Accuracy Improvement of UNet Based on Dilated Convolution

Shengyuan Piao1, * and Jiaming Liu2
1 Beijing University of Posts and Telecommunications, BEIJING, 100876 CHINA
2 School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, BEIJING, 100876 CHINA
*Corresponding author: shengyuan_park@bupt.edu.cn

Abstract. Though previous approach for the satellite image semantic segmentation task has already achieved reliable performance across some different benchmarks, there are still some limitations for the existing methods when faced with the scenarios of the high-resolution images. In this paper, we propose a powerful dilated convolution module and successfully apply it into our improved UNet network structure. The dilated convolution module can directly expand the receptive field of the network without reducing the resolution. Experiments on DeepGlobe Road Extraction Task have shown that the proposed method in this work can achieve the mIoU score of 63.72%, which significantly outperforms the original algorithm.

1. Introduction
Satellite image semantic segmentation task refers to the task of utilizing the image provided by the satellite to perform pixel-level image semantic segmentation. In recent years, the satellite image is widely used in many fields such as road extraction, building inspection, vegetation analysis, soil analysis, agricultural planning, urban planning, and navigation systems, etc. With the widespread use of deep learning technology in the field of computer vision, deep convolutional neural networks reflect its irreplaceability in image semantic segmentation tasks. More specifically, in most image semantic segmentation tasks, the Full Convolutional Network (FCN) has shown its irreplaceable advantages. In addition, the UNet network architecture [2] excels in most medical image segmentation tasks. These two methods have therefore become an important cornerstone of most current semantic segmentation problem solutions.

Comparing with other types of semantic segmentation task, the satellite semantic segmentation task has higher requirements on the fineness of prediction. Satellite imagery has characteristics of ultra-high resolution. Because of the limitations of resolution, major network structures including FCN and Unet in current stage do not achieve a satisfactory level of performance in satellite image segmentation tasks.

In order to meet the needs of higher accuracy in the field of satellite image segmentation, we proposed Dilated Deep Unet, which is a new semantic segmentation network architecture based on encoder-decoder, skip-connection and dilated convolution. The feasibility of the network structure is based on the assumption that the dilated convolution [3] can expand the receptive field of the neural network guaranteeing that the original image pixel-level semantic information is not weaken. As the network deepens, in the semantic segmentation, the spatial information contained in the feature map is reduced, and the semantics is especially critical at this time. The essence of convolution is to extract the edge information of the image, so in general, as the convolution layer deepens, the semantics will be higher. But at this time, the spatial information will be lost, so the purpose of dilated convolution is to preserve the spatial information as much as possible. Based on the results of the experiment, our hypothesis
proven to be correct: the use of dilated convolution instead of pooling can improve the accuracy of the satellite image segmentation prediction to some extent.

We have alleviated the problem of too few data segmentation with correctly labeled satellite imagery by using the method of pre-trained models for transfer learning. In the neural network model involved in this paper, we have adopted the pre-training model of VGG16 [4] to perform weight initialization on the encoder part of the network structure which we designed. In addition, we have also performed data argumentation for the training data to prevent the model proposed from being over-fitting. The data argumentation methods including horizontal flip, vertical flip, color jittering and so on. We used the road extraction data set in the CVPR2018 DeepGlobe Challenge [5] to evaluate the performance and capability for network structure proposed in this work.

The main contributions of our project are listed:
- Analyzed the uniqueness of satellite image semantic segmentation task compared with other types of semantic segmentation tasks and proposed a feasible optimization scheme for the uniqueness of this kind of task.
- Proposed Dilated Deep UNet, a reliable satellite image semantic segmentation neural network architecture, which improves the accuracy of satellite image semantic segmentation tasks to a certain extent.

2. Background

2.1. Related work

2.1.1. Traditional Algorithm. The clustering algorithm is a more general non-deep learning clustering algorithm represented by kmeans [6]. The algorithm has a better classification ability when the similarity within data can be measured by Euclidean distance. However, when the data composition is more complicated, or the Euclidean distance can hardly correctly represent the degree of similarity between the data, the result of the algorithm's division of semantics information is usually unsatisfactory.

