Journal of New Media
DOI:10.32604/jnm.2020.014115
Article

Tech Science Press

Robust Cultivated Land Extraction Using Encoder-Decoder

Aziguli Wulamu[1,2,*], Jingyue Sang[3], Dezheng Zhang[1,2] and Zuxian Shi[1,2]

[1]Department of Computer, School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, China

[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083, China

[*]Corresponding Author: Aziguli Wulamu. Email: aziguli@ustb.edu.cn

Received: 31 August 2020; Accepted: 11 September 2020

Abstract: Cultivated land extraction is essential for sustainable development and agriculture. In this paper, the network we propose is based on the encoder-decoder structure, which extracts the semantic segmentation neural network of cultivated land from satellite images and uses it for agricultural automation solutions. The encoder consists of two part: the first is the modified Xception, it can used as the feature extraction network, and the second is the atrous convolution, it can used to expand the receptive field and the context information to extract richer feature information. The decoder part uses the conventional upsampling operation to restore the original resolution. In addition, we use the combination of BCE and Loves-hinge as a loss function to optimize the Intersection over Union (IoU). Experimental results show that the proposed network structure can solve the problem of cultivated land extraction in Yinchuan City.

Keywords: Semantic segmentation; encoder-decoder; cultivated land extraction; atrous convolution

1 Introduction

With the launch of a large number of high-resolution multi-spectral remote sensing satellites in China, it is more and more convenient to obtain high-resolution multi-spectral remote sensing images. The images obtained by remote sensing satellites are applied to various industries and fields such as land, agriculture, surveying, and water conservancy. The original remote sensing image semantic segmentation is done by manual visual observation. The interpretation process is realized according to the what the interpreter knows. This method is labor intensive, subjectively affected, high in requirements for staff, and difficult to update [1]. Automatic interpretation of remote sensing images is conducive to saving a lot of manpower and material resources; transforming the extracted feature information with remote sensing knowledge into vector files, using machine learning algorithms for horizontal and vertical comparison, can realize change monitoring, discover illegal land use and some natural disasters; mining valuable information in a large amount of interpreted data for scientific planning land use, improving work efficiency and social and economic benefits. Therefore, it is of great significance to explore an efficient and highly accurate method for automatic interpretation of remote sensing images. In this paper, our goal is to solve the cultivated land extraction in three urban areas of Yinchuan City, with a total area of 2,045 square kilometers. These areas have complex terrain, including mountains, alluvial plains, deserts, cultivated land, urban areas, photovoltaic power plants, and woodlands. The shape of cultivated land in different terrain is also different. To this end, we propose a network structure based on encoder-decoder [2] and atrous convolution [3]. At the same time, the final loss function is obtained by combining many loss functions. Through experiments, it can be seen that this method can better solve the problem of cultivated land extraction in the three districts of Yinchuan. The rest of this article is arranged as follows.

The related work introduced in Section 2 is the part of cultivated land extraction. Section 3 is a detailed description of this method, mainly by splitting different modules to explain. Section 4 introduces the experimental data, results and evaluation criteria. The final Section 5 of the article gives the final conclusion.

2 Related Work

The image explanation of remote sensing images is roughly divided into three stages: visual interpretation, human-computer interaction interpretation and fully automatic interpretation.

Visual interpretation is the most basic and intelligent interpretation method. The interpreter combines visual information (color, texture, size, shape, shadow, layout, etc.), interpretation experience (expert knowledge in remote sensing) and various thematic maps with a variety of non-remote sensing information, using the relevant laws of biological geology, perform comprehensive analysis and logical reasoning proceed from the one to the other, from the surface to the centre, eliminates the false and retain the true [4]. Although visual interpretation can achieve good results, it requires a lot of manpower, material resources and financial resources. The visual interpretation request of the interpreter is very high, each interpreter's professional knowledge, interpretation experience is not the same, they cannot guarantee the consistency of the results of different interpreters.

