A Land Use Classification Model Based on Conditional Random Fields and Attention Mechanism Convolutional Networks

Kang Zheng 1,2, Haiying Wang 1,2,3,4,*, Fen Qin 1,2,3,4 and Zhigang Han 1,2,3,4

1 College of Environment and Planning, Henan University, Kaifeng 475004, China; zkang@henu.edu.cn (K.Z.); qinfen@henu.edu.cn (F.Q.); zhghan@henu.edu.cn (Z.H.)
2 Key Laboratory of Geospatial Technology for Middle and Lower Yellow River Regions, Ministry of Education, Kaifeng 475004, China
3 Henan Technology Innovation Center of Spatio-Temporal Big Data, Henan University, Zhengzhou 450046, China
4 Henan Industrial Technology Academy of Spatio-Temporal Big Data, Henan University, Zhengzhou 450046, China
* Correspondence: whyhdgis@henu.edu.cn

Abstract: Land use is used to reflect the expression of human activities in space, and land use classification is a way to obtain accurate land use information. Obtaining high-precision land use classification from remote sensing images remains a significant challenge. Traditional machine learning methods and image semantic segmentation models are unable to make full use of the spatial and contextual information of images. This results in land use classification that does not meet high-precision requirements. In order to improve the accuracy of land use classification, we propose a land use classification model, called DADNet-CRFs, that integrates an attention mechanism and conditional random fields (CRFs). The model is divided into two modules: the Dual Attention Dense Network (DADNet) and CRFs. First, the convolution method in the UNet network is modified to Dense Convolution, and the band-hole pyramid pooling module, spatial location attention mechanism module, and channel attention mechanism module are fused at appropriate locations in the network, which together form DADNet. Second, the DADNet segmentation results are used as a priori conditions to guide the training of CRFs. The model is tested with the GID dataset, and the results show that the overall accuracy of land use classification obtained with this model is 7.36% and 1.61% higher than FCN-8s and BiSeNet in classification accuracy, 11.95% and 1.81% higher in MIoU accuracy, and with a 9.35% and 2.07% higher kappa coefficient, respectively. The proposed DADNet-CRFs model can fully use the spatial and contextual semantic information of high-resolution remote sensing images, and it effectively improves the accuracy of land use classification. The model can serve as a highly accurate automatic classification tool for land use classification and mapping high-resolution images.

Keywords: attention mechanism; convolutional neural networks; conditional random fields; land use classification

1. Introduction

Land use information is fundamental to understanding the dynamic changes on the surface of the earth and socio-ecological interactions [1], and is essential in a number of fields of earth observations, such as urban and regional planning [2], environmental impact assessment, etc., [3,4]. Land use classification is a key way to obtaining land use information, and it is also a necessary means to deepen the interrelationship between human activities and the spatial environment [5–7]. With the continuous development of
Remote sensing technology, the spatial and temporal resolutions of remote sensing images have improved. The Worldview, SPOT and Ikonos series of satellites are capable of acquiring remote sensing images with high spatial resolution. Chinese satellite remote sensing technology has also entered the era of sub-meter level, with for instance the panchromatic band of Gaofen-2 reaching 0.8 m spatial resolution [8]. High-resolution remote sensing images contain rich spatial information and have become an important tool for land use classification research. Under the condition of high-resolution remote sensing images, it is especially important how to obtain high-precision land use information.

Traditional land use classification methods include visual interpretation classification, cluster analysis, and spectral classification. For example, Huang et al. [9] proposed geometric probability-based cluster analysis to improve land classification accuracy, and Tehrany et al. [10] used SPOT5 remote sensing images and an object-oriented k-nearest neighbor algorithm to effectively improve image land use classification. With the improvement of computer and remote sensing image technology, classification and spatial statistical techniques such as fuzzy, artificial neural network, support vector machine (SVM), and geo-detector have emerged [11–17]. For example, Li et al. [14] fused fuzzy and decision tree algorithms to achieve object-oriented land use classification in the Dongjiang River Basin. Talukdar et al. [15] used six machine-learning algorithms methods to assess land use classification and the six methods were random forest, SVM, artificial neural network, fuzzy adaptive resonance theory-supervised predictive mapping, spectral angle mapper, and Mahalanobis distance. Abdi et al. [16] used Sentinel-2 images and machine learning methods to obtain accurate land use classifications. Wang et al. [17] used a geodetector to evaluate the spatial variation of prediction results in order to confirm the validity and applicability of their classification results.

Traditional land use classification methods have reduced labor costs to a certain extent, but their accuracy and timeliness still cannot meet the current demand in the era of big data. Obtaining high-precision land use classification is a serious challenge. Therefore, it is critical to improve the learning ability of land use classification models such that they can use spatial features and contextual semantic information from high-resolution remote sensing images to achieve high-precision land use classification.

Along with the rapid development of artificial intelligence technology, semantic segmentation techniques have gradually replaced traditional land use classification methods. Image semantic segmentation is the process of subdividing a digital image into multiple superpixels with the aim of simplifying or changing the presentation of the image to make it simpler and easier to understand and analyze [18]. Image semantic segmentation techniques are divided into traditional semantic segmentation techniques and deep learning based semantic segmentation techniques. Traditional semantic segmentation techniques include threshold-based image semantic segmentation, edge-based image semantic segmentation, region-based image semantic segmentation, and specific theory-based image semantic segmentation [19]. For example, in 2000, Shi and Malik proposed N-cut (Normalized Cut), a semantic segmentation method based on graph partitioning [20]. Based on this, Microsoft Cambridge Research proposed the famous interactive image semantic segmentation method, Grab Cut [21].

