forests

MDPI

Article

A Novel Multi-Scale Attention PFE-UNet for Forest Image Segmentation

Boyang Zhang, Hongbo Mu, Mingyu Gao, Haiming Ni, Jianfeng Chen, Hong Yang and Dawei Qi *

College of Science, Northeast Forestry University, Harbin 150040, China; zby0624@nefu.edu.cn (B.Z.); mhb-506@163.com (H.M.); gmy2019@nefu.edu.cn (M.G.); nihaiming2013@126.com (H.N.); jianfengchen1212@nefu.edu.cn (J.C.); yh2020@nefu.edu.cn (H.Y.)
* Correspondence: qidw9806@nefu.edu.cn

Abstract: The precise segmentation of forest areas is essential for monitoring tasks related to forest exploration, extraction, and statistics. However, the effective and accurate segmentation of forest images will be affected by factors such as blurring and discontinuity of forest boundaries. Therefore, a Pyramid Feature Extraction-UNet network (PFE-UNet) based on traditional UNet is proposed to be applied to end-to-end forest image segmentation. Among them, the Pyramid Feature Extraction module (PFE) is introduced in the network transition layer, which obtains multi-scale forest image information through different receptive fields. The spatial attention module (SA) and the channel-wise attention module (CA) are applied to low-level feature maps and PFE feature maps, respectively, to highlight specific segmentation task features while fusing context information and suppressing irrelevant regions. The standard convolution block is replaced by a novel depthwise separable convolutional unit (DSC Unit), which not only reduces the computational cost but also prevents overfitting. This paper presents an extensive evaluation with the DeepGlobe dataset and a comparative analysis with several state-of-the-art networks. The experimental results show that the PFE-UNet network obtains an accuracy of 94.23% in handling the real-time forest image segmentation, which is significantly higher than other advanced networks. This means that the proposed PFE-UNet also provides a valuable reference for the precise segmentation of forest images.

Keywords: forest image segmentation; PFE-UNet; PFE; spatial attention; channel-wise attention; DSC unit

Citation: Zhang, B.; Mu, H.; Gao, M.; Ni, H.; Chen, J.; Yang, H.; Qi, D. A Novel Multi-Scale Attention PFE-UNet for Forest Image Segmentation. *Forests* **2021**, *12*, 937. https://doi.org/10.3390/f12070937

Academic Editor: William W. Hargrove

Received: 29 May 2021
Accepted: 13 July 2021
Published: 16 July 2021

1. Introduction

Forest resources are highly relevant to national economic development [1], and play an essential role in many aspects, such as ecotourism, landscape construction, and maintenance of ecological security of the country [2–5]. In recent years, the impact of overexploitation and other reasons has led to a serious shortage of forest resources in China, and the ecological environment has become extremely fragile [6,7]. Thus, the monitoring and management of forest resources is a matter of great value. Forest image segmentation is widely used in forest resource monitoring as a prerequisite for tasks such as forest surveys and timber area statistics. However, traditional image segmentation methods suffer from low accuracy and time consumption, which render it hard to handle segmentation tasks in complex scenes. Therefore, how to handle the segmentation task in complex scenes has been one of the most challenging problems within the field of forest image segmentation.

With the continuous development of deep learning, deep neural networks can be built for end-to-end image semantic segmentation by supervised training at the pixel level. This means that the output of segmentation prediction results of the same size can be achieved by inputting color images of arbitrary size [8,9]. Convolutional neural networks (CNN) are favored by many researchers due to their powerful feature representation in the field of image segmentation [10]. The Fully Convolutional Network (FCN) has attracted much attention for its remarkable pixel-level classification, and this model was first proposed

by Long et al. [11]. The major difference between the FCN and CNN is that the fully connected layer of CNN is replaced by a convolutional layer. Ronneberger et al. [12] proposed the UNet model based on FCN; UNet consists of a typical down-sampling encoder and up-sampling decoder structure and a 'skip connection' between them. Basaeed et al. [13] fused inter-band memory and intra-band information to enhance the spatial spectrum of the image and improve generalization. Kampffmeyer et al. [14] proposed a fully convolutional neural network-based image land-cover-mapping method and achieved precise segmentation of feature targets of different scale sizes on the ISRPS 2D semantic annotation dataset [15]. Additionally, for the ISRPS 2D dataset, Audebert et al. [16] proposed a deep fully convolutional neural network (DFCNN) based on the self-encoder type, the key of which is the introduction of multicore convolutional layers to achieve multi-scale feature extraction. The DeepLab family of networks [17–20] put forward by the Google team has broken through the accuracy in the field of image segmentation time and again. Among them, DeepLabv3+ [20] further used the Xception model and depthwise separable convolution and combined atrous spatial pyramid pooling (ASPP) with a simple decoding module to obtain a larger and stronger encoding–decoding network architecture. Although the segmentation accuracy is enhanced to some extent, the consequent increase in computational effort hinders its widespread use in real-time image segmentation tasks. Therefore, the construction of an efficient and lightweight network is also a critical issue that needs to be addressed.

To solve these problems, researchers are increasingly focusing on the design of efficient network structures, an idea that can reduce computational costs and the number of parameters while maintaining a better segmentation performance. In particular, based on multi-scale feature propagation, the pyramid feature attention network (PFA-Net) [21] maximized the fusion of contextual information, which greatly reduced the number of parameters, by still obtaining sufficient receptive fields and enhancing model learning capability, thus striking a balance between speed and segmentation performance. Furthermore, it is shown that the attention mechanism can effectively capture global contextual information and reduce the computational burden to some extent. Fu et al. [22] proposed a dual attention network (DA-Net) to integrate local features and global dependencies adaptively, using the channel attention module and the position attention module to establish channel dependence and spatial dependence, respectively. In contrast, a deep feature aggregation network (DFA-Net) [23] attached a fully connected attention module to the tail of the backbone network in pursuit of higher precision to retain the maximum receptive field and refine the prediction results. Meanwhile, researchers have focused on the fusion of different depth features in the network and have successively proposed various methods, represented by SegNet [24], ENet [25], ICNet [26], etc.

The above image segmentation methods are highly susceptible to noise and light intensity, are not able to precisely locate forest information and accurately segment forest areas, and still have some shortcomings. Firstly, the feature map after the convolution of the above segmentation method lacks a targeted feature information extraction of the process, for objective factors such as shadows cannot be precisely segmented to confirm whether they belong to the forest area. Secondly, it is easier to ignore the different feature information between high-level and low-level features, leading to problems such as discontinuity and blurred boundaries in the forest area. Finally, with the deepening of the convolutional network, it is prone to overfitting, and there are also challenges such as difficulties in training the network with the increasing number of model parameters.

