Journal of Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Segmentation model based on convolutional neural networks for extracting vegetation from Gaofen-2 images

Chengming Zhang
Jiping Liu
Fan Yu
Shujing Wan
Yingjuan Han
Jing Wang
Gang Wang

SPIE.

Chengming Zhang, Jiping Liu, Fan Yu, Shujing Wan, Yingjuan Han, Jing Wang, Gang Wang, "Segmentation model based on convolutional neural networks for extracting vegetation from Gaofen-2 images," *J. Appl. Remote Sens.* **12**(4), 042804 (2018), doi: 10.1117/1.JRS.12.042804.

Segmentation model based on convolutional neural networks for extracting vegetation from Gaofen-2 images

Chengming Zhang,[a,b,c,*] Jiping Liu,[b] Fan Yu,[b] Shujing Wan,[d] Yingjuan Han,[e] Jing Wang,[e] and Gang Wang[a]

[a]Shandong Agricultural University, College of Information Science and Engineering, Tai'an, China

[b]Chinese Academy of Surveying and Mapping, Beijing, China

[c]Shandong Technology and Engineering Center for Digital Agriculture, Tai'an, China

[d]QuFu Normal University, Network Information Center, Qufu, China

[e]Key Laboratory for Meteorological Disaster Monitoring and Early Warning and Risk Management of Characteristic Agriculture in in Arid Regions, Yinchuan, China

Abstract. Convolutional neural network (CNN) models achieve state-of-the-art performance for natural image semantic segmentation. An approach for extracting vegetation from Gaofen-2 (GF-2) remote sensing imagery based on the CNN model is presented. We constructed a convolutional encoder neural networks (CENN) consisting of two layers. The first layer has two sets of convolutional kernels for extracting the features of farmland and woodland, respectively. The second layer consists of two encoders that use nonlinear functions to encode the learned features and map the encoding results to the corresponding category number. In the training stage, samples of farmland, woodland, and other lands are categorically used to train the CENN. After training is accomplished, the CENN would acquire enough ability to accurately extract farmland and woodland from GF-2 imagery. The CENN was trained on 36 GF-2 images and tested on three other GF-2 images. We compared the proposed method to a deep belief network, a fully convolutional network, and a DeepLab model using the same images. The experiments demonstrate that the proposed approach improves upon the accuracy of existing approaches. The average precision, recall, and kappa coefficient of the proposed approach were 0.91, 0.87, and 0.86, respectively. Thus, the proposed approach is proven to effectively extract vegetation from GF-2 imagery. © *The Authors. Published by SPIE under a Creative Commons Attribution 3.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.* [DOI: 10.1117/1.JRS.12.042804]

Keywords: convolutional neural network; Gaofen-2 remote sensing imagery; remote sensing image segmentation; convolutional encoder neural network; categorical learning; vegetation extraction.

Paper 180223SS received Mar. 17, 2018; accepted for publication Jul. 10, 2018; published online Aug. 2, 2018.

1 Introduction

Image segmentation is the precondition and foundation for the extraction and target identification of high-resolution remote sensing images.[1] In a high-resolution image, spectral confusion is more serious, differentiation is substantially reduced, and the accuracy of the spectral statistics-based segmentation method is reduced.[2,3] Object-oriented image segmentation method can overcome the influence of "salt and pepper" noise and improve accuracy by using object structure and spectral signature. Because this approach must adjust the segmentation scale to obtain an acceptable image segmentation result and it is difficult to determine

*Address all correspondence to: Chengming Zhang, E-mail: chming@sdau.edu.cn

a suitable segmentation scale, development of the object-oriented segmentation method has been slow.[4,5]

With the development of machine learning technology, researchers began to apply algorithms, such as neural networks (NNs)[6,7] and support vector machines (SVM),[8,9] to the segmentation of high-resolution images.[3,10,11] Studies have revealed that image segmentation based on machine learning algorithms can obtain more optimal results compared to traditional statistical and object-oriented methods.[12,13] Both SVMs and NNs are shallow learning algorithms,[14–16] owing to their limited network structures. Shallow learning algorithms have difficulty expressing complex functions effectively. So, when sample size and diversity are increased, shallow learning models cannot adapt to the increasing complexity. [17,18]

Advancements in deep learning enable us to address these problems with deep neural networks (DNNs).[19–22] As one of the most important branches of deep learning, the convolutional neural network (CNN) is commonly applied to image data owing to its superior feature learning ability.[23–25] The CNN is a deep learning network composed of multiple, nonlinear mapping layers with strong learning abilities that obtain excellent results in image segmentation.[26,27] Traditional deep learning methods include deep convolutional neural networks (DCNN)[18,28] and deep deconvolutional neural networks (DeCNN).[29] Since then, many methods of remote sensing image segmentation based on CNN have been developed.[30,31] Many large CNNs with performance that can be scaled depending on the size of training data, model complexity, and processing power have achieved meaningful improvements in the object segmentation of images.[32–39]

A fully convolutional network (FCN) is a deep learning network for image segmentation originally proposed in 2015.[39] Leveraging the advantages of convolutional computation in feature organization and extraction, an FCN establishes a multilayer convolution structure and reasonable sets deconvolution layer to realize pixel-by pixel segmentation.[40–42] Researchers have since developed a series of segmentation models based on convolution, including segNET,[43] UNet,[44] DeepLab,[45] multiscale FCN,[46] and reSeg.[47] Each of these segmentation models has its own strengths and works well with certain selected types of images.

Segmentation models such as FCN are effective because the multilayer structure of these models adeptly handles the rich detail features of images. However, in regions of vegetation in Gaofen-2 (GF-2) imagery, one pixel usually contains several different types of plants or crops. Thus, the information between pixels does not reflect this variety and the image texture is smoother. Although a single tree is larger than most crop plants, a typically sized tree occupies two or three pixels in the GF-2 imagery and smaller trees may only occupy one pixel or less, continuing the problem of little information difference between pixels. Therefore, because the detail features of a vegetation region may be lacking, the effect of deep layer CNN may also be weak and may even pull in greater noise, resulting in poor segmentation accuracy. To obtain accurate segmentation results from GF-2 images, the size of the coverage area of individual plants relative to the spatial resolution of the GF-2 imagery must be considered when designing a CNN.

Based on the above analysis, we constructed convolutional encode neural networks (CENN) to accurately distinguish and extract farmland and woodland from GF-2 imagery. Because the CENN considers the previously described complexities presented by the features of vegetation regions in the GF-2 imagery, such as the smaller coverage area size of a single plant, fewer detail features, and the continuous appearance of vegetation, our approach achieves improved accuracy compared to the existing approaches.

The following summarizes the proposed method for the segmentation of vegetation in GF-2 images and its evaluation.

(1) A network structure consisting of a convolution layer and an encoder layer has been designed based on the features specific to vegetation regions in the GF-2 imagery. The convolution layer is used to extract the features of farmland and woodland respectively, the encoder layer is used to encode the learned features and map the encoded results to the corresponding category number.
(2) In the model training stage, a categorical training method was adopted. To train the CENN and obtain a set of convolutional kernels with which to identify farmland, we use an image sample of farmland as the positive sample and employ other images as the negative sample. Then, to train the CENN to identify woodland, we repeated the

previous process, using a sample of forestland as the positive sample and others as the negative sample. After training, the CENN can accurately extract farmland and woodland from GF-2 images.

(3) Finally, the trained CENN were used to segment GF-2 images. The accuracy of the segmentation results was then evaluated as per a comparative experiment conducted using three existing segmentation models.

