applied sciences

MDPI

Article

Fully Convolutional Neural Network with Augmented Atrous Spatial Pyramid Pool and Fully Connected Fusion Path for High Resolution Remote Sensing Image Segmentation

Guangsheng Chen [1], Chao Li [1], Wei Wei [2], Weipeng Jing [1], Marcin Woźniak [3], Tomas Blažauskas [4] and Robertas Damaševičius [4,*]

[1] College of Information Science and Technology, North East Forest University, Harbin 150040, China; kjc_chen@163.com (G.C.), lichaoluck@outlook.com (C.L.), weipeng.jing@outlook.com (W.J.)
[2] College of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; weiwei@xaut.edu.cn
[3] Institute of Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland; marcin.wozniak@polsl.pl
[4] Department of Software Engineering, Kaunas University of Technology, 51386 Kaunas, Lithuania; tomas.blazauskas@ktu.lt
* Correspondence: robertas.damasevicius@ktu.lt

Received: 26 March 2019; Accepted: 29 April 2019; Published: 1 May 2019

Abstract: Recent developments in Convolutional Neural Networks (CNNs) have allowed for the achievement of solid advances in semantic segmentation of high-resolution remote sensing (HRRS) images. Nevertheless, the problems of poor classification of small objects and unclear boundaries caused by the characteristics of the HRRS image data have not been fully considered by previous works. To tackle these challenging problems, we propose an improved semantic segmentation neural network, which adopts dilated convolution, a fully connected (FC) fusion path and pre-trained encoder for the semantic segmentation task of HRRS imagery. The network is built with the computationally-efficient DeepLabv3 architecture, with added Augmented Atrous Spatial Pyramid Pool and FC Fusion Path layers. Dilated convolution enlarges the receptive field of feature points without decreasing the feature map resolution. The improved neural network architecture enhances HRRS image segmentation, reaching the classification accuracy of 91%, and the precision of recognition of small objects is improved. The applicability of the improved model to the remote sensing image segmentation task is verified.

Keywords: semantic segmentation; remote sensing; dilated convolution; fully convolutional neural network; deep learning

1. Introduction

Remote sensing (RS) image segmentation technology plays a key role in the fields of urban planning [1], RS mapping [2,3], precision agriculture [4,5], landscape classification [6,7], traffic monitoring [8], environmental protection [9], climate change [10] and forest vegetation [11], and therefore provides important decision support for human work and life.

Traditional image segmentation methods are mainly based on spectral statistical features, such as minimum distance, maximum likelihood, and K-means clustering [12,13]. Although these methods achieved good results, with the improvement of RS image resolution, segmentation and recognition accuracy can no longer meet the requirements [14]. Motivated by the ongoing success of deep learning methods in computer vision, the RS image segmentation tasks have been tackled by Convolutional

Neural Networks (CNNs) [15], which greatly improved the precision. In the classical CNN model, the image blocks composed of a pixel point and its adjacent pixels were input into the network to extract the feature, which are used for the classification of each pixel [16]. This approach introduces much redundant computation in the batch operation, and leads to large memory consumption and low partition efficiency. Shelhamer et al. [17] proposed Fully CNN (FCN), which can accept image of any size as input, extract the features by the convolutional layer, followed by deconvolution upsampling, and output a segmentation image with the same size, with accurate target object edges and assigned label. At present, the FCN model has been widely used in image segmentation [18–23]. In addition, since the image resolution is decreasing in the convolution and pooling operations of CNN, the segmentation result generated by the last layer is often low in resolution. Many subsequent models for image segmentation further extend the idea of FCN. The representative models include SegNet [24], U-Net [25], DeepUNet [26], Y-Net [27] and DeepLab [28,29]. Among them, DeepLabv3 network is one of the most excellent methods for image segmentation. It acquires a larger receptive field by migrating learning to initialize the encoder ResNet [12] in ImageNet [30–33] pre-trained weights and frame capture of the Atrous Spatial Pyramid Pooling (ASPP) composed of parallel convolution with different expansion rates. Similarly, in [34], the pyramid pooling module is used to extract feature maps at multiple scales.

Other recent state-of-the art ideas include attention structure to deal with different scale information and select features, akin to learning in the discriminative feature network [35]; the refinement residual block [36], which can aggregate the information across different channels and refine the feature map to improve the recognition ability of each stage; and maximum fusion strategy to combine information from deep and shallow layers to avert the loss of detailed information because of downsampling in FCN [37]; and controlling network training strategies using multi-threading [38]. Gates (or entropy maps), introduced in Gated CNN (GCNN) [39], allow to ascribe adaptive weights to feature maps depending on their importance. As a result, the gates can train the network to target pixels with high uncertainty to enhance the separability of these pixels. The use of pre-trained semantic segmentation network layers can improve the representability of the low-level features, while allowing the effective use of CNNs for smaller datasets [40]. In Intersection of Union (IoU)-Adaptive deformable region-based CNN (IAD R-CNN) [41], the number of dilated convolutions and the IoU threshold of the detectors for training is defined by the IoU value corresponding to a small object, while cascade R-CNN architecture is used to achieve a better overall detection performance. Merging Fast R-CNN, which use a multi-task loss in a single network training stage, with Region Proposal Network (RPN) allowing for the sharing of their convolution features, thus achieving much faster object detection [8].