To deal with the above problems, this paper proposes a novel end-to-end image segmentation method, named the Pyramid Feature Extraction-Unet network (PFE-UNet). Specifically, following the convolution of the encoder–decoder in the traditional UNet, an attention mechanism is introduced, so that it can adaptively derive image features. Meanwhile, irrelevant regions are suppressed and different weight ratios are assigned to different scales, thus ensuring the network could focus on the features related to the specific segmentation task. Then, the Pyramid Feature Extraction module (PFE) is adopted

to obtain high-level features with high receptive fields at multiple scales. The traditional UNet transition layer is replaced with PFE, the channel-wise attention module (CA) is used to focus on the high-level features, and the appropriate scales and receptive fields are selected to generate the prediction result maps. The spatial attention module (SA) is applied to focus on low-level features, and following different attention mechanisms, high-level and low-level features are perceived in a complementary way to generate a feature map. Furthermore, a novel depthwise separable convolutional unit (DSC Unit) is adopted to reduce the model parameters using depthwise separable convolution, which divides the depth convolution into two depth convolutions with cascaded kernel sizes. To integrate contextual information faster, two depth convolutions are inserted separately before each 1×1 convolution to improve the forest image segmentation accuracy by training a deeper network. The proposed PFE-UNet is capable of accurately segmenting forest images, which can provide a beneficial reference value for the research of image segmentation and recognition studies.

2. Related Works

2.1. Attention Mechanism

Nowadays, there are many image segmentation networks using attention mechanisms to improve segmentation accuracy [27–30], which improve the representational power of neural networks by extracting key information relevant to the task [31,32], and have achieved promising breakthroughs. Yu et al. [33] proposed the discriminative feature network (DFN), which added modules such as channel attention and global average pooling to solve the problem of inconsistent features between classes. Jie H et al. [34] proposed an attention module named squeeze-and-excitation block (SE), which automatically obtains the weight of each feature channel. Furthermore, it combined the features of different stages and enhanced the representation of features. Before splicing the image features obtained by convolution of the encoder with the corresponding features in the decoder, Oktay et al. [35] produced an attention gate and readjusted the encoder's output features, which was successfully applied to segmentation tasks.

2.2. Depthwise Separable Convolutions

Many deep learning network architectures [36–38] use depthwise separable convolutions to replace standard convolutional blocks as core units. Compared to previous full rank filters, 3×1 filters and 1×3 filters are used to approach 3×3 filters. This scheme, where many complex filters are divided into small base filters followed by training the network from scratch using weight initialization methods, tends to consume fewer parameters and improve the performance of semantic segmentation compared to conventional convolutional kernels in convolutional operations. At the same time, 3×3 is the smallest size that can capture the pixel-eight neighborhood information by stacking small-size convolutional layers instead of large-size convolutional layers, and the receptive field size remains unchanged. Therefore, we split the depthwise convolution into two depthwise convolutions with cascaded kernel size. To construct an efficient and lightweight network architecture, we factorize the 3×3 depthwise convolutional layer into a 3×1 depthwise convolution and a 1×3 depthwise convolution. Changing the order of the convolutional layers allows the network to capture more contextual information to improve performance.

2.3. DropBlock

Motivated by the successful application of DropBlock in recent computer vision works [39–41], we adopt DropBlock to regularize the network. DropBlock, a structured form of dropout, can effectively prevent over-fitting problems in convolutional networks. It drops the units in adjacent areas of the feature map together, instead of dropping random units. This effectively prevents the overfitting problem in convolutional networks. The DSC Unit is constructed based on the depthwise separable convolution and DropBlock mentioned in Section 2.2. We factorize a 3×3 depthwise convolutional layer to a 3×1

depthwise convolution and a 1 × 3 depthwise convolution; each 3 × 1 layer and the 1 × 3 layers will be followed by a 1 × 1 convolutional layer. Each convolutional layer is followed by a DropBlock, a layer of batch normalization (BN), and a ReLU activation unit, as shown in Figure 1a. The DSC Unit replaces the standard convolution in the UNet network (as shown in Figure 1b) to construct a U-shaped network as the 'backbone' of the overall network architecture, effectively avoiding the overfitting problem and accelerating the convergence of the network.

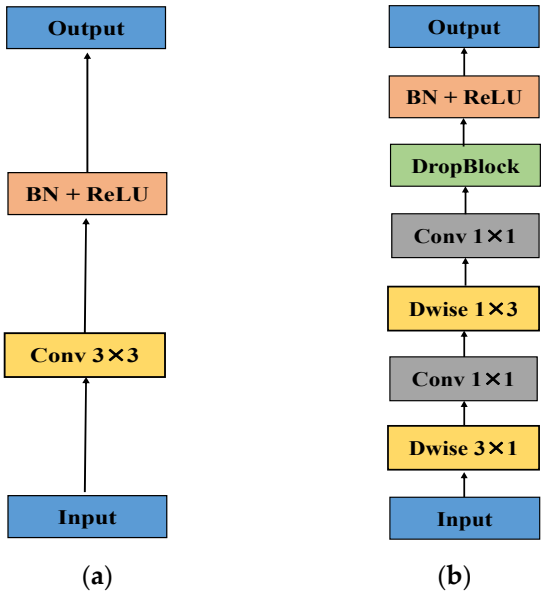


Figure 1. Network units: (**a**) Traditional UNet unit; (**b**) Novel convolutional DSC Unit.

3. Methods

3.1. PFE-UNet Architecture

The PFE-UNet network is composed of two parts, forming a symmetrical structure: the encoder and the decoder part (Figure 2). The encoder part continuously extracts abundant forest features for capturing the complete contextual information in the forest images. However, the decoder is applied to precisely locate the forest area to be segmented in the input image. The whole architecture consists of 8 DSC Units, 4 pooling layers, 4 SA, 4 CA, 5 PFE, and multiple up-sampling layers. We factorize a 3 × 3 depthwise convolutional layer to a 3 × 1 depthwise convolution and a 1 × 3 depthwise convolution; the pooling layer size is 2 × 2, and the input image is 512 × 512. The input image goes through a series of convolution, feature extraction, and feature merging operations to obtain an output image with the same resolution as the input image. The standard convolutional block in UNet is replaced by a DSC Unit, and the convolutional layers are reordered to maintain a better segmentation performance while effectively reducing the computational cost. Meanwhile, the introduced DropBlock trains deeper networks without degradation, while preventing overfitting problems in convolutional networks.

