplabV3+, U-NET, PSP-NET and MACU net networks on remote-sensing datasets is compared. The experimental results show that the segmentation performance of this method is higher than that of other compared networks, and the training time of the network is significantly reduced. It shows that the improved method in this paper can effectively improve the network segmentation ability and reduce the network training cost.

2. Related Research

2.1. Deeplabv3+ Network Model

Since the DeeplabV3+ network was proposed, it has often been used for high-precision image segmentation due to its excellent image segmentation ability [20,21]. The DeeplabV3+ network has been widely used in remote-sensing image segmentation in recent years [22–24]. The DeeplabV3+ network can mine context information on a multi-scale basis, and can also use the method of reconstructing image spatial information to determine the boundary of

objects. For the input characteristic diagram X of ASPP module in the network structure, its output characteristic diagram (y) is:

$$y = \sum_{k=1}^K x[i + r \times k]w[k] \quad (1)$$

where i is the input signal, $w[k]$ is the filter value, R is the expansion rate, and K is the hole convolution length.

Deeplabv3+ network introduces a large number of hole convolutions in the encoder module, which makes the network increase the receptive field without losing information. The receptive field calculation of hole convolution is consistent with the standard convolution kernel, and the value k of the hole convolution receptive field is:

$$K = k + (k - 1)(r - 1) \quad (2)$$

where k is the size of the original convolution kernel; and R is the size of the actual convolution kernel of hole convolution.

The DeeplabV3+ network model structure mainly refers to the proposed coding and decoding structure, which not only improves the network segmentation effect, but also pays attention to the boundary information. The structure can be divided into encoder and decoder. The encoder is used for feature processing, which is divided into three steps: (1) DeeplabV3+ uses the modified aligned Xception feature extraction network for feature extraction on the network architecture, and finally generates two effective feature layers, namely, shallow features and deep features. Because the down-sampling of the deep characteristic layer is more than that of the shallow characteristic layer, the height and width of the shallow characteristic layer are larger than that of the deep characteristic layer. (2) The extracted deep features are input into the ASPP module, in which 1×1 convolution and expansion rates are 6, 12 and 18 respectively. 3×3 convolution and image global average pooling. The hole convolution with different expansion rates is used to improve the receptive field of the network, so that the network has different characteristic receptive conditions. (3) Stack the feature layers through 1×1 convolution to adjust the number of channels, and finally obtain the multi-scale feature of 256 channel digits. Shallow features generated by modified aligned Xception the generated shallow features enter the decoder, and the multi-scale features are up-sampled by 4 times, and then fused with the shallow features, after 3×3 convolution is used for feature extraction. Finally, the output picture is consistent with the input picture through 4 times of up-sampling, and the prediction result is obtained through Softmax function. Figure 1 shows the DeeplabV3+ network structure.

2.2. Modified Aligned Concept Network Model

The modified aligned Xception network is improved based on the Xception network, which is a network model proposed by the Google team in 2016. The Xception network structure consists of three flows, namely, entry flow, middle flow and exit flow [25]. The network has a total of 36 convolution layers. A deep separable convolution is introduced into the Xception network structure. Firstly, each channel is divided into a 3×3 convolution operation, and then a 1×1 convolution operation, and finally the results are merged.

The residual structure is also introduced into the Xception network structure, which plays an important role in the model effect. Suppose the input is x and the expected output is y . In the traditional linear network structure, the learning objective $f(x) = y$; In the residual structure, the input x is directly transferred to the output as the initial result, and the learning objective $f(x) = y - x$, that is, the learning objective is changed from the original y to the difference between the output and the input.

A modified aligned Xception network makes two improvements to the Xception network. One is to reduce the computational complexity by using deep separable convolution to replace the maximum pooling operation in all Xception networks. Second, in every

3 × 3 normalized and ReLU activation functions are used after the convolution operation. Figure 2 shows the modified aligned Xception network structure.