To decrease the complexity of the network, a number of techniques could be employed, such as to use energy-driven sampling to segment the image into homogeneous superpixels thereby decreasing the number of processing units [42,43]. Kussul et al. [4] use self-organizing Kohonen maps (SOMs) as data pre-processing step for image segmentation and restoration of missing data, which are further processed with 1-D CNN with spectral domain convolutions. Ji et al. [44] propose a scale-robust FCN (SR-FCN). The architecture concatenates the multi-layer features extracted in VGG-16 network to the similar scale features in decoding, in a different way from FCN and DeepLab, where features in encoding are not fully assimilated into features in decoding. Cheng et al. [7] use DeconvNet with a novel local smooth regularization that makes segmentation spatially consistent. An integration of edge network with DeconvNet allows for achieving more accurate edge results as compared with common methods. Panboonyuen et al. [45] used a global convolutional network (GCN) with the modification of backbone architecture with more layers for higher resolution RS images, additional channel attention block to select the most discriminative filters (features), and domain-specific transfer learning to deal with the scarcity problem by employing other RS datasets with varying resolutions as pre-trained data. Shrestha and Vanneschi [46] proposed an enhanced fully CNN (EFCN) that uses the exponential linear unit (ELU), to boost the network performance for more accurate prediction, while Conditional Random Fields (CRFs) are used as a last stage to reduce noise and sharpen the edges of objects. Xu et al. [47]

proposed replacing a convolution layer with a deformable convolution layer in order to obtain the Deformable ConvNet network that is able to recognize the RS objects with a complex shape and visual appearance. Zhao et al. [48] integrated FCN with simple linear iterative clustering (SLIC) in order to utilize the superpixel information for accurate identification of semantic information and precise positioning of small edges.

Because of the improvement in resolution of the RS images, the high resolution RS (HRRS) image contains a large amount of information, which expands the application scope of RS image, and the size of recognized objects of interest is relatively smaller. The existing CNN is directly applied for RS image segmentation, which has some problems, such as the poor segmentation effect of small objects and fuzzy boundary. Based on the DeepLabv3 network [28] and the properties of RS image data, we suggest a new method for HRRS image segmentation. The main contributions of this paper are: (1) that the parallel dilated convolution in the ASPP structure uses varying dilation factors, obtains denser sampling, gathers local information at higher level, and improves the segmentation performance for small objects; (2) fully connected (FC) fusion path is added to ensure the information propagation ability of the model, and the information complementary to the full convolution path is employed to capture diversity of information as well as to further improve segmentation accuracy.

The remainder of this manuscript is organized as follows. In Section 2, we introduced atrous convolution and DeepLab v3 network [28] structure, pointing out the problem of existing ASPP structure applied to small object segmentation. Then we proposed improved structure A-ASPP and a fully connected path fusion method, which is proposed to improve the segmentation effect of remote sensing images. Section 3 is the experimental part, wherein we detail the improvement of some experimental processes, and share the experience of training the proposed model, and compare it with the existing model. Finally, in Section 4 we summarize the full text.

2. Materials and Methods

2.1. Dilated Convolution

In the classical CNN, the convolution kernel can obtain a large receptive field by pooling operation. The size of input and output of the RS image segmentation is the same. Therefore, the image with smaller size after pooling needs to be expanded back to the original size by the deconvolution operation, but the loss of image information in the deconvolution process will be too great if the downsampling by the pool is too large. The dilated convolution can obtain different size receptive fields by controlling the expansion rate. Assuming that in two-dimensional cases, for each position i, the corresponding output is y and the weight of the feature is w, the convolution of the input feature layer x is calculated as:

$$y_i = \sum_k x_{[i+r\times k]} \times w_k \tag{1}$$

where k is the size of the convolution kernel, and r is the expansion rate.

In the dilated convolution, the convolution kernel is expanded by the dilation factors, and $r-1$ zeros are placed along the space dimension between the adjacent weights to create a sparse filter. The extended convolution is checked to input feature x for conventional convolution. The convolution of different expansion rates is shown in Figure 1.

Figure 1a shows a standard 3 × 3 convolution, a special form of dilated convolution rate = 1, covering a 3 × 3 size field of view each time; Figure 1b shows a 3 × 3 dilated convolution with rate = 2. The convolution kernel size is still 3 × 3, but the computational field of view of the convolution kernel is increased to 7 × 7, while the actual parameter is still 3 × 3. The size of the receptive field can be expressed as:

$$v = \Big((k+1)\times(r-1)+k\Big)^2 \tag{2}$$

Therefore, by adjusting the expansion rate in diluted convolution, the receptive field can be expanded without adding additional parameters.

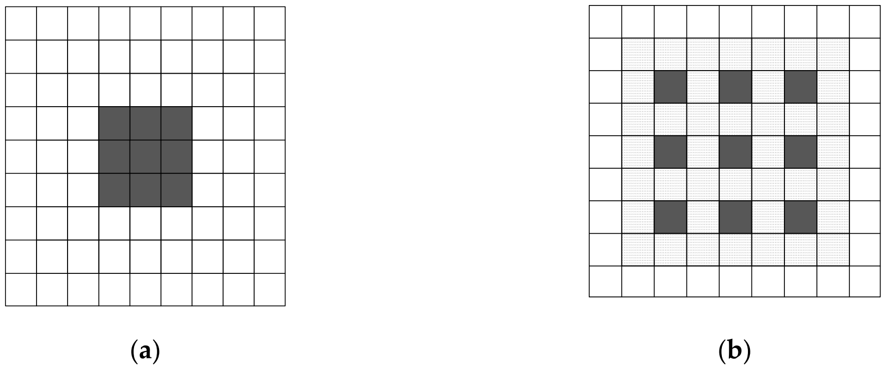


Figure 1. Convolution with kernel size 3 × 3 and different rates. (**a**) Standard convolution corresponds to atrous convolution with rate = 1. (**b**) Employing large value of atrous rate enlarges the model's field-of-view, rate = 2.

2.2. Architecture of DeeplLab v3

The main structure of DeepLabv3 network consists of three parts, as shown in Figure 2. The first part is the basic network, which uses the ResNet architecture, and is trained on the ImageNet as the main feature extraction network, which performs the learning of multi-scale features and the last block in the original ResNet contains atrous convolution with the rate = 2, respectively.