2 Methods

In accordance with established conventions for image segmentation using CNNs, we divide the work into two stages, training and classification, as depicted in Fig. 1. The upper part of Fig. 1 shows the training stage. Together, GF-2 imagery and corresponding pixel-by-pixel artificial classification labels are input to the CENN as training samples. The error between the predicted classification labels and the artificial classification labels is calculated and backpropagated through the network using the chain rule. Then the parameters of the CENN are updated using the gradient descent method. The above process iterates until the error is less than a predetermined threshold. The lower part of Fig. 1 details the classification stage, in which the trained CENN accurately extracts vegetation from input GF-2 imagery.

2.1 *Network Architecture*

The CENN model is divided into four functional groups of components, input, convolution layer, encoder layer, and output, as shown in Fig. 2. In the training stage, the inputs are original images and artificial classification labels. In the classification stage, the inputs are the original GF-2 images, the output is a single-band file, and the content of each pixel in the output is the category number of the corresponding original image pixel. The CENN indicate farmland using category number 100, woodland is denoted by category number 150, and category number 200 distinguishes other land use.

2.1.1 *Convolution layer*

In Sec. 1, we analyzed the characteristics of vegetation areas in GF-2 imagery. Based on this analysis, we constructed a convolution structure in the convolution layer known as "width convolution." With this convolution structure, we can extract more features to improve segmentation accuracy.

The convolutional kernels of the CENN are in $r * c * h$ form, where r denotes the width of the convolutional kernel, c denotes the height of the convolutional kernel, and h denotes the number of channels of the convolutional kernel. In this paper, h is set to 3 because only three channels of GF-2 imagery are employed. As shown in items (b) and (c) of Fig. 2, we

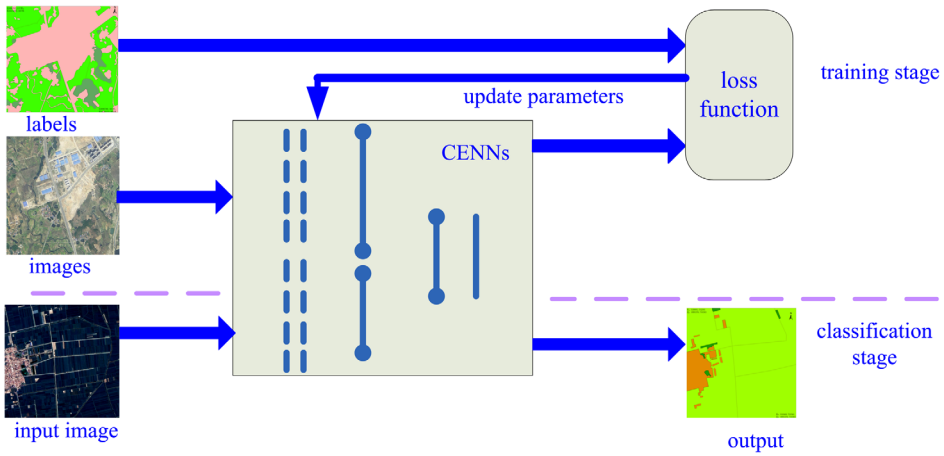

Fig. 1 The training and classification stages of the proposed approach.

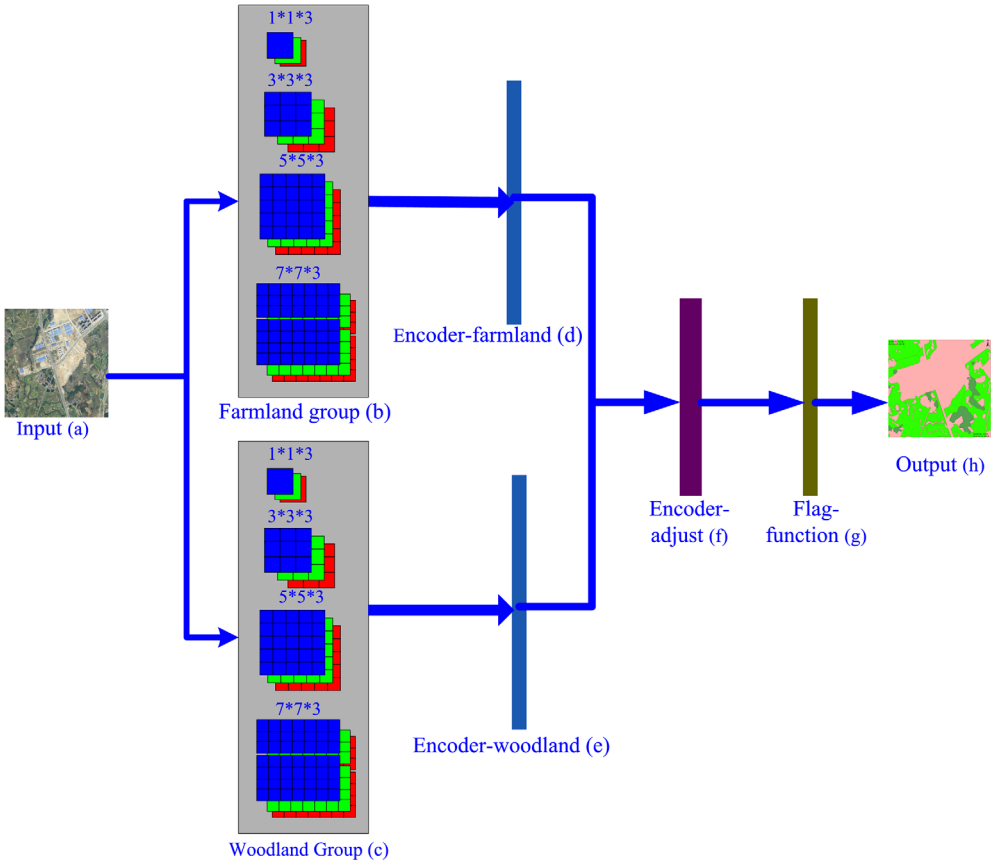

Fig. 2 The network architecture of CENNs.

adopted four types of convolutional kernel in the CENN, which are referred to as A-type, B-type, C-type, and D-type, respectively; r and c of A-type are set to 1, r and c of B-type are set to 3, r and c of C-type are set to 5, r and c of D-type are set to 7. This structure earns the name "width convolution" owing to how all the convolutional kernels use the GF-2 images as direct inputs and process the images parallel to each other.

The primary function of the A-type convolutional kernels is to extract the color features of each pixel. The B-type convolutional kernels are divided into two groups. The first group of convolutional kernels must be trained. They are used to extract the texture features of the central pixels and the surrounding eight pixels. The second group consists of eight convolutional kernels. The values of these convolutional kernels are fixed, and no further training is required. These eight convolutional kernels are used to calculate the absolute value of the color difference between the central pixel and eight adjacent pixels surrounding the central pixel. The roles of the C-type and D-type convolutional kernels are similar to that of the B-type convolutional kernels, and these two types are also composed of two groups. However, the C-type and D-type convolutional kernels have a wider range with which to exploit the features between center pixels and their surrounding pixels.

As shown in Fig. 2, all the convolutional kernels were divided into two groups, a farmland group (b) and a woodland group (c). Both groups contain A-type, B-type, C-type, and D-type convolutional kernels. The purpose of this two-group design is to enable the CENN to better express the characteristics of vegetation, and thus improve their capacity for distinction of different vegetation types.

2.1.2 *Encoder layer*

As shown in Fig. 2, the encoder layer contains two sublayers of encoders to better simulate the nonlinear relationship between features and outputs. There are two encoders in the first encoder

sublayer, the encoder-farmland (d) and the encoder-woodland (e), as observed in Fig. 2. The role of each encoder is to regress and simulate the characteristics of their respective land uses. The second encoder sublayer has a single encoder, the encoder-adjust (f), as depicted in Fig. 2. The role of the encoder-adjust (f) is to adjust the calculations of the upper sublayer so that farmland, woodland, and other lands are distinguished from each other. Then, the flag-function (g) maps the encoded result of encoder-adjust (f) to the appropriate category number.

