



Article

Attention Enhanced U-Net for Building Extraction from Farmland Based on Google and WorldView-2 Remote Sensing Images

Chuangnong Li ¹, Lin Fu ^{1,*}, Qing Zhu ¹, Jun Zhu ¹, Zheng Fang ², Yakun Xie ¹, Yukun Guo ¹ and Yuhang Gong ²

- ¹ Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China; sclcn@sina.com (C.L.); zhuq66@263.net (Q.Z.); zhujun@swjtu.edu.cn (J.Z.); yakunxie@my.swjtu.edu.cn (Y.X.); gyk@my.swjtu.edu.cn (Y.G.)
- ² Sichuan Center of Satellite Application Technology, Sichuan Institute of Land Science and Technology, Chengdu 610041, China; fangzheng5288@163.com (Z.F.); gongyuhang_geo@163.com (Y.G.)
- * Correspondence: vge_fulin@my.swjtu.edu.cn

Abstract: High-resolution remote sensing images contain abundant building information and provide an important data source for extracting buildings, which is of great significance to farmland preservation. However, the types of ground features in farmland are complex, the buildings are scattered and may be obscured by clouds or vegetation, leading to problems such as a low extraction accuracy in the existing methods. In response to the above problems, this paper proposes a method of attention-enhanced U-Net for building extraction from farmland, based on Google and WorldView-2 remote sensing images. First, a Resnet unit is adopted as the infrastructure of the U-Net network encoding part, then the spatial and channel attention mechanism module is introduced between the Resnet unit and the maximum pool and the multi-scale fusion module is added to improve the U-Net network. Second, the buildings found on WorldView-2 and Google images are extracted through farmland boundary constraints. Third, boundary optimization and fusion processing are carried out for the building extraction results on the WorldView-2 and Google images. Fourth, a case experiment is performed. The method in this paper is compared with semantic segmentation models, such as FCN8, U-Net, Attention_UNet, and DeepLabv3+. The experimental results indicate that this method attains a higher accuracy and better effect in terms of building extraction within farmland; the accuracy is 97.47%, the F1 score is 85.61%, the recall rate (Recall) is 93.02%, and the intersection of union (IoU) value is 74.85%. Hence, buildings within farming areas can be effectively extracted, which is conducive to the preservation of farmland.

Keywords: building extraction; farmland range; attention enhancement; U-Net network improvement; multi-source remote sensing image



Citation: Li, C.; Fu, L.; Zhu, Q.; Zhu, J.; Fang, Z.; Xie, Y.; Guo, Y.; Gong, Y. Attention Enhanced U-Net for Building Extraction from Farmland Based on Google and WorldView-2 Remote Sensing Images. *Remote Sens.* **2021**, *13*, 4411. <https://doi.org/10.3390/rs13214411>

Academic Editors: Jiaojiao Tian, Qin Yan, Mohammad Awrangjeb, Beril Sirmacek and Nusret Demir

Received: 21 August 2021

Accepted: 29 October 2021

Published: 2 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Farmland constitutes an important resource for human survival and development. Countries worldwide have each issued corresponding farmland protection policies or measures [1–6]. Farmland protection represents the basic guarantee to maintain the red line of 1.8 billion mu of farmland in China, but the phenomenon of illegal farmland occupation to build houses in rural areas is serious [7,8]. According to statistics, since 1 January 2013, more than 7 million mu of farmland has been occupied for house construction in China, which more than 600,000 mu has been occupied in the southwest region. Therefore, building extraction in farmland is of great significance for farmland protection [9–12].

Traditional field surveys provide high accuracy and reliable results but require considerable manpower, materials and financial resources. With the continuous development of remote sensing technology, the obtained remote sensing images with higher spatial resolution contain abundant building information. Extracting buildings from high-resolution

remote sensing images has become a research hotspot [9,13–16]. Traditional building extraction methods, based on optical remote sensing images, mainly consider low-level semantic features such as color, texture, and shape to extract buildings. Methods of this kind include edge detection, region segmentation, corner detection, threshold segmentation, clustering, etc. [17–21]. However, these methods are affected by lighting conditions, sensor types, and building structures. Even if the high-resolution remote sensing images are rich in details, the complex types of features, pixel mixing, shadows and other problems within the farmland are serious, making the phenomenon of “same subject with different spectra” or “different subject with same spectra” more common. These methods are limited when solving the problem of building extraction under specific data conditions [14,22].

In recent years, deep-learning methods have continued to be developed, and various neural network models have been widely used for intelligent building extraction [23–26]. With complex structural features, single-pixel segmentation may destroy the integrity and structural features of the building. In response to this problem, scholars at home and abroad have further improved various semantic segmentation models and created networks for the extraction of buildings. Good results have been achieved [27–29]. For example, Li et al. proposed a rural building segmentation method based on Mask R-CNN and a histogram threshold, which achieved the high-performance semantic segmentation of buildings via a small number of samples [30]. Zhang et al. combined a neural network and edge detection to conduct building extraction experiments. After pixel classification through the neural network, edge detection was used to complete an accurate segmentation of building boundaries [31]. Wu et al. established a multi-constrained full convolutional neural network architecture based on FCN, and used multiple constraints to optimize the parameters of the middle layer in order to obtain more multi-scale features [32]; Lin et al. combined residual block and expanded convolution to construct a deep network architecture for building extraction, and improved computational efficiency through a certain accuracy loss [33]; Xu et al. combined Resnet and U-net networks to extract buildings from high-resolution remote sensing images, and integrated them using guided filters to eliminate noise, improve accuracy and optimize the output result [34]. Bai et al. proposed an improved faster R-CNN building extraction method, using dense residual network and region of interest alignment methods to solve the problem of regional mismatch, and further improve the effect of building detection [35]. Deng et al. used a new feature extraction method to combine an object suggestion network (MS-OPN) and object detection network (AODN) to construct multi-scale features for building extraction [36]. The above methods mainly focus on urban areas where buildings are dense. For these places, the features on the image are mainly buildings, and the occlusion of buildings is not considered. However, within the range of farmland, there are few buildings and most of the areas comprise non-construction land use. These methods are not suitable for effectively extracting small targets within the farmland, and attention needs to be paid to small target buildings.

The attention mechanism imitates human brain–eye vision, which can more accurately focus on and process the most important details, rather than establishing the whole visual content. Therefore, it is widely used in deep learning to improve the accuracy of target extraction [37,38]. Yang et al. combined a lightweight DenseNet and a spatial attention fusion module to construct a dense attention network for building extraction from remote sensing images [39]. Pan et al. combined spatial and channel attention mechanisms to detect buildings using a U-Net network [40]. Jiang et al. input a global co-attention mechanism, building an attention-guided Siamese network based on a pyramid feature to detect urban building changes and achieved good results [41]. Guo et al. proposed a building extraction structure based on attention and multiple losses, which further improved the sensitivity of the model and the feature extraction ability [42]. However, the building distribution within farmland is sparse, shielding effects such as clouds, rain, fog, and vegetation may occur, and the boundary may be blurred. The above existing methods encounter difficulties regarding the accurate extraction of building information

in farmland. With the requirement of overcoming the problem of building occlusion, it is necessary to fuse multi-source remote sensing images to extract farmland buildings and fuse the extraction results.

