sustainability

MDPI

Article

Extraction of Rural Residential Land from Very-High Resolution UAV Images Using a Novel Semantic Segmentation Framework

Chenggao Sha †, Jian Liu †, Lan Wang, Bowen Shan, Yaxian Hou and Ailing Wang *

College of Resources and Environment, Shandong Agricultural University, Taian 271018, China
* Correspondence: ailingwang@sdau.edu.cn
† These authors contributed equally to this work.

Abstract: Accurate recognition and extraction of rural residential land (RRL) is significant for scientific planning, utilization, and management of rural land. Very-High Resolution (VHR) Unmanned Aerial Vehicle (UAV) images and deep learning techniques can provide data and methodological support for the target. However, RRL, as a complex land use assemblage, exhibits features of different scales under VHR images, as well as the presence of complex impervious layers and backgrounds such as natural surfaces and tree shadows in rural areas. It still needs further research to determine how to deal with multi-scale features and accurate edge features in such scenarios. In response to the above problems, a novel framework named cascaded dense dilated network (CDD-Net), which combines DenseNet, ASPP, and PointRend, is proposed for RRL extraction from VHR images. The advantages of the proposed framework are as follows: Firstly, DenseNet is used as a feature extraction network, allowing feature reuse and better network design with fewer parameters. Secondly, the ASPP module can better handle multi-scale features. Thirdly, PointRend is added to the model to improve the segmentation accuracy of the edges. The research takes a plain village in China as the research area. Experimental results show that the Precision, Recall, F1 score, and Dice coefficients of our approach are 91.41%, 93.86%, 92.62%, and 0.8359, respectively, higher than other advanced models used for comparison. It is feasible in the task of high-precision extraction of RRL using VHR UAV images. This research could provide technical support for rural land planning, analysis, and formulation of land management policies.

Keywords: rural residential land extraction; UAV image; semantic segmentation framework; deep learning

Citation: Sha, C.; Liu, J.; Wang, L.; Shan, B.; Hou, Y.; Wang, A. Extraction of Rural Residential Land from Very-High Resolution UAV Images Using a Novel Semantic Segmentation Framework. *Sustainability* **2022**, *14*, 12178. https://doi.org/10.3390/su141912178

Academic Editors: Wei Gao, Yongsheng Li, Lifei Wei and Gwanggil Jeon

Received: 20 June 2022
Accepted: 23 September 2022
Published: 26 September 2022

1. Introduction

Rapid urbanization coupled with socioeconomic transformation has resulted in substantial urban–rural land changes. Consequently, the conflict between human and land relations in rural areas has become increasingly prominent. For farmers and village collectives, RRL is not only a life-sustaining material but also a valuable asset. Changes in RRL can have ecological impacts [1], and their scale and characteristics can reflect rural human–land relations [2,3]. Grasping the area and distribution of RRL cost-effectively, in a timely manner, and precisely, has great research value. Traditional RRL research relies heavily on methodologies such as interviews and field surveys [4,5], which are both costly and inefficient owing to the cumbersome transportation available in rural areas, and remote sensing has the ability to address the limitations of these factors.

At present, methods for extracting targets from remote sensing images can be broadly classified into two types: pixel-based and object-based methods. Maximum likelihood estimation [6,7], and support vector machines [8–10] are the two most frequently utilized pixel-based approaches. Ensemble learning [11] is generally superior to using a single classifier for classification, and random forests as an ensemble method are adopted by

a large number of researchers in remote sensing image processing [12], demonstrating some utility in the classification of LCLU [13]. However, these methods are shallow algorithms that frequently overlook the relationship between deep-level features and data, are prone to overfitting and underfitting during the training process, and are thus unsuitable for processing high-resolution remote sensing images with a wealth of geometric, textural and spatial features.

Object-based image analysis (OBIA) is increasingly frequently utilized [14,15] in land cover categorization. However, because object-oriented classification algorithms segment and classify objects independently of one another, classification accuracy is extremely dependent on the selection of the picture segmentation scale and the segmentation results. When it comes to complicated object classes, the issues of determining the optimal segmentation size and algorithm selection still remain unresolved. Additionally, feature selection frequently necessitates time-consuming manual determination efforts that are difficult to accommodate in reality. In short, these impacting factors mitigate some of OBIA's classification effects. Therefore, it is vital to develop more precise and efficient algorithms for extracting RRL from high-resolution remote sensing images in complicated situations.

Semantic segmentation is a fundamental task in computer vision and a recent topic of research for deep learning algorithms [16,17], as it is capable of predicting the class labels of pixels based on the semantic information given by the image pixels. As a relatively new technique for remote sensing image interpretation [18], it is being utilized to efficiently detect the types of land use and cover [19–21], as well as to extract ground targets such as buildings [22], roads [23], and water bodies [24]. Recent semantic segmentation models frequently adopt Convolutional Neural Networks (CNNs), which have achieved significant advances in classical issues such as speech recognition, picture recognition, and natural language processing due to their robust feature learning and expression capabilities. Although it has been a research hotspot in remote sensing image interpretation and has yielded some achievements [25–30], the typical CNN model has significant limitations, including complex parameters and redundant calculations induced by the addition of relatively more layers.

The fully convolutional neural network (FCN) is an end-to-end model that eliminates fully connected layers, thereby reducing computational complexity and increasing segmentation efficiency. It is currently used in a variety of applications for semantic segmentation of high-resolution remote sensing images [31–35]. However, FCN classifies using only the high-level features extracted from the final convolutional layer, resulting in a loss of spatial information and finer structures in the classification results.

Numerous improved models based on FCN architectures have been proposed recently, including LANet [36], which was constructed by integrating two attention mechanisms into an FCN; MAP-Net [37], which was constructed by combining attention mechanisms and pyramid pooling as two parallel paths; DTCDSCN [38], which was constructed by combining a change detection network and two semantic segmentation networks; and a novel FCN constructed by combining ResNet and an attention module [39]. In the realm of medical picture segmentation, U-Net has demonstrated higher feature extraction and identification performance, and many of its improved solutions [40–43] have been applied by researchers to solve remote sensing issues. In summary, while deep learning algorithms have been extensively studied for processing remote sensing tasks, few studies have focused on land classification and extraction in rural areas, particularly when using Very-High Resolution images such as those captured by UAVs.

RRL has the characteristics of spatial aggregation and co-existence of buildings in different periods and has heterogeneous forms with high intraclass differences that present a huge scale of divergence under high-resolution remote sensing images. Moreover, the typical semantic segmentation approach is unable to cope with such scenarios, and picture edge segmentation accuracy is very poor. DenseNet [44], a convolutional neural network with densely connected structure, was proposed by Huang et al. in 2017, and some progress

in extracting land use information from remote sensing images has been made, for example, using the improved DenseNet network to extract industrial land information from remote sensing images [45]. Li et al. [46] used high-resolution images of GaoFen-1 (GF-1) in the Weiku area of China and an improved DenseNet for the identification of cotton planting land, which performed better than the four widely used classic CNNs, including ResNet, VGG, SegNet and DeepLab v3+. DenseNet can fully exploit global features and perform better with fewer parameters and computational costs when used as a feature extraction network. By dilation convolution, Atrous Spatial Pyramid Pooling (ASPP) may efficiently increase the receptive field, enabling the model to understand multi-scale features, and with certain achievements in practical applications in the field of remote sensing, for example, it can be used for road extraction from remote sensing images [47,48] and for fast and accurate land cover classification [49] on medium-resolution remote sensing images. The PointRend [50] neural network module treats image segmentation as a rendering problem and performs adaptively selected point-based segmentation predictions at adaptively selected locations using an iterative subdivision algorithm, and it has good performance in remote sensing image instance segmentation; for example, the introduction of PointRend in MaskRcnn has improved the accuracy of the edge, and the accuracy and efficiency in emergency remote sensing mapping have been greatly improved compared with traditional methods [51].

The main goal of our research is to achieve fast and accurate extraction of RRL in complex rural scenes, and solve the problem of the overall inefficiency and low accuracy of edge segmentation when existing deep learning methods process remote sensing images. To this end, image data of our research were collected using UAV, semantic segmentation datasets were constructed by manual segmentation and labeling, and an automatic RRL extraction framework based on DenseNet, ASPP and PointRend neural network modules was proposed. For comparison, we used UNet, VGG19, VGG19_bn, ResNet and the framework without PointRend. Three sets of controlled experiments were conducted on the dataset to verify the rationality of the proposed framework.

2. Research Area and Data

2.1. Overview of the Research Area

Sanniangmiao Village is located in Wenyang Town, Feicheng County, Shandong Province, PRC (see Figure 1), at latitude 35°56′–35°57′ N and longitude 116°53′–116°54′ E. With an average annual temperature of 12.9 °C and an average annual rainfall of 659 mm, it has a moderate continental monsoon climate. Feicheng County is located in the northern part of the Western Shandong Anticline and belongs to the third-level tectonic elements, such as Feicheng-Yiyuan Depression, Xinfushan Uplift, and Dawenkou-Mengyin Depression. There are four relatively large regional faults in Feicheng County, and the stratum belongs to the type of Northern China Stratigraphic Deposition. The village is flat and surrounded by the Dawen River in the east and south. It is a typical agricultural community. The soil texture is light sandy loam, with a total land area of 121 hectares, of which 61.3 hectares are cultivated fields. There are 340 families, a total population of 1117, and a resident population of 700. For a long time, the village has developed agriculture relying on the rich waters of the Wenhe River. In recent years, with the advancement of urbanization in China, some residents have moved to cities for employment and residence, leaving their homesteads idle. At the same time, with the implementation of China's Rural Revitalization Strategy, the village has used farmers' homesteads to build B&B (Bed and Breakfast) and develop rural tourism.

2.2. Data Acquisition and Preprocessing

The data mainly include the UAV images of the village, the current land utilization map, the cadastral map and the construction situations of the homestead. The high-resolution images of the research area captured by the UAV at low altitudes are the main data of the experiment. The data of village homestead utilization, housing construction, and

socioeconomic status were obtained from field surveys and interviews, which were used as a reference for the dataset labeling. The land use status map and cadastral map were provided by the natural resources department to learn about the distribution of homesteads and verify the results of homestead extraction (see Figure 2).