2.2 *Network Training*

2.2.1 *Label sample*

The training dataset comprised a total of 39 GF-2 remote sensing images (size 7300×6900 pixels) of Shandong Province, China. Of these images, 21 images were captured on February 17, 2016, and 18 images were captured on May 12, 2017. The spatial resolution of the panchromatic band was 1 m. The spatial resolution of the multispectral was 4 m. Environment for visualizing images (ENVI) software was used for preprocessing tasks, such as fusion and color stretching. We then selected the 321 band as RGB band to improve visual effects.

Artificial label samples are an important training foundation. Because the CENN use pixels as the primary learning object, they must be accurately labeled. We used ENVI software for labeling and designed a preprocessor to build the mask. The process of artificial labeling is as follows:

(1) Use the region-of-interest (RoI) tool in the ENVI software to select farmland regions, woodland regions, and other regions in the image. Then, the map locations of the pixels in each region are output to different files according to category.
(2) The preprocessor added a band to the image file as the mask band. The spatial resolution, size, and other parameters of the mask band were the same as the original image. Then, the category number of each pixel is written to the mask band according to the map location of the pixel previously output.

We manually labeled all images at the pixel level. Thus, for each image, there exists a 7300×6900 label map, with a row–column indexed pixel-class correspondence. We used 36 images for training and the remaining three images for testing. Figure 3 shows an example of one image-label pair.

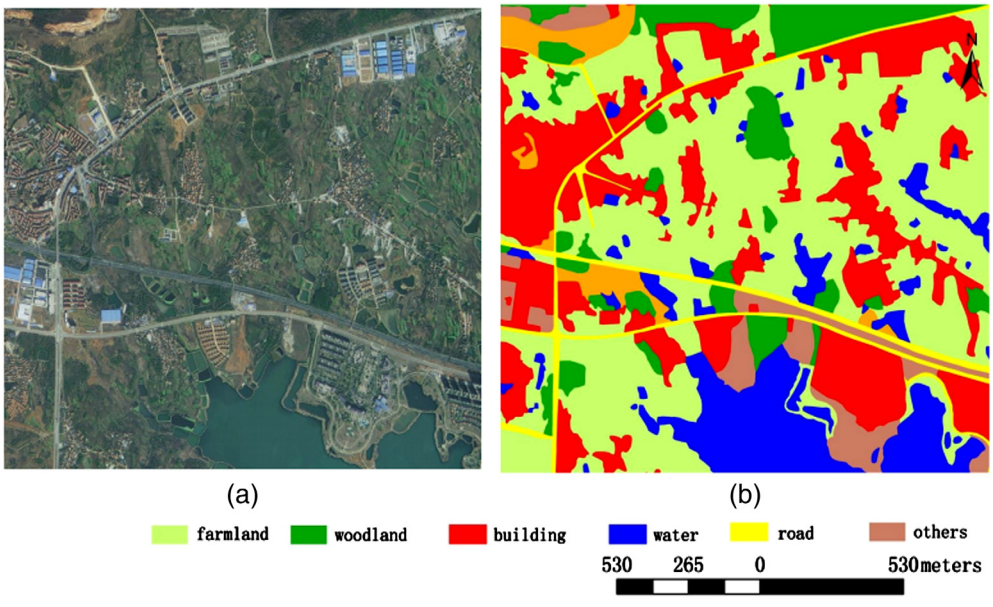

Fig. 3 Image-label pair example: (a) original image and (b) labels.

2.2.2 *Model training*

Images from two different time periods were selected as training data. We select images from different periods to increase the anti-interference abilities of the CENN, mitigating complications, such as the change of seasons, and thus enhancing applicability. The training stage proceeded through the following steps:

(1) Image-label pairs were input into the CENN as training samples.
(2) The cross-entropy loss was calculated and backpropagated.
(3) The network parameters were updated using stochastic gradient descent (SGD)[41,48] with momentum.

Training obtained farmland group convolutional kernels and woodland group convolutional kernels. Each group enhanced the farmland or woodland features, whereas features of all other types were suppressed as much as possible. For example, in the woodland group, woodland features were enhanced while features of all other types were suppressed.

In our training, the SGD method with momentum was used for parameter updates, and the following expression illustrates the SGD[41,48] method with momentum:

$$W^{(n+1)} = W^{(n)} - \Delta W^{(n+1)}, \tag{1}$$

where $W^{(n)}$ denotes the old parameters, $W^{(n+1)}$ denotes the new parameters, and $\Delta W^{(n+1)}$ is the increment for the current iteration. The iteration increment, which is a combination of old parameters, gradient, and historical increment, is calculated as shown below:

$$\Delta W^{(n+1)} = \left[d_w \cdot W^{(n)} + \frac{\partial J(W)}{\partial W^{(n)}} \right] + m \cdot \Delta W^{(n)}, \tag{2}$$

where, $J(W)$ is the loss function, ϑ is the learning rate for step length control, d_w denotes the weight decay, and m denotes the momentum.

2.3 *Segment Using the Trained Network*

After successful training, the CENN can be used to segment the input imagery pixels-by-pixel. According to our design, the output is written into a new band. The benefit of this design is that in saving the segmentation result to a new band, it avoids damaging the original file.

3 Experiments

We designed a set of test experiments and comparative experiments to verify the feasibility of the proposed CENN method. The proposed approach was implemented using Python 2.7 on a Linux Ubuntu 16.04 operating system using an NVIDIA GeForce Titan X Graphics device with 12 GB graphic memory.

The data and classification criteria used were described in Sec. 2.2.1.

3.1 *Learning Ability Indicators*

The primary functions of the CENN are reflected in their feature extraction and encoding abilities. The concentration degree of feature values was used as an index to examine feature extraction capabilities, using distinctions between farmland, woodland, and other lands as an index to examine encoding capacity.

3.2 *Comparison Model*

We chose the deep belief network (DBN) model, the FCN model, and the DeepLab model as the comparison models. A comparative experiment was conducted using methods established in published literature.

3.2.1 *Deep belief network*

The paper of Dawei et al.[2] presented a method of pixel-by-pixel classification for high-resolution images using DBN. Their method calculated the texture features of an image through non-subsampled contourlet transform and used the DBN to classify high-resolution remote sensing images based on spectral-texture features. The training process included two subprocesses: pretraining and fine-tuning. Pretraining was performed in an unsupervised manner. A greedy algorithm was used to perform layer-by-level optimization during training, and the parameters of each restricted Boltzmann machine (RBM) were adjusted individually. After training the upper layer, the output is used as the input to train the RBM of next layer. After completing the pretraining, the last-level backpropagation network was trained in a supervised learning manner. The error was propagated backward through the layers and the weight of the entire DBN network was fine-tuned.

3.2.2 *Fully convolutional network*

For the FCN model, we directly employed the FCN-8s model proposed by Long et al.[40] The architecture of the model was derived from the VGG-16 network. After the upsampling operation, the final prediction was fused from the output of three branches—the primary network, the pool4 layer, and the pool3 layer. In the training phase, the input data and the training parameters for FCN-8s in the comparative experiment were the same as those used to train the proposed model. The testing stage also used the same classification parameters as applied in the proposed approach.

3.2.3 *DeepLab*

For the DeepLab model, we directly employ the DeepLab v3 model proposed by Chen et al.[45] DeepLab was also developed based on the VGG network. Unlike the FCN model, to ensure that the output size would not be not too small without excessive padding, DeepLab changed the stride of the pool4 and pool5 layers of the VGG network from the original 2 to 1, plus 1 padding. To compensate for the influence of the stride change on the receptive field, DeepLab used a convolution method called "Atrous convolution" to ensure that the receptive field after pooling remains unchanged and the output is more refined. Finally, DeepLab incorporated a fully connected conditional random fields (CRF) model to refine the segmentation boundary.