Before the emergence of FCN-based methods, edge detection-based algorithms have been widely used in semantic segmentation tasks. The algorithm is based on the assumption that the image boundary has strong correlation with the semantic information. It has a good segmentation effect in the scene with clear boundary. But for the segmentation of small target and weak boundary target, it does not perform well in the term of its running speed as well as its accuracy.

The Conditional Random Field algorithm is a special case of Markov Random Field. It is a conditional probability distribution model that outputs another set of random variables given a set of random variables. This algorithm relies on the assumption that the output random variable constitutes a Markov Random Field. In a semantic segmentation task, a random variable can be either a single pixel or a super pixel. However, this algorithm is limited by its accuracy and speed so that it is difficult to be widely used.

2.1.2. Semantic Segmentation Neural Network. A semantic segmentation algorithm based on patch classification, which cuts a picture into several patches. These patches are then sent to the neural network for training and these image blocks are then be classified. This method makes it possible that the patches sent to the neural network have the same size. However, this method runs slower and causes a waste of computing resources, so this method is not used in this paper.

The full convolutional neural network is the core of the current mainstream semantic segmentation algorithm. By replacing all the fully connected layers with convolutional layers and merging feature information of different scales, this method is light weighted. In addition, the running speed of the full convolutional network is significantly improved compared to the algorithm based on patch classification. The full convolutional network is not perfect. When performing the down-sampling operation in the full convolutional network, although the receptive field of the full convolutional network can be expanded
to better integrate the context information, it also reduces the resolution of the image, resulting in the clipping of space and detail information.

The Encoder-Decoder architecture [2] is a neural network structure based on FCN improvements. The architecture is mainly composed of two parts, in which the encoder captures deep semantic information through several down-sampling processes; the decoder part gradually restores the space and detail information of the input image through several up-sampling operations. This architecture's neural network model performs well in both high-level and low-level computer vision tasks. The current popular U-Net is one of the encoder-decoder architectures.

PSP-Net [7], which has been added to the pyramid pooling, generates a series of feature maps of different sizes in several layers of down sampling process, and connects them with the up sampled feature map generated by transposition convolution to achieve multi-level feature fusion. This feature fusion method effectively improves the ability of the network model to aggregate contextual semantic information.

2.2. Dilated Convolution
Modern image classification network models typically use a series of pooling operations to obtain multi-scale context information. The pooling operation can effectively increase the receptive field of the neural network, which is beneficial for high-level computer vision tasks (such as classification). However, the pooling operation also has the problem of degrading the resolution of the picture: because the pooling operation is irreversible and will result in the loss of spatial information, which may result in the information of the small object cannot be reconstructed. This has a detrimental effect on the semantic segmentation task, because the semantic segmentation task requires both pixel-level precision and multi-scale context inference.

Yu et al. [9] replaced the original pooling operation by proposing a dilated convolution module. The dilated convolution can expand the receptive field and extract multi-scale context information on the basis of maintaining high-resolution conditions, and generate large-scale feature maps with rich spatial information, which is effectively applied to the field of semantic segmentation.

3. The baseline

3.1. Original Unet
UNet [2] was first released on the 2015 MICCAI and has gradually become the baseline for most medical semantic segmentation tasks. In the current field of computer vision, more and more semantic segmentation and target detection tasks have begun to pay attention to and use the network structure of Unet. Unet is usually used in the medical field, and it is implemented by four consecutive down-sampling operations, four times corresponding upsampling operations, and skip-connection. Because Unet has an input size limit of 572*572, it is not suitable for road segmentation tasks with an input image size of 1024*1024. In addition, the original Unet does not prove that its four-layer down sampling and up sampling processing structure has the best effect in all semantic segmentation tasks. Therefore, this project does not directly use the original Unet architecture but makes a certain degree of modification in the original model, so that it has better performance in the road segmentation task.