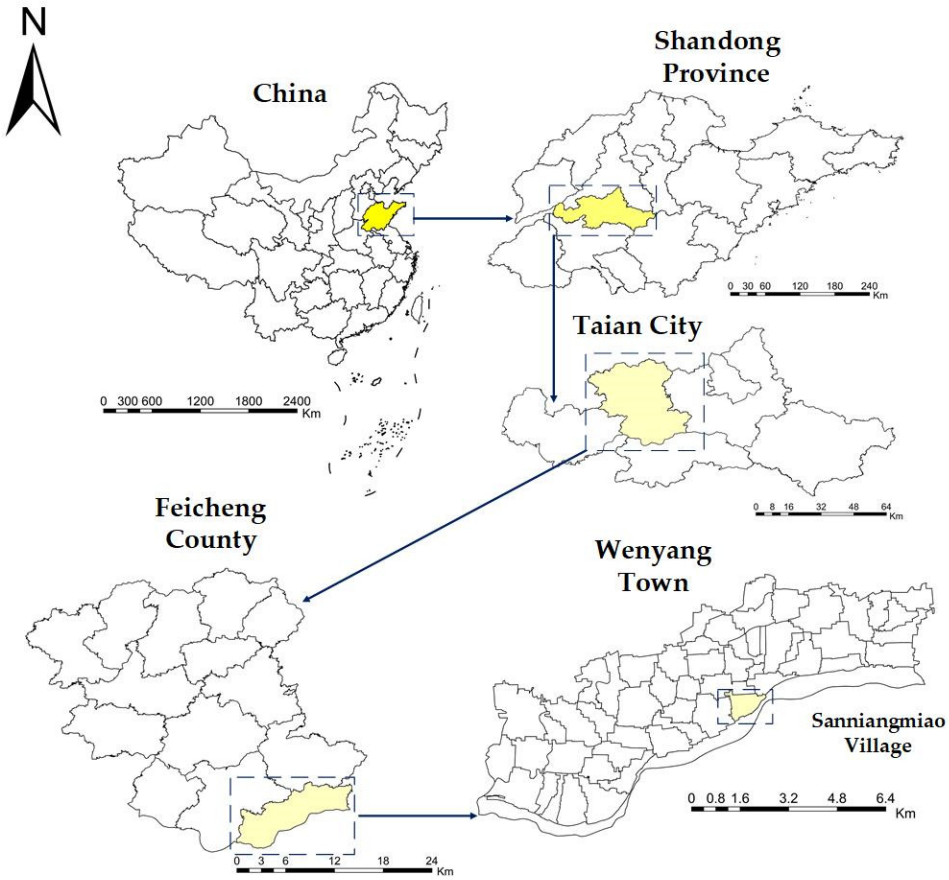


Figure 1. Location of the research area. The pictures are schematic diagrams of China, Shandong Province, Taian City, Feicheng County, and Wenyang Town in order. The research area Sanniangmiao Village is located on the southeast of Wenyang Town.

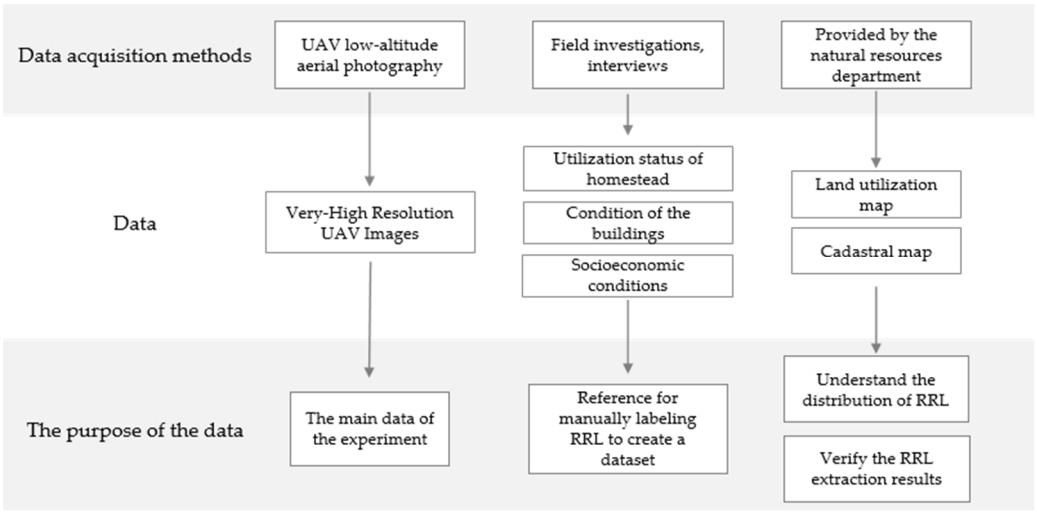


Figure 2. The data, the obtained ways and the purpose. The data were obtained in three different ways and their purposes are represented from top to bottom.

The DJI Phantom 4 Pro V2.0 drone was used to take low-altitude photographs and capture VHR images in Sanniangmiao Village. The weather conditions were considered good and there was no constant wind direction at the time of filming, which was less affected by the environment. The camera lens was kept vertically downwards, the flight altitude was set at 100 m, the course speed was 5 m/s, the photo interval was 2.8 s, the overlap rate was 70% in the side direction and 80% in the heading. There were 635 RGB three-channel, 5472 × 3648 resolution images collected. In addition, Xiaowang Village and Zhangling Village in Wenyang Town were selected as the test areas for the effective verification of the extraction framework.

Due to unavoidable factors such as the unstable flight altitude of the UAV and the optical distortion of the mounted sensors during the acquisition process, pre-processing of the images was required after the acquisition was completed. It included initialization of the image, ortho-rectification, aerial triangulation, point cloud and texture processing to obtain an orthophoto image including all the RRL of the study village.

For the facilitation of training, validation and testing of the model, the residential land was divided into the training area and the test area (see Figure 3). Fully considering the characteristics of rural homestead distribution, building types, pixel categories and other characteristics of the research area, the areas with large and representative pixel category coverage were selected as training areas, and the remaining areas were selected as test areas.

Figure 3. Preprocessed image: Sanniangmiao Village, Xiaowang Village, and Zhangling Village are depicted in the three images, divided into a training area and a test area. The training area is used to construct the semantic segmentation dataset for training the models, and the test area is used to test the performance of the trained models.

2.3. Construct Semantic Segmentation Dataset

RRL is the collective construction land used by farmers for building houses, dwellings and supporting agriculture, and is the main land type in villages. Before constructing the data set, it is necessary to judge whether it belongs to RRL according to the field situation. The land use types of the village land in the study area include homesteads, public land, streets, green spaces and open spaces (see Table 1). Most of the homesteads have been judged as RRLs. However, there are still some areas where the houses have collapsed, and the courtyards have been idle for a long time and are overgrown with plants that no longer have a residential function and are no longer judged as residential land. Public land includes educational land and public facility land. Because of the houses and courtyards, it is indistinguishable from residential land in the image, and against the background of the merger of rural schools, there are few schools in rural areas. Moreover, this study classifies public land as residential land. Furthermore, streets, green spaces, and open spaces are not considered as RRLs. In conclusion, the RRL in this study refers to courtyard land that has buildings and is being used, including public land such as homesteads and village committees in use, while the idle homesteads, streets, green spaces, and open spaces without buildings are not regarded as the RRLs.

Table 1. Land use classification of RRL in the study area.

Classification of Land Use	Whether to Determine as RRL
homestead in use	✔
homestead without residential function	✗
educational land·	✔
public facility land	✔
street land	✗
green spaces	✗
open spaces	✗

The residential buildings in the study area are of different types such as red tile, gray tile, and colorful steel tile buildings as well as concrete roof buildings (see Figure 4) with concrete floors in the courtyard. Therefore, RRL has different semantic features. Due to the fact that there is no public dataset for rural residential scenes, this study identified and labeled the residential land and non-residential land in the images according to the features of the study area. The labeled images are segmented into 256 × 256 pixels, and the images are segmented in steps of 128 pixels along with the X and Y directions, that is, 50% overlap. To maintain the quality of deep learning samples and prevent overfitting during model training, data enhancement is performed by mirroring, rotating and flipping, and a total of 7555 sample images and corresponding labels are obtained to construct the semantic segmentation dataset. The data are randomly selected in the ratio of 8:2 to distinguish the training set and validation set, with the training set used to train the model and the validation set used to evaluate the model.

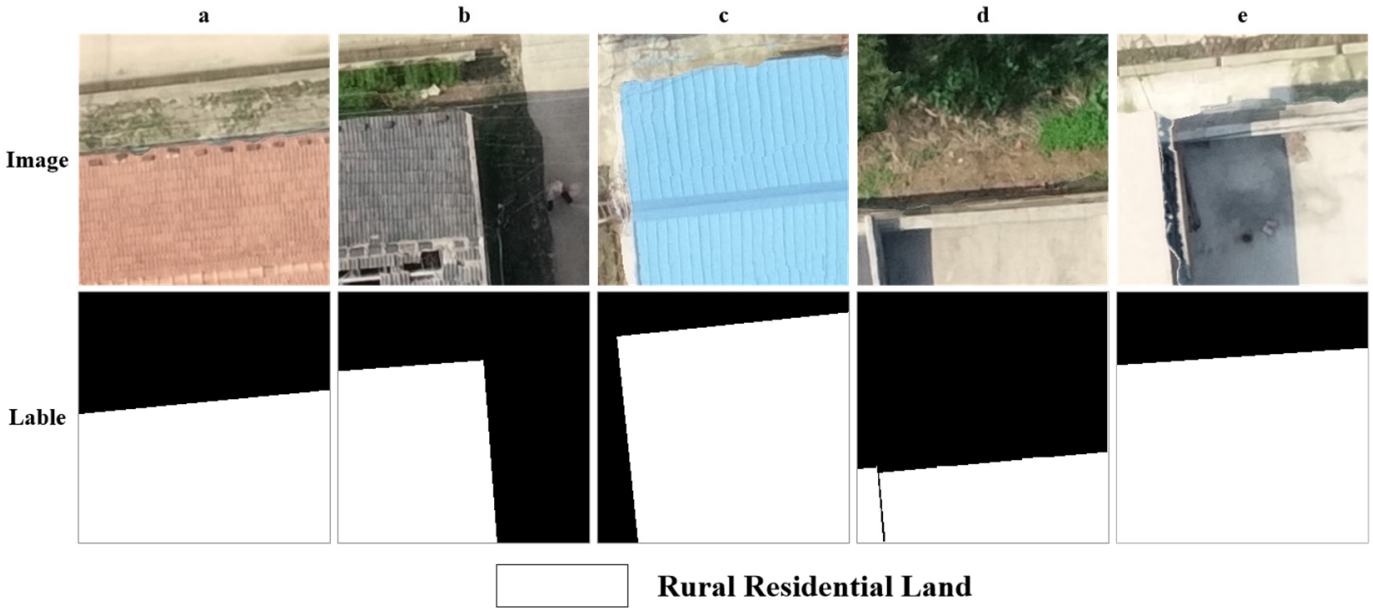


Figure 4. Examples of images and labels for the training set, which is manually segmented and labeled, from left to right: (**a**) buildings with red tiles; (**b**) buildings with grey tiles; (**c**) buildings with painted steel tiles; (**d**) buildings with cement roofs; (**e**) courtyards. The dataset is generated from the training area. There are 7555 sample images and corresponding labels, respectively, and each of them is 256 × 256 pixels.