In addition, to be able to derive forest information from the feature mapping convolved by the encoder, we use a spatially oriented attention module after each DSC Unit to adaptively extract image features. The Pyramid Feature Extraction module is adopted in the transition layer of the network to sense contextual information for capturing high-level features with multi-scale and multi-receptive fields. Before each up-sampling feature fusion operation, a channel-wise attention module is adopted to refine the feature information details. To acquire the output segmentation map, 1 × 1 convolution and sigmoid activation function are obtained at the final layer.

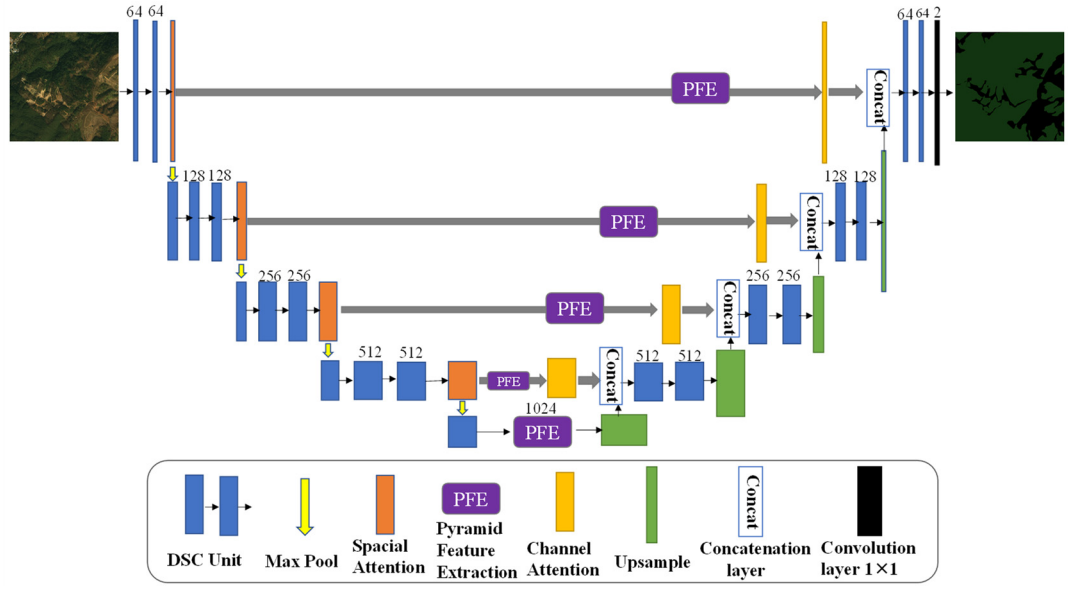


Figure 2. The architecture of the proposed PFE-UNet.

3.2. Pyramid Feature Extraction

Feature extraction is of great importance for image segmentation tasks. Many researchers have proposed convolutional neural network models to perform feature learning in images by simply stacking convolutional and pooling layers, which may not be able to effectively handle these complicated variations. Atrous convolution, in contrast to traditional methods, can be used to obtain features with the same scale but different receptive fields [42]. Specifically, the Pyramid Feature Extraction module is added to the transition layer of the UNet network, as shown in Figure 3. To acquire semantic information for more global and abstract high-level feature maps, atrous convolution with different dilation rates of 3, 5, and 7 is adopted to capture multi-receptive-field contextual information. As the output of the Pyramid Feature Extraction module, the feature mappings from different atrous convolution layers and a 1 × 1 dimension reduction feature are combined by cross-channel concatenation.

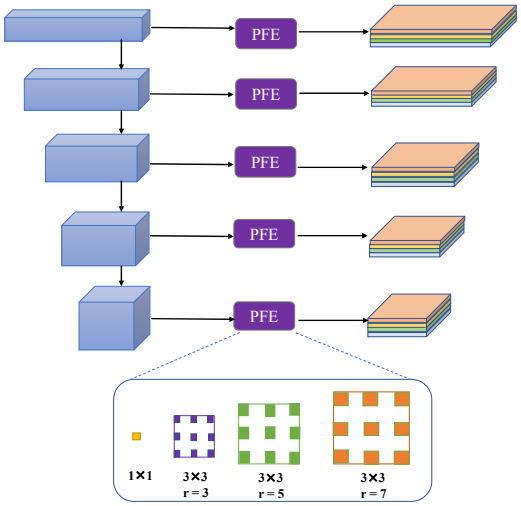


Figure 3. The detailed structure of Pyramid Feature Extraction. A Pyramid Feature Extraction module takes a feature from a side of the SA module as input, and it contains three 3 × 3 convolutional layers with different dilation rates and a 1 × 1 convolution layer.

3.3. Spatial Attention

The spatial attention mechanism, which plays an important role in enhancing the representational power of image segmentation models, is adopted into the model architecture of the network as part of the convolutional attention module that uses the spatial relationships between features to produce spatial attention maps. To calculate the spatial attention, SA is first taken along the channel axis for the average-pooling and max-pooling, respectively, and the feature maps of both are connected to produce efficient feature descriptors, as shown in Figure 4. Input features $F \in R^{H\times W\times C}$ generate features $F^S_{mp} \in R^{H\times W\times 1}$ and $F^S_{ap} \in R^{H\times W\times 1}$ through max-pooling and average-pooling of the channel axes, respectively. Then, the spatial attention map $M^S(F) \in R^{H\times W\times 1}$ is generated by using a convolution layer followed by the sigmoid activation function on the concatenated feature descriptor. In short, the following calculation equation for spatial attention is shown as:

$$\begin{aligned} F^S &= F \cdot M^S(F) \\ &= F \cdot \sigma\left(f^{3\times 3}([MaxPool(F); AvgPool(F)])\right) \\ &= F \cdot \sigma\left(f^{3\times 3}\left(\left[F^S_{mp}; F^S_{ap}\right]\right)\right) \end{aligned} \tag{1}$$

where $f^{3\times 3}(\cdot)$ denotes a convolution operation with a kernel size of 3 and $\sigma(\cdot)$ represents the sigmoid function.