Therefore, in response to the above problems, this paper proposes a method using an attention-enhanced U-Net for building extraction from farmland. First, the Resnet unit is introduced as the basic structure in the coding layer of the U-Net network, and the spatial and channel attention modules are added to the convolutional layer of the U-Net network to enhance the convolution process's attention, given in dispersed small targets. The buildings within farmland are better extracted and the U-Net network is improved. Then, we integrated WorldView-2 and Google remote sensing images (WorldView-2 images are provided by the Sichuan Provincial Bureau of Surveying and Mapping, and Google images are freely downloaded from the Internet). Taking the farmland boundaries given by the third national census dataset as the spatial semantic constraint, the buildings within farmland can be extracted and the influence of interference factors can be reduced. Finally, through morphological operations such as opening and closing operations, the extracted building boundaries are optimized and merged to enhance the accuracy of building extraction in farmland. The main chapters of this paper are arranged as follows: the second part introduces the main technical methods of the paper, including U-Net network improvement, intelligent building extraction from WorldView-2 and Google remote sensing images under boundary constraints, and boundary optimization processing operations. The third part mainly describes the source of the dataset, parameter settings, experimental results and discussion. The fourth part introduces the research conclusions and prospects for future work.

2. Methodology

2.1. Overall Framework

This paper proposes a method of attention enhanced U-Net for building extraction from farmland. The overall research idea is shown in Figure 1. The main content can be divided into two parts: (1) spatial-channel attention-enhanced building extraction: using the Resnet unit as the basic structure, adding a spatial-channel attention mechanism module and a multi-scale fusion module, the U-Net network is improved and the network's attention to small-building targets is enhanced; (2) building boundary optimization and fusion processing: with remote sensing images from WorldView-2 and Google as the input, the building extraction range is narrowed under farmland boundary constraints, an improved U-Net network is employed to extract buildings for farmland, morphological filtering methods are implemented, such as opening/closing operations to optimize the extraction results, and the optimized building extraction results are fused to achieve accurate building extraction results from the farmland.

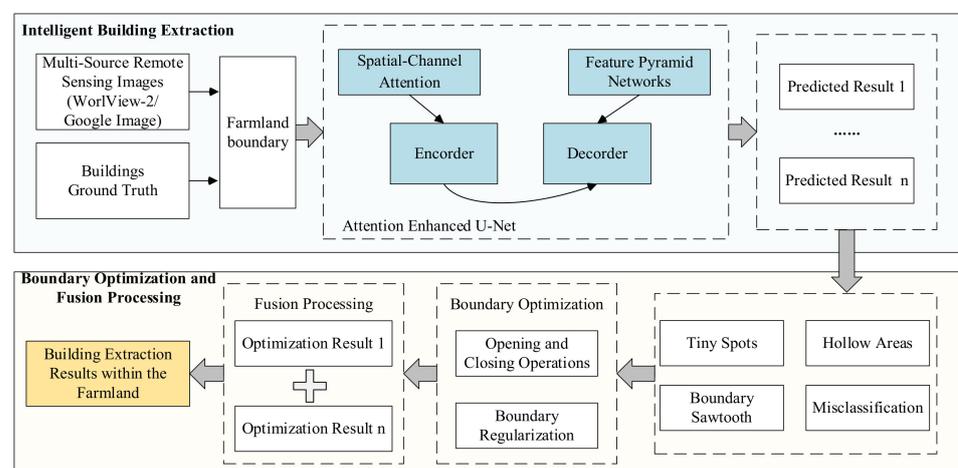


Figure 1. Overall procedural framework.

2.2. Improvement of the Attention-Enhanced U-Net Network

The U-Net network relies on skip connections to force the aggregation of same-scale feature maps of the encoder and decoder sub-networks, and the network performance is good. However, the types of features in the farmland are variable. These comprise not only buildings but also roads, woodlands, etc. Moreover, the building distribution is sparse, and various problems such as clouds, rain, fog, and vegetation shielding may ensue, resulting in an inability to extract small targets from complex backgrounds, incomplete building area extraction, and inaccurate boundaries. To overcome the limitations of existing methods in the extraction of small building targets within the farmland range from a complex background, this paper improves the U-Net network, as shown in Figure 2. The upper part represents the encoding structure, and the lower part represents the decoding structure. In the U-Net network coding stage, Resnet and attention models are added to enhance the network’s attention to small target buildings. Since continuous up-sampling will cause the loss of detail in the building information, this paper adds a multi-scale fusion module in the boundary stage to ensure the local detail characteristics of the buildings are retained.

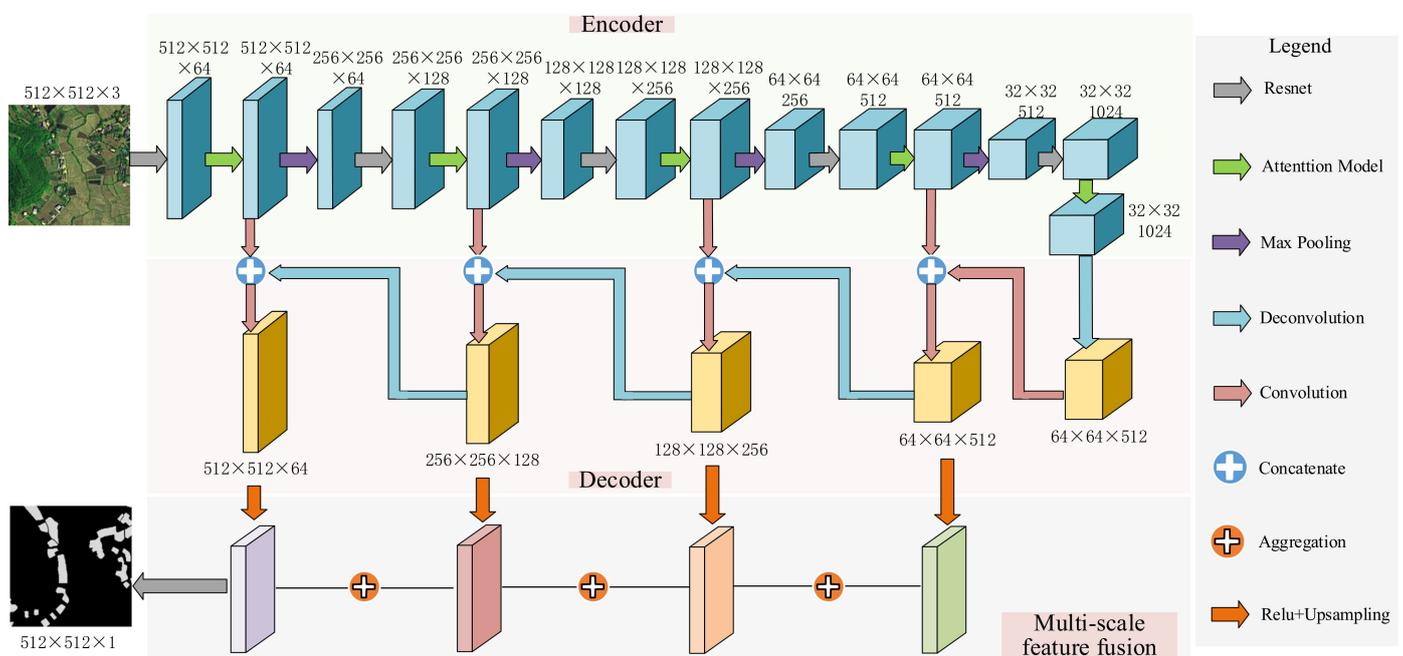


Figure 2. Improvement of the attention-enhanced U-Net network.

The encoding part adopts the Resnet unit as the basic structure, as shown in Figure 3. Compared to a traditional convolution layer, the Resnet residual network achieves convergence more easily and avoids the performance degradation issues caused by an increase in network depth. To prevent overfitting, batch normalization (BN) and rectified linear unit (ReLU) activation function layers are added, based on the residual network, to establish a refined residual network. The nonlinear expression ability of the network model is thus improved, and the features extracted from remote sensing images become richer.