3. Methodology

The extracted object RRL is a complex of houses and courtyards. In the VHR UAV images, the spatial distribution is clustered, the appearance of houses and courtyards is characterized by a rectangular shape, the building materials have regional clustering, while there are complex impermeable layers and natural ground surface, the interference of tree shadows and other complex backgrounds. It exhibits features at different scales with

rich spatial and semantic relationships. It requires the extraction model to have a strong feature-processing capability.

3.1. The Framework for Extraction

3.1.1. Architecture of the Framework

In this study, we constructed a Cascaded Dense Dilated Network (CDD-Net), using DenseNet, which can utilize global features, such as the feature extraction network, the ASPP module to handle multi-scale features, and add the PointRend module to handle edge features on top of that to further improve the extraction accuracy of RRL (see Figure 5). Firstly, the input original image is passed through the feature extraction network DenseNet. Each layer in DenseNet consists of DenseBlock, convolution, and pooling and the output of each layer is stitched together and used as the input of the ASPP module. In ASPP, the input feature maps are sampled in parallel with dilated convolution of different rates, and the obtained results are concatenated together to expand the number of channels, and then the number of channels is reduced to the original size by 1×1 convolution. In PointRend, the edge points are selected so that they can be predicted independently. Then, the output result is achieved.

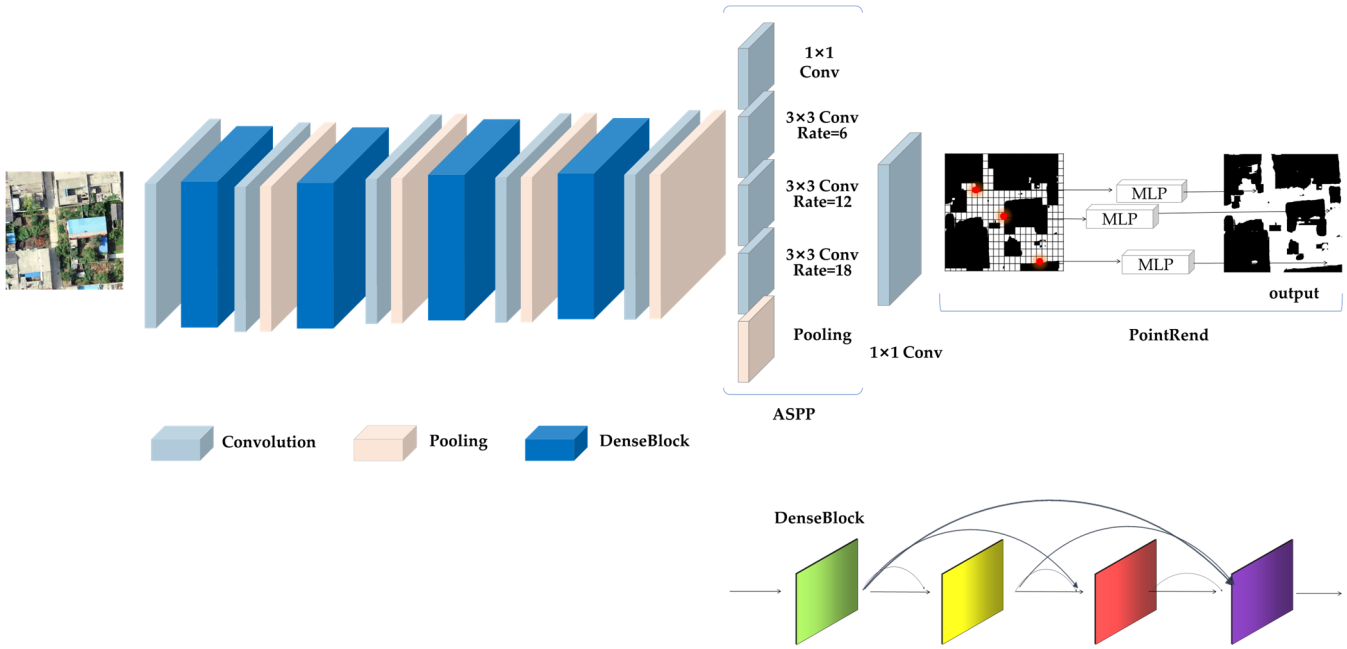


Figure 5. The architecture of CDD-Net. The whole framework is a cascaded structure, followed by DenseNet, ASPP, and PointRend. The DenseNet is the feature extraction network of the framework, consisting of convolutional layer, pooling layer and DenseBlock. Each layer in DenseBlock is connected to all previous layers to deal with global features. There are four dilated convolutions in ASPP to handle multi-scale features. PointRend improves the accuracy of edge segmentation by performing predictions for points at the edge of the image through the independent MLP.

3.1.2. Feature Extraction Network: DenseNet

The DenseNet network has fewer parameters and lower computational costs compared to networks such as ResNet (see Table 2). The computational efficiency is higher for the same computer memory. The DenseNet network structure consists of several DenseBlocks, where each layer within a block is connected to all previous layers and used as input to the next layer, thus enabling feature superposition and improved efficiency. A transition layer exists between every two blocks, and the size of the feature map is changed by convolution and pooling.

Table 2. The number of parameters for different feature extraction networks.

Method	Params (M)
DenseNet121	6
resnet50	23
vgg19_bn	20
vgg19	20
Xception	22

Based on the ImageNet data set.

The conventional feed-forward convolutional neural network Resnet uses the output of the Lth layer as the input to the Lth + 1st layer, which can be expressed as:

$$X_i = H_i(X_{i-1}) + X_{i-1} \tag{1}$$

where H_i denotes a nonlinear transformation. The output of layer i is the sum of the output of layer $i-1$ and the nonlinear transform of the output of layer $i-1$. In contrast, the DenseNet, where any layer is connected to all its subsequent layers, and the input of any layer also contains the outputs of all its previous layers, can be expressed as:

$$X_i = H_i([X_0, X_1, \ldots, X_{i-1}]) \tag{2}$$

where $[X_0, X_1, \ldots, X_{i-1}]$ can be regarded as the stitching of the layer 0, 1, ..., $i-1$ feature map, and H_i denotes the composite function of three consecutive operations, which are, respectively, batch normalization (BN), Rectified linear unit (ReLU), and 3×3 convolution.

3.1.3. Dilated Convolution and ASPP

The differences in scale exhibited by RRL in space at VHR images are extremely large, requiring models with excellent multi-scale feature-processing capabilities.

The pooling layer in a convolutional neural network performs a down sampling operation on the input image. Moreover, continuous pooling reduces the resolution of the output image and loses feature information. Compared with it, the dilation convolution can enlarge the receptive field without increasing the computation and network parameters, which is beneficial to extract multi-scale information. The dilation rate controls the size of the receptive field, and a dilation rate of 1 is equivalent to the standard convolution.

The ASPP module uses convolution kernels of different sizes, that is, sampling with different rates of dilated convolution, and adds a batch normalization layer to better capture the multi-scale features of RRL. The ASPP module contains a total of five branches, each with 256 channels. It includes one 1×1 convolution and three 3×3 dilation convolutions (rates of 6, 12, and 18, respectively), with different rates of dilation convolutions to handle features at different scales, and a global average pooling layer to enhance feature acquisition. Then, the channel dimensions of all output layers are superimposed and 1×1 convolved to obtain the output feature map. Finally, an image of the same size as the actual one is gained by 1×1 convolution.

3.1.4. PointRend Module

The boundary information of RRL is very important for the extraction results. Semantic segmentation models usually suffer from the problem of under sampling and insufficient utilization of superficial features, which leads to too sparse features in the boundary regions and loss of important detail information. To improve the utilization of image edge features, the PointRend module is introduced.

PointRend treats image segmentation as a rendering problem, using an iterative segmentation algorithm to perform point-based segmentation prediction at adaptively selected locations. First, a coarse prediction is performed, that is, points with high class uncertainty are selected and predicted independently for each of these points by a multilayer perceptron (MLP). Then, the accuracy of the classification of individual uncertain points is improved

by continuous internal iteration. The gradual increase in the number of uncertainty points will improve the resolution of the final result.

3.2. Accuracy Evaluation Method

3.2.1. Evaluation Metrics for the Model Training Phase

The extraction of RRL in this study is a binary classification problem. The binary classification problem often uses *Precision*, *Recall*, *F*1 score as evaluation metrics to evaluate the extraction effect. The *Precision* and *Recall* can be calculated as

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

where *F*1 score can be represented as:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

P (Positive), and *N* (Negative) represent the prediction results of the model; *T* (True) and *F* (False) evaluate whether the prediction results of the model are correct, that is, *TP*—True Positive, indicating correct detection as true; *FP*—False Positive, indicating the wrong detection as true (not true); *FN*—False Negative, not detected as true, but true (wrong detection result). Overall Accuracy (OA) can be obtained by

$$\text{OA} = \frac{TP + TN}{TP + +TN + FP + FN} \tag{6}$$

Since the distribution of RRLs in the study area is extremely aggregated, resulting in an uneven distribution of sample categories in the training data set, with far more RRL samples than non-RRL samples, Dice loss is applied in the case of positive and negative sample unevenness and is therefore chosen to be used as the loss function.