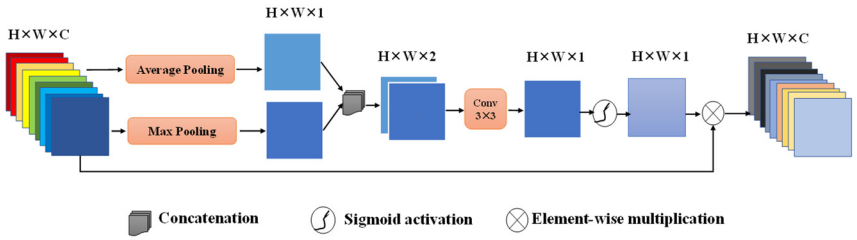


Figure 4. Diagram of the spatial attention module.

3.4. Channel-Wise Attention

The channel attention mechanism in convolutional neural networks can maintain the mapping relationship between different channels. After the Pyramid Feature Extraction, the channel-wise attention module [43] is provided to the transition layer, leading to better learning of the spatial distribution of images to capture multi-scale multi-receptive-field high-level features. We unfold high-level features $f^h \in R^{H\times W\times C}$ as $f^h = \left[f^h_1, f^h_2, \ldots, f^h_C\right]$, where $f^h_i \in R^{H\times W}$ is the i-th slice of f^h and C is the total channel number. As shown in Figure 5, the global average pooling operation is adopted to transform the feature map with the dimensions $H \times W \times C$ into a dimensional shape of $1 \times 1 \times C$ in one step. Connect two fully connected layers to capture channel dependencies. Then, the normalization processing measures are implemented to the encoded channel-wise feature vector mapped to (0,1) using the sigmoid operation:

$$CA = F\left(v^h, W\right) = \sigma_1\left(f_{C_2}\left(\delta\left(f_{C_1}\left(v^h, W_1\right)\right), W_2\right)\right) \tag{2}$$

where W refers to parameters in the channel-wise attention module, σ_1 refers to sigmoid operation, f_C refers to FC layers, and δ refers to the ReLU function.

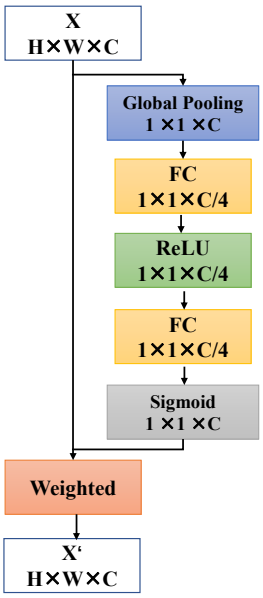


Figure 5. Diagram of the channel-wise attention module, where X and X′ mean weighted feature and weighting feature, respectively.

3.5. Loss Function

Cross entropy is widely used in image segmentation tasks. As shown below:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\hat{P}_i \log P_i + \left(1-\hat{P}_i\right)\log(1-P_i)\right] \tag{3}$$

where P_i refers to the predicted probability of pixel i, while $\hat{P}_i$ represents the truth standard, and N represents the number of samples.

However, in actual forest segmentation, the unbalanced ratio of forested and non-forested areas is likely to cause missed segmentation of detailed areas of the forest. To avoid the above problem, an additional weighting factor w_i^{class}. is introduced to weight the original cross entropy loss function, as shown in the following equation:

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_i^{class}\left[\hat{P}_i \log P_i + \left(1-\hat{P}_i\right)\log(1-P_i)\right] \tag{4}$$

$$w_i^{class} = \frac{N-n_i}{n_i} \tag{5}$$

where n_i presents the number of pixels belonging to class i.

4. Experiment

4.1. Dataset and Implementation

In this experiment, we use the DeepGlobe dataset, a publicly available dataset that provides high-resolution sub-meter satellite imagery, to delineate forest areas. The dataset contains 803 RGB satellite images divided into training/validation/test sets of 803/171/172 images each (corresponding to 70%/15%/15% segmentation). The image space resolution size is 2448 × 2448 and the imagery has 50 cm pixel resolution, collected by Digital Globe's satellite. Each satellite image is paired with a mask image for forest cover annotation. The solution is expected to predict an RGB mask for the input, i.e., a colored image of the same height and width as the input image. The expected result is a forest cover map of the same size in pixels as the input image, where the color of each pixel indicates its class label. Figure 6 shows a few samples of forest images and ground truth. To enhance the training efficiency, the image size is uniformly changed to 512 × 512 as input and the

Labelme image label tool is used for labeling the images. Data augmentation is crucial for the performance evaluation of PFE UNet since there are not enough training samples. The input image and the corresponding labels are simultaneously rotated at an angle or scaled up or down to expand the data set.

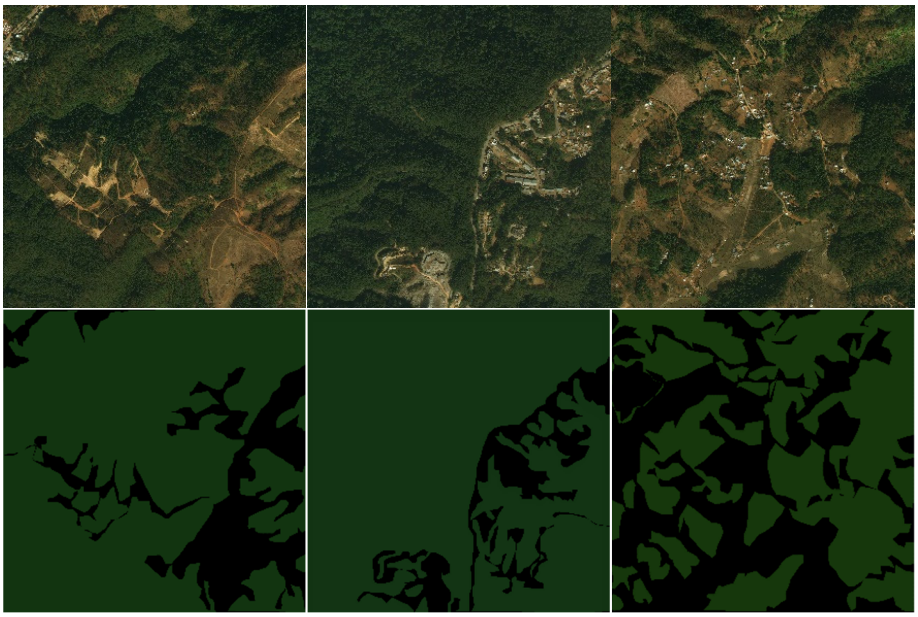


Figure 6. Samples of collected forest images and ground truth.