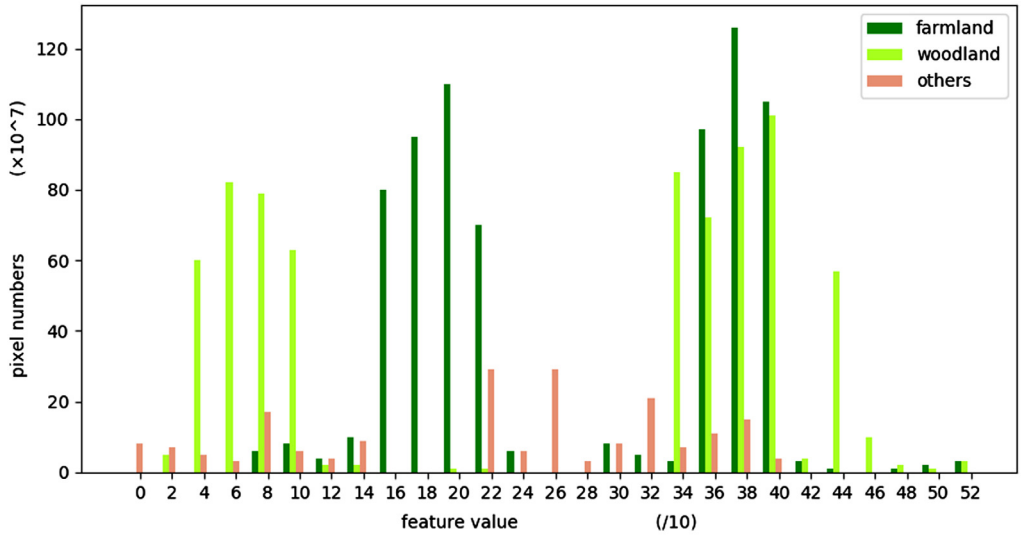

Fig. 4 Learning result of A-type convolutional kernels.

3.3 *Results and Comparison*

3.3.1 *Learning ability of the CENN*

Figure 4 graphs the distribution of the feature values learned from the farmland samples, woodland samples, and other samples using A-type convolutional kernels. Figure 4 reveals that after the convolution operation, the feature values of farmland and woodland are concentrated in two regions while the other land use type is scattered. This is mainly because the different seasons in which the data were collected results in substantial differences in color values. The features' concentration of woodland is less than that of farmland, mainly because the seasonal color change of woodland is more dramatic than that of farmland.

Figure 5 shows the feature values learned from the farmland samples, woodland samples, and other samples using B-type convolutional kernels. As shown in Fig. 5, because the B-type convolutional kernels are primarily used to learn the color difference between adjacent pixels, the features of farmland samples have a better concentration degree, and the dispersion level of

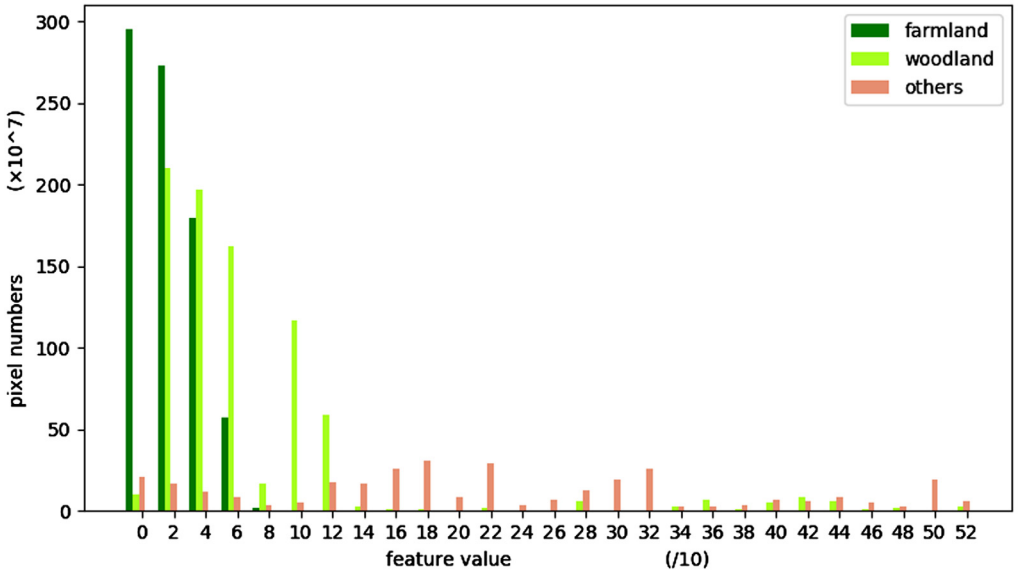

Fig. 5 Learning result of B-type convolutional kernels.

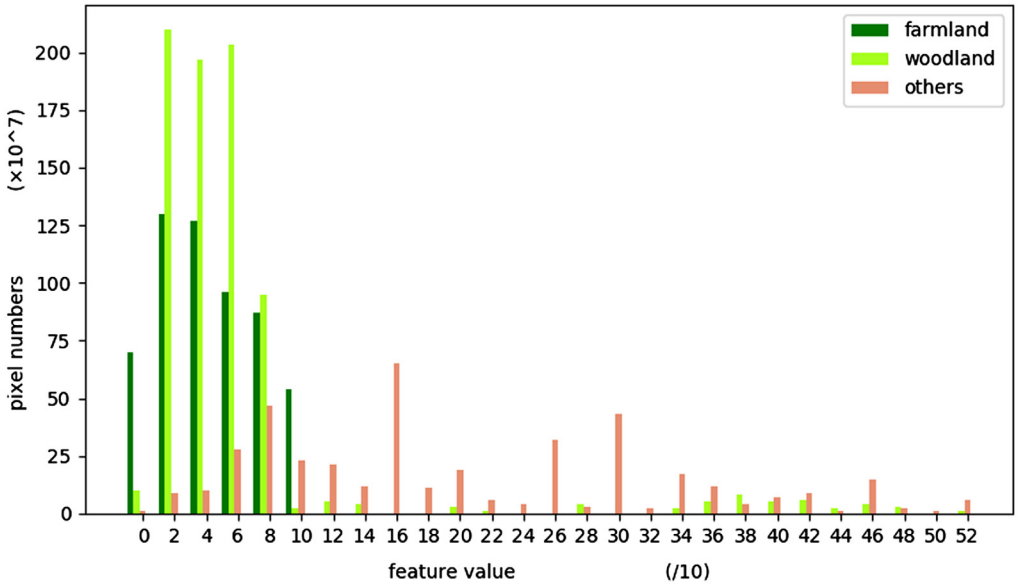

Fig. 6 Learning result of C-type convolutional kernels.

woodland sample is substantially larger. The results reflect the smooth texture of farmland and the rough texture of woodland.

Additionally, for farmland samples, the features' concentration of C-type convolutional kernels and D-type convolutional kernels is both less than that of B-type convolutional kernels. This is because as the field of view expands, more other-type pixels are introduced at the boundary. For woodland samples, the C-type convolutional kernels achieve the best concentration of features. Figure 6 shows the learning results of C-type convolutional kernels.

Based on the learning results, we observe that although multispace convolutional kernels are more suitable for the extraction of farmland and forestland features than deep convolution, it remains necessary to combine multiple features to accurately determine the category to which each pixel belongs.

Figure 7 shows the encoding result of the first encoder sublayer, and Fig. 8 shows the encoding result adjusted by the second encoder sublayer. As demonstrated by the figure, the adjusted encoding result could already be used to segment image.