The *Dice* coefficient is a metric often used for image segmentation. It is a measure of the similarity of a set and is usually used to calculate the similarity of two samples with a value threshold of (0, 1). The *Dice* coefficient and *DiceLoss* are defined by

$$Dice = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{7}$$

$$DiceLoss = 1 - \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{8}$$

3.2.2. Evaluation Metrics for the Model Testing Phase

The OA, Kappa coefficient, is used as an evaluation metric to test the accuracy of the actual results of the model to extract RRLs. The overall classification accuracy indicates the probability of how many pixels are correctly classified in this classification. The Kappa coefficient indicates the percentage reduction of errors generated by the evaluated classification over the completely random classification, which can be calculated as

$$Kappa = \frac{P_{OA} - \frac{\sum_{i=1}^{n} a_i * b_i}{N^2}}{1 - \frac{\sum_{i=0}^{n} a_i * b_i}{N^2}} \tag{9}$$

where: P_{OA} is the overall classification accuracy; a_i is the true sample pixel count of n-class features; b_i is the number of pixels in the sample classified as n-class features; n is the total number of categories of classification results; N is the total number of sample pixels.

3.3. Transfer Learning

Because there are a large number of parameters in the deep learning model that need to be trained, to ensure that the trained model has high accuracy and excellent generalization ability, it is generally necessary to meet a sufficient number of training samples, but when faced with specific problems in a certain field, it is often difficult to obtain the required scale of data, which cannot meet the needs of training models. The ImageNet dataset [52] is used for testing algorithm performance in the fields of classification, recognition, and detection of images [53,54], and has been the most used open large dataset in the field of deep learning in recent years, including more than 14 million images and more than 20,000 categories, of which more than a million images have clear category labeling and object location labeling; it is widely used to perform deep learning model pre-training of deep learning models. Due to the relatively small scale of the RRL dataset constructed for training in this study, the transfer learning strategy is used, that is, the feature extraction network is pre-trained on the ImageNet dataset before training on the RRL dataset to get the weights and parameters of the model.

4. Results and Analysis

4.1. The Results of Training

The training results are evaluated by two metrics, loss and OA. The whole experiment is based on the PyTorch deep learning framework of python 3.8, using NVIDIA GTX1050Ti GPU for the training of the model. All training processes use the same data set. Combining the size of the dataset and the complexity of the model, the epochs are set to 30, and the batch are set to 8. For the 7555 image sample data set, the training set and the validation set are randomly selected at a ratio of 8:2. The loss curve and the accuracy curve during the training process are compared by observing each model for the purpose of being used as the basis for evaluating the performance of the model.

The steps are divided into three parts. Firstly, the ASPP is compared with the classic semantic segmentation model UNet, and the validity of ASPP is verified on the premise that ResNet is used as the feature extraction network. Then, in order to verify the rationality of using DenseNet in the extraction framework, under the premise of using ASPP, the DenseNet, VGG19, VGG19_bn, and ResNet are used as feature extraction networks for training and verification, and the accuracy of DenseNet and the other three models is compared. Finally, the complete extraction framework CDD-Net is compared with the model without using PointRend to check whether PointRend can effectively improve the performance of the model. The specific results of the training phase are as follows:

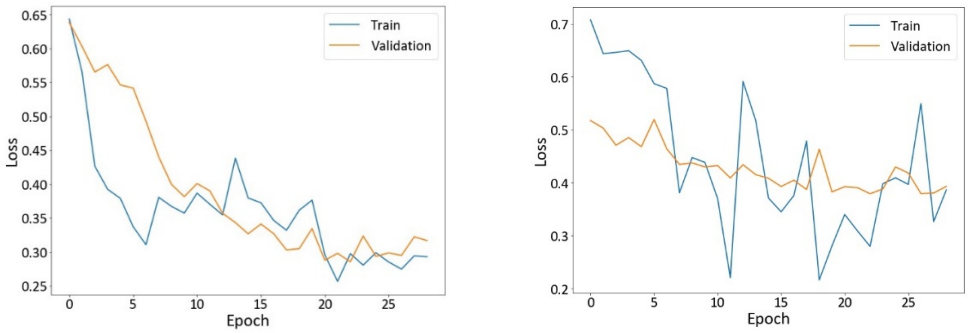
4.1.1. Loss Curves

The overall loss curve and the final loss value both demonstrate the effectiveness of the usage of ASPP in the RRL extraction framework. Figure 6 shows the loss curves of ASPP and U-Net for training and validation, and both models use ResNet as the feature extraction network. The decline of ASPP is relatively large at the beginning of training, indicating that the learning rate is reasonable and in the process of gradient descent. After the 20th epoch, the loss curve declines slowly and the loss change is not as obvious as at the beginning, and the loss curve becomes smooth after about the 25th epoch, which means that the model is no longer improved and the training is completed. The loss curve of U-Net is not stable enough; although the loss curve of the training set and the test set as a whole are still in a decreasing trend, the fluctuation of the loss curve between each epoch is much larger. There may be two reasons for the phenomenon: Firstly, these training parameters are not the best for U-Net. Secondly, it requires more epochs to achieve a similar effect to ASPP. The final training set loss of U-Net is 0.3867 and the validation set loss is 0.3931; the final training set loss of ASPP is 0.2928 and the validation set loss is 0.3167.

The DenseNet, as the feature extraction network, performs better compared with the other methods. Figure 7 shows the loss curves for training and validation using DenseNet, ResNet, VGG19, and VGG19_bn as the extraction networks, respectively. By comparison,

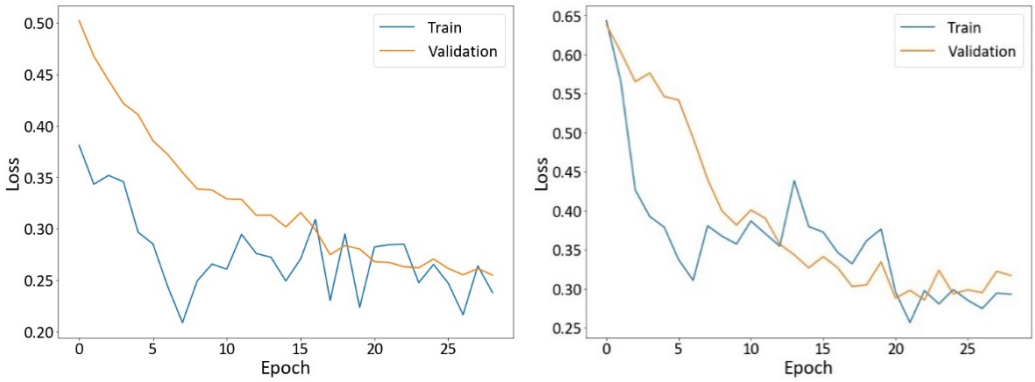
we can find that using DenseNet as the feature extraction network has the best performance, the loss decreases quickly from the beginning of training to before the 10th epoch, and the loss of the validation set is slightly higher than the loss of the training set, indicating that the learning rate is reasonable and in the process of gradient decrease, after which the loss curve decreases slowly, indicating that the model is in the process of slow improvement. About after the 25th epoch, the loss curve becomes smooth, indicating that the training is almost completed. The final losses of DenseNet are 0.2377 for the training set and 0.2549 for the validation set, which are both lower than other models. When using ResNet as the feature extraction network, the final training set loss is 0.2928 and the validation set loss is 0.3167; both loss values are higher than DenseNet and ResNet. When VGG19 is used as the feature extraction network, the final training set loss is 0.4512 and the validation set loss is 0.4646. The effect of using VGG19 after batch normalization is improved, the final training set loss is reduced to 0.3917 and the validation set loss is reduced to 0.4287.

The model is not only easier to train but also performs better after being improved by PointRend. Figure 8 shows the loss curves of the model before and after adding the PointRend module. It can be seen that the loss functions of the two models eventually converge and the models have completed training. PointRend makes the loss curve smoother, indicating that the model performs better in the training process, the final training set loss is reduced from 0.2377 to 0.1466, and the loss in the validation set is reduced from 0.2549 to 0.2320. The convergence status of the loss curves and the loss value are better compared with all the compared methods.



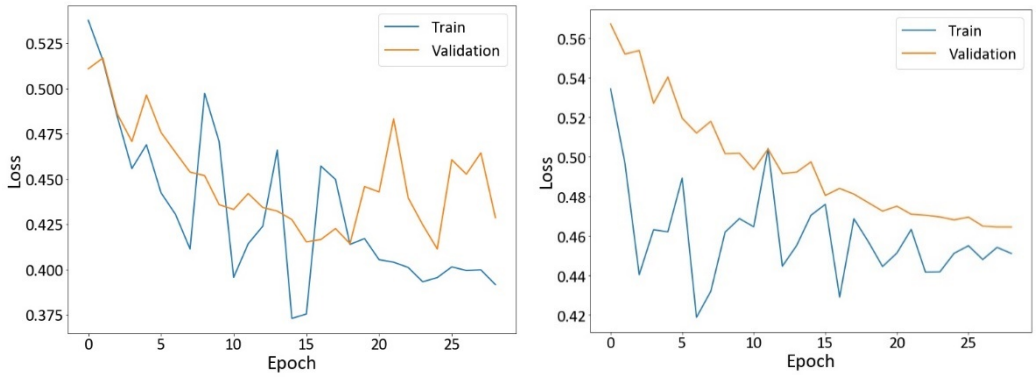
(**a**) (**b**)

Figure 6. Loss curves of ASPP and U-Net: (**a**) ASPP; (**b**) U-Net. The horizontal axis represents the number of epochs, and the vertical axis represents the loss value. The loss curve in (**a**) first decreases rapidly, then the speed becomes slow, and the loss curve in (**b**) is in a downward trend as a whole, and the loss value fluctuates greatly between each epoch.



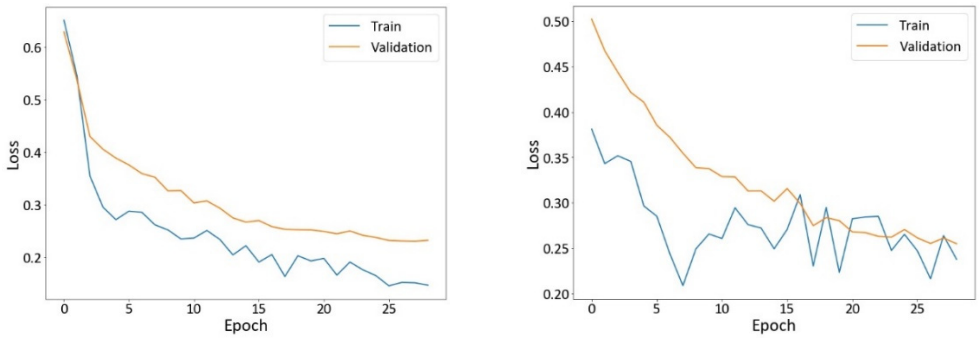
(**a**) (**b**)

Figure 7. *Cont.*



(**c**) (**d**)

Figure 7. The loss curves utilizing several feature extraction networks in the model, from left to right, are: (**a**) DenseNet; (**b**) ResNet; (**c**) VGG19_bn; (**d**) VGG19. The horizontal axis represents the epoch number, and the vertical axis represents the loss value. The loss curves in (**a**,**b**) both decrease rapidly at first and then slow down, but the loss value in (**a**) is smaller. The loss curves of (**c**,**d**) show a certain degree of oscillation.