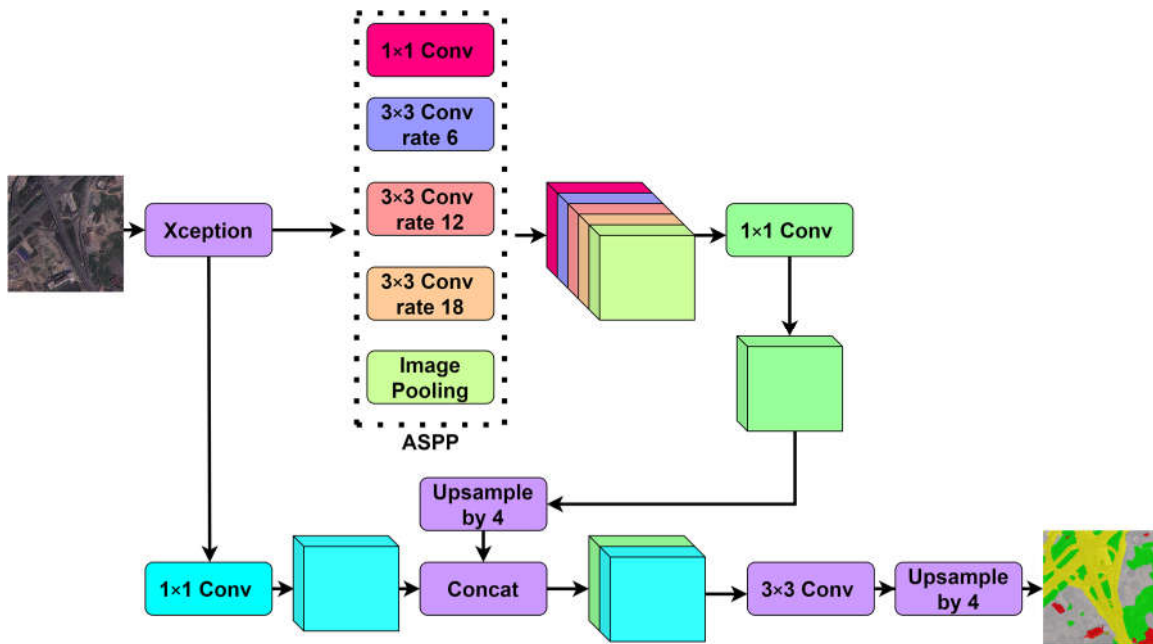


Figure 1. Structure diagram of DeeplabV3+.

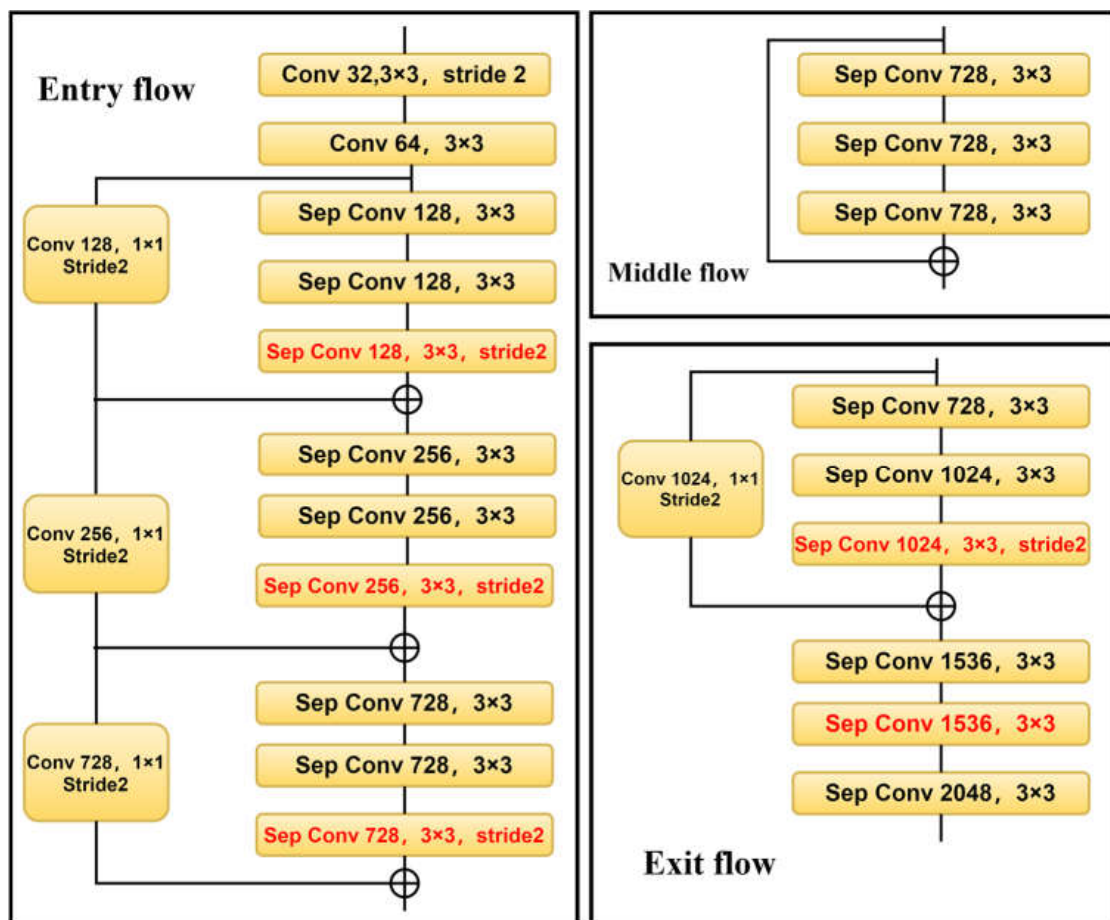


Figure 2. Structure diagram of Modified Aligned Xception.

3. Improved DeeplabV3+ Network Model

In the traditional DeeplabV3+ network model, the modified aligned Xception network is selected for feature extraction. Due to the large number of layers of the network, the large amount of parameters, and the common convolution in the ASPP module after feature extraction, the complexity of the model is large, which increases the difficulty of network training, and the network training speed is slower and the convergence is slower. In order to improve the network segmentation ability and training efficiency, this paper makes three improvements on the network structure of high-resolution remote-sensing image ground-object segmentation based on DeeplabV3+: (1) the modified aligned Xception feature extraction network used for feature extraction is changed to a lighter MobileNetV2 network. (2) In the MobileNetV2 feature extraction network and after the ASPP module, the Ca (Coordinate attention) module [26] is added to improve the segmentation accuracy of the model. (3) Optimize the loss function and introduce focal loss equilibrium loss to reduce the impact on the accuracy of model feature classification due to the large difference in the proportion of feature categories. Figure 3 shows the improved DeeplabV3+ network structure.

