W-net: the Convolutional Network for Multi-temporal High-resolution Remote Sensing Image Arable Land Semantic Segmentation

Linshan Shi1,2, He Huang1*, Yang Shi1, Yimin Hu1

1Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, China
2University of Science and Technology of China, Hefei, Anhui, 230026, China
*Corresponding author’s e-mail: lsd201@126.com

Abstract. Aiming at the problem that manual arable land semantic segmentation is time-consuming, inefficient, segmentation results rely heavily on the staff’s experience and traditional methods cannot meet the requirements of land segmentation, this paper first proposes a new neural network W-net to realize multi-temporal high-resolution remote sensing image arable land semantic segmentation. We use some data augmentation methods, build the W-net neural network, train with the training set, and test the performance of the neural network with the test set. In this paper, the results of manual labeling are used as the reference. In the case of maintaining the same training set, verification set and test set, FCN, SegNet and W-net are used for control-experiments. The experimental results show that the segmentation accuracy of W-net is higher than that of FCN and SegNet, and the edge segmentation is flatter. W-net has higher accuracy of arable land semantic segmentation for multi-temporal high-resolution remote sensing images than single-temporal high-resolution remote sensing images. W-net neural network can effectively fulfill the task of arable land semantic segmentation.

1. Introduction

High-resolution remote sensing images play an important role in agricultural insurance because agricultural insurance covers a large area and cannot meet this requirement with ordinary cameras. In recent years, with the rapid development of remote sensing imaging technology and the advancement of image processing technology, the processing technology for high-resolution remote sensing images has gradually developed into an important research field. The semantic segmentation of arable land is a research direction with high application value. In general, the arable land semantic segmentation technology extracts the important parts of the high-resolution remote sensing images by designing an automatic or semi-automatic segmentation algorithm, and finally the segmentation result is basically consistent with the ground survey results. Arable land semantic segmentation is of great significance in agricultural insurance, because agricultural insurance companies need to rely on the results of arable land semantic segmentation when formulating specific claims plan, otherwise they cannot formulate reasonable claims plan.

2. Related Work

For the semantic segmentation task of arable land, some people adopt the method of manual segmentation. Although this method can complete the task of arable land semantic segmentation, there
are three main drawbacks in this method. The first is that the staff needs to spend a lot of time dividing the arable land, as a result, the efficiency is low. The second is that the manual segmentation result is heavily dependent on the staff's experience, for the same kind of arable land, different staff may get different segmentation results. The third is the high cost of completing the task. To overcome the traditional problem of manual segmentation by staff, some researchers began to study the automated or semi-automated arable land semantic segmentation algorithm to realize the semantic segmentation of high-resolution remote sensing images. The result of semantic segmentation can assist insurance companies to design rational agriculture insurance plan.

In the field of arable land semantic segmentation, there are a large number of models that apply traditional algorithms. However, because the arable land is affected by the complex factors such as water source, topography, geological conditions and land contracting, it has different shapes. So the task of dividing the arable land is very challenging. At present, there are few kinds of research on the automatic segmentation algorithm of arable land at home and abroad, but some researchers have yielded a rich harvest in this field, such as arable land semantic segmentation based on threshold [1], arable land semantic segmentation based on regional splitting and consolidation [2], and arable land semantic segmentation based on edge detection [3]. Although the principles of these methods are not the same, they are basically realized by using the three low-level features of the image, namely, color, texture and shape features. However, the overall segmentation result of these methods is not ideal when encountering complex scenes.

Some people also used the idea of unsupervised learning for reference. According to several observable properties of pixels in the image, the clustering algorithm was used to merge pixels with high similarity into one superpixel [4]. The clustering method can realize the function of dividing the image into super-pixel blocks with uniform size, which provides a good foundation for subsequent processing tasks. However, in the picture of the actual scene, the structure of some objects is more complicated, and the difference between pixels is large. Only using the low-level information such as the color, brightness and texture is not enough to generate a good segmentation effect, or even easily produces the wrong segmentation.