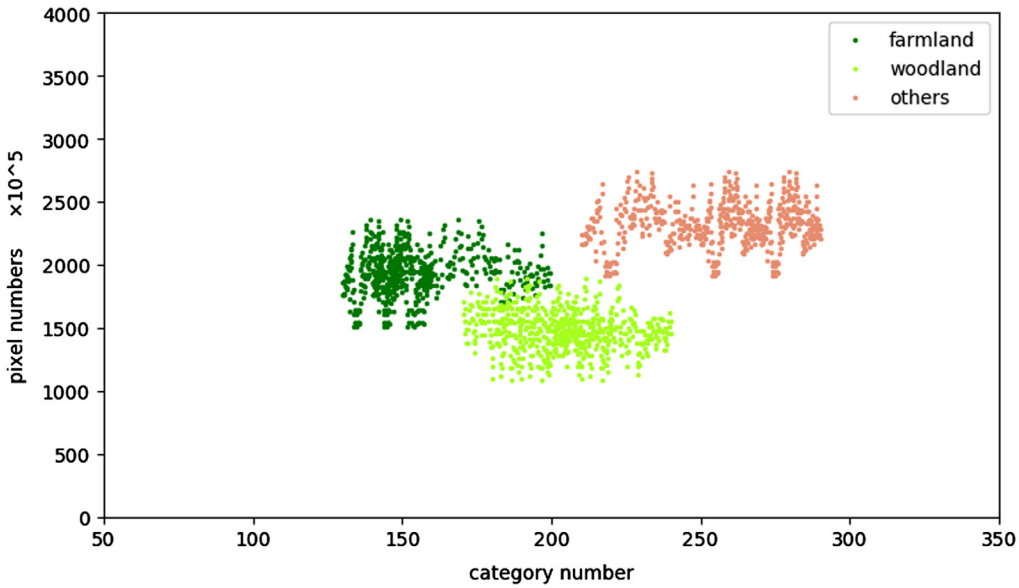

Fig. 7 Encoding result of the first encoder sublayer.

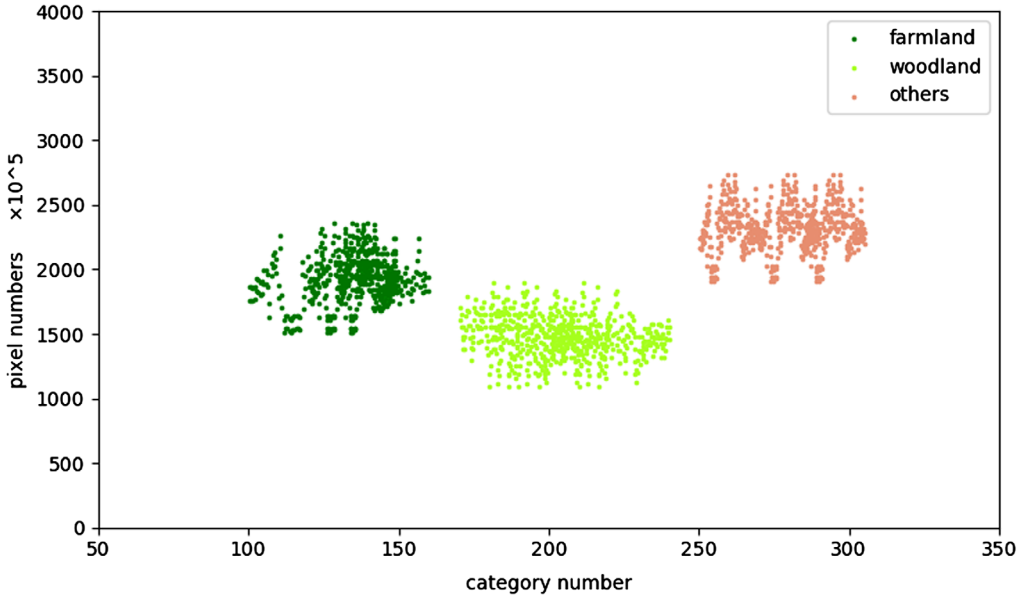

Fig. 8 Encoding result of the second encoder sublayer.

3.3.2 *Experiment result comparison*

In the comparative experiment, we apply our trained model to three GF-2 images for segmentation. All images were equally sized at 7300×6900 pixels. These images were only used for

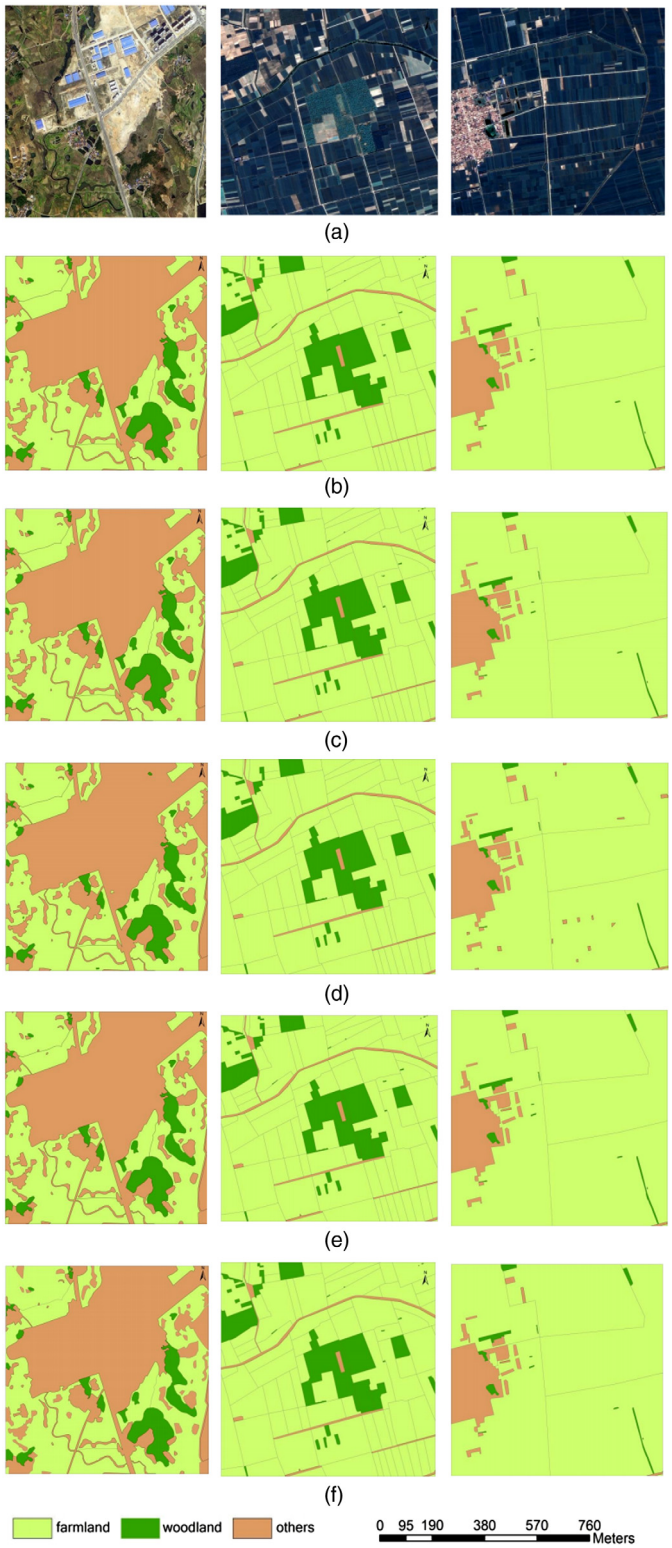

Fig. 9 Segmentation results on GF-2 images: (a) original images, (b) ground truth, (c) our results corresponding to the images in (a), (d) results of DBN, (e) results of FCN, and (f) results of DeepLab.

Table 1 Confusion matrix C of our approach for Fig. 9.