Human-computer interaction interpretation is the mainstream interpretation method. This method is divided into supervised classification and unsupervised classification based on pixel, object-oriented supervised classification and unsupervised classification. The supervised classification method based on pixels mainly includes the shortest distance method [5], the Mahalanobis distance method [6], etc. The unsupervised classification method mainly includes clustering method [7] and iterative self-organizing data analysis method (ISODATA) [8] and so on. The object-oriented interpretation method was first proposed by [9] according to the characteristics of high-resolution multi-spectral images. Image analysis software represented by e Cognition [10] appeared. The typical interpretation steps of the object-oriented interpretation method are divided into: Segmentation image, feature extraction and combination, and classification based on feature classification algorithm. The key step here is to segment the image. The most common is multi-scale segmentation [11], but for different features, there are different optimal segmentation scales. The optimal segmentation scale of a feature requires a lot of repeated experiments. Although the object-oriented interpretation method can improve the pixel-based interpretation method, the determination of the segmentation scale and the cumulative error caused by the segmentation make the method have certain limitations.

We define the automated extraction of cultivated land in remote sensing images as a semantic segmentation task, which is the focus of research in recent years. Deep learning has achieved great success in the field of image recognition [12, 13]. The network architecture based on CNN [14–16] can automatically extract feature maps in remote sensing images, which lays a solid foundation for semantic segmentation. Long et al. [17] proposed FCN, which allows convolutional neural networks to perform dense pixel prediction without the need for a fully connected layer. A classic encoder-decoder architecture was proposed in U-Net network structure by Ronneberger et al. [18]. In this structure, the encoder is used to gradually reduce the spatial size of the pooling layer, while the function of the decoder is to gradually restore the details and spatial size of the object. The reason why the decoder can determine the details of the target more accurately is because there is a shortcut connection between the encoder and the decoder.Subsequently, Chen et al. [19] proposed the DeepLab semantic segmentation model , and improved on this basis, and then proposed DeepLab V3 [20], DeepLab V3+ [21] models, the DeepLab series models have in common use the atrous convolution, and realize the spatial pyramid pooling in the spatial dimension, extracting context information. The difference is that DeepLab V1 and DeepLab V2 regard the output of the convolutional neural network as the input of the fully connected condition random field, which is the purpose of considering global information and improving local information. However in DeepLab V3 and DeepLab V3+ no longer use this structure, use Resnet, Xception [22] and encoder-decoder structure to consider global information, and up to now is the highest accuracy rate in the PASCAL VOC 2012 dataset [23]. Recently, the D-LinkNet proposed by Zhou et al. [24] used dilated

convolution layer with a shortcut to achieve excellent results in remote sensing image road extraction tasks. The Lovasz hinge proposed by Berman et al. [25] further improved IoU in the segmentation problem of 2-value images. This paper absorbs the advantages of these most advanced models and recombines them into a completely new model for the automatic extraction of cultivated land in the Ningxia City.

3 Proposed Method

This section will introduce our network structure and loss function for cultivated land extraction.

3.1 Network Architecture

The network architecture is presented in Fig. 1. We propose a new encoder-decoder architecture. It is divided into two parts: an encoder for extracting feature maps and a decoder for restoring feature map resolution. In the encoder part, the modified aligned Xception is used to extract the features of the input remote sensing image; then the atrous convolution layers with atrous rate of 1, 2, 4, 8 combined with the cascade mode and the parallel mode is used to expand the receptive field and increase the context information, extract more rich feature information. In the decoder part, at first, the feature maps generated by different atrous rates in the encoder are combined, and the combined feature maps are upsampled by 4 times in a bilinear manner, and then the lower level of the same spatial resolution is obtained in the encoding and the combined feature maps. The feature map adjusts the number of channels by 1×1 convolution, so that the spatial resolution and the number of channels of the low-level feature map are exactly the same as the feature map of 4 times of the upsampling, and then the two feature maps are concated. After the connection, a few 3×3 convolutions are used to refine the connected feature maps to make the connected feature maps smoother. Finally, the bilinear method is used to upsample 4 times to restore the original resolution.