In 2006, Hinton and Salakhutdinov discovered that multi-layer feedforward neural networks can be pre-trained layer by layer, and then backpropagation algorithm was used for effective learning, which proved the practicality of deep learning [5]. Since then, deep learning [6] has developed rapidly, and it has become an important method for the arable land semantic segmentation. The deep convolutional neural network and its variants are widely used. They have certain effects on the arable land semantic segmentation task, but have a common problem that they use a region around the pixel as input, which results in a large number of overlapping image blocks, which increases the redundant computation, so this method is less efficient. What’s more, because this method ignores the overall information, it is also less accurate. Some researchers used Fully Convolutional Networks (FCN)[7] to solve this problem. FCN is an end-to-end model that can use an entire picture as the input of the network and use a whole picture as the output of the network [8], thus it can avoid the problems caused by a large number of overlapping image blocks. The reason why it can use the entire picture as the input of the network is that it replaces the fully connected layer with a convolutional layer, allowing the neural network to accept images of any size. In addition, the feature map of the neural network is smaller than the input image. If you want the output image of the neural network to be as large as the original image, then adding the deconvolution layer is useful. We can classify each pixel individually by doing this. FCN has two distinct advantages over the traditional convolutional neural network. First, FCN does not require the size of the input image, and the size of the training image and the test image can be different. Second, it is more efficient because it avoids the redundant computation caused by a large number of overlapping pixel blocks. However, FCN also has two obvious shortcomings. First, the results are rough, the results of the upsampling are vague, and it is not sensitive to the details in the image. Second, when classifying each pixel, the relationship between pixels and pixels is not fully utilized.
In addition, some people use SegNet [9] to fulfill the task of arable land semantic segmentation. The advantage of SegNet is that it can improve edge characterization and reduce training times. Although its structure is elegant, its results are not necessarily better than FCN.

Although some deep learning models have been applied to the arable land semantic segmentation, they are still relatively little in general, and the results of these deep learning models are not ideal in the process of being applied. So this paper first proposes a new neural network W-net to realize multi-temporal high-resolution remote sensing image arable land semantic segmentation.

3. Research Method

3.1. Data Preprocessing

The high-resolution remote sensing images size is particularly large, if you directly train the neural network, the computer's memory can't bear it, so we can't directly use these images to train the neural network. Therefore, we need to cut images randomly to obtain small-sized high-resolution remote sensing images.

In order to eliminate geometric distortion caused by factors such as the sensor, atmosphere and earth curvature, we perform geometric correction on the high-resolution remote sensing images used in the experiment. In addition, we also eliminate the error caused by the sensor itself by radiometric calibration. In order to eliminate the errors caused by atmospheric scattering, absorption, and reflection, we also perform additional atmospheric correction processing.

3.2. Data Augmentation

Since the available annotated high-resolution remote sensing images data is less and difficult to train a deep neural network model, some data augmentation methods are used in this paper to increase the number of available training samples. The data augmentation operation includes randomly rotating the original image and the corresponding label at the same time, mirroring the original image and the corresponding label simultaneously, blurring the original image and the corresponding label simultaneously and adding noise to the original image and the corresponding label. After these data augmentation operations, we get a lot of training samples.

3.3. Modeling

Figure 1. W-net network structure.

Figure 1 shows the W-net network structure, which consists of four parts. From left to right are contracting path 1, expansive path 1, contracting path 2 and expansive path 2. The contracting path 1 and contracting path 2 all follow a typical convolutional neural network architecture. It consists of multiple repeating structures. Each repeating structure has two convolutional layers with a convolution
kernel size of 3*3, and the convolutional layer is followed by a rectified linear unit and a 2*2 max pooling operation with stride 2 for downsampling. In each downsampling operation, we double the number of feature channels. Each step in the expansive path 1 and expansive path 2 first uses a deconvolution that halves the number of feature channels and then stitches the result of the deconvolution with the corresponding feature map in the contracting path.

In general, W-net is downsampled by convolution and pooling, then upsampled by deconvolution, then secondarily downsampled by convolution and pooling, and secondarily upsampled by deconvolution. The output feature map is obtained, and finally the split graph is obtained through the activation function. Convolution can reduce the complexity of the network model and reduce the number of weights. There is no need to select features manually, and the weights are trained to obtain features. The classification effect is good. Pooling can improve the receptive field, reduce parameters, increase nonlinearity and achieve translational rotation scale invariance, but it will also cause information loss, especially some low-level features, so we need to splice some low-level features to restore the low-level information so as to improve segmentation precision. The downsampling process is followed by a symmetrical upsampling process to improve the output resolution. Moreover, the low-level features obtained by downsampling are combined with the upsampled input to refine the boundary information.