The PFE UNet network proposed in this paper is trained using a GPU (GTX 960M 2G), and the experimental environment is shown in Table 1. During training, the learning rate of PFE UNet is set to 0.00008, Adam optimization algorithm is selected, batch-size is set to 16, and the size of the discard blocks of DropBlock is set to 7. There are 100 epochs during the training. The specific parameters for training are shown in Table 2.

Table 1. Experimental hardware and software environment settings.

Hardware Environment		Software Environment	
CPU	Intel(R) Xeon(R) Bronze 3204 CPU 1.90 GHz	Operating System	Windows 10
Graphics Card	NVIDIA GeForce RTX 3090	Python	3.6.5
Memory	128 GB	Opencv Framework	3.4.2 Pytorch-gpu-1.8.1

Table 2. Parameter settings and corresponding introduction in network training.

Related Parameter	Value	Meaning
Batch size	16	Number of pictures per training
Learning rate	0.00008	Initial learning rate
Epoch	100	Training iteration times
CUDA	Enable	Computer unified device architecture
CUDNN	Enable	A GPU acceleration library for deep neural networks

4.2. Evaluation Metrics

To evaluate the performance of the proposed model, the segmentation results are compared with the corresponding ground truth and the comparison results for each pixel are grouped into true positive (TP), false positive (FP), false negative (FN), and true negative

(TN). Then, the performance of the model is evaluated with the F_1-score (F_1), precision, recall, and accuracy (ACC). F_1 represents the harmonic mean of precision and recall. They are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

while

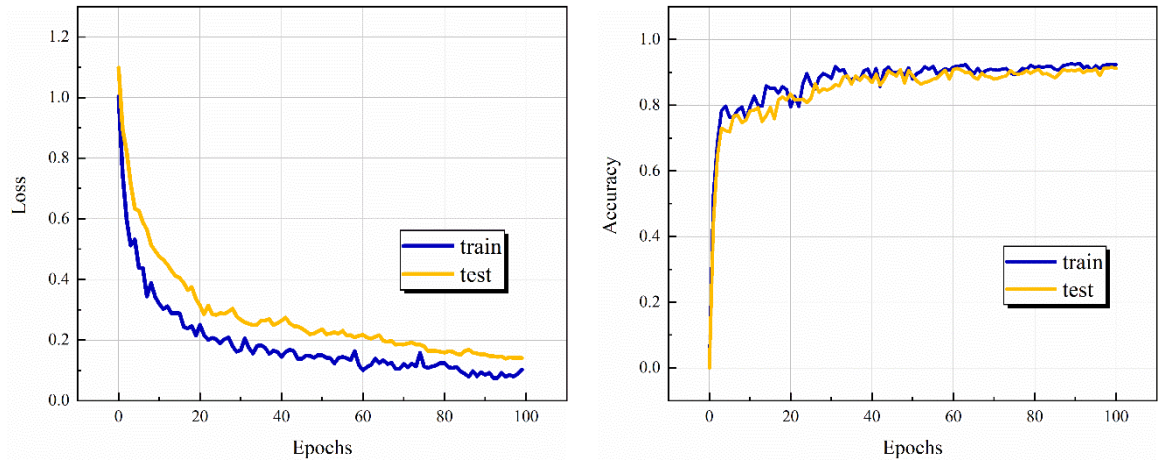
$$precision = \frac{TP}{TP + FP} \tag{8}$$

where TP, TN, FP, and FN denote the amount of true positive, true negative, false positive, and false negative, respectively.

5. Results and Discussion

5.1. Performance and Comparative Analysis

Figure 7a shows the loss of the proposed PFE-UNet on the training and test sets, and it can be observed that the network converges faster, and the trend of both changes is consistent and not much different. It is shown that PFE-UNet has a well-generalized ability to iterate until the loss converges to near 0.1. Figure 7b shows the variation of the accuracy of PFE-UNet on the training and test sets for different iterations. Both values improve at the beginning of network training rapidly, and the improvement slows down with the increase in iterations. After 100 iterations, the model performance remains stable and the accuracy rate reaches 94.23%.



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

Figure 7. The training situation of the dataset in PFE-UNet: (**a**) Loss value; (**b**) Accuracy value.

In the experiments, the segmentation results of UNet [12], DA-Net [22], DFA-Net [23], and our proposed PFE-UNet are quantitatively analyzed, and the segmentation results are shown in Figure 8. It could be seen that the UNet, DA-Net, and DFA-Net segmentation areas occur as over-segmentation or under-segmentation. In contrast, our proposed PFE-UNet model successfully circumvents these errors. Furthermore, the segmentation results can be seriously affected by relatively small buildings, blurred forest boundaries, and large shadow areas. It can be seen that the UNet model performs the worst, not only ignoring the forest area but also leading to the typical under-segmentation errors. DA-Net and DFA-Net, although showing significant advantages in dealing with forest areas, still occur with insufficient segmentation. Regarding the indistinguishable background areas in the forest image, such as intricate shadows, their pixel values as well as shape and

size are very close to those of the forest. The other three models will misclassify some of the shadows as forest, and the segmentation results have a larger coarse error. However, our proposed PFE-UNet model can identify the boundary well and separate the forest from the shadow with great segmentation performance. In addition, the model effectively combines the recognition function of deep convolutional neural networks with the function of channel-wise attention and spatial attention to fuse contextual information, which makes the forest area contour more specific and segments the forest area well. It is experimentally proven that UNet and DA-Net cannot handle segmentation in complex scenes well, and DFA-Net has a better performance compared with UNet and DA-Net. However, it has a few segmentation errors in the border and shaded parts. The PFE-UNet model avoids the above errors to accurately segment the forest area, thanks to the introduction of the PFE module and attention mechanism.