Traditional semantic segmentation techniques tend to use the low-order visual information of the image pixels themselves for segmentation without a training phase, and the results are not satisfactory in more difficult segmentation tasks. The birth of deep learning, with the continuous development of computer technology and hardware, essentially provide algorithms that use artificial neural networks as an architecture for learning representations of information [22]. More recently, models combining deep learning with image semantic segmentation have emerged and can be used for semantic segmentation. These include: The fully convolutional network (FCN) [23], based on a traditional convolutional neural network to solve the semantic segmentation problem at the pixel level; the residual convolutional network (ResNet), which uses residual structure to make the network deeper and converge faster compared to FCN-8s, effectively preventing gradient
explosion and solving the problem of training difficulties as the number of network layers increases [24]; the dense convolutional network (DenseNet), which improves the computational efficiency and compresses the number of parameters compared to ResNet, solves the problem of large number of parameters and slow computational efficiency of ResNet, and achieves performance comparable to ResNet [25]; the dual attention network (DANet), which proposes a spatial location attention module and a channel attention module to capture the rich semantic information of images [26].

At present, it is difficult to further improve the accuracy of land use classification in the process of changing from traditional semantic segmentation techniques to deep learning-based semantic segmentation techniques. Sun et al. [27] showed that combining LiDAR data and high-resolution image data with a deep convolutional neural network can improve land use classification. Wang et al. [28] proposed a multi-scale neural network model based on a deep convolutional neural network using 0.5 m resolution optical aerial remote sensing images of Zhejiang Province as the data source, and compared it with traditional FCN and SVM methods. They showed that the multi-scale neural network model has higher accuracy and stronger overall classification results than traditional methods. Huang et al. [29] used Nanshan in Shenzhen City and Kowloon Peninsula, northern Hong Kong Island and Shatin, New Territories in Hong Kong City as study areas, using worldview-2 and worldview-3 as sample datasets, and the semi-transfer deep convolutional neural network (STDCNN) classification method. The STDCNN method can directly classify using multispectral image data, thus solving the problem of many classical networks which can only input three channels. Although semantic segmentation techniques based on deep learning have been widely used in the field of land use classification, due to the lack of constraints on image context information, only the local information of the image is used. Without the full spatial and contextual information of the image, the segmentation results are too smooth, other phenomena occur, and classification accuracy cannot be sufficiently improved.

To overcome the above difficulties, we propose a land use classification model combining attention mechanism and conditional random fields (CRFs). These models focus on spatial and contextual features, combining multi-scale semantic information of features, and improving the generalization ability of the model. CRFs introduce contextual information, fully consider the semantic linkage between pixels, refine the segmentation effect of edges, and improve the classification accuracy of land use. This study will help to improve the accuracy and applicability of land use classification models and provide a more effective modeling tool for land use classification and mapping research.

2. DADNet-CRFS

2.1. General Framework of the Model

The DADNet-CRFS model combines the DADNet and CRFs modules (Figure 1). DADNet is the application of the dense convolutional module of DenseNet to the encoding-decoding structure of U-Net model [30], and incorporates the channel attention mechanism and spatial attention mechanism atrous spatial pyramid pooling (ASPP) [31] method to enhance the model’s ability to learn contextual semantics and spatial features. Deep separable convolution is used to reduce the number of parameters and improve the training efficiency of the model. The input of the CRFs module is the result of DADNet, thus further optimizing the model results for edge segmentation and improving the overall continuity of the segmentation results.
2.2. DADNet

The DADNet model is a convolutional neural network built based on DenseNet and a channel and spatial location attention mechanism. The former features reduces the number of parameters and improves the computational efficiency of a model; the latter is able to solve the semantic segmentation of objects at different scales, as well as the fusion of similar features. The model framework uses the classical full convolutional network UNet with the addition of DenseNet, a two-way attention module, and ASPP in the encoder and decoder, as shown in Figure 2.

Figure 1. General framework of DADNet-CRFs model.

Figure 2. DADNet network structure.

The encoder of DADNet includes the densely connected block and transition layer of DenseNet, as well as the spatial attention module (SAM) and the channel attention module (CAM), in which the sizes of the convolution kernels are $32 \times 32$, $64 \times 64$, $128 \times 128$, $256 \times 256$. 

The decoder of DADNet includes the dense block and transition layer of DenseNet, as well as the channel and spatial attention module. The sizes of the convolution kernels are $32 \times 32$, $64 \times 64$, $128 \times 128$, $256 \times 256$. 

---

Remote Sens. 2022, 14, 2688

4 of 19
The feature maps are input to SAM and CAM respectively. The size of the feature maps is kept consistent before and after the application of the module, and then the two features are fused by summing. The decoder includes Transpose Convolution and CAM, and the fused feature maps are upsampled using Transpose Convolution, before and after the jump-connected tandem, combining the high-level semantic features from the decoder and the high-level semantic features from the encoder for the corresponding size feature maps. Finally, the segmentation result is output using the softmax activation function.