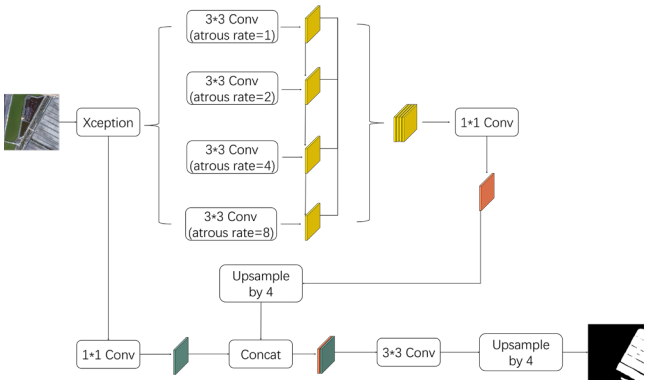

Figure 1: Our network architecture

3.2 Loss Function

Cultivated land extraction is an unbalanced task. In some places, the cultivated land may occupy the entire image. At the same time, in some places, the background occupies the entire image. There are two ways to solve these problems: one is the balance of data sets; the other is to solve the problem of insufficient data sets. Here, the method we use to improve model performance is the second.

Pixel-level cross-entropy loss is a common semantic segmentation loss function. After checking each pixel one by one, it compares the encoding target vector and class prediction. The reason why we can treat each pixel in the image equally is because the cross-entropy loss function evaluates the classification prediction of each pixel vector separately and then averages it. But the problem is that the categories of our data set are extremely unbalanced, in which background or farmland with too many training times will cause the final training results to tend to them. The loss function based on Dice coefficient is another

popular semantic segmentation loss function, and the calculation of the degree of overlap between two samples is the essence of this function. The Dice coefficient calculation formula is as follows:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

where $|A \cap B|$ represents the number of common elements of the set A and B, $|A|$ represents the number of elements of the set A, and $|B|$ represents the number of elements of the set B. It can be seen from equation (1) that the intersection between the prediction and the target mask is where the molecules are concentrated, and the activation amount in the mask is related to the denominator. The resulting effect is to normalize the loss based on the size of the target mask, so that the learning class can be learned from images with lower distribution. We noticed that the evaluation index IOU proposed by Berman et al. [3] to optimize the segmentation task uses Lovasz-Softmax loss. And we combine the dice coefficient, binary cross entropy (BCE) and Lovasz-Hinge loss as our final loss function, which is: $L_f = (1 - \alpha) * (L_{BCE} + L_{DC}) + \alpha * L_{LH}$ (2)

The BCE loss is represented by L_{BCE}, the loss function (based on the Dice coefficient) is represented by L_{DC}, and the loss Lovasz-Hinge is represented by L_{LH}, and the weights from different losses are also controlled by it.

4 Experimental

This section introduces our experience. We will introduce the data set at the beginning, then explain the experimental details, explain the evaluation criteria, and finally show our experimental results.

4.1 Datasets

The satellite image we used in the experiment comes from Satellite Gaofen 2, which is used to extract cultivated land. It covers three districts of Yinchuan City, namely Jinfeng District, Xingqing District and Xixia District. The total area covered by it reaches 2045km^2, and 100cm/pixel is the ground resolution of the image pixel. Extracting arable land from the above three regions is the goal we want to achieve. In the data set that has been marked, 3968 images with a resolution of 512 × 512 come from the Gaofen-2 satellite, covering 1041 km^2 in total. There are three types of labeled data sets, the first is the training set, with 2778 images, corresponding to a 70% division; the second is the validation set, with 396 images, corresponding to a 10% division; the third is the test set, there are 794 pictures, corresponding to a 20% division. Its application scenarios are Jinfeng District, Xingqing District and Xixia District of Yinchuan City.