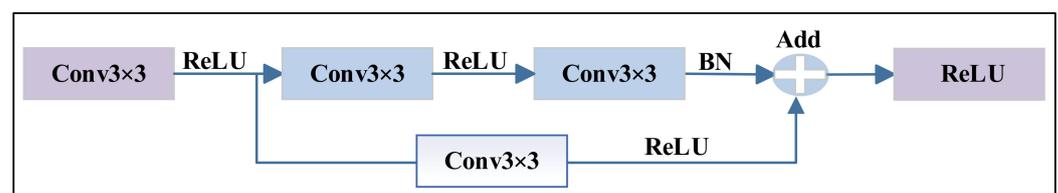


Figure 3. Schematic diagram of the Resnet unit.

Buildings are scattered throughout farmland. To enable the model to focus on scattered buildings, in the coding stage, the small target buildings within the farmland are focused upon, and the details of important targets are paid more attention to, according to the patterns of human vision. Meanwhile, the attention-mechanism module is adopted between each Resnet unit and the maximum pooling layer, which mainly includes two parts: channel and spatial attention. Channel attention is based on the relationships among the extracted features, and the channel weight is modeled to determine what the target object is. Spatial attention uses the spatial relationship of features to remodel the weight of spatial location pixels and determine which location represents the information. The local and global information can be aggregated by the attention-mechanism module, the relationship between buildings and the background can be captured, the feature weight of each channel and spatial location can be adaptively adjusted, and the network feature-extraction capacity with regard to small building targets and the ability to better grasp complex scenes can be enhanced.

As shown in Figure 4, the attention-mechanism model designed in this paper can be used to locate the area where small building targets are scattered in a given remote sensing image and suppress useless information. First, channel attention is modeled for input feature mapping, the global average pool and maximum pool layers are processed to obtain an input feature map, and a multi-layer perceptron is constructed. The weight of each channel is automatically obtained via self-learning and is then multiplied by the input feature map to obtain a channel attention feature map. Second, spatial attention modeling is conducted, and the above channel attention feature map is applied as an input to perform the spatial convolution operation and obtain the corresponding attention weights at different spatial positions on the feature map. Then, the feature map processed by the spatial and channel mixed-attention mechanism is multiplied by the input feature map, for adaptive feature refinement. In the process of training and prediction, the model can better focus on the most important feature channels and spatial positions in remote sensing images, which thus improves the model's detection performance.

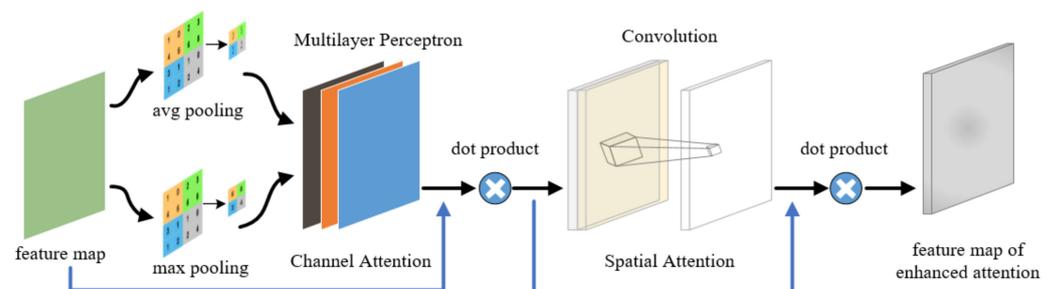


Figure 4. Attention mechanism model.

In the decoding structure, the feature map is reconstructed through the deconvolution layer, the feature map after the deconvolution layer is skip-connected with the attention-enhanced feature map included at the encoding stage, the depth of the feature map is reduced relying on a conventional convolution layer, and the size of the feature map is gradually expanded. Finally, to further improve the network performance, which entails the perception ability of multi-scale buildings in the farmland, especially small buildings and building edge information, and considering that shallow features have high resolution, but rich details and deep features have low resolution but offer rich semantic information, multi-scale feature fusion is carried out, to acquire both deep and shallow features after up-sampling and nonlinear processing. This strategy can ensure that the network will not ignore texture, edge, and other image details, while extracting global semantic information to obtain building texture, shape and spatial context features and provide more precise building segmentation results.

2.3. Building Extraction from Multi-Source Remote Sensing Images under Boundary Constraints

Buildings within the farmland may be obscured, the building extraction range may require narrowing, and the extraction accuracy may necessitate improvement. To resolve the above problems, this paper fully employs the advantages of satellite images and proposes an intelligent building extraction method from multi-source remote sensing images under boundary constraints, considering different satellite remote sensing images of the same area, as shown in Figure 5. WorldView-2, Google, and other image series can be used to assess whether buildings are located within the boundary range, through the boundary constraints of farmland. If buildings are not located within the range, they are directly discarded. With regard to buildings within the boundary range, an improved U-Net neural network model with an enhanced attention mechanism is adopted to intelligently extract building contours and the relative position information.

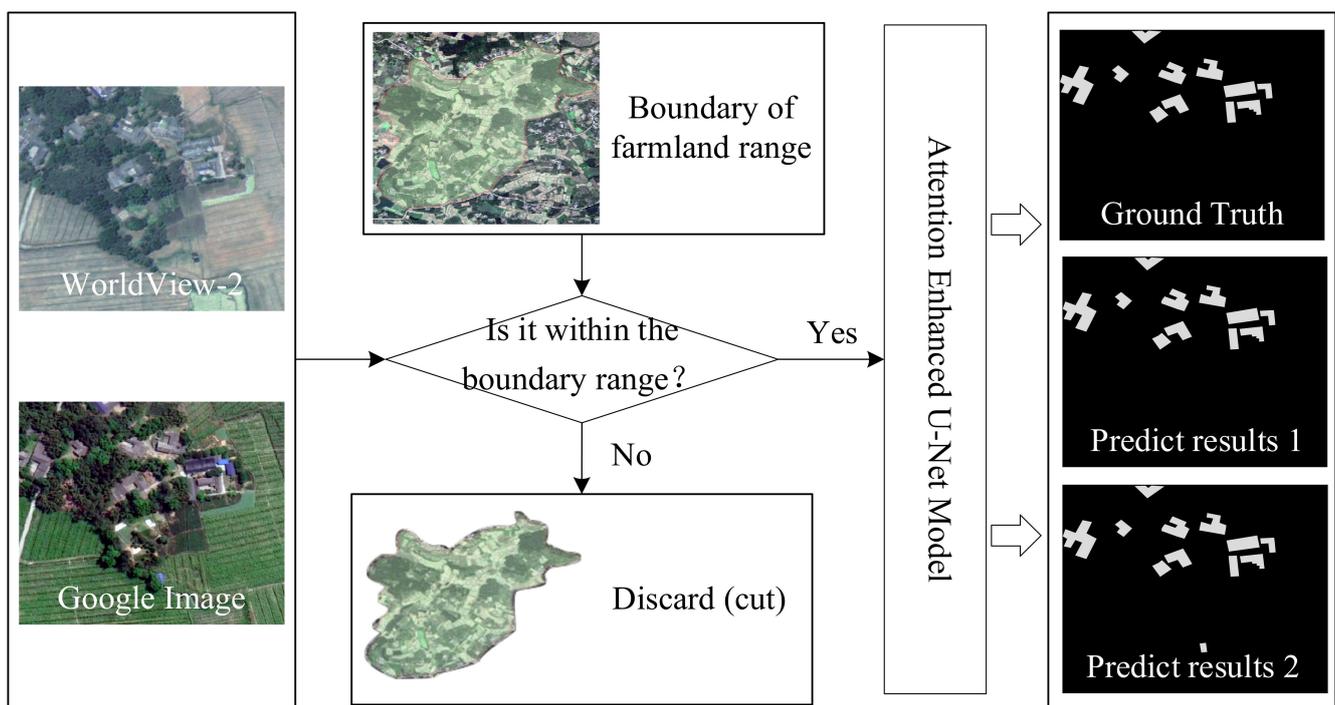


Figure 5. Building extraction based on the boundary constraints of farmland range.