4. The Optimization

4.1. Paralleled dilated convolution module structure
According to the results of a study by He et al. in ResNet [10], a shortcut connection is added between the network layers to implement identity mapping of input and output within the neural network solving the gradient explosion problem and the network degradation caused by the deep depth of the network.

In the work of this paper, we design a dilated convolution module with parallel structure by using the shortcut mapping according to the characteristics of the residual network. This module can combine the multi-scale semantic information to improve the performance and performance of the network model.
under the premise of using the dilated convolution to increase the receptive field of the neural network while maintaining the high resolution.

![Fig. 1 Paralleled dilated convolution module structure](image)

The parallel dilated convolution structure can extract different levels of semantic feature maps from the original image and fuse the feature maps from different levels in a stacking manner to obtain better multi-level feature fusion results. By adding a shortcut connection between the parallel hole convolutional structures, we avoid the gradient explosion and network degradation caused by the network layer stacking too deep in a similar way to ResNet [10]. By using a parallel dilated convolution structure, we can apply deeper networks in the algorithm without causing degradation in feature extraction performance. This structure is applied to the network model proposed in this paper to help deep dilated Unet achieve better semantic segmentation performance.

### 4.2. **Encoder of Deep Dilated UNet**

The improved Deep Dilated UNet model inherits three main features of the UNet network structure: the down sampling process, the up-sampling process, and the skip-connection. The down sampling operation can increase the robustness of the model to some weak noise of the input image, such as image translation, rotation, etc., reducing the risk of overfitting, reducing the amount of computation, and increasing the size of the receptive field. The biggest effect of up sampling is to re-decode the abstract features to the size of the original image, and finally get the segmentation result.
Fig. 2 Overall structure diagram of the Deep Dilated Unet

The figure shows the overall structure of the Deep Dilated Unet we designed. The overall architecture of the network consists of an encoder, a parallel dilated convolution module, a decoder, and a shortcut connection. The difference between Deep Dilated Unet and original UNet is that a parallel hole convolution module is added to the bottom layer of UNet. We use a parallel dilated convolution module to connect the encoder part and the decoder part of the network structure. The parallel dilated convolution module is placed at the bottom of the UNet network to effectively adjust the depth of the convolution layer without being affected by network degradation.

The Encoder section follows the classification paradigm of convolutional neural networks, which contains six convolution operations, each of which is followed by an activation function to increase the nonlinearity of the network. Among them, every two convolution operations constitute an operation phase, we use the maximum pooling layer at the end of each down sampling stage to achieve the down sampling of the operation phase, and the size of the feature map is changed to the original one-half.

4.2.1. VGG 16. Because the dataset with good annotations that can be used for training contains fewer images, VGG16 [4] trained to convergence [8] is used in the work of this paper as the pre-training model of the network model proposed in this paper. In order to adapt this pre-training mode to the requirements of the network model proposed in this paper, we have deleted the last three layers of fully connected layers in the VGG network. Specifically, the weight values of the first, third, sixth, eighth, and eleventh pre-training models of VGG16 are respectively applied to the first to sixth layers of the model proposed in the present work. Using a pre-training model will help us reduce the time required for the training model to converge, while also reducing the over-fitting of the network model to some extent.

4.2.2. The decoder of Unet. Prior to the feature merging operation, the transposed convolutional layer is used to double the size of the feature map to implement an up-sampling operation to achieve size matching of different levels of network feature maps. This is done by using a deconvolution layer instead of a bilinear interpolation layer because the deconvolution layer has better performance for restoring features from the output of the convolutional layer. The model uses three deconvolution layers to amplify the size of the feature map by 2, 4, and 8 times, respectively. At the same time, the feature maps generated during the three down sampling processes are respectively connected to the feature maps output by the three deconvolution layers by the ‘concat’ operation. In our implementation, we added a padding layer on top of the original Convolution-Relu structure in the decoder section, which was converted to a Convolution-Padding-Relu structure, which helps us maintain the size of feature map not
changed during the convolution process. In this way, we restore the feature map of the same size as the input image for the final pixel classification.