The second part is the Atrous Spatial Pyramid Pool (ASPP) structure, which uses four kinds of dilated convolution with different expansion rate to perform the convolution operation on the output result of the previous layer, in order to obtain the multi-scale information and perform the upsampling to restore the correct dimension size. In addition, global average pooling adds more global context information.

In the last part, the features of each branch are merged into a single feature map by a concentration operation. Then we use a 1 × 1 convolution filter to get the result of fine tuning and get the final segmentation logic of 5 channels (the ground truth objects are divided into 5 categories).

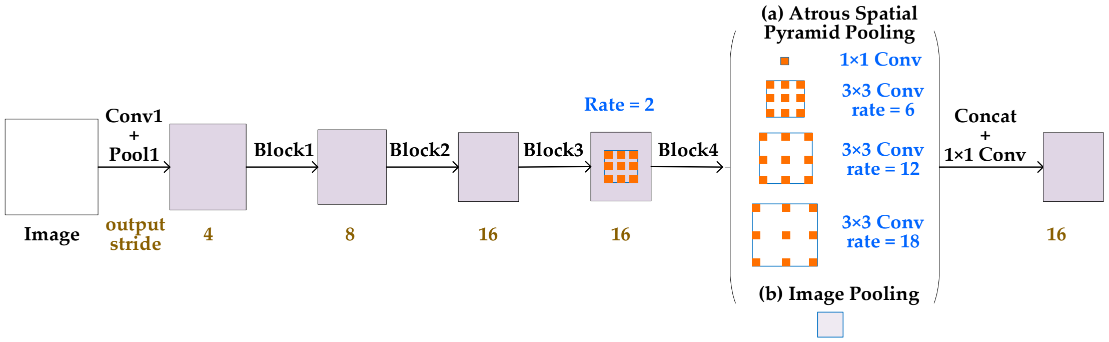


Figure 2. Overview of DeepLab v3 architecture consisting of three stages: basic network, Atrous Spatial Pyramid Pooling (ASPP) module and post-processing stage [28].

2.3. Proposed Model

As pointed out in [31], context information is important for detecting minute objects. Context information, such as roads, cars or other buildings, helps to recognize objects. A higher resolution is also important. In low resolution images, fine details can be over-segmented into a single mask, or missed altogether. Figure 3 presents an architecture of the proposed segmentation network: basic network model, Augmented ASPP (A-ASPP) module, fully convoluted (FC) fusion path and post processing. Each module has a different role. The structure of the basic network and the post-processing module are the same as in the DeepLab v3 architecture. The basic network module model is designed to extract basic features, while the post-processing model is designed to combine the characteristics of each branch into a single feature map through a concentration operation. The A-ASPP layer is designed to make calculations more intensive and enhance the learning of small object features, and thus, firstly to cover large context, calculate more intensive features, gradually increase the dilation factors; then decrease the dilation factors to aggregating local features scattered by the increased dilation factors. The fully connected (FC) fusion path is designed to capture different views for FCN path to further improve the result.

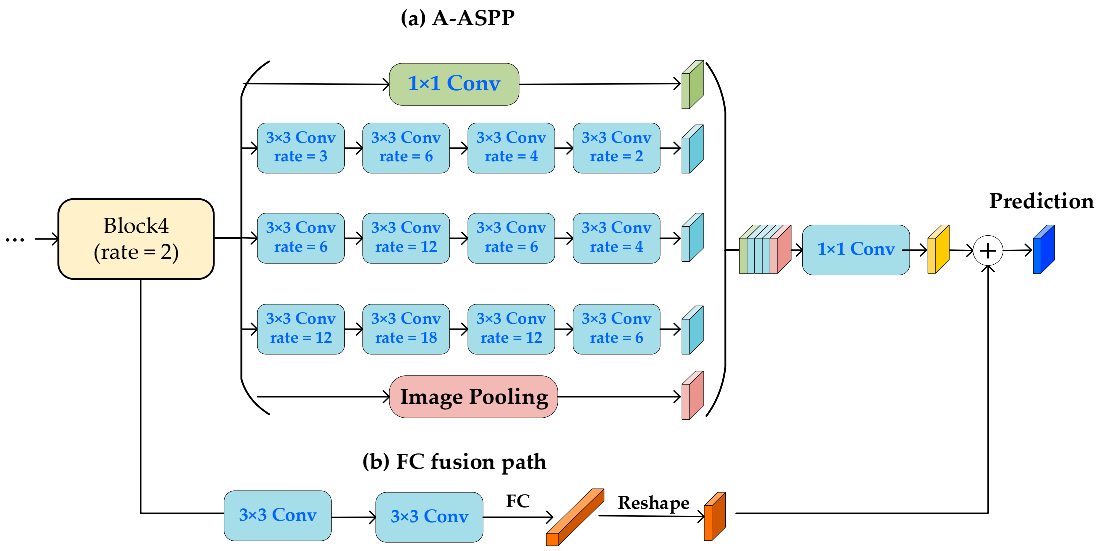


Figure 3. Schematic representation of the proposed architecture: Augmented ASPP network with Fully Convoluted fusion path, and post-processing stage. (**a**) Augmented Atrous Spatial Pyramid Pooling (A-ASPP) module with three parallel dilated convolution branches, each branch consists of four different expansion rate dilated convolution layers. (**b**) Fully connected (FC) fusion branch module consists of two convolution layers and one FC layer, our method fuse predictions from A-ASPP and FC outputs for better prediction.