2.2.1. DenseNet

DenseNet is a densely connected network improved on ResNet. The combinatorial function is assumed to be \( H_l(\cdot) \), \( l \) refers to the number of layers, and the output of each layer is defined as \( x_l \). The conventional convolutional network connects the \( l \) layer and the \( l+1 \) layer by forward propagation as follows.

\[
x_l = H_l(x_{l-1})
\]  

ResNet adds jump connections, which short-circuit the \( l \)th layer to the previous 2 to 3 layers. The connections are summed at the element level, and the output is as in Equation (2). This solves the gradient explosion to a certain extent and enables feature sharing, but the connection by summation may lead to a loss of network information flow.

\[
x_l = H_l(x_{l-1}) + x_{l-1}
\]  

DenseNet aims to improve the utilization of information flow in the network. It uses dense connections, and the input of the \( l \)th layer is the output feature map of all the previous layers through forward feedback, which is directly connected using a series connection, as follows. This significant improvement of DenseNet over ResNet enables feature reuse and improves utilization efficiency.

\[
x_l = H_l[x_0, x_1, \ldots, x_{l-2}, x_{l-1}]
\]  

The DenseNet network structure is mainly composed of densely connected blocks and transition layers, as shown in Figure 3. The feature maps of each layer of the densely connected blocks are of the same size and can be connected in series over the channels. The combined function structure is a 3 × 3 Separable Convolution layer + Batch Normal layer + Rectified Linear Units activation function layer.
The transition layer is used to connect two adjacent densely connected blocks and to reduce the feature map size. In DADNet, the structure of the encoder transition layer is $1 \times 1$ Separable Convolution + $2 \times 2$ MaxPooling layer, and in decoder it is $1 \times 1$ Separable Convolution + $3 \times 3$ Transpose Convolution layer.

2.2.2. SAM and CAM

For forward propagating convolutional neural networks, the convolutional block attention module (CBAM) is a simple and effective attention mechanism [32]. In this paper, embedding SAM and CAM, CBAM solves the global dependency problem, fuses similar features of objects of different sizes to some extent and avoids the labeling of large object features affecting inconspicuous small objects.

CAM is used to specify useful features and to fuse these features with each other as shown in Figure 4. The whole module has no convolution layer, and the input feature map $A$ is subjected to reshape, multiply, transpose, softmax, sum, and multiply operations to complete the output feature map $E$. Specifically, Equations (4) and (5). $A$ and $E$ represent the input and output feature maps, respectively, $ij$ is the number of rows and columns, $\beta$ is a self-learnable parameter with an initial value of 0, and $x$ is obtained by doing softmax operations on each row.

$$x_{ij} = \frac{\exp(-A_i A_j)}{\sum_{l=1}^{N} \exp(-A_i A_j)}$$

(4)

$$E_j = \beta \sum_{i=1}^{C} (x_{ij} A_j) + A_j$$

(5)

SAM can make local features link contextual semantics for feature enhancement purposes, and similar features are fused after module computation. This is conducive to
semantic segmentation, as shown in Figure 5, Equations (6) and (7). Unlike the channel attention module, the input feature map is first subjected to a convolution operation to obtain three feature maps $B$, $C$ and $D$. Then a series of calculations such as reshape, transpose, and phase multiplication are performed to output the feature map $E$, which is consistent with the shape of $A$. The parameter $S_i$ is a softmax activation function after multiplying $B$ and $C$. It can be understood that the greater the similarity between $B$ and $C$, the greater the value of $S$, where the two are symmetrically related.

$$S_i = \frac{\exp(B_i \cdot C_j)}{\sum_{k=1}^{n} \exp(B_i \cdot C_k)}$$

(6)

Finally, the feature map $E$ is output by linear combination with $D$. $\alpha$ is also a self-learning parameter with an initial value of 0. The formula is as follows.

$$E_j = \alpha \sum_{t=1}^{c} (S_{ji}D_j) + A_j$$

(7)

![Figure 5. SAM schematic [18].](image)

2.2.3. ASPP

In high-resolution images, different scale features have obvious boundaries. The ASPP structure is used to fuse multi-scale information in order to increase the recognition ability of the model for multi-scale features. It learns the features of different scales by a set of dilated convolutions [33] with different sampling rates to improve the multi-scale feature segmentation accuracy. The ASPP structure is shown in Figure 6, and the specific calculation process is as follows.

$$Y = \text{Concat} \left( H_{1,3}(X), H_{6,3}(X), H_{12,3}(X), H_{18,3}(X), I_{\text{pooling}}(X) \right)$$

(8)

$\text{Concat}(\cdot)$ denotes the stitching operation on feature results of different scales, $H_{r,n}(X)$ is the convolution of holes with different sampling rates $r$ and convolution kernel size $n$, and $I_{\text{pooling}}(X)$ is the pooling operation on the input feature map.
2.3. CRFs

Lafferty et al. [34] proposed CRFs in 2001 as a family of discriminative probabilistic models that incorporates the features of maximum entropy models and Markov models. These models are mainly used for sequence labeling tasks and have been commonly used in recent years for semantic segmentation with good results. For example, Krähenbühl et al. [35] used improved fully connected CRFs for effective image segmentation, and Chunjiao et al. [36] performed semantic segmentation of remote sensing images by deep fusion network CRFs for accurate recognition of edge contours.