The end of the W-net model uses a 1*1 convolutional layer to reduce the number of feature maps to 2, and uses the sigmoid function to process the final output so that each pixel value in the network output corresponds to a range of 0 to 1. The corresponding value of each pixel indicates the type of point. To enhance the display, the final map maps 1 to 50.

If the activation function is not used, then the output is a linear combination of inputs, in this case, the expressive power of the neural network is limited. In order to improve the expressive power of the neural network, a nonlinear activation function is generally added after the convolutional layer, so that the output of the neural network is no longer a linear combination of linear inputs. Commonly used nonlinear activation functions are sigmoid function, tanh function, ReLu function and maxout function. Among them, ReLu function is widely used in various network models.

The gradients of sigmoid and tanh are very gentle in the saturation region, which easily causes the gradient to disappear and reduces the convergence speed. And the deeper the network is, the more obvious gradient disappears will be. The gradient of Relu is a constant, the convergence speed is fast, and there is no problem that the gradient disappears. In addition, it is more in line with the characteristics of biological neurons, the effect is often better. Therefore, this paper uses the ReLu function as the activation function.

The traditional CNN-based segmentation method is as follows: in order to classify a pixel, an image block around the pixel is used as input to the CNN for training and prediction. This method has several disadvantages. First, the storage overhead is large. For example, if the size of an image block used for each pixel is 15*15, the required storage space is 225 times that of the original image. Second, calculation efficiency is low. Adjacent pixel blocks are essentially repetitive, and the convolution is calculated one by one for each pixel block. This calculation is also largely repetitive. Third, the size of the pixel block limits the size of the perceived area. Usually the size of a pixel block is much smaller than the size of the entire image, and only some local features can be extracted, resulting in the limited performance of the classification.

Unlike traditional CNN-based segmentation methods, the end-to-end semantic segmentation method is to input the entire image and output a segmentation graph of the same size. For the arable land semantic segmentation task of this paper, if the end-to-end segmentation method is adopted, then only one image should be input to obtain the complete segmentation result, and the time consumed by cutting the image in advance can be reduced. Therefore, the end-to-end segmentation method is adopted in this paper. The segmentation method based on FCN is a common end-to-end segmentation method. When it processes the feature map, it mainly adopts the method of superimposing the feature maps of different layers. Different from the segmentation method based on FCN, the segmentation method based on W-
net extracts the feature map obtained by convolution and the feature map obtained by deconvolution when the feature map is processed and connected by jump, so that the ability to capture image edge information can be improved.

### 3.4. Training Model

After dividing the dataset, the forward propagation operation is performed on each sample first. The formula of the forward propagation operation is as follows:

$$ net^{(l+1)} = W^{(l)} a^{(l)} + b^{(l)} $$ \hspace{1cm} (1)

where $net^{(l+1)}$ is the input weighted sum of $l+1$ layer, $W^{(l)}$ is the join weight between $l$ layer and $l+1$ layer, $a^{(l)}$ is activation of $l$ layer, $b^{(l)}$ is the bias term of $l+1$ layer.

$$ a^{(l+1)} = f\left(net^{(l+1)}\right) $$ \hspace{1cm} (2)

Where $f(\cdot)$ denotes activation function

Using the above forward propagation formula, the activation values of the second layer, the third layer, ... up to the output layer can be obtained. In order to find the parameters $W$ and $b$ that can make the cost function $L(W, b)$ the smallest, we first calculate the residual of the output layer $L^n$:

$$ \delta^{(n)} = \frac{\partial}{\partial net^n} L(W, b) $$ \hspace{1cm} (3)

Where $net^n$ is the input weighted sum of the output layer

We calculate residuals of $l = n - 1, n - 2, n - 3, ..., 2$ layers according to the following formula.

$$ \delta^{(l)} = \left(\left(W^{(l)}\right)^T \delta^{(l+1)}\right) \cdot f'\left(net^{(l)}\right) $$ \hspace{1cm} (4)

Then, calculate the partial derivative of the cost function of a single sample:

$$ \nabla_{W^{(l)}} L(W, b; x, y) = \delta^{(l+1)} \left(a^{(l)}\right)^T $$ \hspace{1cm} (5)

$$ \nabla_{b^{(l)}} L(W, b; x, y) = \delta^{(l+1)} $$ \hspace{1cm} (6)