2.4. A-ASPP Module

As the main structure of ASPP, dilated convolution is important for the segmentation task. Although it is useful with regards to resolution and background context, it is not friendly to small object segmentation in HRRS images. The role of the A-ASPP module is to solve problems of the ASPP module. Specifically, aggressive application of dilated convolution in ASPP causes two problems, (1) dilation factors that are too big lead to a sparse convolution kernel, and a large amount of computational information is lost. (2) The consistency of adjacent spaces becomes weak, and local information is lost at the upsampling layer.

Firstly, in order to solve the sparsity problem caused by diluted convolution, more intensive computation is needed, and the dilation factors are increased. For example, in the one-dimensional volume set with convolution kernel of 3, the combination of expansion rate of 3 and expansion rate of 6, compared with only expansion rate of 6, there are seven eigenvalues in the convolution of combinatorial expansion rate. Only three eigenvalues are involved in the computation of convolution with expansion rate of 6.

Therefore, the large expansion can be calculated from the small expansion, which makes the computation and sampling denser, thus allowing for obtaining more detailed context information. Therefore, the A-ASPP structure first uses a gradual expansion rate.

To handle the second problem, we propose decreasing the dilation factor. The idea is that the main cause of the problems is an increasing dilation factor. If we attach structure with decreasing dilation factor after increasing one, information pyramids of neighboring units can be connected again. Thus, decreasing the structure gradually recovers consistency between neighboring units and extracts local structure in higher layer.

The structure of A-ASPP is shown in the (a) section of Figure 3. The A-ASPP structure uses the dilation factors that expands first and then decreases to maintain the advantage of multi-scale information acquisition and enhance the learning ability. First, the dilution factors are gradually expanded to make the receptive field larger and denser, and more detailed contextual information is obtained, and then the feature extraction of small objects is enhanced through the aggregation of local information by decreasing the dilation factors.

2.5. FC Fusion Path

Each node of the FC layer is connected to all the nodes in the previous layer, which synthesizes the previously extracted features. The output of the FC layer can be obtained by the weighted sum of the output of the previous layer and the response of the activation function:

$$u_l = w^l x^{l-1} + b^l \tag{3}$$

where u_l is the FC layer, which is weighted and biased by the output x^{l-1} of the previous layer, and w^l, b^l is the weight coefficient and the bias coefficient, respectively.

When compared with FCN, the FC layer has different characteristics. The FCN predicts each pixel based on the local receiving domain and shares parameters in different spatial locations. On the contrary, the FC layer is location-sensitive. The prediction of different spatial positions is realized by different sets of parameters. As a result, they can adapt to the different spatial positions. At the same time, the prediction of each spatial position is based on the global information of the whole image, which helps to distinguish different objects.

Based on the different properties of the FC layer and convolution layer, we introduce a FC fusion path with the FC layer to fuse with the FCN path. Due to the difference between the natural image and the target remote sensing image in the ResNet, the addition of the FC layer can achieve better model performance and ensure the transfer of the model representation ability.

As seen in Figure 3, the main path is an FCN structure, which consists of basic network blocks and dilated convolution structure. Each image pixel is independently predicted to decouple the segmentation and classification. By creating a short path from the basic network module to the FC layer through two 3 × 3 convolution layers, the second convolution layer cuts the number of channels by half to reduce computation costs.

Since the size of RS image is larger and easy to calculate, the RS image is cut into 512 × 512 sub-images and the output size is 32 × 32 after four residuals with output step of 16, so the FC layer produces a vector of 5120 × 1 × 1. The vector is reshaped to output the same size as that of the FCN structure. Only one layer of FC path is used to avoid collapsing hidden feature maps into a short feature vector to lose spatial information.

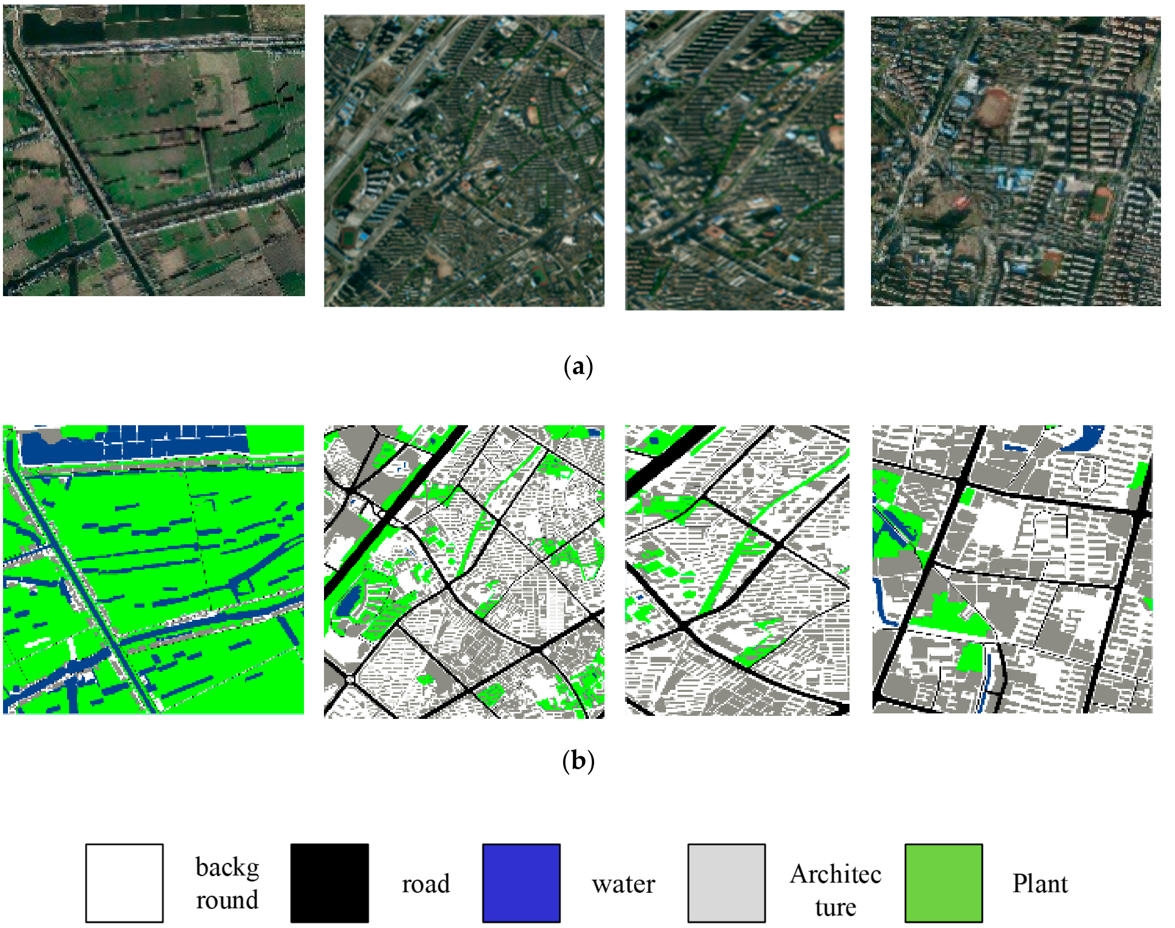
3. Experimental Results, Analysis and Discussion