(**a**) (**b**)

Figure 8. Loss curves before and following the addition of PointRend, respectively: (**a**) after the addition; (**b**) before the addition. The horizontal axis represents the number of epochs, and the vertical axis represents the loss value. The loss curves all show a decreasing trend and both of them fall rapidly, and then the downward trend slows. The loss curve of (**a**) is smoother.

4.1.2. OA

CDD-Net obtained the best accuracy rate, indicating that DenseNet, ASPP and PointRend can effectively improve the OA of the model. With the progress of model training, the accuracy rates of all curves are in an increasing trend, and they all improve rapidly at first and then slow down (see Figure 9).

The two curves of the graph in part (a) express the OA changes of ASPP and U-Net, and it can be seen that after the 5th epoch, ASPP is more effective than U-Net for OA enhancement. When the training proceeds to the 30th epoch, the OA of ResNet+ASPP rises to 87.27%, which is 2.87% higher than that of ResNet+U-Net under the same condition. This shows that the improvement of accuracy by dilation convolution and pyramidal pooling is significant. Part (b) characterizes the accuracy curves of the four different feature extraction networks during training. It can be found that the effect of the difference of feature extraction networks on the improvement of OA starts to become significant from the third epoch, at which time the accuracy of DenseNet improves to 76.97%, which is 1.44%, 2.03% and 5.28% higher than the accuracy of ResNet, VGG19_bn and VGG19, respectively, and the accuracy of DenseNet is higher than that of the remaining three networks. At the 8th epoch, the accuracy of DenseNet improved to 83.75%, which was 1.32%, 5.87% and 10.2% higher than that of ResNet, VGG19_bn and VGG19 under the same conditions,

respectively. After the eighth epoch, the growth of accuracy starts to slow down, and the final OA of DenseNet is boosted to 89.19%, which is 1.92%, 7%, and 11.74% higher than ResNet, VGG19_bn, and VGG19, respectively. The accuracy improvement proves that DenseNet can effectively improve the performance of the model. Part (c) expresses the changes in the accuracy of the model before and after the improvement using PointRend. After the 10th epoch, the improvement of OA of the model by PointRend starts to become significant. At this point, the OA of the model improves by more than 1% compared to that of the model without PointRend. The final OA is 90.51%, and PointRend improves the OA predicted by the model by 1.32%.

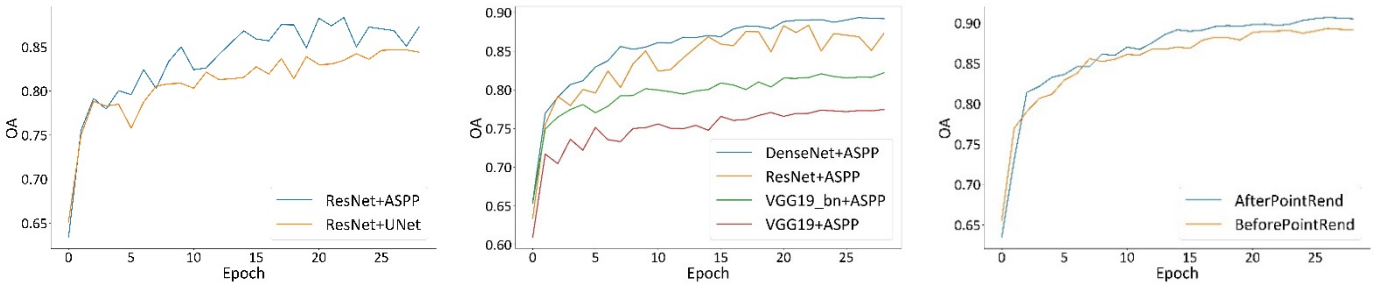


Figure 9. The OA curves for each model in the three sets of comparative tests, from left to right, are as follows: (**a**) ASPP and U-Net; (**b**) models utilizing four different feature extraction networks; (**c**) models before and following the addition of PointRend. The horizontal axis represents the number of epochs, and the vertical axis represents OA. The curves of all models increase rapidly and then tend to slow down until they become stable. The model using PointRend in (**a**) achieves the higher OA, the model using DenseNet in (**b**) has the highest OA at every Epoch, and the model using ASPP in (**c**) has higher OA than the classical structure UNet.

CDD-Net achieved the best extraction accuracy in distinguishing RRL and background, indicating that DenseNet, ASPP and PointRend are all effective in improving the OA of the model, further reflecting the rationality of the CDD-Net design (see Table 3).

Table 3. The ultimate OAof all models.

Method	OA (%)
CDD-Net	90.51
DenseNet + ASPP	89.19
ResNet + ASPP	87.27
VGG19_bn + ASPP	77.45
VGG19 + ASPP	82.19
ResNet + UNet	84.40

4.2. Comparison of the Framework with Other Approaches

We compared the framework with other methods through qualitative comparison and quantitative comparison. Qualitative comparison refers to evaluation by discussing the visual performance of different scenes, and quantitative comparison refers to evaluation by accuracy metrics, such as *Precision*, *Recall*, *F*1 score, and Dice coefficient.

4.2.1. Qualitative Comparison

Among all the methods, CDD-Net could not only easily extract large-sized regions but also correctly distinguish RRL and the background in small-sized regions. Figure 10 presents the visual representations of four different scenes under high-resolution UAV images.

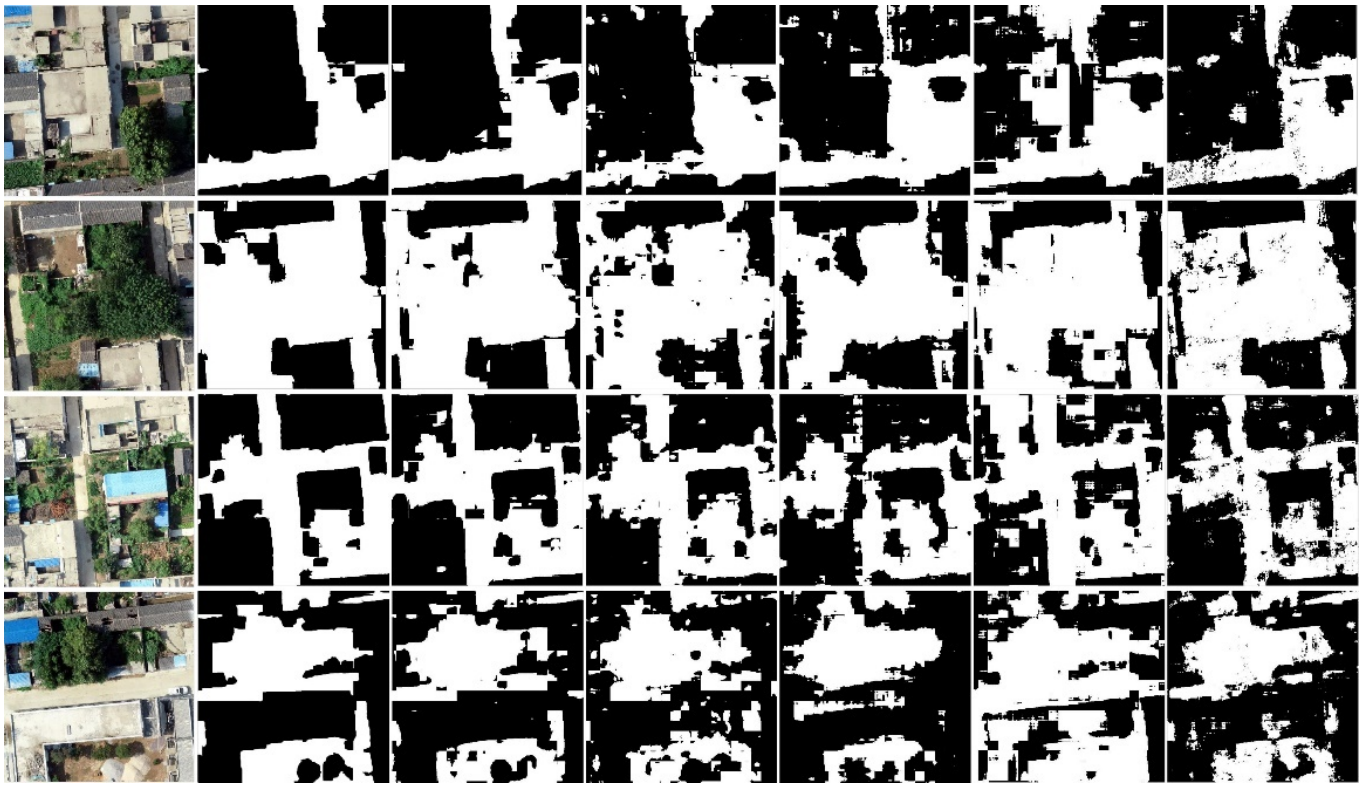


Figure 10. Examples of extraction results of different models; black represents RRL and white represents background. Four different scenarios are represented from the top to the bottom: the first row is the RRL with a large-scale range. The second row represents the scene with a large range of plants in the target. The third is the RRL in the case of large internal differences and serious planting disturbance. The fourth is the RRL in the case where the differences between classes are very small, and the target image contains visually highly similar cement roads and buildings. From left to right are: (**a**) original images; (**b**) CDD-Net; (**c**) DenseNet + ASPP; (**d**) VGG19_bn + ASPP; (**e**) VGG19 + ASPP; (**f**) ResNet + ASPP; (**g**) ResNet + UNet.

By comparing Figure 10f,g, we can find that the effect of ASPP is better than that of U-Net, and the number of noise points is much reduced compared to U-Net. Although U-Net recognizes most of the RRL pixels, the results have problems such as blurred boundaries, unclear classification of RRL and background, etc. In the VHR UAV images, the lack of processing capability of multi-scale features leads to a large number of disconnections and noises in the extraction results, which also confirms the characteristics of ASPP. Owing to the feature extraction network Resnet cannot use global features, the extraction effect of RRL targets in the face of large-scale is not satisfactory, and in the case of small differences between categories, RRL cannot be well distinguished from the background, and more RRL are incorrectly categorized as background.