4.3. Training
All of the experiments based on the DeepGlobe road extraction dataset covered in this article were performed on four NVIDIA GTX1080TI GPU. Because of the limitations of memory, in the experiment, each graphics card was assigned with a batchsize of 8. The learning rate of the adam optimizer used in the experiment was set to 2e-4. The training of the network model is based on the training set in the DeepGlobe road extraction dataset [5]. The network model is trained until its performance is no longer improved.

5. Evaluation

5.1. dataset and metrics
The dataset used in the work of this paper is the DeepGlobe road extraction dataset, which is derived from data extracted from satellite images. The DeepGlobe road extraction dataset contains 6226 training images, 1243 verification images and 1101 test images with a ground resolution of approximately 0.5m/pixel. All of these images have size of size 1024*1024. These images are from Thailand, India, Indonesia, and contain many complex scenes such as cities, villages, beaches, and rainforests. All of the experiments involved in the work of this paper were based on this dataset.

In the work involved in this paper, we use the pixel-wise Intersection over Union (IOU) metric to evaluate the predictive power of the models proposed in this paper. IOU can be defined as Equation 1 [5]

\[
IOU_i = \frac{TP_i}{TP_i + FP_i + FN_i}
\]

Where TP is defined as the number of road pixels that are correctly predicted, FP is defined as the number of non-road pixels that are erroneously predicted as road pixels, and FN is defined as the number of road pixels that are erroneously predicted as non-road pixels.

For n test picture samples, the model’s comprehensive predictive power mIOU can be expressed as the arithmetic mean of the IOU vs. n. mIOU can be defined as Equation 2.

\[
mIOU = \frac{1}{n} \sum_{i=1}^{n} IOU_i
\]

5.2. Experimental Results

Table 1. Results on validation set on DeepGlobe Road Extraction Task

| Network Structure                                | mIOU Score (%) |
|--------------------------------------------------|----------------|
| Original UNet                                    | 58.29          |
| Deep Dilated UNet                                | 62.94          |
| Deep Dilated UNet with Paralled Dilated Convolution Module | 63.72          |

From the experimental results, it can be found that Deep Dilated UNet can obtain a certain degree of accuracy improvement over than the original UNet, but there are still many samples of recognition errors in the semantic segmentation results. For example, there exists some rivers incorrectly identified as a road.

In order to further improve the performance of the network model, we settle the Paralled Dilated Convolution Module proposed in this work of paper at the bottom of the Deep Dilated UNet, and effectively improve the performance of the network model by further expanding the receptive field of the network model. According to the test results, Deep Dilated UNet with Paralled Dilated Convolution Module will further improve the semantic segmentation accuracy to 63.72% mIOU score based on Deep Dilated UNet, which proves that we propose to apply Paralled Dilated Convolution Module to neural network. The method of improving the receptive field of the network is effective for the satellite semantic segmentation task. Compared with the original UNet structure, our proposed network model
effectively improves the mIoU score of 5.43%, and successfully improves the effectiveness of existing algorithms in satellite image semantic segmentation tasks.

6. Conclusion

In the work of this paper, in order to meet the needs of higher precision satellite image segmentation tasks, we proposed Dilated Deep Unet. The network structure proposed in this paper is designed based on a full convolution architecture, and its central part does not contain any pooling operations. This design guarantees the integrity of semantic information to the greatest extent possible. The architecture makes the results of semantic segmentation of satellite images clearer and more accurate through a redesigned neural network structure. We use VGG16 as a pre-training model to transfer learning our proposed network model to cope with the problem of too small training set. We put the parallel convolution structure at the bottom of the network model to apply the deeper network to our model, and use the long connections in the network structure to effectively couple the semantic information of different levels, effectively improving the performance of the model in the semantic segmentation of satellite images. In the future, we will continue to focus on the application of combining different levels of semantic information in multiple image processing directions.

References

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[2] Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” arXiv preprint arXiv:1312.6229, 2013.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[5] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raska, “Deepglobe 2018: A challenge to parse the earth through satellite images,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2018, pp. 172–17209.

[6] Kmeans: MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 2881–2890, 2017.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[9] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: International Conference on Learning Representations, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.