Then calculate the sum of the partial derivatives of the cost function of all samples:

$$ \Delta W^{(l)} := \Delta W^{(l)} + \nabla_{W^{(l)}} L(W, b; x, y) $$ \hspace{1cm} (7)

$$ \Delta b^{(l)} := \Delta b^{(l)} + \nabla_{b^{(l)}} L(W, b; x, y) $$ \hspace{1cm} (8)

Last update weight parameter:

$$ W^{(l)} := W^{(l)} - \alpha \left[\frac{1}{m} \Delta W^{(l)} + \lambda W^{(l)}\right] $$ \hspace{1cm} (9)

where $\alpha$ denotes learning rate, $\lambda$ denotes weight attenuation parameter

$$ b^{(l)} := b^{(l)} - \alpha \left[\frac{1}{m} \Delta b^{(l)}\right] $$ \hspace{1cm} (10)

Repeat these iterative steps to reduce the value of the cost function $L(W, b)$, until the training of the model is completed.
4. Experiments

Figure 2. A technology roadmap for multi-temporal high-resolution remote sensing image arable land semantic segmentation.

Figure 2 is a technology roadmap for multi-temporal high-resolution remote sensing image arable land semantic segmentation. Because some arable land has different crops cultivated in different seasons or has a fallow period, if only one season of remote sensing images is used to train the neural network, the neural network may not be able to recognize such arable land due to seasonal reasons, and high-resolution remote sensing images can provide more features than low-resolution remote sensing images. Therefore, this paper uses multi-temporal high-resolution remote sensing images to train neural networks. We use high-resolution remote sensing images from different periods to synthesize multi-temporal high-resolution remote sensing images dataset, which lays the foundation for subsequent experiments. After synthesizing the dataset, we first perform data preprocessing operations such as geometric correction, radiometric calibration, and atmospheric correction. There are 50,000 samples in the dataset. We randomly divide 60% of the entire dataset into the training set, 20% into the verification set, 20% into the test set. And there are no overlapping areas between these three parts. The training set is used to train the model, the validation set is used to adjust the parameters of the model to optimize the model, and the test set is used to test the segmentation ability of the trained model. After dividing the dataset, we perform random rotation, mirroring, blurring, and noise enhancement on the training set to ensure the segmentation accuracy and generalization ability of the neural network, and then build FCN, SegNet, and W-net neural networks. The training samples are
input into the neural network for training. After the training is finished, the test set is used to test the segmentation ability of each neural network.

Table 1. The statistical results of the accuracy of the semantic segmentation based on W-net and other neural networks.

| Neural networks | Accuracy |
|-----------------|----------|
| FCN             | 86.22%   |
| SegNet          | 82.81%   |
| W-net           | 90.96%   |

In this paper, the results of manual labeling are used as the reference. In the case of maintaining the same training set, verification set and test set, FCN, SegNet and W-net are used for control-experiments. As the number of iterations increases, the accuracy of W-net is getting higher and higher, the loss value is getting lower and lower, and eventually it tends to be stable, and W-net converges. The statistical results of the accuracy of the semantic segmentation based on W-net and other neural networks are shown in Table 1. As can be seen from the above table, W-net is the highest and SegNet is the lowest in terms of segmentation accuracy.

![Figure 3](image_url)

Figure 3. The partial result of arable land semantic segmentation of W-net and other neural networks. (a)The original image; (b)The result of arable land semantic segmentation of FCN; (c)The result of arable land semantic segmentation of SegNet; (d)The result of arable land semantic segmentation of W-net.

In order to visualize the experimental results, Figure 3 shows the partial results of the semantic segmentation of different neural network models. Through the comparative analysis of the experimental results, it can be found that although the FCN can also segment the arable land, the segmented edges are not flat. The result of arable land semantic segmentation of SegNet is more rough and incomplete, what's more, some edges are not segmented. Compared with the above two neural networks, W-net can segment relatively complete arable land and its result has a relatively flat edge. W-net is downsampled by convolution and pooling to get low-level features. The combination of the low-level features obtained by downsampling and the upsampled inputs can serve to refine the boundary information.