3.1. Hardware and Software

Our experiments were based on the open source deep learning framework Tensorflow. The GPU (Graphics Processing Unit) is Tesla K20m, the main CPU frequency is 2.10 GHz, and the memory is 128 GB. In the experiment, the training set is randomly sampled with a sub-image of 512 × 512, the batch size is 10, and the initial learning rate is set to 0.0001.

3.2. Experiment Dataset

The data set used in the experiment was derived from the "CCF Satellite Image AI Classification and Recognition Competition", which is a high-resolution remote sensing image of a region in southern China in 2015, and the data type is aerial imagery. The resolution of the image is sub-meter, which contains three bands of RGB, and the types of features are divided into five categories: background, vegetation, road, building and water body. An example of images in the dataset is shown in Figure 4a, the corresponding category diagram is shown in Figure 4b, and the annotation is given in Figure 4c. The size of the pictures from left to right is 5664 × 5142, 7969 × 7939 and 3357 × 6116, 4096 × 4096, respectively. The first three groups of the HRRS images from left to right were selected as the training set and the fourth group was selected as the test set.



(**a**)

(**b**)

(**c**)

Figure 4. Sample images from dataset: (**a**) example of training data; (**b**) ground truth; (**c**) image annotation of feature category.

3.3. Evaluation Index

We evaluate the segmentation result using visual inspection and the quantitative index. From the visual inspection, good segmentation requires maintaining edge characteristics and having good regional consistency. As quantitative indicators we use the classification accuracy and mean intersection over union (mIOU) as an evaluation index.

The higher the accuracy is, the training model has a stronger ability to extract and segment the image features. The accuracy is defined as follows:

$$Acc = \frac{1}{m}\sum_{i=1}^{m}[f_i == y_i] \tag{4}$$

where m is the number of samples, f_i, y_i are the true and predicted pixel label values, and $[\cdot]$ is the Iverson bracket operator, which evaluates to 1, when the labels match, and to 0, when labels are inconsistent.

The Mean Intersection-Over-Union (mIoU)) is a common evaluation metric for semantic image segmentation, which first computes the IoU for each semantic class and then computes the average over classes. IoU is defined as follows:

$$IoU = \frac{\text{true positives}}{\text{true positives} + \text{false positives} + \text{false negatives}} \tag{5}$$

where the values are the number of predictions accumulated in a confusion matrix.

3.4. Experiment and Analysis

Our experiments included three axes: the choice of the number of changes in the rate of expansion, the effectiveness of varying the rate of expansion, and the use of the FC path fusion module (with or without). First, in terms of the number of expansion rate changes, a different type of change based on the DeepLab model is used. Then, set the verification effect of increasing and decreasing expansion rate. We will use DeepLab v3 [28] as the baseline for our experiments. Finally, we evaluate the effectiveness of the FC path fusion.

3.4.1. Ablation Studies on A-ASPP

We first choose the number of atrous rate changes without changing the number of parameters of the ASPP module, because if there are too many changes, each change rate corresponds to a small number of layers, which is not conducive to extracting more abstract features. As shown in Table 1, the effect of selecting four different expansion rates is better than the basic version. Note that the effect of continuously increasing and decreasing the expansion rate is significantly better than the basic experiment.

Table 1. Employing different parameter method for A-ASPP with different number of layers at output stride = 16. The best model performance is shown in bold.

Structure	Parallel Conv1	Parallel Conv2	Parallel Conv3	Road	Water	Arch	Plant	Back	mIoU	Acc
ASPP	(6)	(12)	(18)	40.4	47.5	83.6	74.2	75.4	64.2	77.2
Two layers	(3,6)	(6,12)	(12,18)	42.1	48.2	84.7	73.7	73.9	64.5	79.9
Three layers	(2,3,6)	(4,6,12)	(6,12,18)	44.5	50.9	83.9	74.3	75.1	65.7	83.2
Four layers	(2,3,4,6)	(4,6,6,12)	(6,12,12,18)	49.2	53.3	83.6	74.2	76.4	67.3	85.7
Five layers	(2,3,4,5,6)	(4,6,6,6,12)	(6,12,12,12,18)	48.7	51.9	81.5	73.8	76.3	66.4	83.4
Increase	(2,3,4,6)	(4,6,6,12)	(6,12,12,18)	49.2	53.3	83.6	74.2	76.4	67.3	85.7
Decrease	(6,4,3,2)	(12,6,6,4)	(18,12,12,6)	49.1	51.9	**84.9**	74.1	**77.3**	67.5	86.2
Increase-Decrease	**(3,6,4,2)**	**(6,12,6,4)**	**(12,18,12,6)**	**50.7**	**53.7**	84.4	**75.3**	76.1	**68.0**	**88.9**

Finally, the network parameters of A-ASPP structure are designed based on the number of selected changes and the method of increasing and then decreasing the expansion rate. Our best model is the case where (rate1, rate2, rate3, rate4) = ((1), (3,6,4,2), (6,12,6,4), (12,18,12,6)).