2.4. Building Boundary Optimization and Fusion Processing

There are certain problems in the extraction results obtained with neural network-based models, such as very small spots, hollow areas, boundary sawtooth features and misclassification issues. Moreover, the building extraction results based on multi-source remote sensing images are varied. Therefore, this paper proposes a building boundary optimization and fusion processing method that is based on morphological filtering, as shown in Figure 6, to improve the accuracy of building extraction.

First, in terms of the neural network extraction results containing fine patches, a filter based on geometric operations is applied to execute the opening operation. As expressed in Equation (1), isolated points and burrs in tiny spots are removed in this manner. Regarding the extraction results including hollow areas, the closing operation of a morphological filtering operation is implemented, as expressed in Equation (2), to fill any cracks or hollow areas in the extraction results:

$$I \circ S = (I \ominus S) \oplus S \quad (1)$$

$$I \bullet S = (I \oplus S) \ominus S \quad (2)$$

where I denotes the original image extracted by the network, and S denotes the structural element of the filter.

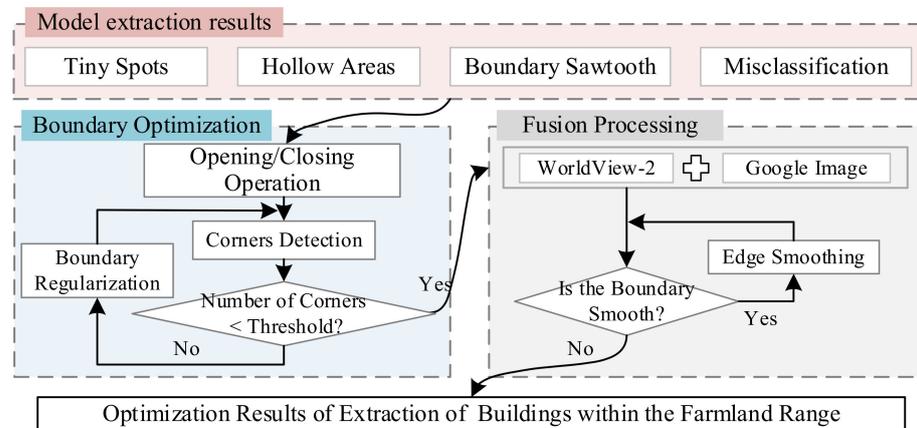


Figure 6. Building boundary optimization and fusion processing.

Then, based on the results of the morphological filtering operation, refer to the method proposed by Xie et al. to judge whether the number of building contour corners is smaller than the set threshold. If the set conditions are met, the multi-source remote sensing image extraction results are fused. Conversely, the building boundary should be optimized until the corner number conditions are satisfied [43]. Finally, the extraction results of the WorldView-2 and Google images after boundary optimization are compared pixel by pixel, and the pixels representing the building at the current location are merged as the prediction results, so as to realize the fusion processing of the extraction results of WorldView-2 and Google remote sensing images and solve the problem of misclassification in the extraction results. Gridlines are created, with the considered building outlines as boundaries. The results are further assessed to establish whether the building outline boundary requires further smoothing. If the conditions are met, the building extraction results for farmland are output. Otherwise, the boundary should be smoothed until the smoothing conditions are satisfied. The sawtooth effect in the extraction results is mitigated so that the extraction results are more precise and accurate.

3. Case Experiment Analysis

3.1. Case Area and Dataset

To verify the proposed method, the considered experimental data included five WorldView-2 satellite remote sensing image datasets, covering Qionglai city, Meishan city, Dayi County and Pujiang County of Sichuan Province, and Google images of the same areas, as shown in Figure 7. The image spatial resolution reached 0.5×0.5 m, with three bands, i.e., red, green and blue bands. The building pixel value was 1, and the pixel values of other features were set to 0. Due to the limited memory size of the adopted graphics card, each image was cropped with a sliding window exhibiting a size of 512×512 pixels.

Considering that the buildings in the images are relatively small, meaningless background slices in the training set were eliminated. The 5 satellite remote sensing images were randomly divided into a training set, validation set, and test set at a ratio of 6:2:2, and the data augmentation method was applied to expand the training dataset and improve the generalization ability of the model. Through comparative experiments, it was found that the addition of Gaussian noise and color perturbation during image data processing did not encourage model accuracy improvement. Therefore, the image processing procedure only involved rotation and flip operations, so that more morphological building features could be recognized, as shown in Figure 8. Finally, 6238 training sets, 2301 verification sets, and 1637 test sets were obtained, and no overlap occurred between the training data and test data.

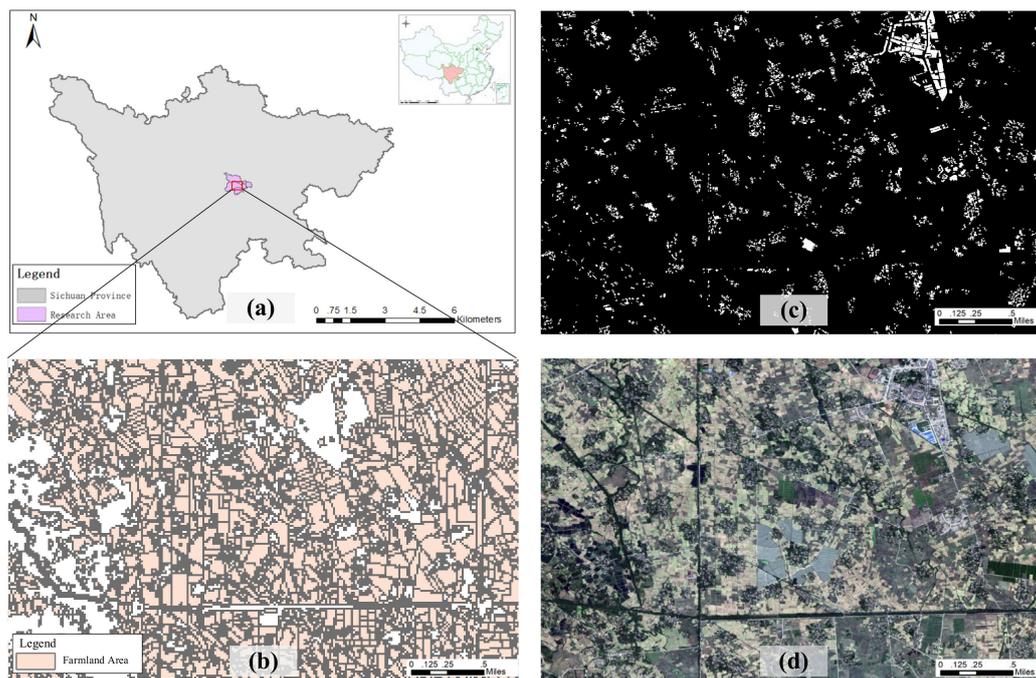


Figure 7. Case study experiment area. (a): Case area; (b): farmland area; (c): ground truth; (d): original images.