Table 2. The accuracy of arable land semantic segmentation for single-temporal and multi-temporal high-resolution remote sensing images based on W-net.

| Temporal type | Accuracy |
|---------------|----------|


The accuracy of arable land semantic segmentation for single-temporal and multi-temporal high-resolution remote sensing images based on W-net is shown in Table 2. It can be seen that W-net has higher accuracy of arable land semantic segmentation for multi-temporal high-resolution remote sensing images than single-temporal high-resolution remote sensing images.

|                | Single-temporal | Multi-temporal |
|----------------|-----------------|----------------|
| Accuracy (%)   | 85.01%          | 90.96%         |

The experimental results reveal that the performance of arable land semantic segmentation on the edge for single-temporal high-resolution remote sensing images based on W-net is acceptable, but it still cannot accurately identify the arable land, while the performance of arable land semantic segmentation on the edge for multi-temporal high-resolution remote sensing images based on W-net is excellent, and it can accurately identify the arable land.

In general, W-net works best for the arable land semantic segmentation, followed by FCN and SegNet. Moreover, the W-net for multi-temporal high-resolution remote sensing images is better than the W-net for single-temporal high-resolution remote sensing images.

5. Conclusion

Aiming at the problem that manual arable land semantic segmentation is time-consuming, inefficient, segmentation results rely heavily on the staff's experience and traditional methods cannot meet the requirements of land segmentation, this paper first proposes a new neural network W-net to realize multi-temporal high-resolution remote sensing image arable land semantic segmentation. The multi-temporal dataset is constructed by using high-resolution remote sensing images from different periods in Yizheng City, Jiangsu Province. W-net has the ability to process multi-temporal high-resolution remote sensing images. The W-net neural network is trained to automatically extract the characteristics of arable land. After training, W-net neural network is used for arable land semantic segmentation with multi-temporal high-resolution remote sensing images. The experimental results show that method of arable land semantic segmentation for multi-temporal high-resolution remote sensing images based on W-net has the advantages of high accuracy, segmentation edge flatness, good segmentation effect and strong generalization ability. What’s more, segmentation accuracy can reach 90.96%. The experimental results show that method of arable land semantic segmentation for multi-temporal high-resolution remote sensing images based on W-net has the advantages of high accuracy, segmentation edge flatness, good segmentation effect and strong generalization ability. What’s more, the method can quickly extract the characteristics of arable land from a large amount of data and automatically segment the arable land. It is efficient and feasible. It can effectively realize the arable land semantic segmentation. In addition, this method can be implemented under a small dataset, which is especially important to expensive high-resolution remote sensing images. The W-net neural network is also an end-to-end image segmentation method that allows the network to make pixel-level predictions and directly derive predicted images. In addition to providing reliable technical support for agricultural insurance, the method of this paper combines W-
net neural network with multi-temporal high-resolution remote sensing images and improve the value of multi-temporal high-resolution remote sensing images in practical applications. W-net neural network has strong learning ability. In addition to the arable land semantic segmentation, it can also be used for road semantic segmentation, river semantic segmentation, etc.

However, we can also see that the results of W-net neural network segmentation of arable land are still flawed, and some of the very close arable land edges will be mistakenly merged together. Therefore, the next step will be to solve this problem.

Acknowledgments
This work was supported by National Natural Science Foundation of China (31671586).

References
[1] Chen X J, Dan L. (2010) Medical Image Segmentation Based on Threshold SVM. In: International Conference on Biomedical Engineering & Computer Science. Wuhan, pp. 1-3.
[2] Pan Y, Liu Q, Lu Z, et al. (2009) Regional arable land consolidation based on agricultural land classification and gradation. Transactions of the Chinese Society of Agricultural Engineering, 25: 260-266.
[3] Nobis M, Hunziker U. (2005) Automatic thresholding for hemispherical canopy-photographs based on edge detection. Agricultural & Forest Meteorology, 128(3): 243-250.
[4] Liu M Y, Tuzel O, Ramalingam S, et al. (2011) Entropy rate superpixel segmentation In: CVPR. Colorado, pp. 2097-2104.
[5] Hinton g, Salakhutdinov R.. (2006) Reducing the dimensionality of data with neural networks. Science, 313(5786): 504-507.
[6] YANN LeCun, YOSHUA Bengio, GEOFFREY Hinton. (2015) Deep learning. Nature, 521(7663): 436-444.
[7] Long J , Shelhamer E , Darrell T. (2014) Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(4): 640-651.
[8] FU G,LIU C J,ZHOU R,et al. (2017) Classification for high resolution remote sensing using a fully convolutional neural network. Remote Sensing, 9(498): 1-21.
[9] Badrinarayanan V, Kendall A, Cipolla R. (2017) Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(12): 2481-2495.