Experiment	GT/predicted	Farmland	Woodland	Others
Experiment-1	Farmland	0.93	0.06	0.04
	Woodland	0.07	0.91	0.07
	Others	0.05	0.03	0.89
Experiment-2	Farmland	0.92	0.05	0.03
	Woodland	0.04	0.93	0.06
	Others	0.05	0.04	0.91
Experiment-3	Farmland	0.91	0.06	0.03
	Woodland	0.03	0.92	0.03
	Others	0.04	0.05	0.90

testing and were not involved in training. Figure 9 illustrates the results obtained from the comparison methods and the proposed method.

Table 1 lists the confusion matrix C of our classification results. From the table, we can see that our approach achieves higher classification performance. In the above example, our recall for farmland and woodland is 0.855. The average proportions of "farmland" wrongly classified as "woodland" and "others" are 0.053 and 0.033, respectively. The average proportions of "woodland" wrongly classified as "farmland" and "others" are 0.047 and 0.053, respectively.

We employ precision, recall, and kappa coefficient as indicators to evaluate our approach. These indexes are calculated from confusion matrix C. Precision denotes the average proportion of pixels correctly classified to one class from the total retrieved pixels. Precision is calculated as follows:

$$\text{Precision} = \frac{1}{3}\sum_{i} \mathrm{C}_{ii} \Big/ \sum_{j} \mathrm{C}_{ij}, \tag{3}$$

Table 2 Comparison between approaches using DBN, FCN, DeepLab, and CENN.

Approach	Index	Experiment-1	Experiment-2	Experiment-3
DBN	Precision	0.69	0.74	0.76
	Recall	0.61	0.63	0.62
	Kappa	0.58	0.69	0.71
FCN	Precision	0.79	0.81	0.76
	Recall	0.72	0.75	0.69
	Kappa	0.71	0.73	0.64
DeepLab	Precision	0.84	0.86	0.79
	Recall	0.77	0.79	0.73
	Kappa	0.79	0.81	0.77
CENN	Precision	0.89	0.91	0.92
	Recall	0.85	0.86	0.89
	Kappa	0.84	0.86	0.88

where C_{ii} denotes the number of pixels of category i that are correctly classified, and C_{ij} denotes the number of pixels of category j that misclassified into category i.

Recall represents the average proportion of pixels that are correctly classified in relation to the actual total pixels of a given class. Recall is computed as follows:

$$\text{Recall} = \frac{1}{3}\sum_i C_{ii} \Big/ \sum_i C_{ij}. \tag{4}$$

The kappa coefficient measures the consistency of the predicted classes with the artificial labels. The indicator values are listed in Table 2.

4 Discussion

This paper presents a classification approach, which extracts vegetation from GF-2 images using the CENN. Compared with the three typical deep learning-based approaches, the proposed method substantially improves classification accuracy. In the following sections, we discuss the reasons for the improvement and the benefits of using the proposed approach to classify land use.

4.1 *Deep Belief Network versus the Proposed Approach*

In the method of pixel-by-pixel segmentation based on DBN, the texture features of an image are first calculated. The obtained two-dimensional texture features are converted into one-dimensional vectors. Then, three channel values of RGB are added to the vectors, and they are merged into a single vector. Finally, the DBN network is constructed using each component value of the vector as an independent input to classify the pixels. Although texture features are completely different from spectral values, this method uses them in combination, resulting in logical confusion. If only the value of the texture feature is used, it cannot indicate spatial relations represented by texture, which results in information loss in the texture extraction. Therefore, like traditional spectral-based methods, this method utilizes only the spectral characteristics of the pixel itself and effectively ignores the spatial relationship between pixels, which makes it easy to generate the incorrect classifications.

Unlike the DBN method, the CENN model makes full use of the advantages of convolution in information aggregation and uses A-type convolutional kernels to extract the common features of the original spectral value. Three kinds of convolutional kernels: B-type, C-type, and D-type are used to extract textural features in three sizes. The CENN use the two-stage encoder to simulate nonlinear equations and encode the features. These strategies effectively improve the classification accuracy.

In GF-2 images, it is easy to confuse tall crops and small, dense trees owing to the relatively small differences in texture and spectrum. As observed in Fig. 10, the proposed method is substantially more effective than the DBN method.

4.2 *Fully Convolutional Network versus the Proposed Approach*

The advantage of the FCN model is that it maximizes the information available in the rich details of GF-2 imagery using deep convolution. This advantage is obvious when extracting a target object that covers many pixels, but if the target object covers fewer pixels, even pixels that contain several target objects, the effect of deep layer convolution drastically weakens, and even greater noise may be pulled in, resulting in poor segmentation effect. When FCN is used to extract farmland and woodland from GF-2 images, although the farmland or woodland may cover many pixels, the advantages of FCN cannot be exploited because of the small differences between pixels.

Unlike the FCN model, which expands the view through deep convolution, the proposed CENN expand the view by using three convolutional kernels of B-type, C-type, and D-type.

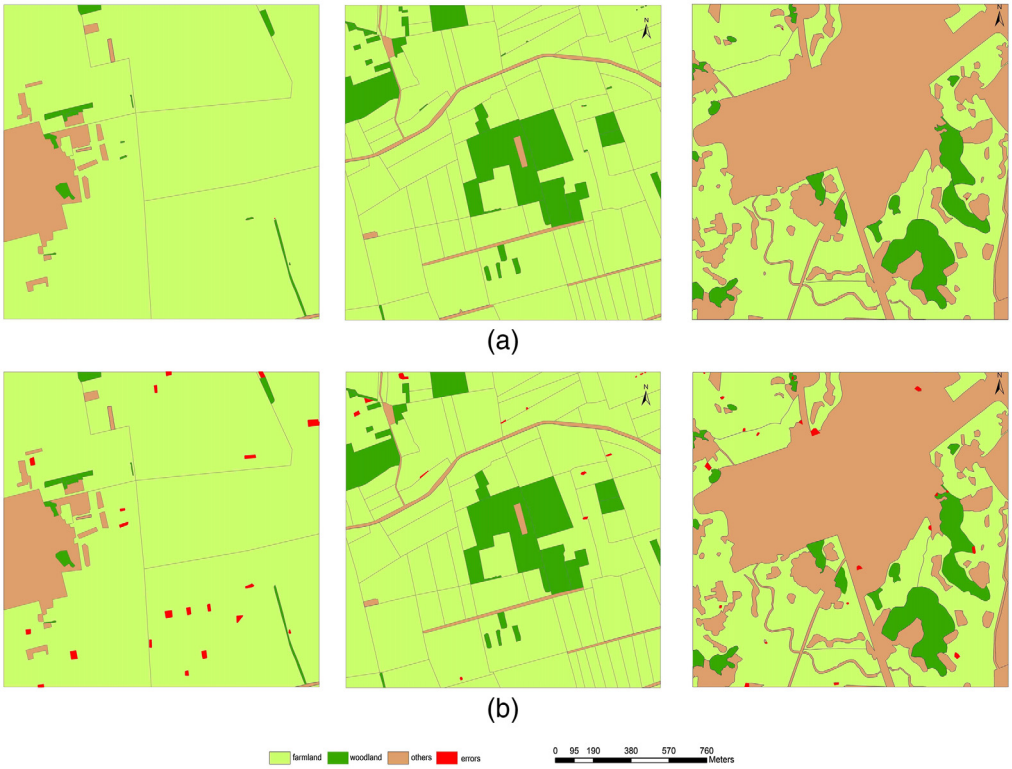

Fig. 10 Segmentation errors: (a) errors of ours and (b) errors of DBN.

By combining three convolutional kernels, the maximum observed area size is about 49 m^2, which is fully capable of covering most of the canopy. The CENN not only make full use of the pixels' own features but also fully exploit the spatial features between pixels, adeptly accounting for the continuous appearances of crops and trees. Additionally, the CENN also fully consider the natural characteristics of farmland and woodland, providing further advantages in identifying corner pixels.