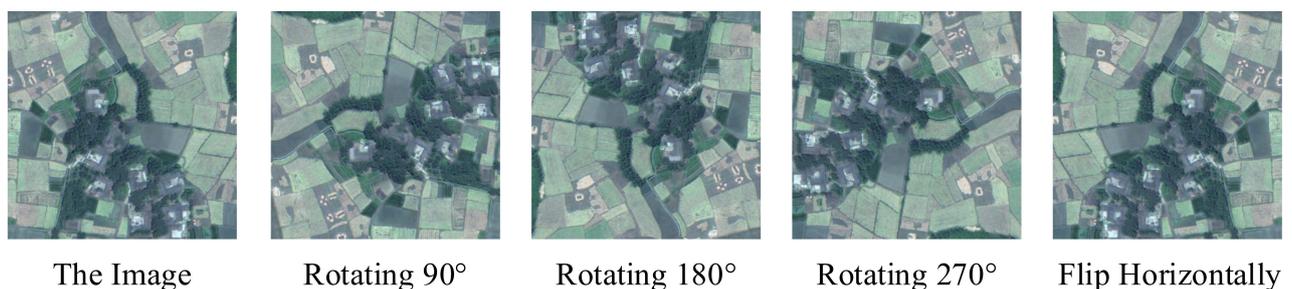


Figure 8. Image and label data.

3.2. Experimental Environment and Parameter Setting

All training and testing operations in this paper were performed on a Windows 10 system with an Intel (R) Core (TM) i9-10920x CPU @ 3.50 GHz processor, 64 GB of memory and an NVIDIA GeForce RTX 3090 GPU graphics card. The initial learning rate during model training was set at 0.001. By monitoring the loss value, the learning rate was reduced to the original value of 0.5 when the performance did not improve after 10 epochs. The experiment was trained using the built-in TensorFlow framework in Python version 3.6, the training batch size was 4 and a total of 60 epochs were iterated with a duration of 12 h. To verify the accuracy of the extraction results obtained using the method proposed in this paper, this paper chose the accuracy, F1, recall, and an intersection of union (IoU) to evaluate the extraction results, where the accuracy is given by the proportion of correctly predicted pixels among the total pixels, as expressed in Equation (3). F1 is a comprehensive index to measure the model, as defined in Equation (4). Recall represents the proportion of correctly predicted buildings among the total buildings, as expressed in Equation (5), and the IoU represents the ratio of the intersection zone (the intersection between the predicted and true values) to the union zone (the union between the predicted and true values), as defined in Equation (6).

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

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (6)$$

where the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) rates evaluate the pixel classification results through comparison of the extracted building pixels to the available ground-truth points. TP denotes the number of correctly extracted building pixels, FP denotes the number of erroneously detected building pixels, TN denotes the number of correctly extracted non-building pixels, and FN denotes the number of missed building pixels.

3.3. Experimental Results and Analysis

As shown in Figure 9, the results for the buildings extracted using the attention-enhanced U-Net model proposed in this paper showed that our model can effectively extract buildings within the scope of farmland. However, the buildings in the dataset are scattered and staggered in the farmland, which leads to the problems of unclear edges and holes at certain map locations, as shown in Figure 9c. With the application of boundary optimization and data fusion methods, the building extraction results are optimized, as shown in Figure 9d. Comparing the red boxes in Figure 9c,d, we can observe that the building pattern in Figure 9d contains no small holes and that the boundary is smooth. In particular, the optimization process can eliminate the very small spots previously found in the extraction results and smooth the boundary, so as to obtain a more complete extracted shape and clearer boundary, respectively.

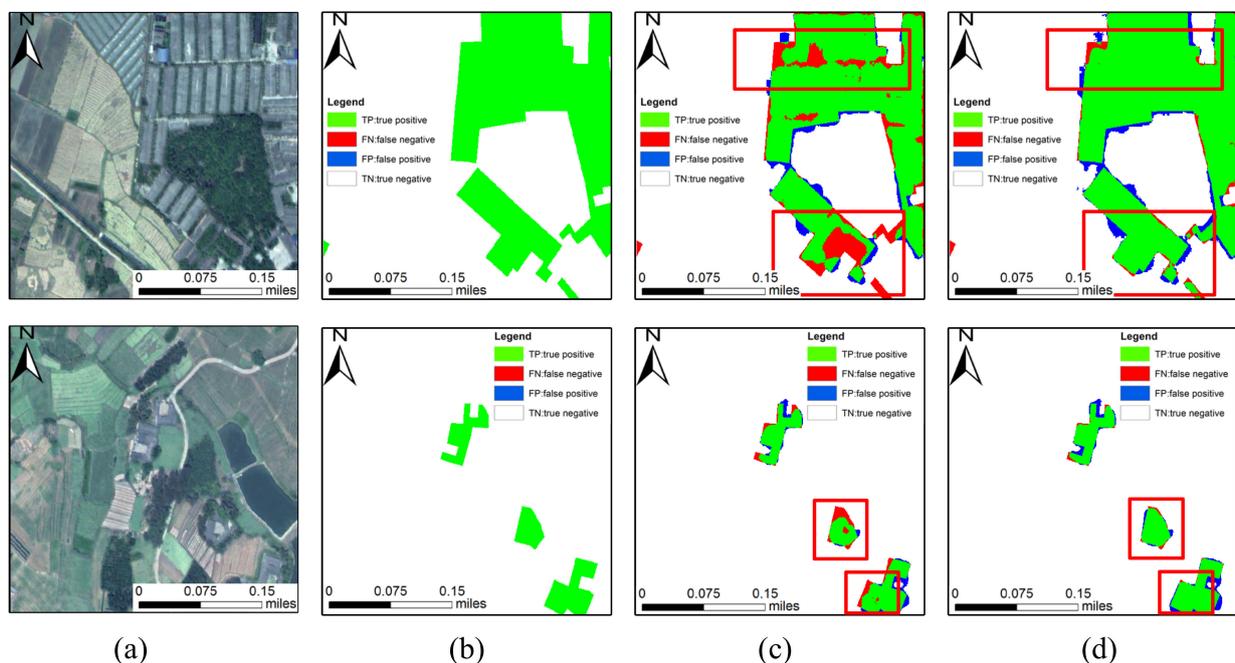


Figure 9. Experimental results. (a): Original images; (b): ground truth; (c): ours model extraction results; (d): post processing results.

To achieve the accurate extraction of building targets, boundary optimization and fusion processing of the building patterns extracted with the network model are executed. The experimental results are listed in Table 1. The IoU value after optimization reaches 74.85%, and the F1 score reaches 85.61%, which are 6.13% and 4.14% higher, respectively, than the values without optimization, thus demonstrating that the post-processing op-

timization method designed in this paper can effectively improve building extraction accuracy.

Table 1. Quantitative comparative analysis of case experiment results.

| Method | Accuracy | F1 | Recall | IoU |
|-----------------|----------|--------|--------|--------|
| Our model | 96.96% | 81.47% | 82.72% | 68.72% |
| Post-processing | 97.47% | 85.61% | 93.02% | 74.85% |

3.4. Discussion

To verify the accuracy and applicability of the proposed method, several groups of comparative experiments were established to compare this method with FCN8, U-Net and Attention_UNet, DeepLabv3+ [44] and other models qualitatively and quantitatively, where Attention_UNet includes a convolutional block attention module (CBAM) module based on the U-Net model [38].