CRFs are commonly used in the field of image semantic segmentation to compare the contextual information of the original image and the predicted labeled image simultaneously, and then model the class validation probabilities based on the relationship between pixels. Based on the proof of Hammersley–Clifford theorem, in equation denotes a given high-resolution remote sensing image, and \( X \) is the posterior probability of the corresponding labeled image, which obeys a Gibbs distribution as follows.

\[
P(X|I) = \frac{1}{Z(I)} \exp \left( - \sum_{c \in C} \varphi_c(X_c|I) \right)
\]

(9)

\( Z(I) \) is the normalization factor and \( \varphi_c \) is the conditional potential energy function on the group \( c \). Depending on the number of variables contained in \( c \), the potential energy function is divided into first-order potential, second-order potential, and higher-order potential, as expressed below.

\[
E(x) = \sum_i \Phi_i(x_i) + \alpha \sum_{i,j} \Phi_{i,j}(x_i, x_j)
\]

(10)

where \( \Phi_i(x_i) \) is a first-order potential function and \( \Phi_{i,j}(x_i, x_j) \) is a second-order potential function of the following form.

\[
\Phi_{i,j}(x_i, x_j) = u(x_i, x_j) \sum_{m=1}^{K} w_m k^m(x_i, x_j)
\]

(11)

In the above equation, if \( x_i \neq x_j \) then \( u(x_i, x_j) = 1 \), otherwise 0. \( w_m \) is the weight of the Gaussian kernel function and \( k^m(x_i, x_j) \) is the Gaussian kernel function. The model is a linear combination of two Gaussian kernel functions as second-order potential functions, one Gaussian kernel function depends on both position and spectral information, and the other depends on position information only. The linear combination forms a
model of CRFs to learn spatial contextual relationships, thus refining edge contours and improving segmentation accuracy. The details are as follows.

\[
k = w_1 \exp \left( - \frac{\| P_i - P_j \|^2}{2\sigma^2_a} - \frac{\| I_i - I_j \|^2}{2\sigma^2_g} \right) + w_2 \exp \left( - \frac{\| P_i - P_j \|^2}{2\sigma^2_y} \right)
\]  

(12)

2.4. Accuracy Evaluation

Here, the semantic segmentation results are evaluated according to both qualitative and quantitative criteria. The qualitative evaluation is based on subjective judgment, such as whether the classification result is complete, the edge contour is clear, and the edges are consistent. For the quantitative evaluation, \textit{Recall}, \textit{Precision}, \textit{F1-score}, overall accuracy (\textit{OA}), and mean intersection over union (\textit{MIoU}) are mainly used. \textit{TP} is the predicted positive sample among positive samples, \textit{FN} is the predicted negative sample among positive samples, \textit{FP} is the predicted positive sample among negative samples, and \textit{TN} is the predicted negative sample among negative samples.

\textit{Recall} is the ratio of the number of correctly classified positive samples to the number of positive samples and is given by the following formula.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]  

(13)

\textit{Precision} is the ratio of the number of correctly classified positive samples to the number of all positive samples in the classification result, as follows.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]  

(14)

\textit{F1 score} is the index proposed to evaluate the model based on \textit{Precision} and \textit{Recall}, and to evaluate \textit{Precision} and \textit{Recall} as a whole, as defined below.

\[
\text{F1score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}
\]  

(15)

\textit{OA} is the ratio of the number of correctly classified samples to the number of all samples and is defined as follows.

\[
\text{OA} = \frac{TP + TN}{TP + FN + FP + TN}
\]  

(16)

\textit{MIoU} is used to calculate the ratio of the intersection and concatenation of the two sets of true and predicted values. It results in a global evaluation of the semantic segmentation results, as defined below.

\[
\text{MIoU} = \frac{1}{k + 1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP}
\]  

(17)

The kappa coefficient is also used as a measure of classification accuracy. In practical classification problems, the kappa coefficient is often used as an indicator of the “bias” of the penalty model instead of \textit{OA} when there is a poor balance between the samples. The more unbalanced the classification result, the lower the kappa coefficient value. The number of true samples in each category is assumed to be \(a_1, a_2, \ldots, a_c\), the number of predicted samples for each category is \(b_1, b_2, \ldots, b_c\), the total number of samples is \(n\), and \(p_e\) is the chance consistency error.

\[
p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \ldots + a_c \times b_c}{n \times n}
\]  

(18)

\[
Kappa = \frac{OA - p_e}{1 - p_e}
\]  

(19)
The significance test performed in this paper is Independent Samples Z-Test. The Independent Samples Z-Test is a statistical test used to determine if there is a significant difference between two different groups of samples. The samples should be continuous and satisfy a normal distribution. There is also a similar distribution between the samples.

3. Experiments and Results

3.1. Experimental Data

GID (see Table 1) is a high-resolution remote sensing image dataset for land use/land cover (LULC) classification, collected from December 2014 to October 2016 [37]. It consists of two parts. One part is a large-scale classification set, which includes 150 high resolution GaoFen-2 remote sensing images from more than 60 different cities in China, covering a geographical area of more than 50,000 km². The individual image size is 6800 × 7200 pixels, with a spatial resolution of 1 m, in RGB form and NIR+RGB form. The land types are Built-up, Farmland, Forest, Meadow and Water, a total of five categories (see Figure 7); the other part is the secondary land use type, comprising 25 HIS-2 images. The secondary land use types are industrial land, urban residential, rural residential, traffic land, paddy field, irrigated land, dry cropland, garden land, arbor forest, shrub land, natural meadow, artificial meadow, river, lake and pond, totaling 15 categories.