4.2 Evaluation Criteria

The pixel accuracy (PA) and Jaccard index of semantic segmentation are the evaluation indicators of our experiment. PA is to calculate the ratio of correctly classified pixels to the total number of pixels. Mean Intersection over Union (MIoU) is a standard metric for semantic segmentation. It calculates the ratio of intersection and union between two sets. In image segmentation, it is the two sets of true and predicted values, which can be converted to the intersection of the two and divided by the union of the two to calculate the intersection ratio within each class, and then calculate the mean.

If we assume that the number of road pixels that are correctly predicted is TP, the number of pixels that are predicted to be incorrect is FP, the number of pixels that are not road pixels but are correctly predicted is TN, and the number of pixels that are not road pixels and the number of predicted error pixels is represented by FN. Then the formula of PA is:

$$\text{PA} = \frac{\text{TP+TN}}{\text{FP+FN}} \quad (3)$$

The definition of mIoU is the mean value of IoU of all images, where n represents the number of all images:

$$\mathrm{mIoU} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{IoU}_i \quad (4)$$

4.3 Implementation Details

In the training stage, we use some common methods to expand the data set to avoid overfitting problems, such as random offset, flip, scale, and Gaussian blur. The learning strategy we used for the 50-period network training is poly, 0.008 is its initial learning rate, 5e-4 is the weight decay value, and 0.9 is the momentum setting value. The α in the loss function is set to 0.1.

In the testing and application phase, we use the test time increase (TTA) method to enhance the robustness of model prediction. The method includes vertical flip, test diagonal flip and horizontal flip, and consists of 8 predicted averages.

4.4 Experimental Results

The model we selected comes from the prediction results of cultivated land extraction from different terrains, and it is random, as shown in Fig. 2. As we have seen, our model distinguishes cultivated land from other green vegetation. As can be seen from the Fig. 2(b), our model has achieved good results for the extraction of large-area cultivated land and small-area cultivated land. For the Field ridge in the cultivated land has a certain ability to distinguish (see Figs. 2(a), 2(b), 2(d)), the developed area of the cultivated land has the ability to distinguish (see Fig. 2(c)).

We compared the current model with the network architecture, and the details are shown in Table 1.U-net is slightly inferior in all models due to its too simple structure, but it provides a very good idea for building a network model. Compared to our model and D-LinkNet, our model has been improved by 0.28% on mIoU, which proves that the modified xception we adopted and the loss function introduced specifically for iou are effective. Compared to our model and Deeplabv3+, our model has been improved by 1.12% on mIoU, which proves that we are using the combine cascade mode and parallel mode of atrous convolution slightly better than the ASPP module in the Yinchuan dataset, so in the simple semantic segmentation task, we can use the combine atrous convolution.

Table 1: Comparison with U-Net, D-LinkNet, DeepLabV3+, ours network on the test datasets

Model	Pixel Accuracy	mIoU
U-Net	0.833	0.672
D-LinkNet	0.841	0.710
DeepLabv3+	0.840	0.704
Ours	0.842	0.712

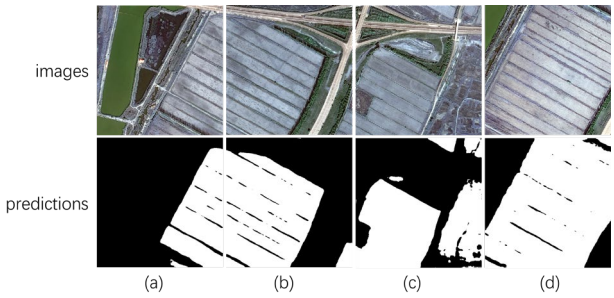

Figure 2: Example results of our method on cultivated land extraction in the Yinchuan City

5 Conclusion

In this paper, we present a new network of cultivated land extraction. In terms of network architecture, we reassemble the atrous convolution, the modified aligned Xception, and the classic

encoder-decoder architecture to expand the network's receptive field, enabling the network to obtain richer context information for more advanced feature mapping. In terms of the loss function, we introduce a lovasz loss function optimized for IoU in the conventional loss function, which can achieve a slightly better extraction effect. Experiments show that the network architecture can solve the cultivated land extraction task in Yinchuan City and has certain robustness, which lays a foundation for agricultural automation solutions.