Figure 11 demonstrates that the FCN model and the CENN have the same segmentation accuracy when identifying the interior regions of farmland and woodland. However, when identifying the pixels in the corner region, the FCN model had many errors, whereas the CENN has almost no errors.

4.3 *DeepLab versus the Proposed Approach*

Compared to the FCN model, there are two important improvements in the DeepLab model: (1) the deconvolution part is improved and (2) the network uses the fully connected CRF model to refine the segmentation boundary. According to existing literature, when identifying large objects such as buildings, the segmentation accuracy at the boundary of DeepLab is better than that of the FCN. This is because DeepLab makes better use of the details of the image and spatial correlation of pixels. However, when DeepLab was used to identify woodland and farmland, because the information between pixels does not noticeably change and the image texture is smoother, the advantages of DeepLab are lost.

As explained in Sec. 4.2, the CENN not only make full use of the pixels' own features but also make fully exploit the spatial features between conjoint pixels and account for the continuous appearance of crops and trees. Therefore, the CENN can effectively avoid the defects of the DeepLab approach and ensure the segmentation accuracy. As shown in Fig. 12, the segmentation accuracy of the CENN at the boundary and corners is much better than that of DeepLab approach.

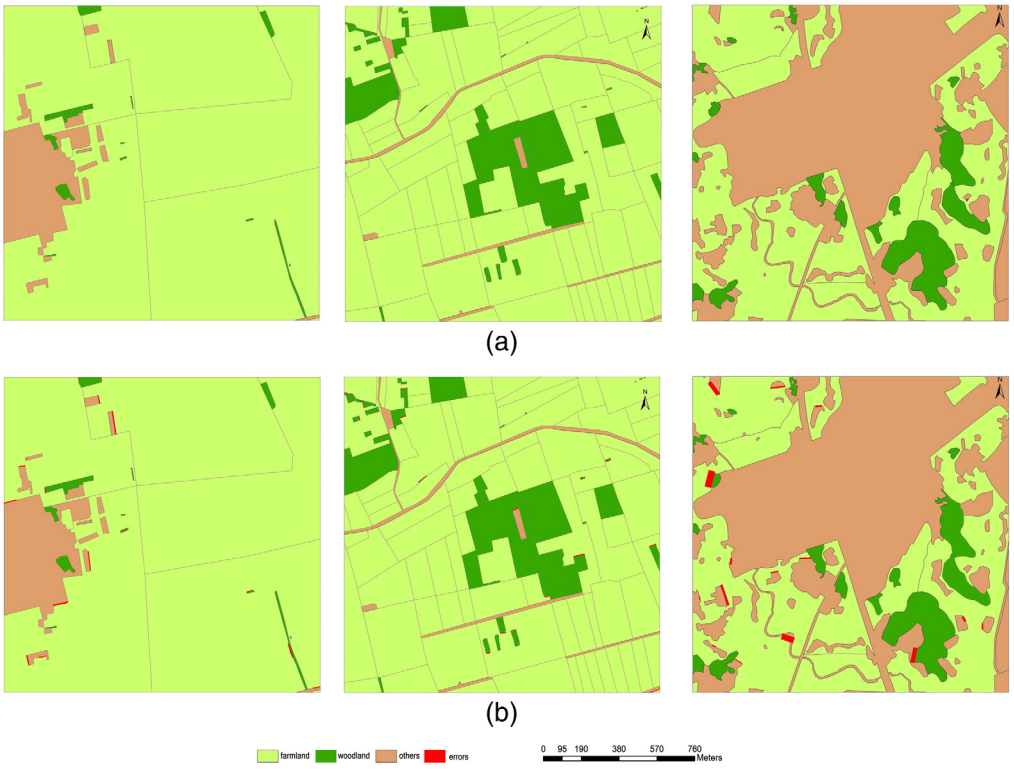

Fig. 11 Segmentation errors: (a) errors of ours and (b) errors of FCN.

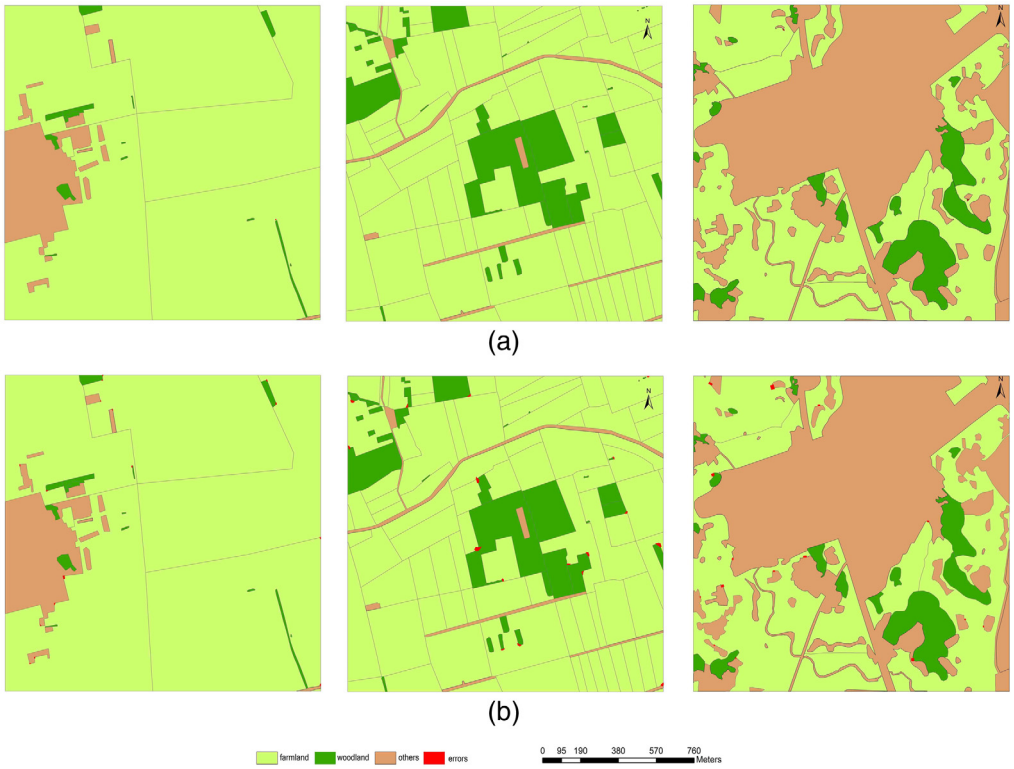

Fig. 12 Segmentation errors: (a) errors of ours and (b) errors of DeepLab.

4.4 *Benefits of Using the Proposed Approach to Classify Land Use*

Accurate land use classification results play an important role in scientific research and agricultural production. The use of remote sensing data to produce land use classification results is becoming more common. In the GF-2 image, it is difficult to accurately distinguish between forest land and farmland using traditional methods. Because our approach can better solve this problem, it has played an important role in agricultural surveys and improved the efficiency of agricultural surveys. Our approach has been applied in the Meteorology Bureau of Shandong Province, China.

5 Conclusions

This paper presents a CENN model that extracts vegetation in farmlands and woodlands from GF-2 images. Compared to the traditional DBN model, FCN model, and DeepLab model, the proposed CENN fully consider the characteristics of farmland and woodland in the GF-2 images. According to the characteristics of the model, categorical training was implemented to enable the model to effectively discriminate farmland and woodland from other land types and extract vegetation from GF-2 images with high accuracy. The paper also provides a software-based method of using RoI for sample annotation, which can reduce the manual workload and enhance the efficiency of marking.

The main limitation of our approach is that the accuracy of the extraction results is greatly reduced when the CENN is applied to submeter level images, which results in a limited application scope of the model. In the following work, we will try to use multilayer convolutions, to further enhance its applicability.