![Image](image_url)

**Figure 7.** (Left): The geographical locations of GID images. This experiment selects 120 training images and 30 validation test images, which are marked with green and red respectively. (Right): Examples of GID images and their corresponding label. The RGB of each feature in labels: built-up (255,0,0), farmland (0,255,0), forest (0,255,255), meadow (255,255,0), water (0,0,255), background (0,0,0).

**Table 1.** Introduction of experiment image data.

<table>
<thead>
<tr>
<th>Payload</th>
<th>Bands</th>
<th>Band Range (µm)</th>
<th>Spatial Resolution (m)</th>
<th>Width (km)</th>
<th>Side-Swing Capability</th>
<th>Orbital Period (day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panchromatic multi-spectral camera</td>
<td>1</td>
<td>0.45–0.90</td>
<td>1</td>
<td>45</td>
<td>±35°</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.45–0.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.52–0.59</td>
<td>4</td>
<td>45</td>
<td>±35°</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.63–0.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.77–0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We used GID large-scale data for our experiment, this is because the GID large-scale data basically covers the major cities and basic landscapes in China, and its land use category is based on the Chinese Land Use Classification Criteria (GB/T21010-2017) as a reference. The experiment randomly selected 120 images from 150 remote sensing images as the training set, 15 images from the remaining 30 images as the validation set, and other images as the test set. All images were cropped to the size of 224 × 224 pixels. A large-scale experimental dataset was obtained, then put into the model for training and validation, and finally, the accuracy of the cropped images was compared after prediction. The details are shown in Table 2.

Table 2. Number of experiment data categories.

<table>
<thead>
<tr>
<th></th>
<th>built-Up</th>
<th>Farmland</th>
<th>Forest</th>
<th>Water</th>
<th>Meadow</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>6.19 × 10^8</td>
<td>1.71 × 10^9</td>
<td>1.85 × 10^8</td>
<td>5.92 × 10^8</td>
<td>1.37 × 10^8</td>
<td>2.63 × 10^9</td>
</tr>
<tr>
<td>Validation</td>
<td>8.70 × 10^7</td>
<td>2.54 × 10^8</td>
<td>1.79 × 10^7</td>
<td>9.97 × 10^7</td>
<td>5.74 × 10^4</td>
<td>2.76 × 10^8</td>
</tr>
<tr>
<td>Test</td>
<td>5.38 × 10^7</td>
<td>2.55 × 10^8</td>
<td>7.10 × 10^7</td>
<td>8.80 × 10^7</td>
<td>7.70 × 10^6</td>
<td>2.59 × 10^8</td>
</tr>
<tr>
<td>Total Percentage</td>
<td>10.72%</td>
<td>31.37%</td>
<td>3.87%</td>
<td>11.01%</td>
<td>2.05%</td>
<td>40.98%</td>
</tr>
</tbody>
</table>

3.2. Experimental Results and Analysis

To verify the superiority of DADNet-CRFs, it was compared with the model without fused CRFs, as well as with the two models, FCN-8s and Bilateral Segmentation Network (BiSeNet) [38] in terms of accuracy and segmentation details. The classical semantic segmentation model, FCN-8s, achieves semantic segmentation at the pixel level, and its adoption of jumping structure is widely used in other algorithmic models. BiSeNet is a model proposed in recent years which adopts feature fusion and attention optimization modules and retains sufficient spatial and visual field domain information while improving calculation speed.

3.2.1. Visual Analysis of Classification Results

To further compare the classification results of different models, four cropped images were selected for analysis. The results are shown in Figure 8.

In Group I, the classification results of FCN-8s were more fragmented than those of the BiSeNet method for Farmland, and the boundary between Farmland and Water appeared obviously incorrect. The overall segmentation effect of DADNet-CRFs Farmland was not only continuous, but the model also distinguished well Water, Farmland, and Background.

In Group II: all three methods showed over-segmentation of the Built-up area. FCN-8s had an obvious jagged segmentation result, and DADNet-CRFs segmented the Built-up area fringes better than the first two.

In Group III: BiSeNet segmented the areas better than FCN-8s, but the Built-up area and Forest contours were too smooth, and DADNet-CRFs preserved the spatial patterns of land types well.

In Group IV: FCN-8s and BiSeNet showed obvious classification errors. FCN-8s could not distinguish Water, Meadow, and Farmland well. BiSeNet identified a large amount of Farmland as Meadow, but DADNet-CRFs had a better overall classification recognition effect. Overall, the DADNet-CRFs model was better than both FCN-8s and BiSeNet in terms of classification effect.
3.2.2. Analysis of the Accuracy of Classification Results

The classification results of the test set were statistically compared with the semantic segmentation accuracy evaluation results of the three models, see Table 3. In the recall comparison, DADNet-CRFs performed best and had the highest recall of all types. In the precision comparison, FCN-8s had the highest precision in the Forest category, BiSeNet had the highest precision in the Water category, and DADNet-CRFs had the highest accuracy in the three categories of Built-up, Farmland and Meadow. In the F1-score comparison, DADNet-CRFs had the highest F1-score in all categories, except for the Water category of BiSeNet model, which had the highest F1-score. In the OA comparison, the OA of FCN-8s was 80.62% and that of BiSeNet was 86.37%, and the OA of DADNet-CRFs improved by 7.36% and 1.61% compared to FCN-8s and BiSeNet, respectively. In the MIoU comparison, DADNet-CRFs had the highest MIoU of 69.47%, while the MIoU of FCN-8s and BiSeNet were 57.52% and 67.66%, respectively. The kappa coefficient of DADNet-
CRFs reached 84.70%, which was 9.35% and 2.07% higher than that of FCN-8s and BiSe-Net, respectively.