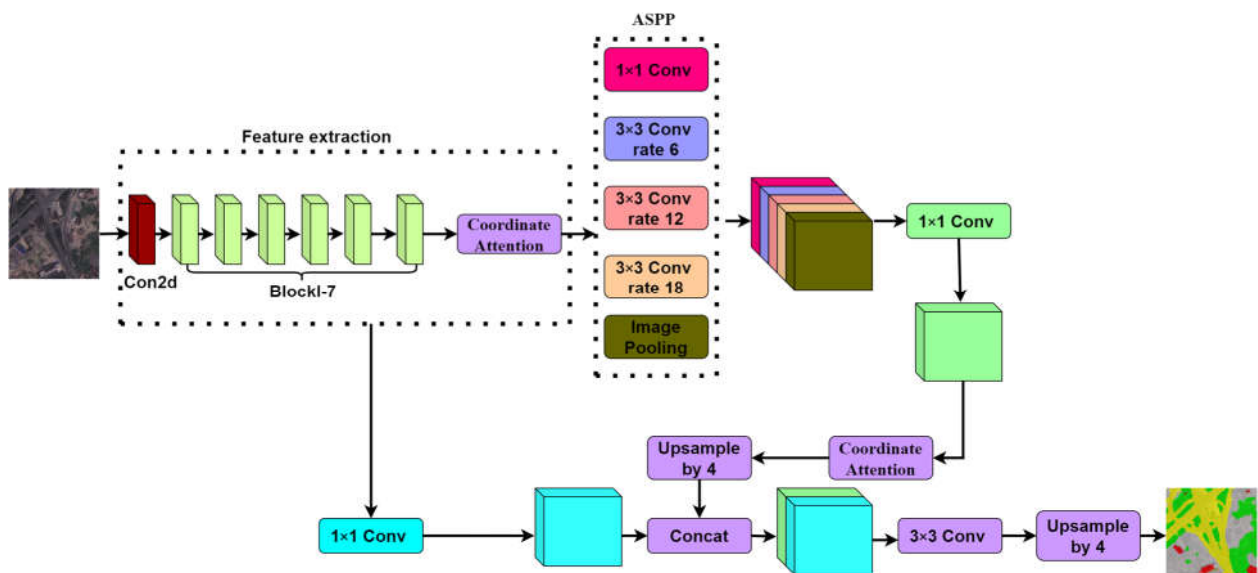


Figure 3. Improved structure diagram of improved DeeplabV3+.

3.1. Optimized Feature Extraction Module

In the coding layer, the original network used for feature extraction is changed to a more lightweight MobileNetV2 network, which is a lightweight network model proposed by the Google team in 2018 [27]. Compared with the modified aligned Xception network, the MobileNetV2 network has fewer parameters, a simpler model structure, and faster network training speed. MobileNetV2 network uses linear bottleneck structure and reverse residual structure [28].

In the feature extraction operation, the neural network extracts the useful information of the target, which can be embedded in the low dimensional subspace. The traditional network structure is normalized by convolution containing the ReLU activation function, but using the ReLU activation function in the low dimensional space will lose more useful information. In the linear bottleneck structure, the ReLU activation function is changed to the linear function to reduce the loss of useful network information.

The reverse residual structure of MobileNetV2 network application consists of three parts. Firstly, 1×1 convolution to increase the dimension of input features, and then use 3×3 depth separable convolution for feature extraction, and then 1×1 convolution for dimension reduction. The specific network structure is shown in Table 1.

Table 1. MobileNetV2 network structure.

Input	Network	Expansion Multiple of Input Channel	Number of Output Channels	Module Repetitions	Step
$256 \times 256 \times 3$	Conv2d	-	32	1	2
$128 \times 128 \times 32$	Bottleneck	1	16	1	1
$128 \times 128 \times 16$	Bottleneck	6	24	2	2
$64 \times 64 \times 24$	Bottleneck	6	32	3	2
$32 \times 32 \times 32$	Bottleneck	6	64	4	2
$32 \times 32 \times 64$	Bottleneck	6	96	3	1
$16 \times 16 \times 96$	Bottleneck	6	160	3	2
$8 \times 8 \times 160$	Bottleneck	6	320	1	1

3.2. Add CA Module

The attention mechanism in deep learning selects from a large set of information the information that is more critical for the task at hand. The performance of image segmentation tasks can be better improved by combining the attention mechanism with the fast convolution [29]. The common attention mechanisms include channel attention mechanism [30] and spatial attention mechanism [31]. This paper adds a CA module in the MobileNetV2 feature extraction network and after the ASPP module. The attention mechanism added to the encoder is equivalent to the feature extraction process to obtain the input attention representation. Adding an attention mechanism to the decoder input can improve the situation that the encoder output is a single long tensor and cannot store too much information.

The CA module first uses two one-dimensional global pooling operations to aggregate the vertical and horizontal input features into two independent direction-aware feature maps. Then the two feature maps embedded with direction-specific information are encoded into two attention maps, which capture the long-distance dependence of the input feature map in the spatial direction. Therefore, the position information can be saved in the generated attention map. Finally, two attention graphs are applied to the input feature graph by multiplication to emphasize the representation of interest.

Channel attention is globally encoded by global pooling, but it is difficult to retain location information, which is the key to obtain spatial structure in visual tasks. In order to enable the attention module to capture remote interaction in space by using accurate location information, the global pooling is decomposed, and paired one-dimensional feature coding is performed according to Formula (3).