3.4.2. Ablation Studies on FC Fusion Path

We investigate the performance with different ways to instantiate the FC fusion path. We consider two aspects, i.e., the ResNet block to start the new branch and the way to fuse predictions from the new branch and DeepLab v3. We experiment with creating new paths from block1, block2, block3 and block4, respectively, "max", "sum" and "product" operations are used for fusion. We take our re-implemented DeepLab v3 network as the baseline. The results are presented in Table 2.

Table 2. Ablation studies on FC fusion path on HRRS data in terms of mean Intersection of Union (IoU) (mIou) and Acc. The best model performance is shown in bold.

FC-Block	Road	Water	Architecture	Plant	Background	mIoU	Acc
baseline	40.4	47.5	83.6	74.2	75.4	64.2	77.2
block1	41.1	46.9	83.0	74.4	76.9	64.5	79.0
block2	43.2	48.4	84.3	74.2	75.4	65.1	83.2
block3	43.7	49.5	83.1	75.3	76.8	65.7	85.6
block4	**44.2**	**49.9**	**83.9**	**75.1**	**77.2**	**66.1**	**87.1**
PROD	43.4	49.1	83.6	74.9	75.8	65.4	85.7
SUM	**44.2**	**49.9**	**83.9**	**75.1**	**77.2**	**66.1**	**87.1**
MAX	42.9	48.8	83.4	74.8	76.9	65.4	84.8

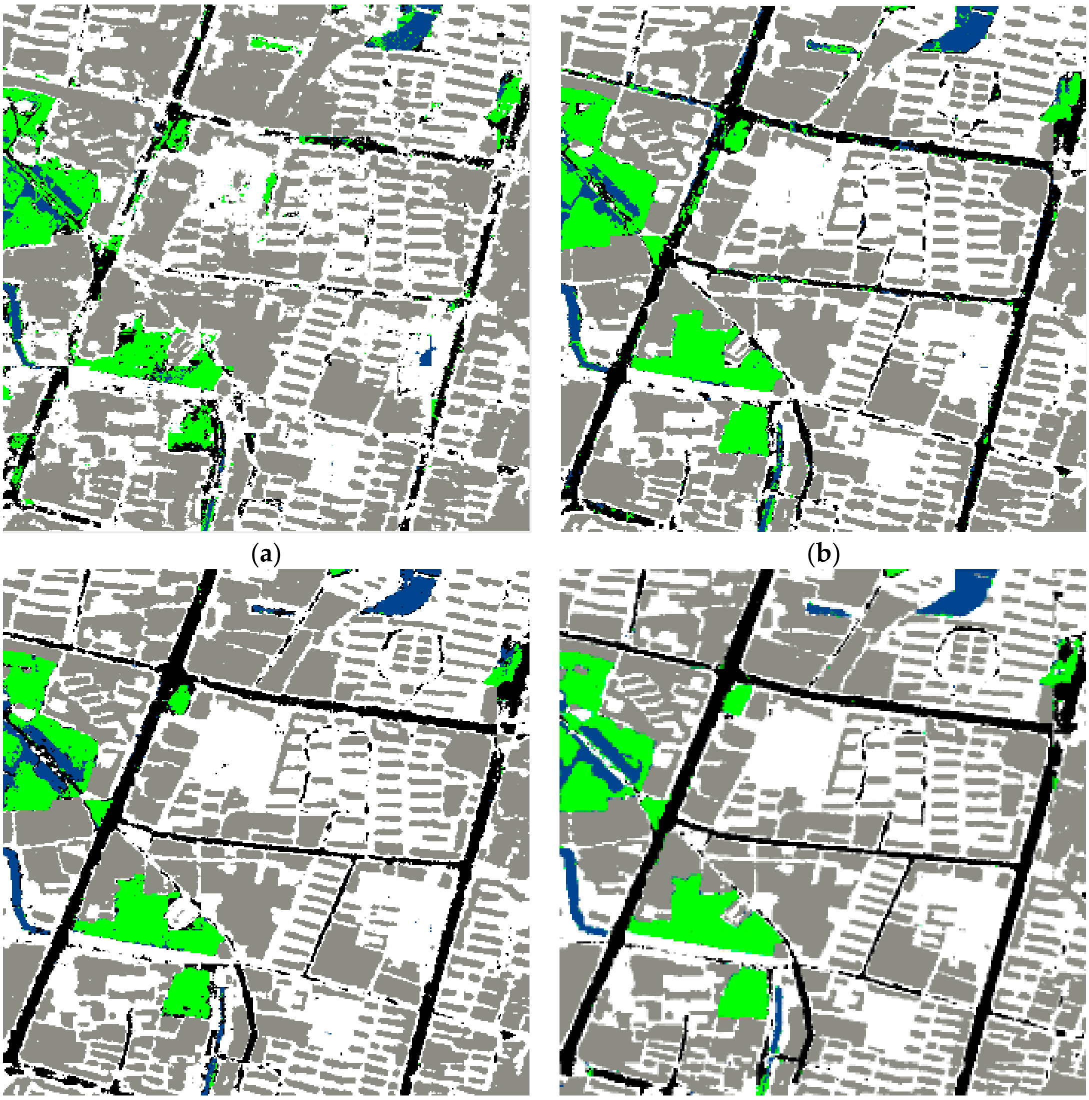
3.4.3. Experiments on HRRS Dataset

We compared the state-of arts in Table 3 from two aspects of accuracy and mean IoU (mIoU). According to the HRRS data training, the accuracy of the A-ASPP is almost the same as that of the method with FC fusion path, and the overall accuracy of the second method is improved by at least 10% when compared with the baseline method.