3.4.1. Comparative Experiments of Building Extraction

Figure 10 shows the results obtained by the proposed method, FCN8, U-Net and Attention_UNet, DeepLabv3+ and the other considered models based on WorldView-2 images, where green indicates the positive extraction area, red indicates the missing extraction area and blue indicates the false extraction area. As shown in Figure 10b,c, the building boundaries extracted with the U-Net and FCN8 models are relatively fuzzy, and the building omission phenomenon is obvious with the FCN8 model, while incorrect extraction obviously occurs with the U-Net model. As shown in Figure 10d, the extraction results obtained with the U-Net model containing the attention mechanism are notably better than those obtained with the original U-Net and FCN8 models, but the extraction results remain insufficiently fine. Furthermore, Figure 10e shows that there are many holes, missing extraction and incorrect extraction results among the results obtained with the DeepLabv3+ model. The black box in Figure 10 reveals that the building pixels extracted with the network model designed in this paper are closer to the real image, and compared to the other four network models, the positive extraction area accounts for the majority of the image, which demonstrates that the proposed model can finely extract scattered small buildings from farmland images.

To quantitatively analyze the building extraction results, we considered the building labels drawn by manual visual interpretation as a reference, and the four evaluation indicators of Accuracy, Recall, IoU, and F1 were adopted to evaluate the building extraction results for farmland protection, as indicated in Table 2. Compared with U-Net, FCN8, Attention_UNet, and DeepLabv3+, the model proposed in this paper achieves the highest accuracy in building extraction from remote sensing images, with the IoU value reaching 68.72% and the F1 score reaching 81.47%. The IoU value obtained with our model is 33.73%, 25.84%, 14.71% and 5.33% higher, respectively, than that obtained with the other four models. The F1 score is 29.63%, 13.29%, 11.33% and 3.87% higher, respectively. The results indicate that the shape of the building block extracted with the model developed in this paper is closer to the actual building block shape.

Table 2. Comparison of the building extraction results obtained with the different methods.

| Method | Accuracy | F1 | Recall | IoU |
|----------------|----------|--------|--------|--------|
| U-Net | 88.99% | 51.84% | 73.31% | 34.99% |
| FCN8 | 95.70% | 68.18% | 57.02% | 42.88% |
| Attention_UNet | 94.42% | 70.14% | 81.07% | 54.01% |
| DeepLabv3+ | 96.60% | 77.60% | 72.78% | 63.39% |
| Our model | 96.96% | 81.47% | 82.72% | 68.72% |

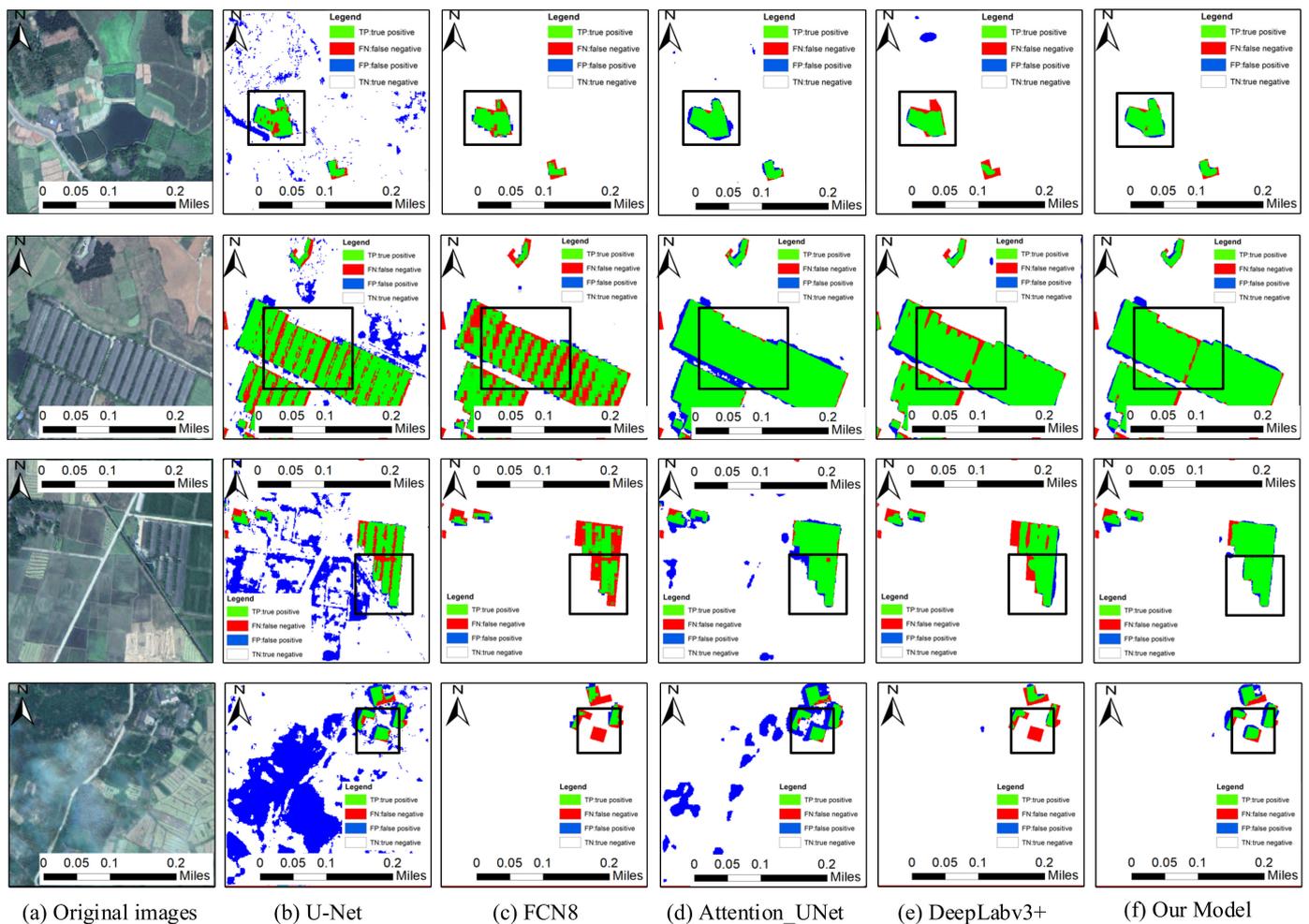


Figure 10. Model comparison results. (a): Original images; columns (b–e): represent extracted results by U-Net, FCN8, Attention_UNet, DeepLabv3+; (f): represent our model extracted results.

3.4.2. Comparative Experiments of Boundary Optimization and Fusion

The results of building extraction using the method proposed in this paper, FCN8, U-Net, Attention_UNet and DeepLabv3+ are shown in Figure 11. According to Figure 11a, there are a large number of small spots showing wrong detection in the U-Net fusion result, and the missed detection is serious. According to Figure 11b–d, the FCN8 fusion presents obvious areas of missed detection. Meanwhile, the fusion results of Attention_UNet and DeepLabv3+ demonstrate obvious false detection areas around the boundary of the building's block. According to Figure 11e, compared with the other four network models, the fusion processing results of building patches extracted by this method account for the most positive inspection areas, and the building patches are more complete; that is, the fusion processing can eliminate the fine patches in the results and smooth the edges of the patches.

As shown in Table 3, compared with the other four network models, the model presented in this paper offers the highest accuracy after fusion processing; the IoU score is 74.85% and the F1 score is 85.61%, 36.69% and 32.06% higher than U-Net, respectively. Compared with FCN8, it is 5.33% and 7.55% higher, respectively. Compared with Attention_UNet, it is 17.71% and 13.55% higher, respectively. Compared with DeepLabv3+, it is 7.06% and 5.04% higher, respectively. These results show that the fusion processing of building blocks extracted by this model can greatly improve the accuracy of building extraction.