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j) \quad (3)$$

With the given input x , each channel is encoded along the horizontal and vertical coordinates using a pooled core of size $(h, 1)$ or $(1, w)$. Therefore, the output with height h of channel C can be expressed as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(j, w) \quad (4)$$

Similarly, the output with width W of channel C can be expressed as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq i < H} x_c(j, w) \quad (5)$$

Through the above transformation, the features are aggregated along two directions, and two aggregation feature graphs are obtained. This section will stack the two generated

aggregate characteristic graphs, and then use 1×1 convolution transform function f_1 change it:

$$f = \delta \left(F_1 \left[z^h, z^w \right] \right) \quad (6)$$

where $[\cdot, \cdot]$ represents the concatenation operation along the spatial dimension, δ is a nonlinear activation function, f is an intermediate feature map encoding spatial information in the horizontal and vertical directions, γ is the reduction ratio of the control module size. f is decomposed into two independent tensors $f^h \in R^{C/\gamma \times H}$, $f^w \in R^{C/\gamma \times W}$. Using the other two 1×1 convolution transforms F_h and F_w and converts F_h and F_w into tensors with the same number of channels to input X , and obtains:

$$g^h = \sigma \left(F_h \left(f^h \right) \right) \quad (7)$$

$$g^w = \sigma \left(F_w \left(f^w \right) \right) \quad (8)$$

σ is the sigmoid activation function. In order to reduce the computational overhead and complexity of the model, an appropriate reduction ratio is used γ to reduce the number of channels of f . Then, the output g^h and g^w are extended and used as attention weights respectively. Finally, the output Y of the CA module can be expressed as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (9)$$

3.3. Optimizing the Loss Function

For a deep-learning model, the neural network weights in the model are trained through loss back propagation [23]. Because the high-resolution remote-sensing image acquired has a large ground range and the proportion of different types of ground objects is different, the network training will be dominated by larger objects, and the classifier will classify other types of objects as larger objects, which will affect the performance of the segmentation network. Therefore, for the purpose of balancing the loss, this paper uses focal loss to deal with the problem of unbalanced proportion of classified objects. The calculation formula of focal loss is shown in Equation (10).

$$FL = -a_c(1 - p_c)^\gamma \log(p_c) \quad (10)$$

where, through weight a_c to balance the uneven proportion of samples, a_c 's range is $[0, 1]$, γ is a hyper parameter, γ 's range is $[0, 5]$, When the value of γ is 0, focal loss is the traditional cross entropy loss, and p_c is the prediction probability of different categories. The lower the value of p_c the harder it is to classify, the higher the $(1 - p_c)^\gamma$ is, and the smaller the value of $(1 - p_c)^\gamma$ the easy it is to classify. For simple and easy samples, the larger the corresponding prediction probability, the smaller the corresponding weight. The smaller the prediction probability, the greater the weight of the complex samples.

4. Experiment and Result Analysis

4.1. Experimental Data

This paper uses the Wuhan dense annotation dataset (WHDL) and CCF BDCI to test the effectiveness of the improved DeeplabV3+ network. WHDL contains 4940 images taken in Wuhan by gaofen-1 and ziyuan-3 satellite sensors 256×256 pixel RGB image with a spatial resolution of 2 m. The categories of image annotation in the dataset are divided into six categories: bare soil, buildings, sidewalks, water bodies, vegetation and roads. CCF BDCI is the dataset of the 2020 Baidu independent competition entitled "remote-sensing image parcel segmentation", which contains 140,000 pieces of 256×256 pixel RGB image with a spatial resolution of 2 m. The categories of image annotation in the dataset are divided into seven categories: buildings, cultivated land, forest land, water, roads, grasslands and others.

4.2. Experimental Environment and Evaluation Criteria

The CPU used in this experiment is Intel i7-10700, the operating system is windows10, the GPU is NVIDIA Geforce RTX2060S, the video memory is 8 g, and the development environment is pytorch1.9.0 and python3.

In this study, the segmentation performance of high-resolution remote-sensing image datasets is evaluated. The average intersection union ratio (mIoU), average pixel accuracy (mPA) and average recall rate (mRecall) are selected as the indicators to evaluate the network segmentation performance. mIoU formula is shown in Equation (11), mPA formula is shown in Equation (12), and mRecall formula is shown in Equation (13).

$$\text{mIoU} = \frac{1}{n} \sum_{i=0}^n \frac{p_{ii}}{\sum_{j=0}^n p_{ij} + \sum_{j=0}^n p_{ji} - p_{ii}} \quad (11)$$

$$\text{mPA} = \frac{1}{n} \sum_{i=0}^n \frac{p_{ii}}{\sum_{i=0}^n \sum_{j=0}^n p_{ij}} \quad (12)$$

$$\text{mRecall} = \frac{1}{n} \sum_{i=0}^n \frac{p_{ii}}{p_{ii} + \sum_{i=0}^n p_{ji}} \quad (13)$$

where n is the total category, p_{ij} represents the number of pixels that predict class I as class J , p_{ji} represents the number of pixels that predict class J as class I , p_{ii} represents the number of pixels that predict class I as class I .

4.3. Experimental Process

For the dataset WHDLD, training images were selected randomly from 60% of the images, 20% of the images as the verification set, and the remaining 20% of the images as the test set. A total of 14000 images were randomly selected for the dataset CCF BDCI, of which 60% were used as the training set, 20% as the verification set, and the remaining 20% as the test set. All models are implemented with Pytorch. The learning rate is 0.0005 and the training iteration number was 1000 times. In order to improve the accuracy of model segmentation, the MobileNetV2 pre-training weight is loaded before model training, and the focal loss function is used. When the control variables are the same, it is obtained through multiple experiments that the segmentation accuracy of the model is the highest when γ is 2, a_c is 0.25.