Table 3. Evaluation table of classification results. The bolded numbers are the best performing models in terms of accuracy.

<table>
<thead>
<tr>
<th>Model</th>
<th>Category</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
<th>OA</th>
<th>MIoU</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-8s</td>
<td>built-up</td>
<td>83.38%</td>
<td>71.08%</td>
<td>76.74%</td>
<td>80.62%</td>
<td>57.52%</td>
<td>75.35%</td>
</tr>
<tr>
<td></td>
<td>farmland</td>
<td>75.33%</td>
<td>87.02%</td>
<td>80.75%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>forest</td>
<td>82.33%</td>
<td><strong>88.17%</strong></td>
<td>85.15%</td>
<td>80.62%</td>
<td>57.52%</td>
<td>75.35%</td>
</tr>
<tr>
<td></td>
<td>water</td>
<td>88.49%</td>
<td>85.70%</td>
<td>87.07%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>meadow</td>
<td>75.32%</td>
<td>40.14%</td>
<td>52.37%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiSeNet</td>
<td>built-up</td>
<td>85.20%</td>
<td>72.56%</td>
<td>78.37%</td>
<td>86.37%</td>
<td>67.66%</td>
<td>82.63%</td>
</tr>
<tr>
<td></td>
<td>farmland</td>
<td>83.00%</td>
<td>85.89%</td>
<td>84.42%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>forest</td>
<td>90.43%</td>
<td>85.42%</td>
<td>87.85%</td>
<td>86.37%</td>
<td>67.66%</td>
<td>82.63%</td>
</tr>
<tr>
<td></td>
<td>water</td>
<td>90.16%</td>
<td><strong>90.93%</strong></td>
<td>90.54%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>meadow</td>
<td>76.00%</td>
<td>44.67%</td>
<td>56.27%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DADNet-CRFs</td>
<td>built-up</td>
<td>87.72%</td>
<td><strong>74.18%</strong></td>
<td><strong>80.38%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>farmland</td>
<td><strong>85.32%</strong></td>
<td><strong>88.11%</strong></td>
<td><strong>86.69%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>forest</td>
<td>90.89%</td>
<td>86.62%</td>
<td><strong>88.70%</strong></td>
<td><strong>87.98%</strong></td>
<td><strong>69.47%</strong></td>
<td><strong>84.70%</strong></td>
</tr>
<tr>
<td></td>
<td>water</td>
<td>92.17%</td>
<td>87.85%</td>
<td>89.96%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>meadow</td>
<td>79.37%</td>
<td>47.40%</td>
<td>59.36%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on the prediction results, we conducted further significance tests for the FCN-8s, BiSeNet, and DADNet-CRFs models by taking Independent Sample Z-tests. The results show that the significance levels among the models are all less than 0.05, and there are significant differences. Then, we performed a pre- and post-CRF optimization analysis on the experimental models.

3.2.3. Analysis of Model Performance Differences before and after Optimization Using CRFs

The maximum number of iterations of CRFs was adjusted, a test set image was randomly selected, and the number of iterations with the highest OA and MIoU accuracy was chosen to optimize the model classification results. Differences in model performance before and after optimization using CRFs were used to analyze the effectiveness of including CRFs in classification models.

We performed a visual judgment analysis on Figure 9.

In Group I: Farmland showed obvious holes in the DADNet classification results, while this category remained more intact in the DADNet-CRFs classification results.

In Group II: the DADNet and DADNet-CRFs classifications were clearer for the Water segmentation, and the feature integrity was better maintained. DADNet-CRFs were smoother on the edges of Water and Built-up.

In Group III: DADNet and DADNet-CRFs showed clearer classification for Forest and Built-up area, but the Built-up area was enlarged to different degrees. Moreover, there was a gap between Forest and Built-up in DADNet classification results, while DADNet-CRFs did not have this problem and showed smoother performance on the edges.

In Group IV: both DADNet and DADNet-CRFs classification results misclassified Farmland into Meadow in the upper right and bottom regions of the sample area, but DADNet-CRFs still effectively reduced the effect of noise, and the classification results were smoother. Overall, DADNet-CRFs had better classification results than DADNet.
To further demonstrate the classification effect after adopting CRFs, the FCN-8s, BiSeNet and DADNet models were combined with CRFs, and a single test set image was randomly selected for accuracy comparison before and after the combination. Table 4 shows that the classification accuracy of land use types was improved after combining FCN-8s, BiSeNet, and DADNet models with CRFs. Among them, the Forest classification accuracy of FCN-8s-CRFs model improved the most, reaching 2.12%; the Farmland classification accuracy of BiSeNet-CRFs model improved the most, reaching 3.64%; the Forest classification accuracy of DADNet-CRFs model improved the most, reaching 1.13%.
The overall classification accuracies of FCN-8s-CRFs, BiSeNet-CRFs, and DADNet-CRFs were 90.70%, 92.44%, and 93.04%, respectively, higher by 0.70%, 1.23%, and 0.42% respectively than before combining with CRFs. Among all models, the overall accuracy of DADNet-CRFs was the highest, except for Farmland. The classification accuracy of all other types was the highest, including 93.09% for Built-up and 97.30% for Water.