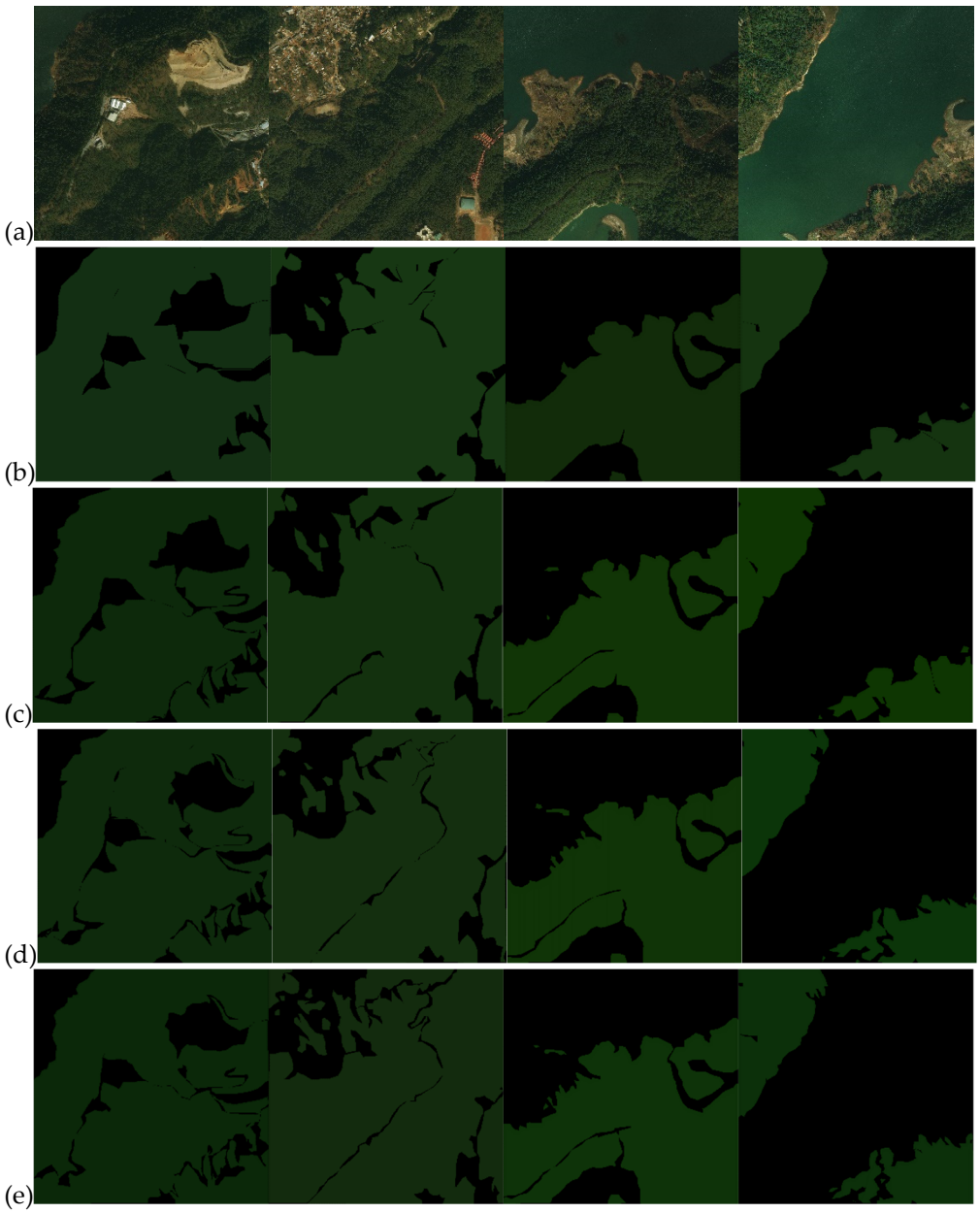


Figure 8. *Cont*.

(f) 

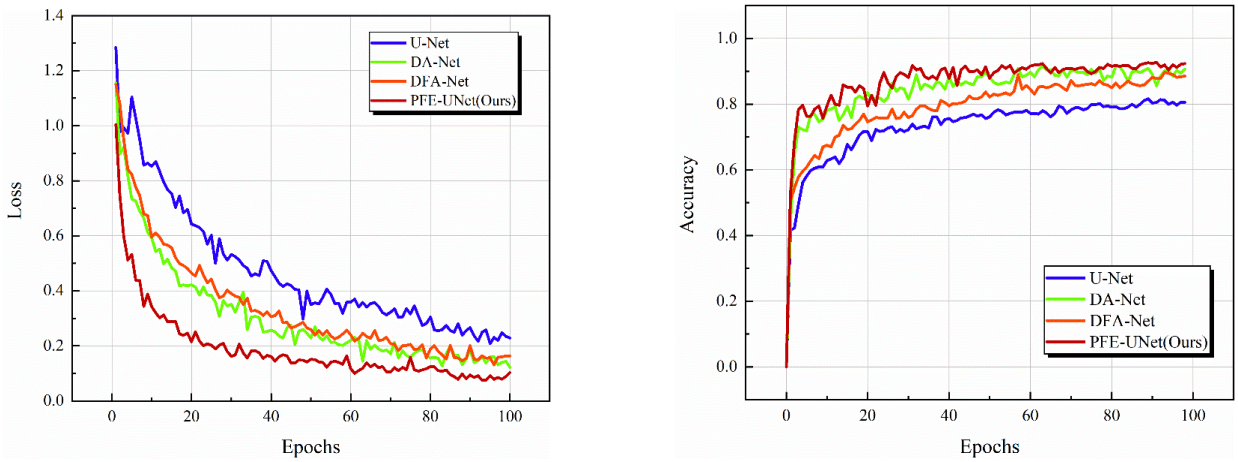
Figure 8. The segmentation result of forest images by different models: (**a**) Input image; (**b**) UNet; (**c**) DA-Net; (**d**) DFA-Net; (**e**) PFE-UNet (ours); (**f**) Ground truth.

According to the above parameter settings, the trained model is used to segment the test. With the purpose of testing the performance of the model, four sets of comparison experiments were performed: the first group experimented with the segmentation performance of UNet on forest images; the second group experimented with the segmentation performance of DA-Net; the third group experimented with the segmentation performance of DFA-Net; and the fourth group experimented with the performance of the proposed PFE-UNet. Table 3 shows the segmentation results for each of these four groups of experiments. Among them, the accuracy of PFE-UNet is improved by 5.48% compared with UNet, and 3.41% and 2.67% compared with DA-Net and DFA-Net, respectively. For other parameters, such as recall and F_1, the recall of PFE-UNet is 93.86%, which is 4.13% higher than DA-Net and 2% higher than DFA-Net. On the one hand, it is due to the introduction of the attention mechanism, focusing on the forest information while suppressing irrelevant forest regions, which enables accurate prediction of the class of each pixel in the original image. On the other hand, the introduction of different dilation rates of convolution in the Pyramid Feature Extraction module can capture multi-receptive-field contextual information, which can restore the information of the original image to the maximum extent, which makes PFE-UNet well able to handle segmentation problems in complex scenes such as small forest areas, blurred forest edges, and shadow-obscured areas. Therefore, the PFE-UNet network can precisely locate the forest area and accurately delineate the forest information.