4.4. Ablation Experiment

The effectiveness of changing the original network used for feature extraction into MobileNetV2 network, adding a CA module after the feature extraction module and ASPP module, and setting up four different schemes for experiments. The table below shows the experimental results.

Scheme 1 based on the traditional DeeplabV3+ network structure, the feature extraction network is replaced by MobileNetV2 network.

Using scheme 1 as a basis for scheme 2, add CA module to MobileNetV2 network;

Using scheme 1 as a basis for scheme 3, add CA module after ASPP module;

Using scheme 1 as a basis for scheme 4, enter the CA module after the ASPP module in the MobileNetV2 network neutralize the ASPP module and then enter the CA module.

The initial learning rate is 0.0005 and the training iteration is 1000 steps. Table 2 shows the experimental results of different schemes on the WHDLD dataset, and Table 3 shows the experimental results of different schemes on the CCF BDCI dataset. Figures 4 and 5 show the results of the ablation experiment. It can be seen that the network performance of scheme 4 exceeds that of other schemes. It shows that adding CA module in the feature extraction module and after ASPP module can effectively improve the accuracy of model segmentation, but adding CA module in the feature extraction module and after ASPP module can improve the effect better.

Table 2. WHDLD Results of ablation experiment.

Method	mPA (%)	mRecall (%)	mIoU (%)
Traditional			
DeeplabV3+	72.52	71.74	60.52
Scheme 1	74.34	73.78	62.67
Scheme 2	76.14	75.45	64.22
Scheme 3	75.72	75.07	63.84
Scheme 4	76.75	76.12	64.76

Table 3. CCF BDCI Results of ablation experiment.

Method	mPA (%)	mRecall (%)	mIoU (%)
Traditional			
DeeplabV3+	71.24	70.44	59.23
Scheme 1	74.09	73.58	62.47
Scheme 2	75.91	75.31	64.07
Scheme 3	75.46	74.87	63.64
Scheme 4	76.57	75.95	64.58

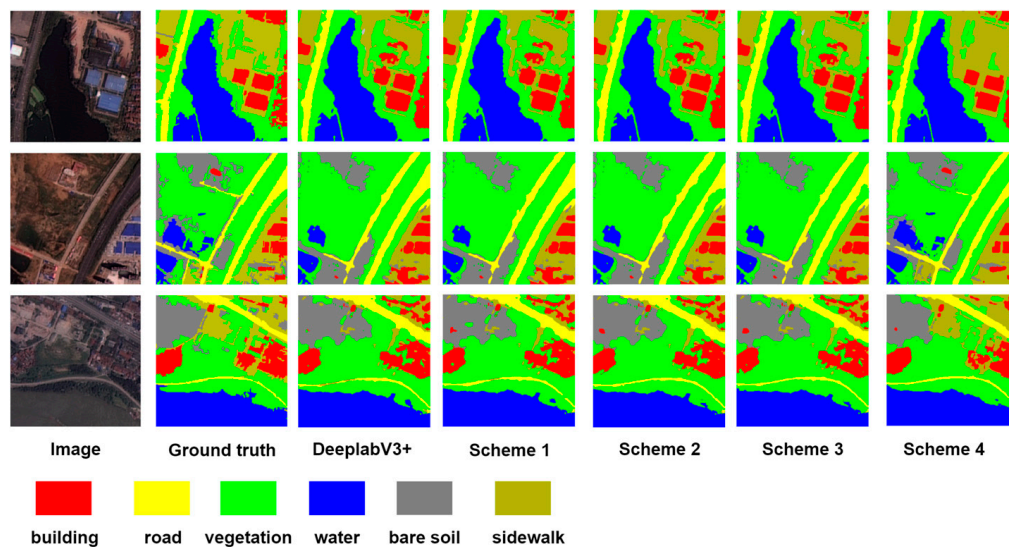


Figure 4. WHDLD Comparison of segmentation effect.

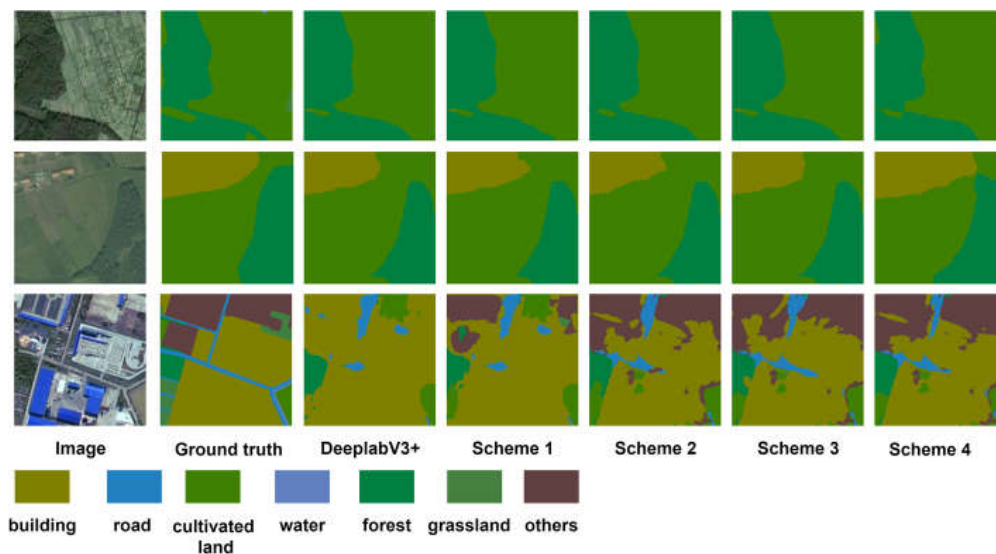


Figure 5. CCF BDCI Comparison of segmentation effect.