By comparing Figure 10c–f, we also find that when DenseNet is used as the feature extraction network, the extraction results show stronger recognition accuracy for RRLs and backgrounds, and almost all large-size RRL pixels are correct and completely detected, which leads to the conclusion that DenseNet can enhance global feature perception capability. However, the accuracy of small-scale RRL extraction is slightly lacking, and the recognition ability of boundaries and contours is poor.

By comparing visual performance before and after the utilization of PointRend, we also find that the improvement of the boundary segmentation effect of PointRend is significant. The contours of RRL have been shown more correctly in the result, which proves the effectiveness of using PointRend in the framework.

Overall, the CDD-Net, combined with ASPP, DenseNet and PointRend, has achieved superior visual results. Although other methods can reasonably extract some RRL information, there are still a considerable number of incorrect extractions of RRL and misidentifications of the background.

4.2.2. Quantitative Comparison

Figure 11 displays the evaluation metrics of all methods. In comparison to other approaches, CDD-Net has shown excellent results across the board and has the highest RRL extraction accuracy.

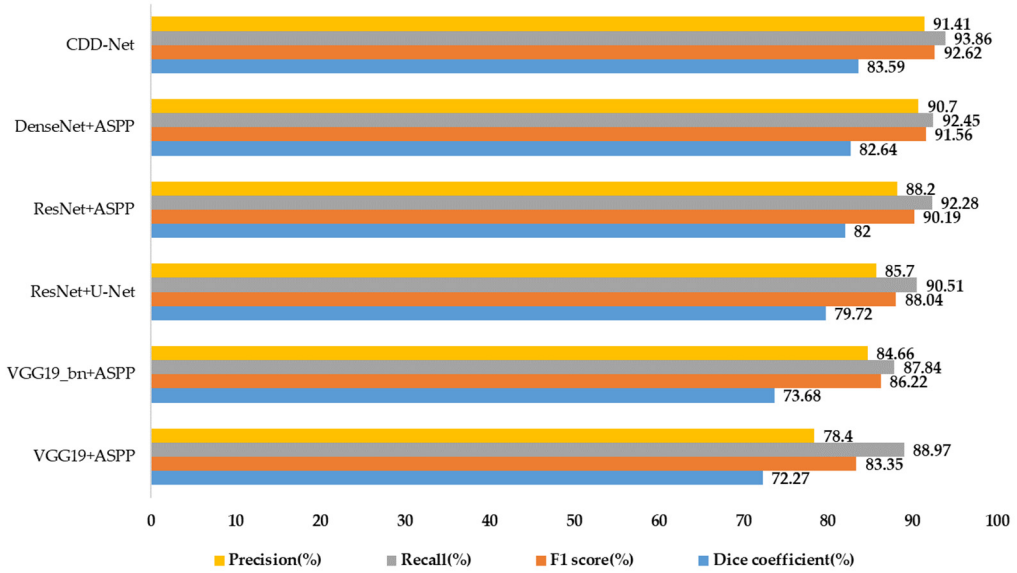


Figure 11. Using the Precision, Recall, F1 score, and Dice coefficient as the evaluation metrics of the model training results. The CDD-Net and DenseNet + ASPP express the differences between the models before and after utilizing PointRend. The DenseNet + ASPP, ResNet + ASPP, VGG19_bn + ASPP, and VGG19 + ASPP express the differences for models constructed using various feature extraction networks. The ResNet + ASPP, and ResNet + U-Net express the differences using ASPP in the model.

Under the conditions of using the same feature extraction network ResNet, if the experimental environment and parameters are the same, by comparing the accuracy evaluation results of ASPP and the referencing model U-Net, we find that compared with the traditional segmentation network U-Net, ASPP has 2.50%, 1.77%, 2.15%, and 2.28% higher Precision, Recall, F1 score, and Dice coefficient, respectively. This is attributed to the model performance improvement due to the expansion of receptive field by dilated convolution and the ability to process multi-scale features, which proves the effectiveness of ASPP.

By comparing the differences in the accuracy metrics of the different feature networks, we also find that when DenseNet is used as a feature extraction network, all evaluation metrics have the highest values. The Precision is 90.70%, which is 2.5%, 6.04%, and 12.3% higher than ResNet, VGG19_bn, and VGG19, respectively. The Recall is 92.45%, which is 0.17%, 4.61%, and 3.48% higher than ResNet, VGG19_bn, and VGG19, respectively. The F1 score is 91.56, which is 1.37%, 5.34%, and 8.21% higher than ResNet, VGG19_bn, and VGG19, respectively. The Dice coefficient is 82.64%, which is 0.64% 8.96%, and 10.37% higher than ResNet, VGG19_bn, and VGG19 respectively. They have all proved that the feature reuse capability of DenseNet is effective in improving model performance.

Interestingly, by comparing differences in the accuracy metrics before and after the improvement of PointRend, we find that when PointRend is added to the model, all metrics show a certain degree of improvement. The Precision increases by 0.71%, the Recall increases by 1.41%, the F1 score increases by 1.06%, and the Dice coefficient increases by

2.95%, which confirms that PointRend can not only optimize boundary segmentation but also improve the model accuracy.

4.3. The Results of Testing

Three research areas, Sanniangmiao Village, Xiaowang Village, and Zhangling Village, were chosen as test areas to verify the universality of this method. As shown in Figure 12b, Figure 13b, Figure 14b, new RRL distribution maps were drawn based on the utilization status of homestead and condition of the buildings obtained by the field investigations. The qualitative analysis was completed by comparing the visual performance of the original images, field investigation results and the results extracted by different methods. For the quantitative analysis part, accuracy evaluation points were created in the test area by a stratified random method, and the OA and Kappa coefficients of the different methods were calculated using the field investigation results as the evaluation standard.

Figure 12. The original images, the results of the field investigation and the extraction results of different methods in Sanniangmiao Village, respectively: (**a**) original image; (**b**) results of field investigation; (**c**) extraction results of CDD-Net; (**d**) extraction results of ResNet + ASPP; (**e**) extraction results of ResNet + UNet.

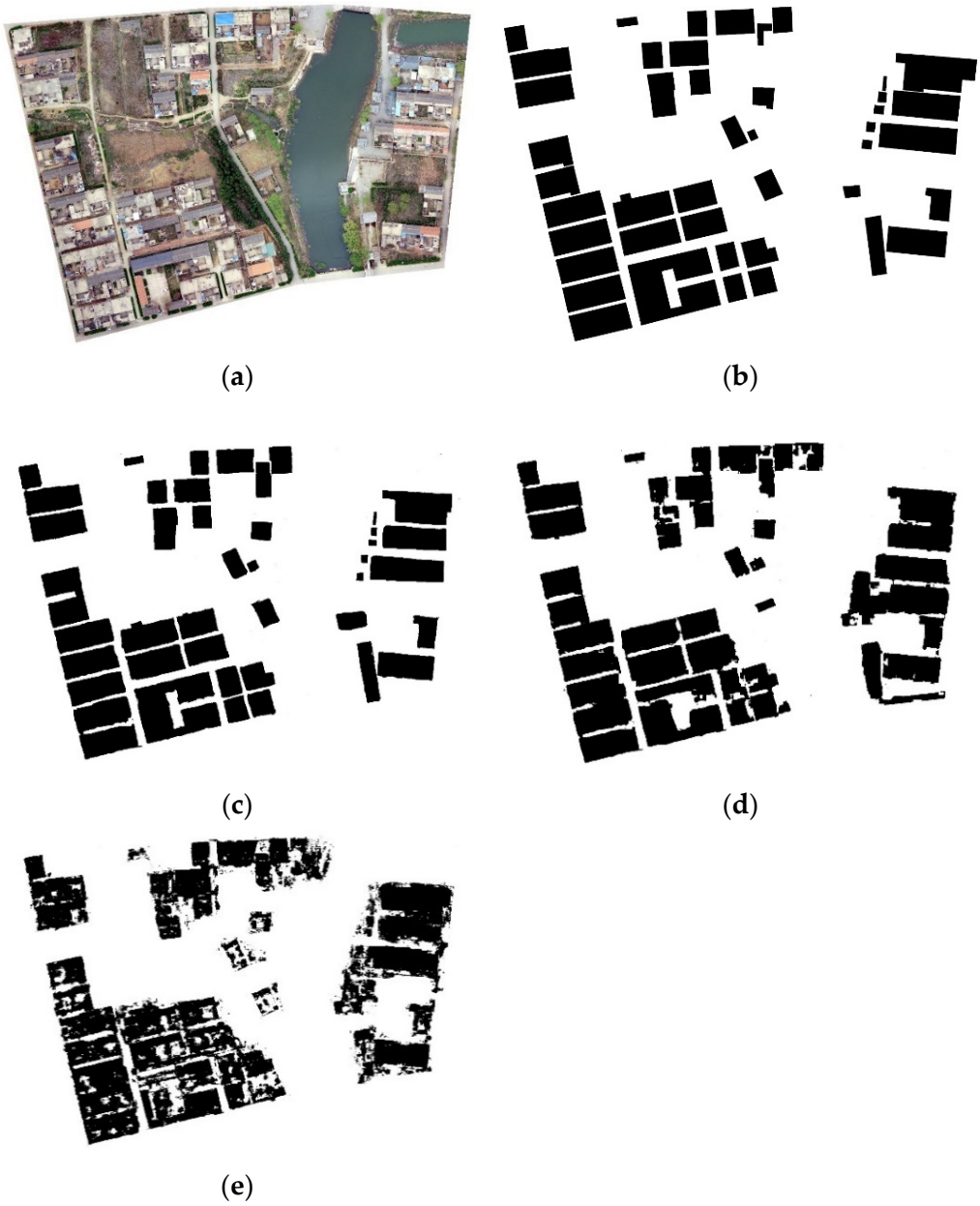


Figure 13. The original images, the results of the field investigation and the extraction results of different methods in Xiaowang Village, respectively: (**a**) original image; (**b**) results of field investigation; (**c**) extraction results of CDD-Net; (**d**) extraction results of ResNet + ASPP; (**e**) extraction results of ResNet + UNet.