Table 3. The evaluation results of different deep learning models on the DeepGlobe dataset.

Models	F_1	Precision	Recall	Accuracy
UNet	0.8769	0.8971	0.8823	0.8875
DA-Net	0.9023	0.9035	0.8973	0.9082
DFA-Net	0.9004	0.9103	0.9186	0.9156
PFE-UNet (Ours)	**0.9328**	**0.9418**	**0.9386**	**0.9423**

Figure 9 shows the loss and segmentation accuracy of four deep learning models on the DataGlobe dataset. The loss result in Figure 9a shows that PFE-UNet can rapidly be converged and stabilized around 0.103. The loss trends of UNet, DA-Net, and DFA-Net are consistent and not much different. However, the network converges slowly and does not adapt well to the dataset. Figure 9b presents iterations with a different accuracy of each network. The results demonstrate that the accuracy of the PFE-UNet model is up to 94.23%, while the accuracy of the UNet, DA-Net, and DFA-Net models is 88.75%, 90.82%, and 91.56%, respectively, and the segmentation accuracy does not achieve the expected standard. In contrast, the PFE-UNet model outperforms other deep learning models, showing relatively significant robustness in handling complex segmentation problems such as small forest areas, discontinuous forest areas, and blurred forest boundaries.



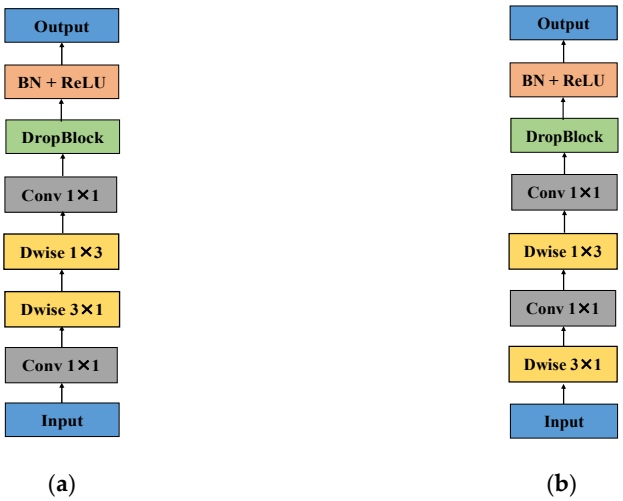
(**a**) (**b**)

Figure 9. The overall loss and accuracy assessment of 4 deep learning models on the DeepGlobe dataset: (**a**) Loss value; (**b**) Accuracy value.

5.2. Ablation Study

5.2.1. Reorder the Convolutional Layers or Not

In the DSC Unit, we factorize a 3 × 3 depthwise convolution layer to a 3 × 1 depth convolution and a 1 × 3 depthwise convolution. Figure 10a shows the DSC Unit before reordering and Figure 10b shows the DSC Unit after reordering. We change the order of the convolutional layers to a 3 × 1 depthwise convolution layer, a 1 × 1 convolution layer, a 1 × 3 depthwise convolution layer, and a 1 × 1 convolution layer. Each convolutional layer is followed by a DropBlock, a layer of batch normalization (BN), and a ReLU activation unit. Adjusting the order of the convolutional layers enhances the flow of information through the convolutional layers, applying a single filter to an input channel, convolving point by point, and finally combining the output results. If no reordering of the DSC Unit is done, the semantic information can only flow from one channel to another between the 1 × 3 depthwise convolutional layer and the 3 × 1 depthwise convolutional layer, destroying the principle of information flow. After reordering the DSC Units, a 1 × 1 convolutional layer is inserted between each 3 × 1 convolution and 1 × 3 convolution to facilitate combining information from all output channels of the depthwise convolutional layers. Reordering of the convolutional layers improves the performance of the network.



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

Figure 10. Before and after reordering in a DSC Unit: (**a**) Before reordering; (**b**) After reordering.

5.2.2. Different Components Combinations

To investigate the effect of different modules on model performance, a comparison of different module combinations was conducted. From Table 4, it can be concluded that the PFE-UNet model (containing Pyramid Feature Extraction (PFE), spatial attention (SA), and channel-wise attention (CA)) achieves the best performance, which demonstrates that all components are necessary for the PFE-UNet to obtain the best segmentation results. The network with a DSC Unit replacing the standard convolutional blocks is adopted as our basic model (backbone) with an ACC of 0.8932. First, ACC is enhanced by adding PFE to the basic model. Furthermore, the added SA yields an increase of 4.43% in the ACC relative to the backbone. On this basis, after adding CA, ACC increased by 5.49% over the basic model, and the best results were obtained, which proves the effectiveness of multi-scale pyramid feature fusion.

Table 4. Ablation study using different component combinations. Backbone means using the DSC Unit to replace the network of standard convolutional blocks, PFE means using Pyramid Feature Extraction after backbone, SA means using spatial attention after PFE, and CA means using channel-wise attention after SA.

Backbone	PFE	SA	CA	ACC
✓				0.8932
✓	✓			0.9156
✓	✓	✓		0.9328
✓	✓	✓	✓	**0.9423**

5.3. Training Time and Prediction Time

As is shown by the operation times of the five deep learning models in Table 5, the UNet model obtained better results. DFA-Net performed in 3756.23 s, which is 271 s faster than DA-Net, while the PFE-UNet model proposed in this paper performed in 3667.15 s. However, due to the attention mechanism, it can be seen that the training times of DA-Net, Attention UNet [35], and PFE-UNet are all longer than that of UNet. Furthermore, the prediction times for single images show no significant differences between each model, whereas the PFE-UNet model has a slightly shorter prediction time than most models, performing for 1.58 s.

Table 5. Training time and prediction time of various segmentation models.