4.5. Comparison of Segmentation Performance of Different Methods

In order to further verify the segmentation performance of the improved DeeplabV3+ model proposed in this paper, this method is compared with the traditional WHDL, U-Net, PSP-Net and MACU-Net networks.

Table 4 shows the experimental results of different methods on the WHDL dataset, and Table 5 shows the experimental results of different schemes on the CCF BDCI dataset. It can be seen that the performance of the method proposed in this paper exceeds that of other networks. In the results of WHDL dataset, the mIoU of the proposed method is 6.44% higher than that of u-net and 2.39% higher than that of MACU-Net. In the results of CCF BDCI dataset, the method proposed in this paper is 5.35% higher than the traditional DeeplabV3+ network mIoU. The results show that the traditional DeeplabV3+ network has better segmentation performance than U-Net and PSP-Net in the task of high-resolution remote-sensing image segmentation. Also, it proves that the improved method proposed in this paper improves the segmentation performance of the network in high-resolution remote-sensing image segmentation.

Table 4. WHDL Comparison of segmentation methods.

Method	mPA (%)	mRecall (%)	mIoU (%)
Traditional			
DeeplabV3+	72.52	71.74	60.52
U-Net	70.12	69.62	58.32
PSP-Net	67.23	66.56	55.46
MACU-Net	74.28	73.67	62.37
Paper Method	76.75	76.12	64.76

Table 5. CCF BDCI Comparison of segmentation methods.

Method	mPA (%)	mRecall (%)	mIoU (%)
Traditional			
DeeplabV3+	71.24	70.44	59.23
U-Net	70.01	69.41	58.11
PSP-Net	66.90	66.32	55.23
MACU-Net	73.95	73.36	62.07
Paper Method	76.57	75.95	64.58

Table 6 shows the comparison of the time spent by different methods to train 1000 steps on the WHDL dataset and the number of parameter quantity, from which it can be seen that the training time and parameters of the model proposed in this paper are significantly reduced compared with the traditional DeeplabV3+ network. Compared with MACU-Net, this method has less training time and more parameters. This is because the MACU-Net model has fewer parameters, but its model structure is more complex, which increases the model training time. In general, the MobileNetV2 network is proposed as the feature extraction network, and the CA module is added after the feature extraction module and ASPP module, which greatly reduces the amount of parameters and training time of the model.

Table 6. Comparison of training time and number of parameters in WHDL.

Method	Training Time/Epoch (S)	Parameter Quantity (M)
Traditional DeeplabV3+	331	209.71
U-Net	245	10.86
PSP-Net	211	9.31
MACU-Net	317	5.12
Paper method	265	22.51

Figure 6 shows the prediction results of the proposed method and the traditional DeeplabV3+network. On the left is the result of the WHDLD dataset, and on the right is the result of the CCF BDCI dataset. From this, it can be seen that the method in this paper is better than the traditional DeeplabV3+network segmentation. The main purpose of this method is to improve the segmentation effect of roads and buildings more. Because the categories of roads and buildings account for less in the dataset, the traditional DeeplabV3+network has poor segmentation accuracy in the categories that account for less. The results prove the necessity of adding CA module and introducing focal loss function.

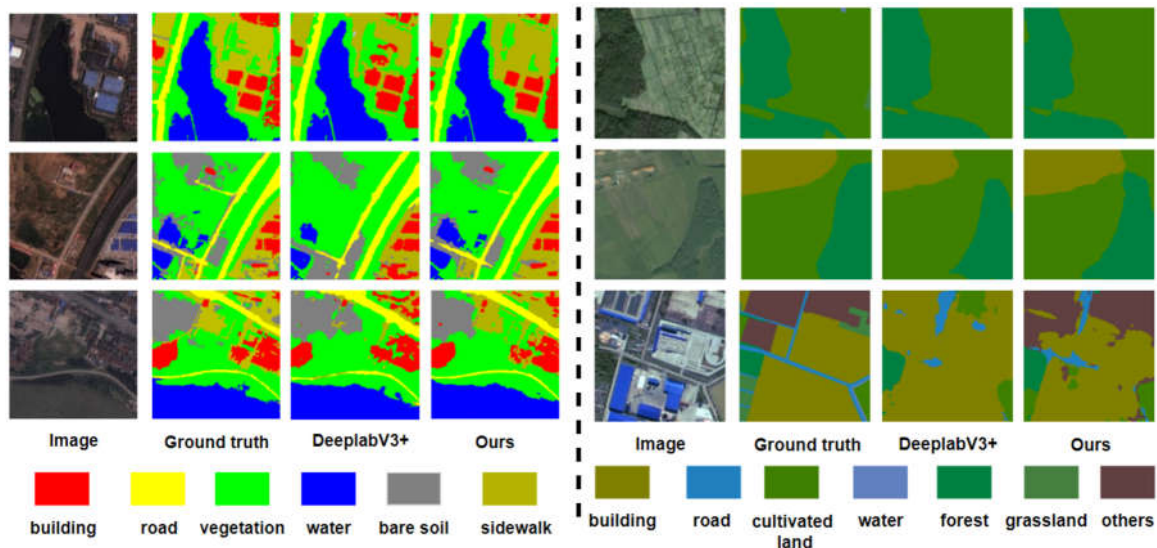


Figure 6. WHDLD and CCF BDCI Comparison of segmentation effect.