In the future, further subdivision of cultivated land can achieve precise agricultural automation, which is of great significance. In addition, because large-scale manual labeling of remote sensing images is a very time-consuming and difficult task to achieve, consider using a weakly supervised network to improve the network's training capabilities.

Funding Statement: The main sources of support for this work are as follows: Ningxia Hui Autonomous Region Key Research and Development Program Project: Research and demonstration application of key technologies for intelligent monitoring of spatial planning based on high-scoring remote sensing (Project No. 2018YBZD1629).

Conflicts of Interest: No conflicts of interest to report on this research.

References

[1] B. Du, W. Zhang. "Research on object-oriented high resolution remote sensing image classification technology," *Western Resources*, no. 5, pp. 135-138, 2016; *FNM Surname* (2018), vol. 10, no. 3, pp. 1-10.

[2] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention,* pp. 234-241, 2015.

[3] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2016.

[4] J. J. Pu, "Principles and methods of visual interpretation of remote sensing images," *Science and Technology of China Press,* 1992.

[5] D. T. Feng, G. Chen, K. L. Xiao, W. Y. Du and X. Y. Wu, "Remote sensing image classification based on minimum distance method," *Journal of North China Institute of Aerospace Engineering*, no. 3, pp. 1–2, 2012.

[6] X. Zhao, Y. Li, Q. Zhao, "Mahalanobis distance based on fuzzy clustering algorithm for image segmentation," *Digital Signal Processing*, vol. 43, no. C, pp. 8–16, 2015.

[7] J. Zheng, Z. Z. Cui, A. Liu and et.al, "A K-means remote sensing image classification method based on AdaBoost," in *Fourth Int. Conf. on Natural Computation*, *IEEE*, vol. 4, pp. 27–32, 2007.

[8] N. Memarsadeghi, D. M. Mount, N. S. Netanyahuand and J. L. Moigne "A fast implementation of the isodata clistering algorithm," *International Journal of Computational Geometry & Applications*, vol. 17, no. 1, pp. 71–103, 2007.

[9] R. Kettig and D. Landgrebe, "Classification of multispectral image data by extraction and classification of homogeneous objects," *Geoscience Electronics IEEE Transactions on*, vol. 14, no. 1, pp. 19–26, 1975.

[10] eCognition. [Online]. Available: http://www.ecognition.com/.

[11] M. Baatz and A. Schäpe, "Multiresolution segmentation: An optimization approach for high quality multiscale image segmentation," *Angewandte Geographische Informations-Verarbeitung XII*, pp. 12–13, 2000.

[12] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[13] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] Y. Lecun, B. Boser, J. S. Denkerand, R. Howard, W. Hubbard *et al,* "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[16] K. He, X. Zhang, S. Ren and S. Jian, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[17] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2014.

[18] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[19] L. C. Chen, G Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *Computer Science*, no. 4, pp. 357–361, 2014.

[20] L.C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[21] L.C. Chen, Y. Zhu and G. Papandreou, F. Schroff and H. Adam "Encoder-Decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, pp. 833–851, 2018.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[23] M. Everingham, "The PASCAL visual object classes challenge, (VOC2007) results," *Lecture Notes in Computer Science*, vol. 111, no. 1, pp. 98–136, 2007.

[24] L. Zhou, C. Zhang, M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, *IEEE*, 2018.

[25] M. B. A. R. T. Matthew and B. Blaschko, "The Lov\\'asz-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.