In summary, combining CRFs with FCN-8s, BiSeNet and DADNet models can improve the classification accuracy of the models to some extent, and the DADNet-CRFs model provides the best accuracy, from both visual and quantitative comparisons.

Table 4. Comparison of accuracy before and after combining the land classification models with CRFs. The bolded numbers are the best performing models in terms of accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Built-Up</th>
<th>Farmland</th>
<th>Forest</th>
<th>Water</th>
<th>Meadow</th>
<th>OA</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-8s</td>
<td>88.29%</td>
<td>67.63%</td>
<td>80.02%</td>
<td>96.59%</td>
<td>0.00%</td>
<td>90.00%</td>
</tr>
<tr>
<td>FCN-8s-CRFs</td>
<td>89.24%</td>
<td>69.01%</td>
<td>82.14%</td>
<td>96.60%</td>
<td>0.00%</td>
<td>90.70%</td>
</tr>
<tr>
<td>BiSeNet</td>
<td>89.65%</td>
<td>64.17%</td>
<td>87.44%</td>
<td>96.64%</td>
<td>0.00%</td>
<td>91.21%</td>
</tr>
<tr>
<td>BiSeNet-CRFs</td>
<td>91.50%</td>
<td>67.81%</td>
<td>89.16%</td>
<td>96.84%</td>
<td>0.00%</td>
<td>92.44%</td>
</tr>
<tr>
<td>DADNet</td>
<td>92.44%</td>
<td>60.90%</td>
<td>88.03%</td>
<td>97.25%</td>
<td>0.00%</td>
<td>92.62%</td>
</tr>
<tr>
<td>DADNet-CRFs</td>
<td>93.09%</td>
<td>61.02%</td>
<td>89.16%</td>
<td>97.30%</td>
<td>0.00%</td>
<td>93.04%</td>
</tr>
</tbody>
</table>

3.3. Time Consumption

About the time consumption of DADNet-CRFs model, tests were performed on the small sample GID dataset, and the training times of different models were compared (Table 5), with time units in seconds. From the table, we can find that the time consumption of the model in this paper is approximately equal to that of the BiSeNet model. Compared with than the FCN-8s model, the computation time is more than 1000 s faster, and the average time consumed per step is a few 10 s faster.

Table 5. Training time comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>DADNet-CRFs</th>
<th>FCN-8s</th>
<th>BiSeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>time(s)</td>
<td>3158.00</td>
<td>4285.00</td>
<td>3288.00</td>
</tr>
<tr>
<td>average time(s)</td>
<td>45.11</td>
<td>61.21</td>
<td>46.97</td>
</tr>
</tbody>
</table>

4. Discussion

Classical deep learning semantic segmentation models are not very effective in land use classification due to the loss of spatial and contextual semantic features, since they use only local information in images. To solve this problem, we fully integrated the advantages of attention mechanisms and CRFs and proposed a new land use classification model, DADNet-CRFs. Both visual analysis and evaluation metrics showed that DADNet-CRFs have better classification accuracy than FCN-8s and BiSeNet models [23,38]. The overall accuracy was improved by 7.36% and 1.61% compared to FCN-8s and BiSeNet models, respectively. CRFs were also found to enhance the accuracy of classification results for FCN-8s, BiSeNet, and DADNet models. The results were tested by randomly selecting single test set images and found to be improved by 0.70%, 1.23%, and 0.42%, respectively, compared with those before combining CRFs. Among the three combined models, DADNet-CRFs had the highest overall classification accuracy of 93.04%. Under the same conditions of environment, the DADNet-CRFs model has shorter training time and higher classification efficiency compared with the FCN-8s and BiSeNet models. The DADNet-CRFs model can effectively enhance the accuracy of land use classification results and improve the classification effect.

First, to improve the model’s ability to learn contextual semantic features and spatial connections, Hou Biao et al. [39] modified the DeepLab V3+ network by adding a path for upsampling, thus realizing the union of feature maps at different scales. Compared to the
research method in their study, DADNet uses the SAM module to fuse the pixel with the location information of every other pixel, associating similar spatial features with each other, thereby preventing the occurrence of feature classification errors, and cooperating with ASPP to learn features at different scales in space. Li Rui et al. [40] developed the Attention Aggregation Feature Pyramid Network (A2-FPN) by adding an attention aggregation module to the FPN pyramid. Although the attention aggregation module can enhance the learning of multi-scale features, compared to A2-FPN, DADNet-CRFs incorporates both spatial location attention and channel attention mechanisms [33]. In the model, the spatial location attention module is used for multi-scale feature learning and the channel attention module, by constructing interdependencies between channel feature maps to improve the semantic-specific feature representation.