Method	Training Time (s)	Prediction Time (s)
UNet	**3390.82**	**1.49**
Attention UNet	3931.44	1.62
DA-Net	4027.65	2.01
DFA-Net	3756.23	1.76
PFE-UNet (Ours)	3667.15	1.58

6. Conclusions

In this paper, a novel end-to-end image segmentation network—Pyramid Feature Extraction-UNet (PFE-UNet)—is proposed. Pyramid Feature Extraction, spatial attention, channel-wise attention, and depthwise separable convolution are added to the traditional UNet. Among them, the SA module can adaptively extract forest features and simultaneously suppress irrelevant areas, so that the network can aggregate low-level features of the forest regions. The PFE module is considered to be applied in the transition layer of the network, which contains multi-scale atrous convolution. The CA module is used to pay attention to the high-level features of the network so that the higher-level features are perceived complementarily with the lower-level features. To improve the network training speed and reduce the computational cost, the standard convolution block is replaced by a novel DSC Unit, integrating depthwise separable convolution, DropBlock, BN, and

ReLU. Furthermore, the flexibility and robustness of our proposed network are verified on the DeepGlobe dataset. The introduction of the attention mechanism, Pyramid Feature Extraction, and depthwise separable convolution leads to significant advantages of the PFE-UNet model in dealing with small forest areas, discontinuous forest areas, and fuzzy forest boundaries. The experimental results show that the accuracy of the PFE-UNet network is as high as 94.23%, which is higher than the accuracy of other image segmentation networks. Therefore, the PFE-UNet model proposed in this paper can precisely locate forest information and accurately segment the forest area, which has potential application value in forest image segmentation.

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

Funding: This work was supported by the Fundamental Research Funds for the Central Universities (No.2572020BC07) and the National Natural Science Foundation of China (No.31570712).

Acknowledgments: We are highly grateful to the anonymous reviewers and the handling editor for their insightful comments, which greatly improved an earlier version of this manuscript.

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

References

1. Roy, D.P.; Kovalskyy, V.; Zhang, H.K.; Vermote, E.F.; Yan, L.; Kumar, S.S.; Egorov, A. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *J. Remote Sens. Environ.* **2016**, *185*, 57–70. [CrossRef]
2. Gabrielle, F.P.; Marcos, H.C. Deforestation causes different subregional effects on the Amazon bioclimatic equilibrium. *J. Geophys. Res. Lett.* **2013**, *40*, 3618–3623.
3. Boers, N.; Marwan, N.; Barbosa, H.; Kurths, J. A deforestation-induced tipping point for the South American monsoon system. *J. Sci. Rep.* **2017**, *7*, 41489. [CrossRef] [PubMed]
4. Angela, L.; Stefan, E.; Douglas, K.; Paul, M.; Marco, H. Understanding Forest Health with Remote Sensing-Part I—A Review of Spectral Traits, Processes and Remote-Sensing Characteristics. *J. Remote Sens.* **2016**, *8*, 1029.
5. Schulze, K.; Malek, Ž.; Verburg, P. Towards better mapping of forest management patterns: A global allocation approach. *J. For. Ecol. Manag.* **2019**, *432*, 776–785. [CrossRef]
6. Amigo, I. When will the Amazon hit a tipping point. *Nature* **2020**, *578*, 505507. [CrossRef] [PubMed]
7. Curtis, P.G.; Slay, C.M.; Harris, N.L.; Tyukavina, A.; Hansen, M.C. Classifying drivers of global forest loss. *Science* **2018**, *361*, 1108–1111. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
10. Peng, C.; Li, Y.; Jiao, L.; Chen, Y.; Shang, R. Densely Based Multi-Scale and Multi-Modal Fully Convolutional Networks for High-Resolution Remote-Sensing Image Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 2612–2626. [CrossRef]
11. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lect. Notes Comput. Sci.* **2015**, *9353*, 234–241.
13. Basaeed, E.; Bhaskar, H.; Al-Mualla, M. Supervised remote sensing image segmentation using boosted convolutional neural networks. *Knowl. Based Syst.* **2016**, *99*, 19–27. [CrossRef]
14. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016.
15. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D.; Breitkopf, U.; Jung, J. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *13*, 293–298. [CrossRef]
16. Audebert, N.; Saux, B.L.; Lefèvre, S. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. *Lect. Notes Comput. Sci.* **2017**, *10111*, 180–196.
17. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Comput. Sci.* **2014**, *4*, 357–361.

18. 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.* **2018**, *40*, 834–848. [CrossRef]
19. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
20. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Lect. Notes Comput. Sci.* **2018**, *11211*, 833–851.
21. Zhao, T.; Wu, X. Pyramid Feature Attention Network for Saliency detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR, Long Beach, CA, USA, 15–20 June 2019; pp. 3085–3094.
22. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 20 June 2019; pp. 3146–3154.
23. Li, H.; Xiong, P.; Fan, H.; Sun, J. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2020; pp. 9522–9531.
24. 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]
25. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147,.
26. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. *Lect. Notes Comput. Sci.* **2018**, *11207*, 418–434.
27. Xie, D.; Cheng, D.; Hao, W.; Chao, L.; Tao, D. Semantic Adversarial Network with Multi-Scale Pyramid Attention for Video Classification. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 9030–9037. [CrossRef]
28. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A. Context Encoding for Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
29. Li, H.C.; Xiong, P.F.; An, J. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.
30. Roy, A.G.; Nav, A.N.; Wachinger, C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks. *Lect. Notes Comput. Sci.* **2018**, *11070*, 421–429.
31. Zhu, Y.; Zhao, C.; Guo, H. Attention Couplenet: Fully convolutional attention coupling network for object detection. *IEEE Trans. Image Process.* **2018**, *28*, 113–126. [CrossRef]
32. Bo, Z.; Xiao, W.; Feng, J.S. Diversified Visual Attention Networks for Fine-Grained Object Classification. *IEEE Trans. Multimed.* **2017**, *19*, 1245–1256.
33. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a Discriminative Feature Network for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
34. Jie, H.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023.
35. Oktay, O.; Schlemper, J.; Folgoc, L.L. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
36. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1804.03999.
37. Howard, A.; Zhmoginov, A.; Chen, L.C. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *arXiv* **2018**, arXiv:1801.04381.
38. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
39. Guo, C.; Szemenyei, M.; Pei, Y.; Yi, Y.; Zhou, W. SD-Unet: A Structured Dropout U-Net for Retinal Vessel Segmentation. In Proceedings of the IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 28–30 October 2019.
40. Ghiasi, G.; Lin, T.Y.; Le, Q.V. DropBlock: A regularization method for convolutional networks. *arXiv* **2018**, arXiv:1810.12890.
41. Guo, C.L.; Szemenyei, M.; Yi, Y.; Xue, Y.; Zhou, W.; Li, Y. Dense Residual Network for Retinal Vessel Segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
42. Fisher, Y.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the ICLR, San Juan, Puerto Rico, 2–4 May 2016; pp. 1–9.
43. Chen, L.; Zhang, H.W.; Xiao, J. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA 21–26 July 2017; pp. 6298–6306.