Figures 12–14 show the real conditions of several test areas and the extraction results of different methods. It can be seen from the local situation that due to the neat arrangement and close connection of the houses, the semantic segmentation model treats these plots as a whole. In the three different test areas, the CDD-Net's extraction performance is the best, almost all RRL ranges can be identified. The phenomena of missing or wrong extraction can be avoided, and the boundary is also the most accurate. In comparison, ResNet + UNet has the worst integrity and the most noise in the results. Although the integrity of ResNet + ASPP is better than that of ResNet + UNet, there are omissions in some detection inside RRLs, and the extraction of contours is not accurate enough. The main problem is that the courtyards and trees among several buildings cannot be accurately classified. In conclusion, the CDD-Net has given the most complete and accurate building extraction results.

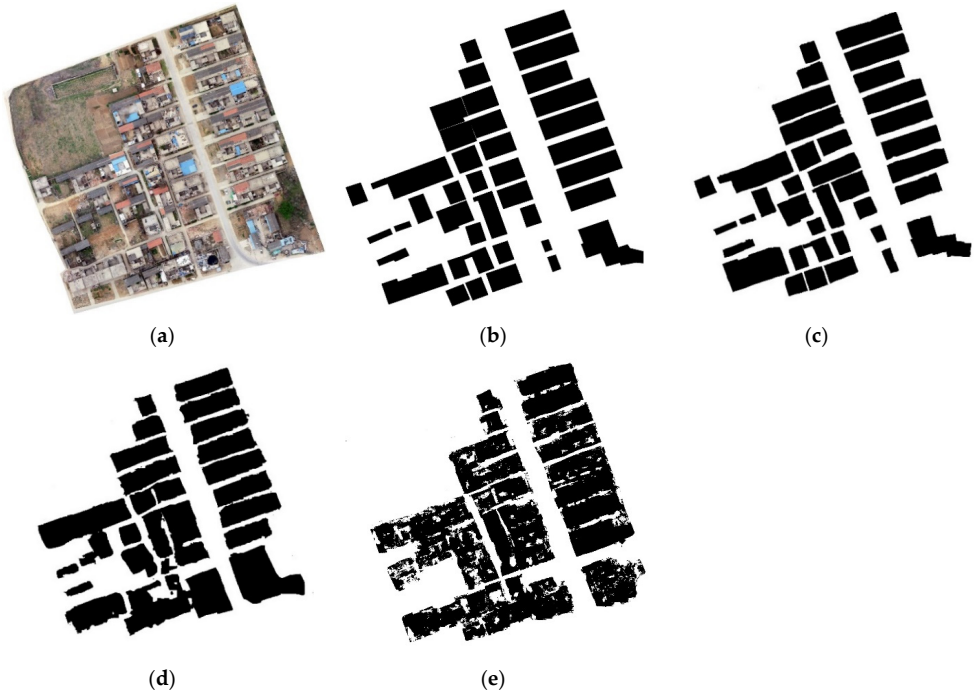


Figure 14. The original images, the results of the field investigation and the extraction results of different methods in Zhangling Village, respectively: (**a**) original image; (**b**) results of field investigation; (**c**) extraction results of CDD-Net; (**d**) extraction results of ResNet + ASPP; (**e**) extraction results of ResNet + UNet.

Figures 15–17 show the quantitative evaluation indicators of the extraction results of different models in the three test areas. It can be seen intuitively that the OA value of CDD-Net in all areas is higher than 95%, and the Kappa coefficient in all areas is higher than other methods. In Sanniangmiao Village, compared with ResNet + ASPP and ResNet + UNet, the OA value of CDD-Net is 1.9% and 3.6% higher, and the Kappa coefficient is 4.9% and 8.7% higher. In Xiaowang Village, the CDD-Net has a 0.6% and 2.3% higher OA value than the other two methods, and the Kappa coefficient is 3.4% and 6.8% higher. In Zhangling Village, the OA of the CDD-Net is 1.8% and 3.9% higher than the other two methods, and the Kappa coefficient is 4.5% and 6.8% higher. In conclusion, the results in different test areas show that the method proposed in this study has a significant performance, which further proves the universality of the method.

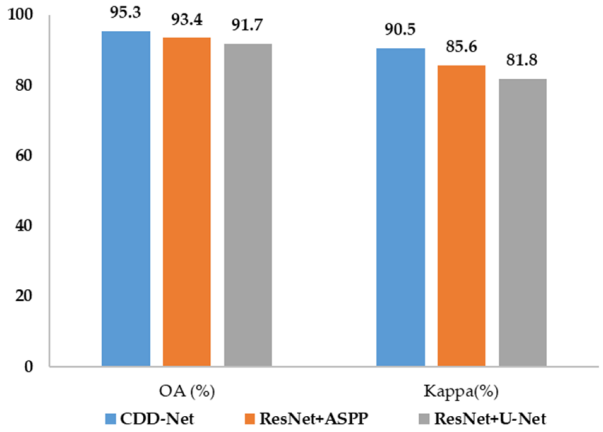


Figure 15. Effect evaluation of different methods in the test area of Sanniangmiao Village, expressed by OA and Kappa coefficients.

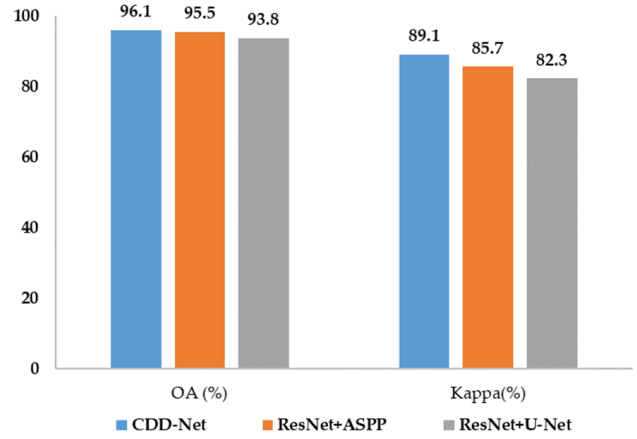


Figure 16. Effect evaluation of different methods in the test area of Xiaowang Village, expressed by OA and Kappa coefficients.

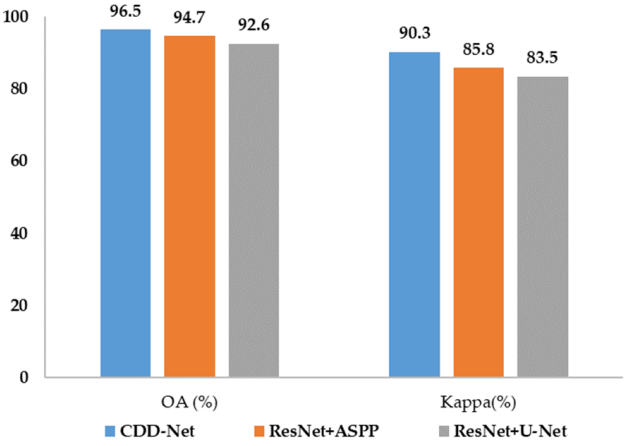


Figure 17. Effect evaluation of different methods in the test area of Zhangling Village, expressed by OA and Kappa coefficients.

5. Discussion

It is of great value to use remote sensing images to carry out urban and rural land surveys. In comparison to satellite remote sensing platforms, the UAV photography platforms can acquire high-resolution aerial photographs more efficiently and affordably. Additionally, it can be equipped with a variety of sensors and allows for controlled flight paths, which facilitates the acquisition of multi-source and multi-temporal data, which is critical for tasks requiring rapid response. In rural areas, UAV remote sensing images can discern minute details such as buildings, cars, and farmers' production equipment. Thus, the benefits of UAV photography technology must be effectively exploited to address the shortcomings of time-consuming and expensive data collection in rural land surveys.

Being different from typical land use classifications, the concept of RRL is more abstract and difficult to detect in remote sensing images. On the one hand, residential land frequently exhibits distinct morphologies as a result of rural areas' disparate architectural styles and building policies; on the other hand, residential land is primarily composed of buildings and impervious surfaces that share characteristics with other types of rural land such as roads and squares, limiting the accuracy of remote sensing interpretation. As a result of these problems, there is an urgent need for efficient and accurate extraction methods of RRL. Therefore, in this research, we have combined Dense-Net, ASPP, and PointRend to propose a new extraction framework to address this problem.

There are few open-source datasets for complex rural land use classification. Although our research adopts the data enhancement methods of mirroring, turning, and rotating to expand the training data, which is suitable for RRL extraction in a certain area, if the research target is a large-scale area, it is difficult for different regions, different buildings and different architectural and planning styles of RRL as the size of the dataset is still small. Subsequent research needs to further expand the amount of data and experimental scenarios. Additionally, at present, a large number of farmers have moved to cities to work and live, and "hollow villages" have gradually formed in rural areas. By the usage of night lights, it is possible to determine whether a village is inhabited or not and to learn about status of use and the idle condition of homesteads. In future, the application of night light remote sensing in urban and rural developments could be further strengthened.

6. Conclusions

Mastering the area and distribution of RRL is significant to the refined management of rural land, territorial spatial planning, and the implementation of the rural revitalization strategy. In this research, we took RRL as the research object and used UAV to quickly acquire VHR images, and an automatic RRL extraction framework, CDD-Net, was proposed. The framework used DenseNet to achieve feature reuse to capture global semantics information, perceive multi-scale semantic information through the ASPP module, and improve the extraction accuracy of the boundary through the PointRend module. After comparison and analysis, the model training and test results of the framework were good, and the framework passed the effect evaluation, realizing the rapid and accurate identification and extraction of RRL in complex rural scenes. The research could provide a theoretical basis and technical support for rural land planning, analysis, and the formulation of land management policies. The main conclusions were as follows:

(1) The CDD-Net is a RRL extraction framework with remarkable effects and high precision, which can identify boundary features more accurately, and perform well in extracting both small-scale and large-scale targets. It can better face the interference of complex backgrounds and building shadows in VHR UAV images.
(2) Comparisons from different perspectives have all shown that the ASPP module can better handle multi-scale features in VHR images. As a feature extraction network, DenseNet can realize the reuse of features and achieve the goals of higher accuracy with fewer parameters, and has better applicability in RRL extraction tasks. The PointRend neural network module can better handle edge features. The improved model using PointRend is not only easier to train, but can also output more accurate object boundaries, which has further improved the extraction accuracy of RRL.
(3) The proposed framework outperforms other advanced semantic segmentation algorithms with better performance and solves the problems of low segmentation accuracy and the too smooth output boundaries of existing semantic segmentation models. Subsequent research needs to further expand the amount of data and number of experimental scenarios to achieve RRL identification and extraction in a large area. In future, the application of night light remote sensing data could be used in the research to understand the utilization and idle conditions of RRL.