Table 3. Results on HRRS data in terms of mean IoU (mIou) and Acc. The best model performance is shown in bold.

Method	Road	Water	Architecture	Plant	Background	mIoU	Acc
FCN-8s	23.4	37.5	53.2	52.2	55.1	44.3	61.4
Unet	36.1	41.9	66.1	62.1	57.2	52.7	67.5
SegNet	39.2	47.8	70.1	64.4	65.3	57.4	71.7
PSPNet	42.4	50.1	72.6	73.2	72.8	62.2	74.5
RefineNet	41.3	49.7	76.7	72.4	73.5	62.7	76.2
DeepLabv3	40.4	47.5	83.6	74.2	75.4	64.2	77.2
A-ASPP	50.7	53.7	84.4	75.3	76.1	68.0	88.9
ASPP + FC	44.2	49.9	83.9	75.1	77.2	66.1	87.1
A-ASPP + FC	**52.5**	**54.2**	**84.9**	**76.1**	**77.8**	**69.1**	**91.4**

From the IoU indicator, the segmentation effect improved by FC and A-ASPP is different. The effect of FC is relatively average in each category. The A-ASPP is more effective in improving the size of small-sized objects, such as road water. Because FC provides more underlying information flow through hopping connections, the A-ASPP structure uses deeper and more detailed deep information to enhance the network's expressive ability. The two methods complement each other, so the effect is improved when they work together. The results suggest that the improved model can obviously enhance the feature extraction ability of the model for RS images and thus improve the segmentation accuracy. The method with both A-ASPP and FC fusion path achieves the highest accuracy when compared with the A-ASPP and FC fusion methods. Our method is further improved, as the small object segmentation effect is obviously improved, the road IoU is 14% higher than achieved using DeepLab v3, the water IoU is increased by 7%, and the overall effect is the best, the precision reaches 91.4%, mIoU reached 69.1%. By pre-training on ImageNet, we scored significantly better than DeepLab v3 with the same settings. Therefore, both improvements complement each other and enhance the prediction result. The visual results are shown in Figure 5.



(**a**) (**b**)

(**c**) (**d**)

Figure 5. Classification results by different models on testing image data: (**a**) DeepLabv3; (**b**) FC fusion path; (**c**) A-ASPP; (**d**) our network.

From Figure 5, we can see that the results of the baseline DeepLab v3 network model [28] are very rough, the edges are fuzzy, the labeling of roads and water is obviously wrong, and the visual effect is the worst. With the addition of the FC fusion path and the improvement of the ASPP structure, the edges of objects became clearer, although there are still some miss-classifications. When two improved models (A-ASPP and FC fusion path) are added at the same time, the visual effect is the best.

4. Conclusions

In this paper, we have presented a novel segment module based on dilated convolution to precisely segment small objects in remote sensing imagery, and designed a simple and yet effective component to enhance information propagation and provide additional contextual information. In particular, we did many experiments to verify the effectiveness of this module. Our method shows good effectiveness for segmentation of small objects. Finally, the proposed network architecture can be applied beyond the remote sensing imagery tasks and is also expected to be effective to applications in image processing where small objects are of importance, such as the segmentation of cells in the biomedical domain.

Author Contributions: Data curation, C.L. and W.J.; Formal analysis, M.W.; Funding acquisition, T.B.; Investigation, G.C.; Methodology, C.L. and W.W.; Resources, G.C. and C.L. Software, G.C. and C.L.; Supervision, W.W.; Validation, W.W. and W.J.; Visualization, C.L. and W.J.; Writing—original draft, G.C. and W.W.; Writing—review and editing, T.B. and R.D.

Funding: This job is supported by the National key R&D Program of China under Grant No. 2018YFB0203900 and National Natural Science Foundation of China (31770768), the Natural Science Foundation of Heilongjiang Province of China (F2017001), Heilongjiang Province Applied Technology Research and Development Program Major Project (GA18B301) and China State Forestry Administration Forestry Industry Public Welfare Project (201504307).

Acknowledgments: The authors gratefully acknowledge the financial supports by the National key R&D Program of China under Grant No. 2018YFB0203900 and National Natural Science Foundation of China (31770768), the Natural Science Foundation of Heilongjiang Province of China (F2017001), Heilongjiang Province Applied Technology Research and Development Program Major Project (GA18B301) and China State Forestry Administration Forestry Industry Public Welfare Project (201504307).