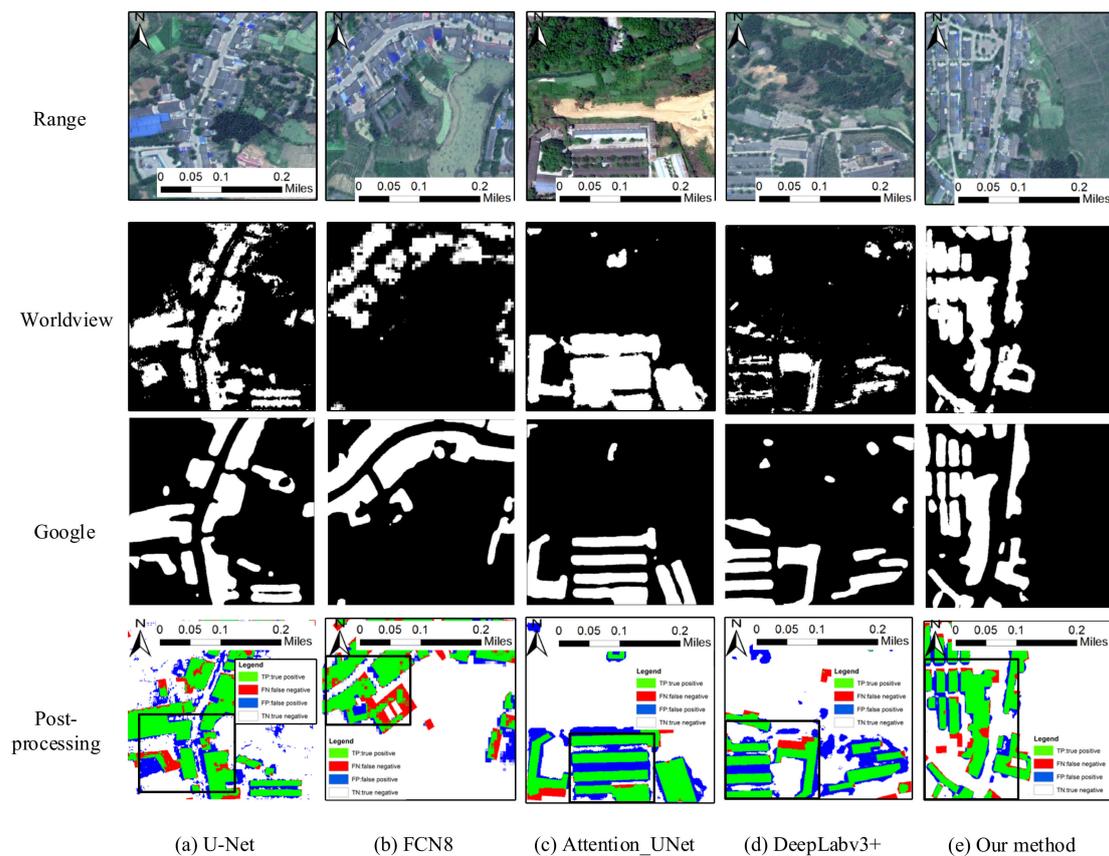


Figure 11. Boundary optimization and fusion comparison results. Columns (a–d) represent extracted results by U-Net, FCN8, Attention_UNet, DeepLabv3+; (e) represent extracted results by our method.

Table 3. Quantitative comparative analysis of fusion results.

| Method | Accuracy | F1 | Recall | IoU |
|----------------|----------|--------|--------|--------|
| U-Net | 89.56% | 53.55% | 87.22% | 38.16% |
| FCN8 | 96.88% | 80.28% | 84.27% | 67.30% |
| Attention_UNet | 94.40% | 72.06% | 89.78% | 57.14% |
| DeepLabv3+ | 96.88% | 80.57% | 86.94% | 67.79% |
| Our model | 97.47% | 85.61% | 93.02% | 74.85% |

4. Conclusions and Future Work

Considering the problems of low extraction accuracy and unclear building boundaries when using existing methods for farmland, a method of attention enhanced U-Net for building extraction from farmland based on Google and WorldView-2 remote sensing images is proposed. The selected farmland range under test covers Qionglai city, Meishan city, Dayi County and Pujiang County of Sichuan Province, and case experiments were performed. The experimental results reveal the following: the accuracy is 97.47%, the F1 score is 85.61%, the recall rate is 93.02%, and the IoU value is 74.85%. All accuracy evaluation indicators are better than those obtained with U-Net, FCN8, Attention_UNet, DeepLabv3+ and other models, which verifies that the method proposed in this paper can effectively extract buildings on farmland. The main contributions of this paper are as follows: first, Resnet is adopted as the U-Net infrastructure, and a spatial and channel attention mechanism module, as well as a multi-scale fusion module, are added to improve the U-Net network and enhance the focus of attention on small building targets in the farmland. Secondly, the method developed uses WorldView-2 and Google remote sensing images to limit farmland boundaries, narrowing the extraction range of buildings and improving extraction accuracy. Finally, a building boundary optimization method based

on morphological filtering is proposed, the extraction results are judged pixel by pixel, and the extraction results of the two images are merged. The method in this paper can effectively solve the problems offered by low accuracy of building extraction results, blurred boundaries, and so on, which can be attributed to the complex types of features, the sparse distribution of buildings, and building occlusion within farmland. Meanwhile, the method provides scientific and technical support for the investigation of buildings within the subject of farmland preservation, which is of great significance to maintaining farmland.

Despite the above achievements, the method presented in this paper also has certain limitations. For example, the buildings in mountainous areas are mostly low-rise bungalows and are relatively old, the boundary between the building and the surrounding ground objects is more blurred, and the method is adversely affected by clouds, rain, fog, and vegetation all the year round. Thus, it is difficult to accurately extract data regarding the buildings in these places. Meanwhile, if several buildings are adjacent and the boundaries are fuzzy, the method outlined in this paper finds it difficult to accurately determine the adjacent relationship of the buildings, and the method is likely to identify them as a single building. The boundary between adjacent buildings is not fully considered, resulting in several adjacent buildings being regarded as one complete building. The further development of remote sensing and interferometric synthetic aperture radar (InSAR) technology could overcome the influence of cloudy and rainy weather conditions and yield higher-resolution remote sensing images. Therefore, in future research, we will continue to integrate additional and higher-resolution remote sensing images, further develop generalization and complex-building abstraction methods, and study the processing methods for neighboring relationships between buildings, to improve building extraction accuracy in farmland.

Author Contributions: Conceptualization, C.L. and L.F.; methodology, C.L.; software, L.F.; validation, C.L., Y.X. and Y.G. (Yukun Guo); formal analysis, C.L. and L.F.; investigation, Y.G. (Yuhang Gong); resources, Z.F.; data curation, Y.X.; writing—original draft preparation, C.L. and L.F.; writing—review and editing, Q.Z.; visualization, C.L. and L.F.; supervision, Q.Z. and J.Z.; project administration, L.F.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Project of Department of Natural Resources of Sichuan Province (grant number KJ-2020-4), Sichuan Science and Technology Program (grant 2020JDTD0003, 2020YFG0083).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chen, Y.M.; Yao, M.R.; Zhao, Q.Q.; Chen, Z.J.; Jiang, P.H.; Li, M.C.; Chen, D. Delineation of a basic farmland protection zone based on spatial connectivity and comprehensive quality evaluation: A case study of Changsha City, China. *Land Use Policy* **2021**, *101*, 105145. [\[CrossRef\]](#)
2. Connell, D.J. The Quality of Farmland Protection in Canada: An Evaluation of the Strength of Provincial Legislative Frameworks. *Can. Plan. Policy Aménage. Polit. Can.* **2021**, *1*, 109–130.
3. Perrin, C.; Clément, C.; Melot, R.; Nougaredes, B. Preserving farmland on the urban fringe: A literature review on land policies in developed countries. *Land* **2020**, *9*, 223. [\[CrossRef\]](#)
4. Perrin, C.; Nougaredes, B.; Sini, L.; Branduini, P.; Salvati, L. Governance changes in peri-urban farmland protection following decentralisation: A comparison between Montpellier (France) and Rome (Italy). *Land Use Policy* **2018**, *70*, 535–546. [\[CrossRef\]](#)
5. Epp, S.; Caldwell, W.; Bryant, C. Farmland preservation and rural development in Canada. In *Agroubanism*; Gottero, E., Ed.; GeoJournal Library; Springer: Cham, Switzerland, 2019; Volume 124, pp. 11–25.
6. Wu, Y.Z.; Shan, L.P.; Guo, Z.; Peng, L. Cultivated land protection policies in China facing 2030: Dynamic balance system versus basic farmland zoning. *Habitat Int.* **2017**, *69*, 126–138. [\[CrossRef\]](#)