4.6. Comparison with Existing Surface Feature Classification Methods of High-Resolution Remote-Sensing Images

At present, there are few research results on remote-sensing image ground feature classification based on depth neural network. The literature [32–35] proposed the research on high-resolution remote-sensing image ground feature classification based on support vector machine, Relief algorithm, genetic algorithm, tabu search algorithm, respectively, and an adaptive neuro-fuzzy inference system. The comparison between the ground feature classification method based on the depth neural network proposed in this paper and the existing ground feature classification methods is shown in Table 7.

Table 7. Comparison between this paper and the existing ground object classification methods.

Literature	Dataset	Feature Extraction	Feature Extraction	Feature Extraction	Feature Extraction
Literature [32]	WorldView2 High resolution satellite remote-sensing Image data	0.5 m and 1.8 m	Paper	Support vector machine	Manual extraction
Literature [33]	Self made dataset	1 m; 0.5 m and 0.2 m	Paper	Relief F Algorithm, genetic algorithm and support vector machine	Manual extraction
Literature [34]	WorldView-2 & QuickBird	0.5 m and 0.6 m	Paper	Tabu search algorithm, genetic algorithm and support vector machine	Manual extraction
Literature [35]	WHDLD	2 m	image	adaptive neuro-fuzzy inference system	Automatic extraction
Paper	WHDLD and CCF BDCI	2 m	image	Deep neural network	Automatic extraction

Based on Table 7, the following improvement is observed compared with the feature classification method in reference [32–35]:

- (1) The ground feature classification methods adopted in references [32–34] need to manually select the spectrum, texture, geometry, shadow, background and geoscience auxiliary features of the ground feature categories of remote-sensing images, and the segmentation results depend on the advantages and disadvantages of the image feature selection; In this paper, the use of depth neural networks can automatically extract the features of images, which can give full play to the advantages of depth neural networks in image segmentation, and provide a new idea for the study of ground object classification of remote-sensing images.
- (2) The advantage of the method adopted in literature [35] is that only a small number of images are needed as the model training set, and the RGB value of each pixel in the image and an additional feature related to the local entropy information are extracted. The final classification accuracy is 0.06% different from that of the deep learning model (FCN) that requires a lot of image training. The classification rules of the method in literature [35] are explained in the paper. When the RGB value of a pixel value in the image of the WHDL D dataset has a high value, it is classified as a building class. The classification results also classify roads and sidewalks as building classes. The method proposed in literature [35] takes a long time to train on a small number of image datasets, and the classification rules of the model limit its inability to accurately classify labels with similar RGB values. In contrast, although the deep-learning method requires a large number of images as the training set, it can be more effectively segment the image dataset with more labels.
- (3) The attention mechanism is applied to the remote-sensing image ground object classification model based on the depth neural network, and the lightweight MobilenetV2 network is used as the backbone feature extraction network, which improves the ground-object classification accuracy of the model, results in a reduction of model parameters and a lower training cost.

5. Conclusions

Based on the traditional DeeplabV3+ network, this paper uses the MobileNetV2 network as the backbone feature extraction network, and adds a CA module after the feature extraction module and the ASPP module. In addition, in view of the different proportion of different types of ground objects in the high-resolution remote-sensing image dataset, the focal loss equalization loss is introduced for high-resolution remote-sensing semantic segmentation. Experiments on WHDL D datasets and CCF BDCI datasets have fully confirmed that the proposed method outperforms the benchmark method.

The later work of this paper mainly includes four aspects: one is to improve the segmentation accuracy of the network, mainly aiming at the problem that the segmentation accuracy of the network is not high in road and building categories. Only by improving the categories with low segmentation accuracy can we effectively improve the segmentation accuracy of the network for high-resolution remote-sensing images. The second is to select more high-resolution remote-sensing images and more types of ground objects for further research and experiments. In addition to the network itself, there is also interference in the image that affects the network segmentation accuracy. So the third is to study how to reduce the interference of noise, shadow and other uncertain factors in the image to the network segmentation accuracy. The fourth is to continue to improve the network model and improve the segmentation accuracy of the network.

Author Contributions: Conceptualization, J.F. and X.Y.; Data curation, J.F.; Formal analysis, J.F.; Investigation, J.F. and G.W.; Methodology, J.F.; Project administration, X.Y., G.W., L.M. and P.W.; Resources, J.F. and K.E.K.; Software, J.F.; Supervision, X.Y.; Validation, J.F.; Visualization, J.F.; Writing—original draft, J.F.; Writing—review & editing, J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: I choose to exclude this statement.