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

References

1. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [CrossRef]
2. Henry, C.; Azimi, S.M.; Merkle, N. Road segmentation in SAR satellite images with deep fully convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
3. Li, Y.; Guo, L.; Rao, J.; Xu, L.; Jin, S. Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 613–617. [CrossRef]
4. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
5. Zhang, C.; Gao, S.; Yang, X.; Li, F.; Yue, M.; Han, Y.; Fan, K. Convolutional neural network-based remote sensing images segmentation method for extracting winter wheat spatial distribution. *Appl. Sci.* **2018**, *8*, 1981. [CrossRef]
6. Buscombe, D.; Ritchie, A.C. Landscape classification with deep neural networks. *Geosciences* **2018**, *8*, 244. [CrossRef]
7. Cheng, D.; Meng, G.; Cheng, G.; Pan, C. SeNet: Structured edge network for sea-land segmentation. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 247–251. [CrossRef]
8. Azimi, S.M.; Fischer, P.; Korner, M.; Reinartz, P. Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, 1–19. [CrossRef]
9. Song, N.; Wang, N.; Lin, W.; Wu, N. Using satellite remote sensing and numerical modelling for the monitoring of suspended particulate matter concentration during reclamation construction at Dalian offshore airport in China. *Eur. J. Remote Sens.* **2018**, *51*, 878–888. [CrossRef]
10. Salzano, R.; Salvatori, R.; Valt, M.; Giuliani, G.; Chatenoux, B.; Ioppi, L. Automated Classification of Terrestrial Images: The Contribution to the Remote Sensing of Snow Cover. *Geosciences* **2019**, *9*, 97. [CrossRef]
11. Wei, W.; Polap, D.; Li, X.; Woźniak, M.; Liu, J. Study on remote sensing image vegetation classification method based on decision tree classifier. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 2292–2297. [CrossRef]

12. 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, 26 June–1 July 2016; pp. 770–778.
13. Luo, B.; Zhang, L. Robust Autodual Morphological Profiles for the Classification of High-Resolution Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 1451–1462. [CrossRef]
14. Zheng, P.; Qi, Y.; Zhou, Y.; Chen, P.; Zhan, J.; Lyu, M.R.-T. An Automatic Framework for Detecting and Characterizing the Performance Degradation of Software Systems. *IEEE Trans. Reliab.* **2014**, *63*, 927–943. [CrossRef]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012; pp. 1097–1105.
16. Guo, Z.; Shao, X.; Xu, Y.; Miyazaki, H.; Ohira, W.; Shibasaki, R. Identification of Village Building via Google Earth Images and Supervised Machine Learning Methods. *Remote Sens.* **2016**, *8*, 271. [CrossRef]
17. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
18. Audebert, N.; Saux, B.L.; Lefèvre, S. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. In *Computer Vision—ACCV 2016*; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer: Cham, Switzerland, 2016; Volume 10111, pp. 180–196.
19. Yuan, J. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. *arXiv* **2016**, arXiv:1602.06564.
20. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
21. Mboga, N.; Georganos, S.; Grippa, T.; Lennert, M.; Vanhuysse, S.; Wolff, E. Fully Convolutional Networks and Geographic Object-Based Image Analysis for the Classification of VHR Imagery. *Remote Sens.* **2019**, *11*, 597. [CrossRef]
22. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]
23. Kulikajevas, A.; Maskeliūnas, R.; Damaševičius, R.; Misra, S. Reconstruction of 3D Object Shape Using Hybrid Modular Neural Network Architecture Trained on 3D Models from ShapeNetCore Dataset. *Sensors* **2019**, *19*, 1553. [CrossRef]
24. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
25. 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 (MICCAI), Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
26. Li, R.; Liu, W.; Yang, L.; Sun, S.; Hu, W.; Zhang, F.; Li, W. DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 3954–3962. [CrossRef]
27. Li, Y.; Xu, L.; Rao, J.; Guo, L.; Yan, Z.; Jin, S. A Y-net deep learning method for road segmentation using high-resolution visible remote sensing images. *Remote Sens. Lett.* **2019**, *10*, 381–390. [CrossRef]
28. Chen, L.C.; Papandreou, G.; Schroff, F.; Hartwig, A. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
29. 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.* **2016**, *40*, 834–848. [CrossRef]
30. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
31. Hu, P.; Ramanan, D. Finding Tiny Faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
32. Wei, W.; Yang, X.L.; Zhou, B.; Feng, J.; Shen, P.Y. Combined energy minimization for image reconstruction from few views. *Math. Probl. Eng.* **2012**, *2012*, 154630. [CrossRef]

33. Ke, Q.; Zhang, J.; Song, H.; Wan, Y. Big Data Analytics Enabled by Feature Extraction Based on Partial Independence. *Neurocomputing* **2018**, *288*, 3–10. [CrossRef]
34. Yu, B.; Yang, L.; Chen, F. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 3252–3261. [CrossRef]
35. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a Discriminative Feature Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1804.09337.
36. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters—Improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1743–1751.
37. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
38. Połap, D.; Woźniak, M.; Wei, W.; Damaševičius, R. Multi-threaded learning control mechanism for neural networks. *Future Gener. Comput. Syst.* **2018**, *87*, 16–34. [CrossRef]
39. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated Convolutional Neural Network for Semantic Segmentation in High-Resolution Images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
40. Pan, B.; Shi, Z.; Xu, X.; Shi, T.; Zhang, N.; Zhu, X. CoinNet: Copy initialization network for multispectral imagery semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 816–820. [CrossRef]
41. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
42. Ren, Y.; Zhu, C.; Xiao, S. Deformable Faster R-CNN with Aggregating Multi-Layer Features for Partially Occluded Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2018**, *10*, 1470. [CrossRef]
43. Lv, X.; Ming, D.; Chen, Y.; Wang, M. Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification. *Int. J. Remote Sens.* **2019**, *40*, 506–531. [CrossRef]
44. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2018**, 3308–3322. [CrossRef]
45. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning. *Remote Sens.* **2019**, *11*, 83. [CrossRef]
46. Shrestha, S.; Vanneschi, L. Improved Fully Convolutional Network with Conditional Random Fields for Building Extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
47. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable ConvNet with Aspect Ratio Constrained NMS for Object Detection in Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
48. Zhao, W.; Zhang, H.; Yan, Y.; Fu, Y.; Wang, H. A Semantic Segmentation Algorithm Using FCN with Combination of BSLIC. *Appl. Sci.* **2018**, *8*, 500. [CrossRef]