Author Contributions: Conceptualization, A.W., C.S. and J.L.; methodology, C.S. and J.L.; software, C.S.; validation, C.S., J.L. and L.W.; formal analysis, C.S. and J.L.; investigation, C.S., J.L., L.W., B.S. and Y.H.; resources, A.W.; data curation, C.S. and J.L.; writing—original draft preparation, C.S., J.L. and A.W.; writing—review and editing, C.S., J.L., L.W. and A.W.; visualization, C.S. and J.L.; supervision, A.W.; project administration, C.S., J.L., L.W., B.S., Y.H. and A.W.; funding acquisition, A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Natural Science Foundation, grant number (ZR2019MD014); Shandong Natural Science Foundation, grant number (ZR2021MD018); and Shandong Natural Science Foundation, grant number (ZR2013DM006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: The authors would like to thank the editors and reviewers for their helpful comments.

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

References

1. Chen, Z.; Li, Y.; Liu, Y.; Liu, X. Does rural residential land expansion pattern lead to different impacts on eco-environment? A case study of loess hilly and gully region, China. *Habitat Int.* **2021**, *117*, 102436. [CrossRef]
2. Tao, Z.; Guanghui, J.; Guangyong, L.; Dingyang, Z.; Yanbo, Q. Neglected idle rural residential land (IRRL) in metropolitan suburbs: Spatial differentiation and influencing factors. *J. Rural Stud.* **2020**, *78*, 163–175. [CrossRef]
3. Yanbo, Q.; Guanghui, J.; Yuting, Y.; Qiuyue, Z.; Yuling, L.; Wenqiu, M. Multi-scale analysis on spatial morphology differentiation and formation mechanism of rural residential land: A case study in Shandong Province, China. *Habitat Int.* **2018**, *71*, 135–146. [CrossRef]
4. Tao, Z.; Guanghui, J.; Wenqiu, M.; Guangyong, L.; Yanbo, Q.; Yingying, T.; Qinglei, Z.; Yaya, T. Dying villages to prosperous villages: A perspective from revitalization of idle rural residential land (IRRL). *J. Rural Stud.* **2021**, *84*, 45–54. [CrossRef]
5. Van Vliet, J.; de Groot, H.L.; Rietveld, P.; Verburg, P.H. Manifestations and underlying drivers of agricultural land use change in Europe. *Landsc. Urban Plan.* **2015**, *133*, 24–36. [CrossRef]
6. Glanz, H.; Carvalho, L.; Sulla-Menashe, D.; Friedl, M.A. A parametric model for classifying land cover and evaluating training data based on multi-temporal remote sensing data. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 219–228. [CrossRef]
7. Jay, S.; Guillaume, M. A novel maximum likelihood based method for mapping depth and water quality from hyperspectral remote-sensing data. *Remote Sens. Environ.* **2014**, *147*, 121–132. [CrossRef]
8. Adugna, T.; Xu, W.B.; Fan, J.L. Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images. *Remote Sens.* **2022**, *14*, 574. [CrossRef]
9. Noi, P.T.; Kappas, M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors* **2018**, *18*, 18. [CrossRef]
10. Xi, Y.B.; Tian, J.; Jiang, H.L.; Tian, Q.J.; Xiang, H.X.; Xu, N.X. Mapping tree species in natural and planted forests using Sentinel-2 images. *Remote Sens. Lett.* **2022**, *13*, 544–555. [CrossRef]
11. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
12. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
13. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [CrossRef]
14. Pande-Chhetri, R.; Abd-Elrahman, A.; Liu, T.; Morton, J.; Wilhelm, V.L. Object-based classification of wetland vegetation using very high-resolution unmanned air system imagery. *Eur. J. Remote Sens.* **2017**, *50*, 564–576. [CrossRef]
15. Matikainen, L.; Karila, K.; Hyyppä, J.; Litkey, P.; Puttonen, E.; Ahokas, E. Object-based analysis of multispectral airborne laser scanner data for land cover classification and map updating. *ISPRS J. Photogramm. Remote Sens.* **2017**, *128*, 298–313. [CrossRef]
16. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [CrossRef]
17. Guo, Y.M.; Liu, Y.; Oerlemans, A.; Lao, S.Y.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
18. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
19. Chen, B.Y.; Xia, M.; Huang, J.Q. MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover. *Remote Sens.* **2021**, *13*, 731. [CrossRef]
20. Ji, S.P.; Wang, D.P.; Luo, M.Y. Generative Adversarial Network-Based Full-Space Domain Adaptation for Land Cover Classification From Multiple-Source Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3816–3828. [CrossRef]
21. Li, R.R.; Liu, W.J.; Yang, L.; Sun, S.H.; Hu, W.; Zhang, F.; Li, W. DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3954–3962. [CrossRef]
22. Xu, Y.Y.; Wu, L.; Xie, Z.; Chen, Z.L. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
23. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]

24. Xia, M.; Cui, Y.C.; Zhang, Y.H.; Xu, Y.M.; Liu, J.; Xu, Y.Q. DAU-Net: A novel water areas segmentation structure for remote sensing image. *Int. J. Remote Sens.* **2021**, *42*, 2594–2621. [CrossRef]
25. Zhao, L.J.; Zhang, W.; Tang, P. Analysis of the inter-dataset representation ability of deep features for high spatial resolution remote sensing image scene classification. *Multimed. Tools Appl.* **2019**, *78*, 9667–9689. [CrossRef]
26. Srivastava, V.; Biswas, B. CNN-based salient features in HSI image semantic target prediction. *Conn. Sci.* **2020**, *32*, 113–131. [CrossRef]
27. Shi, C.P.; Zhao, X.; Wang, L.G. A Multi-Branch Feature Fusion Strategy Based on an Attention Mechanism for Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 1950. [CrossRef]
28. Feng, J.; Chen, J.T.; Liu, L.G.; Cao, X.H.; Zhang, X.R.; Jiao, L.C.; Yu, T. CNN-Based Multilayer Spatial-Spectral Feature Fusion and Sample Augmentation With Local and Nonlocal Constraints for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1299–1313. [CrossRef]
29. Broni-Bediako, C.; Murata, Y.; Mormille, L.H.B.; Atsumi, M. Searching for CNN Architectures for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4701813. [CrossRef]
30. Ahmed, A.A.; Darwish, S.M. A Meta-Heuristic Automatic CNN Architecture Design Approach Based on Ensemble Learning. *IEEE Access* **2021**, *9*, 16975–16987. [CrossRef]
31. Chen, G.Z.; Tan, X.L.; Guo, B.B.; Zhu, K.; Liao, P.Y.; Wang, T.; Wang, Q.; Zhang, X.D. SDFCNv2: An Improved FCN Framework for Remote Sensing Images Semantic Segmentation. *Remote Sens.* **2021**, *13*, 4902. [CrossRef]
32. Li, L.Y.; Li, X.Y.; Liu, X.; Huang, W.W.; Hu, Z.Y.; Chen, F.S. Attention Mechanism Cloud Detection With Modified FCN for Infrared Remote Sensing Images. *IEEE Access* **2021**, *9*, 150975–150983. [CrossRef]
33. Shao, Z.F.; Zhou, W.X.; Deng, X.Q.; Zhang, M.D.; Cheng, Q.M. Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]
34. Tian, T.; Chu, Z.Q.; Hu, Q.; Ma, L. Class-Wise Fully Convolutional Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3211. [CrossRef]
35. Zhu, Y.T.; Long, L.H.; Wang, J.J.; Yan, J.W.; Wang, X.Q. Road Segmentation from High-Fidelity Remote Sensing Images using a Context Information Capture Network. *Cogn. Comput.* **2020**, *14*, 780–793. [CrossRef]
36. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 426–435. [CrossRef]
37. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6169–6181. [CrossRef]
38. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 811–815. [CrossRef]
39. Wang, J.; Shen, L.; Qiao, W.; Dai, Y.; Li, Z. Deep Feature Fusion with Integration of Residual Connection and Attention Model for Classification of VHR Remote Sensing Images. *Remote Sens.* **2019**, *11*, 1617. [CrossRef]
40. Li, X.; Yang, X.F.; Li, X.T.; Lu, S.J.; Ye, Y.M.; Ban, Y.F. GCDB-UNet: A novel robust cloud detection approach for remote sensing images. *Knowl. Based Syst.* **2022**, *238*, 107890. [CrossRef]
41. Priyanka; Sravya, N.; Lal, S.; Nalini, J.; Reddy, C.S.; Dell'Acqua, F. DIResUNet: Architecture for multiclass semantic segmentation of high resolution remote sensing imagery data. *Appl. Intell.* **2022**. [CrossRef]
42. Wang, H.Y.; Miao, F. Building extraction from remote sensing images using deep residual U-Net. *Eur. J. Remote Sens.* **2022**, *55*, 71–85. [CrossRef]
43. Yan, C.; Fan, X.S.; Fan, J.L.; Wang, N.Y. Improved U-Net Remote Sensing Classification Algorithm Based on Multi-Feature Fusion Perception. *Remote Sens.* **2022**, *14*, 1118. [CrossRef]
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
45. Sha, Y.C. An Information Extraction Method of Suburban Industrial Land Using Improved Densenet Network in Remote Sensing Images. *Fresenius Environ. Bull.* **2021**, *30*, 1190–1200.
46. Li, H.L.; Wang, G.J.; Dong, Z.; Wei, X.K.; Wu, M.J.; Song, H.H.; Amankwah, S.O.Y. Identifying Cotton Fields from Remote Sensing Images Using Multiple Deep Learning Networks. *Agronomy* **2021**, *11*, 174. [CrossRef]
47. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
48. Wulamu, A.; Shi, Z.X.; Zhang, D.Z.; He, Z.Y. Multiscale Road Extraction in Remote Sensing Images. *Comput. Intell. Neurosci.* **2019**. [CrossRef]
49. Zhang, W.; Tang, P.; Zhao, L.J. Fast and accurate land cover classification on medium resolution remote sensing images using segmentation models. *Int. J. Remote Sens.* **2021**, *42*, 3277–3301. [CrossRef]
50. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. PointRend: Image Segmentation As Rendering. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9796–9805. [CrossRef]
51. Qiu, T.Q.; Liang, X.J.; Du, Q.Y.; Ren, F.; Lu, P.J.; Wu, C. Techniques for the Automatic Detection and Hiding of Sensitive Targets in Emergency Mapping Based on Remote Sensing Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 68. [CrossRef]

52. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. In Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
53. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
54. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U. Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 105–109. [CrossRef]