In addition, when processing details in the segmentation results, such as image edge optimization and overall continuity, the Edge-FCN network introduces a holistically nested edge detection (HED) network to detect edge information. This significantly improves edge segmentation [41], although the problem of mis-segmentation between significantly different features such as forest and water arise. The graph convolutional network (GCN) is a new type of deep neural network that can use the structural information of graphs and work directly on images [42], with significant nulling problems and insufficient edge smoothing in the classification results with the GID dataset. Compared to GCN, DADNet-CRFs uses DADNet results as a priori condition. CRFs further optimize the classification results by encouraging similar pixels to be assigned the same labels, while pixels with large differences are assigned different labels. This results in more precise refinement of edge contours, less problems with voids, and a more continuous classification effect overall [35].

To a certain extent, the DADNet-CRFs model still results in misclassification of similar features, especially misclassifying Background as Built-up and Farmland as Meadow. The ability of the model to learn feature characteristics is not perfect, especially in distinguishing features that are very similar. The GID dataset has an excessive amount of Background, a very small number of Meadow and Forest pixels, and it is extremely unbalanced in terms of types [43,44]. This can easily lead to misclassification of the corresponding land cover types and lower classification accuracy. In addition, the GID dataset may also have inaccurate data for some of the labels, and the research method in this paper does not adopt relevant strategies like Updating Iteration Strategy [39] for this situation. The above two types of problems will aggravate the misclassification by models such as DADNet-CRFs, leading to a decrease in classification accuracy. The overall accuracy improvement was relatively limited after combining FCN-8s, BiSeNet, and DADNet with CRFs. The classification accuracy of FCN-8s-CRFs, BiSeNet-CRFs, and DADNet-CRFs was dependent on the classification performance of the original model. The classification accuracy of Farmland was generally at a low level with more misclassifications compared to other land use types (Table 4). The large difference with the classification accuracy of Farmland in the test set (Table 4) may be related to the label quality of the single test data in the GID dataset.

Going forward, we suggest further optimizing the method of coupling land classification models with CRFs and deep learning models to maximize the learning migration capability of the model and applies to land use classification in different spatial and temporal resolution situations. At the same time, data enhancement, expansion of the dataset, and elimination of inaccurate data could all improve the classification accuracy of the model. The DADNet model parameters have a larger number of parameters and longer training time compared to the other two models. Therefore, the DADNet model needs to be improved and optimized to reduce the parameters as much as possible and improve the training efficiency of the model without losing its classification ability.
5. Conclusions

This paper proposed DADNet-CRFs, a land use classification model integrating attention mechanism and CRFs. At the beginning, this paper draws on the UNet network structure and combines Dense Convolution, ASPP, SAM, and CAM to complete the construction of the DADNet model. Then, the DADNet model classification results are input into CRFs and used to guide the CRFs training. In this paper, experiments are conducted on the basis of GID large-scale dataset to compare the DADNet-CRFs model with FCN-8s and BiSeNet models. The results show that DADNet-CRFs can improve the model’s ability to learn spatial features of land use features, enhance the semantic links between contexts, and achieve high-precision land use classification.

In future research, we will further optimize the method in terms of model structure and data set to improve the model training efficiency and classification accuracy. At the same time, we will enhance the generality and learning migration capability of the model so that it can be applied to land use classification and mapping applications at different spatial and temporal resolutions.

Author Contributions: Conceptualization, K.Z.; methodology, H.W. and K.Z.; formal analysis, Z.H.; investigation, K.Z., H.W. and F.Q.; data curation, K.Z.; writing—original draft preparation, K.Z.; writing—review and editing, H.W.; visualization, K.Z.; supervision, H.W. and F.Q. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge support by the National Nature Science Foundation of China, Grant/Award Number: 41871316, 41401457; Young Key Teacher Training Plan of Henan, Grant/Award Number: 2020GGJS028; Natural Science Foundation of Henan, Grant/Award Number: 202300410096; Natural Resources Science and Technology Innovation Project of Henan Province, Grant/Award Number:202016511; Key Scientific Research Project Plans of Higher Education Institutions of Henan, Grant/Award Number: 21A170008; Technology Development Plan Project of Kaifeng, Grant/Award Number: 2003009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement:

Acknowledgments: The research was also supported by the National Earth System Science Data Sharing Infrastructure, National Science & Technology Infrastructure of China-Data Center of Lower Yellow River Regions (http://henu.geodata.cn accessed on 19 September 2021).

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

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRFs</td>
<td>conditional random fields</td>
</tr>
<tr>
<td>DADNet</td>
<td>dual attention dense network</td>
</tr>
<tr>
<td>N-cut</td>
<td>normalized cut</td>
</tr>
<tr>
<td>FCN</td>
<td>fully convolutional network</td>
</tr>
<tr>
<td>ResNet</td>
<td>residual convolutional network</td>
</tr>
<tr>
<td>DenseNet</td>
<td>dense convolutional network</td>
</tr>
<tr>
<td>DANet</td>
<td>dual attention network</td>
</tr>
<tr>
<td>STDCNN</td>
<td>semi-transfer deep convolutional neural network</td>
</tr>
<tr>
<td>ASPP</td>
<td>atrous spatial pyramid pooling</td>
</tr>
<tr>
<td>SAM</td>
<td>spatial attention module</td>
</tr>
<tr>
<td>CAM</td>
<td>channel attention module</td>
</tr>
<tr>
<td>CBAM</td>
<td>convolutional block attention module</td>
</tr>
<tr>
<td>OA</td>
<td>overall accuracy</td>
</tr>
<tr>
<td>MIoU</td>
<td>mean intersection over union</td>
</tr>
</tbody>
</table>
References