7. Shao, Z.F.; Li, C.M.; Li, D.R.; Altan, O.; Zhang, L.; Ding, L. An accurate matching method for projecting vector data into surveillance video to monitor and protect cultivated land. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 448. [[CrossRef](#)]
8. Li, C.X.; Gao, X.; Xi, Z.L. Characteristics, hazards, and control of illegal villa (houses): Evidence from the Northern Piedmont of Qinling Mountains, Shaanxi Province, China. *Environ. Sci. Pollut. Res.* **2019**, *26*, 21059–21064. [[CrossRef](#)] [[PubMed](#)]
9. Shao, Z.F.; Tang, P.H.; Wang, Z.Y.; Saleem, N.; Yam, S. BRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [[CrossRef](#)]
10. Xie, J.L. *Research on Key Technologies of Rural Building Information Extraction Based on High Resolution Remote Sensing Images*; Southwest Jiaotong University: Chengdu, China, 2019.
11. Ji, S.P.; Wei, S.Q.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [[CrossRef](#)]
12. You, Y.F.; Wang, S.Y.; Ma, Y.X.; Chen, G.S.; Wang, B. Building detection from VHR remote sensing imagery based on the morphological building index. *Remote Sens.* **2018**, *10*, 1287. [[CrossRef](#)]
13. Guo, H.N.; Shi, Q.; Du, B.; Zhang, L.P.; Wang, D.Z.; Ding, H.X. Scene-Driven Multitask Parallel Attention Network for Building Extraction in High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4287–4306. [[CrossRef](#)]
14. Liao, C.; Hu, H.; Li, H.F.; Ge, X.M.; Chen, M.; Li, C.N. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049. [[CrossRef](#)]
15. Yang, L.; Wang, H.; Yan, K.; Yu, X.Z. Building extraction of multi-source data based on deep learning. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 5–7 July 2019; pp. 296–300.
16. Sun, G.Y.; Huang, H.; Zhang, A.Z.; Li, F.; Zhao, H.M. Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images. *Remote Sens.* **2019**, *11*, 227. [[CrossRef](#)]
17. Cheng, G.; Han, J.W. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [[CrossRef](#)]
18. Liasis, G.; Stavrou, S. Building extraction in satellite images using active contours and colour features. *Int. J. Remote Sens.* **2016**, *37*, 1127–1153. [[CrossRef](#)]
19. Ghaffarian, S.; Ghaffarian, S. Automatic building detection based on Purposive FastICA (PFICA) algorithm using monocular high resolution Google Earth images. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 152–159. [[CrossRef](#)]
20. Liu, Z.J.; Wang, J.; Liu, W.P. Building extraction from high resolution imagery based on multi-scale object oriented classification and probabilistic Hough transform. In Proceedings of the 2005 International Geoscience and Remote Sensing Symposium (IGARSS'05), Seoul, Korea, 29 July 2005; pp. 2250–2253.
21. Lin, C.G.; Nevatia, R. Building detection and description from a single intensity image. *Comput. Vis. Image Underst.* **1998**, *72*, 101–121. [[CrossRef](#)]
22. Zhang, H.; Zhao, H.; Zhang, X. High-resolution Image Building Extraction Using U-net Neural Network. *Remote Sens. Inf.* **2020**, *35*, 3547. [[CrossRef](#)]
23. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
24. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Munich, Germany, 5–9 October 2015, Springer: Cham, Switzerland, 2015; pp. 234–241.
26. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
27. Yi, Y.N.; Zhang, Z.J.; Zhang, W.C.; Zhang, C.R.; Li, W.D. Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. *Remote Sens.* **2019**, *11*, 1774. [[CrossRef](#)]
28. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
29. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
30. Li, Y.; Xu, W.P.; Chen, H.H.; Jiang, J.H.; Li, X. A Novel Framework Based on Mask R-CNN and Histogram Thresholding for Scalable Segmentation of New and Old Rural Buildings. *Remote Sens.* **2021**, *13*, 1070. [[CrossRef](#)]
31. Zhang, L.L.; Wu, J.S.; Fan, Y.; Gao, H.M.; Shao, Y.H. An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN. *Sensors* **2020**, *20*, 1465. [[CrossRef](#)] [[PubMed](#)]
32. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.W.; Shibasaki, R. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sens.* **2018**, *10*, 407. [[CrossRef](#)]
33. Lin, J.; Jing, W.; Song, H.; Chen, G. ESNet: Efficient Network for Building Extraction from High-Resolution Aerial Images. *IEEE Access* **2019**, *7*, 54285–54294. [[CrossRef](#)]
34. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [[CrossRef](#)]

35. Bai, T.; Pang, Y.; Wang, J.C.; Han, K.N.; Luo, J.S.; Wang, H.Q.; Lin, J.Z.; Wu, J.; Zhang, H. An Optimized faster R-CNN method based on DRNet and RoI align for building detection in remote sensing images. *Remote Sens.* **2020**, *12*, 762. [[CrossRef](#)]
36. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [[CrossRef](#)]
37. Ghaffarian, S.; Valente, J.; Voort, M.V.D.; Tekinerdogan, B. Effect of Attention Mechanism in Deep Learning-Based Remote Sensing Image Processing: A Systematic Literature Review. *Remote Sens.* **2021**, *13*, 2965. [[CrossRef](#)]
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision*; Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 3–19.
39. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building Extraction in Very High Resolution Imagery by Dense-Attention Networks. *Remote Sens.* **2018**, *10*, 1768. [[CrossRef](#)]
40. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. *Remote Sens.* **2019**, *11*, 917. [[CrossRef](#)]
41. Jiang, H.W.; Hu, X.Y.; Li, K.; Zhang, J.M.; Gong, J.Q.; Zhang, M. PGA-SiamNet: Pyramid Feature-Based Attention-Guided Siamese Network for Remote Sensing Orthoimagery Building Change Detection. *Remote Sens.* **2020**, *12*, 484. [[CrossRef](#)]
42. Guo, M.Q.; Liu, H.; Xu, Y.Y.; Huang, Y. Building extraction based on U-Net with an attention block and multiple losses. *Remote Sens.* **2020**, *12*, 1400. [[CrossRef](#)]
43. Xie, Y.K.; Zhu, J.; Cao, Y.G.; Feng, D.J.; Hu, M.J.; Li, W.L.; Zhang, Y.H.; Fu, L. Refined Extraction of Building Outlines From High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1852–1855. [[CrossRef](#)]
44. Chen, L.C.; Zhu, Y.K.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*; Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 833–851.