Acknowledgments

All the works were supported by Science Foundation of Shandong (Grant Nos. ZR2017MD018 and ZR2016DP01), the National Science Foundation of China (Grant Nos. 41471299 and 41671440), open research project of key laboratory on meteorological disaster monitoring, early warning, and risk management in characteristic agricultural areas of arid area (Grant No. CAMF-201701).

References

1. D. Zhipeng, W. Mi, and L. I. Deren, "A high resolution remote sensing image segmentation method by combining superpixels with minimun spanning tree," *Acta Geod. Cartographica Sin.* **46**(6), 734–742 (2017).
2. L. Dawei, H. Ling, and H. Xiaoyong, "High spatial resolution remote sensing image classification based on deep learning," *Acta Opt. Sin.* **36**(4), 0428001 (2016).
3. C. Liu et al., "Fusion of pixel-based and multi-scale region-based features for the classification of high-resolution remote sensing image," *J. Remote Sens.* **19**(2), 228–239 (2015).
4. J. Jin, Z. Zou, and C. Tao, "Compressed texton based high resolution remote sensing image classification," *Acta Geod. Cartographica Sin.* **43**(5), 493–499 (2014).
5. Z. Wu et al., "On combining spectral, textural and shape features for remote sensing image segmentation," *Acta Geod. Cartographica Sin.* **42**(1), 44–50 (2013).
6. D. M. Miller, E. J. Kaminsky, and S. Rana, "Neural network classification of remote-sensing data," *Comput. Geosci.* **21**, 377–386 (1995).
7. J. Mas and J. Flores, "The application of artificial neural networks to the analysis of remotely sensed data," *Int. J. Remote Sens.* **29**, 617–663 (2008).
8. G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.* **43**, 1351–1362 (2005).
9. G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: a review," *ISPRS J. Photogramm. Remote Sens.* **66**, 247–259 (2011).

10. F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.* **113**(6), 1276–1292 (2009).
11. X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.* **51**(1), 257–272 (2013).
12. Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization," *IEEE Trans. Cybern.* **46**, 2966–2977 (2016).
13. Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.* **27**, 1279–1289 (2016).
14. Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.* **2**(1), 1–127 (2009).
15. H. Larochelle et al., "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.* **10**(1), 1–40 (2009).
16. N. Jones, "The learning machines," *Nature* **505**(7842), 146–148 (2014).
17. T. Nguyen, J. Han, and D. C. Park, "Satellite image classification using convolutional learning," in *Proc. of the AIP Conf.*, Albuquerque, New Mexico, pp. 2237–2240 (2013).
18. J. Wang et al., "Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine," *Int. J. Remote Sens.* **36**, 3144–3169 (2015).
19. R. Taormina and K. W. Chau, "Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and extreme learning machines," *J. Hydrol.* **529**, 1617–1632 (2015).
20. Z. Liang et al., "Fuzzy prediction of AWJ turbulence characteristics by using typical multiphase flow models," *Eng. Appl. Comput. Fluid Mech.* **11**, 225–257 (2017).
21. S. A. I. Bellary et al., "Application of computational fluid dynamics and surrogate-coupled evolutionary computing to enhance centrifugal-pump performance," *Eng. Appl. Comput. Fluid Mech.* **10**, 171–181 (2016).
22. J. Zhang and K. W. Chau, "Multilayer ensemble pruning via novel multi-sub-swarm particle swarm optimization," *J. Univ. Comput. Sci.* **15**, 840–858 (2009).
23. W. C. Wang et al., "Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition," *Water Resour. Manage.* **29**, 2655–2675 (2015).
24. S. Zhang and K. W. Chau, "Dimension reduction using semi-supervised locally linear embedding for vegetation leaf classification," *Lect. Notes Comput. Sci.* **5754**, 948–955 (2009).
25. C. Wu, K. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," *J. Hydrol.* **389**, 146–167 (2010).
26. M. Castelluccio et al., "Land use classification in remote sensing images by convolutional neural networks," 2015, arXiv:1508.00092 (01 August 2015).
27. F. Hu et al., "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.* **7**, 14680–14707 (2015).
28. S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *J. Imaging Sci. Technol.* **60**, 10402-1–10402-9 (2016).
29. H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 1520–1528 (2015).
30. S. Paisitkriangkrai et al., "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Boston, Massachusetts, pp. 36–43 (2015).
31. G. Papandreou, I. Kokkinos, and P. A. Savalle, "Untangling local and global deformations in deep convolutional networks for image classification and sliding window detection," 2014, arXiv:1412.0296 (30 November 2014).
32. V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labeling," 2015, arXiv:1505.07293 (27 May 2015).

33. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," 2015, arXiv:1502.03167 (02 March 2015).
34. J. Liu, B. Liu, and H. Lu, "Detection guided deconvolutional network for hierarchical feature learning," *Pattern Recognit.* **48**, 2645–2655 (2015).
35. M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 1–9 (2015).
36. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556 (10 April 2015).
37. T. Panboonyuen et al., "An enhanced deep convolutional encoder-decoder network for road segmentation on aerial imagery," *Adv. Intell. Syst. Comput.* **566**, 191–201 (2017).
38. A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," arXiv:1511.02680 (2016).
39. L. C. Chen et al., "Semantic image segmentation with deep convolutional nets and fully connected CRFs," arXiv:1412.7062 (2016).
40. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015).
41. G. Fu et al., "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sens.* **9**, 498 (2017).
42. J. Dolz, "3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study," *Neuroimage* **170**, 456–470 (2018).
43. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017).
44. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," arXiv:1505.04597v1 [cs.CV] (2015).
45. L. C. Chen et al., "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," arXiv:1606.00915 (2017).
46. H. Lin, Z. Shi, and Z. Zou, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," *Remote Sens.* **9**, 480–501 (2017).
47. F. Visin et al., "Reseg: a recurrent neural network-based model for semantic segmentation," in *Proc. of the IEEE Conf.on Computer Vision and Pattern Recognition Workshops*, pp. 41–48 (2016).
48. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2017).

Chengming Zhang is currently a professor working at the College of Information Science and Engineering of Shandong Agricultural University. His main research areas are remote sensing and geographic information system in land use monitoring and evaluation, presided over a number of agricultural remote sensing projects by Ministry of Science and Technology and Shandong Province. Currently, he is mainly engaged in the research of remote sensing technology in agriculture and environment.

Jiping Liu is currently a professor at the Chinese Academy of Surveying and Mapping. His main research interests are the application of remote sensing in urban geographic information acquisition and the study of remote sensing data. He has participated in key projects in many countries and achieved excellent results. Currently, he focuses on the research of intelligent processing of remote sensing and big data.

Fan Yu is currently an associate professor at the Beijing University of Civil Engineering and Architecture. His main research areas are remote sensing image processing, he has published several academic papers in this area. He also takes part in many research projects founded by Ministry of Science and Technology of China. Currently, he is mainly engaged in the research of remote sensing technology in urban management and city information acquisition.

Shujing Wan is currently a engineer at the College of Network Information Center of Qufu Normal University. Her main research areas are remote sensing and digital image processing. Currently, she is mainly engaged in the research of information construction and remote sensing technology in agriculture and the environment.

Yingjuan Han is a senior engineer at the Ningxia Institute of Meteorology. Her main research area is remote sensing of ecology and agriculture. Her main research area is the application of remote sensing of ecology and agriculture in arid and semiarid areas. Currently, she is carrying out remote sensing extraction of crop planting information and monitoring of ecological environment elements by remote sensing.

Jing Wang is a senior engineer at the Ningxia Meteorological Science Institute. Her main research area is remote sensing in agriculture. Currently, she is mainly engaged in the research of fruit tree classification and planting area extraction by remote sensing technology.

Gang Wang is currently a junior studying at the College of Information Science and Engineering of Shandong Agricultural University. His main research areas are remote sensing and geographic information system. Currently, he is mainly engaged in the research of remote sensing technology in classification and LUCC.