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

References

1. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [[CrossRef](#)]
2. Huang, X.; Wang, Y. Investigating the effects of 3D urban morphology on the surface urban heat island effect in urban functional zones by using high-resolution remote sensing data: A case study of Wuhan, Central China. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 119–131. [[CrossRef](#)]
3. Chen, Y.; Fan, R.; Yang, X.; Wang, J.; Latif, A. Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning. *Water* **2018**, *10*, 585. [[CrossRef](#)]
4. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 173–190.
5. Cao, X.; Li, T.; Li, H.; Xia, S.; Ren, F.; Sun, Y.; Xu, X. A Robust Parameter-Free Thresholding Method for Image Segmentation. *IEEE Access* **2019**, *7*, 3448–3458. [[CrossRef](#)] [[PubMed](#)]
6. Li, J.; Cheng, X.; Wu, Z.; Guo, W. An Over-Segmentation-Based Uphill Clustering Method for Individual Trees Extraction in Urban Street Areas From MLS Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2206–2221. [[CrossRef](#)]
7. Pan, S.; Tao, Y.; Nie, C.; Chong, Y. PEGNet: Progressive Edge Guidance Network for Semantic Segmentation of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 637–641. [[CrossRef](#)]
8. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. University of California, Los Angeles, Los Angeles, CA, USA et al. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542.
9. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 9 November 2017; pp. 6230–6239.
10. Evan, S.; Jonathan, L.; Trevor, D. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
11. Yang, F.; Li, X.; Shen, J. MSB-FCN: Multi-Scale Bidirectional FCN for Object Skeleton Extraction. *IEEE Trans. Image Process.* **2021**, *30*, 2301–2312. [[CrossRef](#)] [[PubMed](#)]
12. Yu, Y.; Rashidi, M.; Samali, B.; Mohammadi, M.; Nguyen, T.N.; Zhou, X. Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm. *Struct. Health Monit.* **2022**, *21*, 2244–2263. [[CrossRef](#)]
13. Fan, T.; Sun, T.; Xie, X.; Liu, H.; Na, Z. Automatic Micro-Crack Detection of Polycrystalline Solar Cells in Industrial Scene. *IEEE Access* **2022**, *10*, 16269–16282. [[CrossRef](#)]
14. Li, R.; Duan, C.; Zheng, S.; Zhang, C.; Atkinson, P.M. MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
15. Olaf, R.; Philipp, F.; Thomas, B. 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.
16. Li, C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2015**, arXiv:1412.7062.
17. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
18. Li, C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *Computer Vision and Pattern Recognition. arXiv* **2017**, arXiv:1706.05587.
19. Li, C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 15th European Conference, Proceedings, Part VI, Munich, Germany, 8–14 September 2018; pp. 801–818.
20. Wang, C.; Du, P.; Wu, H.; Li, J.; Zhao, C.; Zhu, H. A cucumber leaf disease severity classification method based on the fusion of DeepLabV3+ and U-Net. *Comput. Electron. Agric.* **2021**, *189*, 106373. [[CrossRef](#)]
21. Azad, R.; Asadi-Aghbolaghi, M.; Fathy, M.; Escalera, S. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 251–266.
22. Li, Z.; Wang, R.; Zhang, W.; Hu, F.; Meng, L. Multiscale Features Supported DeepLabV3+ Optimization Scheme for Accurate Water Semantic Segmentation. *IEEE Access* **2019**, *7*, 155787–155804. [[CrossRef](#)]

23. da Cruz, L.B.; Júnior, D.A.D.; Diniz, J.O.B.; Silva, A.C.; de Almeida, J.D.S.; de Paiva, A.C.; Gattass, M. Kidney tumor segmentation from computed tomography images using DeepLabv3+ 2.5 D model. *Expert Syst. Appl.* **2022**, *192*, 116270. [[CrossRef](#)]
24. Du, S.; Du, S.; Liu, B.; Zhang, X. Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Int. J. Digit. Earth* **2021**, *14*, 357–378. [[CrossRef](#)]
25. Koh, J.C.O.; Spangenberg, G.; Kant, S. Automated machine learning for high-throughput image-based plant phenotyping. *Remote Sens.* **2021**, *13*, 858. [[CrossRef](#)]
26. 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.
27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
28. Zhou, D.; Hou, Q.; Chen, Y.; Feng, J.; Yan, S. Rethinking bottleneck structure for efficient mobile network design. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 680–697.
29. Luo, H.; Chen, C.; Fang, L.; Zhu, X.; Lu, L. High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3492–3507. [[CrossRef](#)]
30. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the 15th European Conference, Proceedings, Part VI, Munich, Germany, 8–14 September 2018; pp. 286–301.
31. Chu, Q.; Ouyang, W.; Li, H.; Wang, X.; Liu, B.; Yu, N. Online Multi-object Tracking Using CNN-Based Single Object Tracker with Spatial-Temporal Attention Mechanism. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 25 December 2017; pp. 4846–4855.
32. Deng, Z.; Li, D.; Ke, Y.; Wu, Y.; Li, X.; Gong, H. An improved SVM algorithm for high spatial resolution remote sensing image classification. *Remote Sens. Land Resour.* **2016**, *3*, 12–18.
33. Zhou, Y.; Zhang, R.; Wang, S.; Wang, F. Feature Selection Method Based on High-Resolution Remote Sensing Images and the Effect of Sensitive Features on Classification Accuracy. *Sensors* **2018**, *18*, 2013. [[CrossRef](#)] [[PubMed](#)]
34. Shi, L.; Wan, Y.C.; Gao, X.J.; Wang, M. Feature Selection for Object-Based Classification of High-Resolution Remote Sensing Images Based on the Combination of a Genetic Algorithm and Tabu Search. *Comput. Intell. Neurosci.* **2018**, *2018*, 6595792. [[CrossRef](#)] [[PubMed](#)]
35. Castellano, G.; Castiello, C.; Montemurro, A.; Vessio, G.; Zaza, G. Segmentation of remotely sensed images with a neuro-fuzzy inference system. In Proceedings of the WILF 2021: 13th International Workshop on Fuzzy Logic and Applications, Vietri sul Mare, Italy, 20–22 December 2021. Available online: ceur-ws.org/Vol-3074/paper15.pdf (accessed on 1 January 2020).