remote sensing

MDPI

Article

Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters

Yongyang Xu [1], Liang Wu [1,2], Zhong Xie [1,2,*] and Zhanlong Chen [1]

[1] Department of Information Engineering, China University of Geosciences, Wuhan 430074, China; yongyangxu@cug.edu.cn (Y.X.); wuliang@cug.edu.cn (L.W.); chenzhanlong2005@126.com (Z.C.)
[2] National Engineering Research Center of Geographic Information System, Wuhan 430074, China
* Correspondence: xiezhong@cug.edu.cn

Received: 19 December 2017; Accepted: 16 January 2018; Published: 19 January 2018

Abstract: Very high resolution (VHR) remote sensing imagery has been used for land cover classification, and it tends to a transition from land-use classification to pixel-level semantic segmentation. Inspired by the recent success of deep learning and the filter method in computer vision, this work provides a segmentation model, which designs an image segmentation neural network based on the deep residual networks and uses a guided filter to extract buildings in remote sensing imagery. Our method includes the following steps: first, the VHR remote sensing imagery is preprocessed and some hand-crafted features are calculated. Second, a designed deep network architecture is trained with the urban district remote sensing image to extract buildings at the pixel level. Third, a guided filter is employed to optimize the classification map produced by deep learning; at the same time, some salt-and-pepper noise is removed. Experimental results based on the Vaihingen and Potsdam datasets demonstrate that our method, which benefits from neural networks and guided filtering, achieves a higher overall accuracy when compared with other machine learning and deep learning methods. The method proposed shows outstanding performance in terms of the building extraction from diversified objects in the urban district.

Keywords: building extraction; deep learning; guided filter; very high resolution

1. Introduction

Remote sensing images with very high resolution (VHR) are widely used in many applications including land cover mapping and monitoring [1], multi-angle urban classification analysis [2], automatic road detection [3], as well as the identification of tree species in forest management [4]. Several of the practical applications are based on VHR remote sensing imagery classification at the pixel level [5–8], also defined as semantic segmentation. Semantic segmentation of remote sensing imagery aims to classify every pixel into a given category, and it is an important task for understanding and inferring objects [9,10] and the relationships between spatial objects in a scene [11].

Automatic semantic annotation of urban areas plays an important role in many photogrammetry and remote sensing applications, such as building and updating a geographical database, land cover change, and extracting thematic information. In recent years, the development of computing hardware and sensor technologies has made high resolution sampling available with a ground sampling distance (GSD) of 5–30 cm [12] so that objects such as roof tiles, cars, buildings, and individual branches of trees, are distinguishable, which has increased the interest to perform semantic segmentation in urban areas.

In the past several years, spatial and spectral features have been used to improve the performance of VHR semantic segmentation based on pixel-wise analysis. Spatial contextual information like the grey level co-occurrence matrix (GLCM) has been employed to obtain a more accurate classification

map [13]. A novel mean shift (MS)-based multiscale method was used in urban mapping [14]. Morphological profiles (MP) were utilized into the spatial-spectral classification [15]. Conditional random fields and machine learning, such as SVM and random forest, were also introduced to solve the classification of remote sensing images [6,16]. In addition, encouraged by deep neural network features that have been shown to have an outstanding capacity in visual recognition [17], object detection [18] and semantic segmentation [19–21], deep learning was introduced to resolve the old problems in remote sensing [22]. Deep neural networks have been successfully used to class and densely label high resolution remote imagery [23]. It can be used in various remote sensing tasks: Detection, classification, or data fusion [24]. A deep learning framework was proposed to detect buildings in high-resolution multispectral imageries (RGB and near-infrared) [25]. Multi-scale convolutional neural networks (CNNs) combined with the conditional random fields (CRFs) were used for dense classification in street scenes [26]. An end-to-end trainable deep convolutional neural network (DCNN) was built to improve semantic image segmentation with boundary detection [12].

Studies have shown that remote sensing image classification results cannot be conclusive [27]. The reason for this is that although the resolution of remote sensing images have improved, which has been helpful to detect and distinguish various objects on the ground, these improvements have made it more difficult to separate some objects, especially spectrally similar classes, due to the increase of the intra-class variance of objects, such as building, streets, shades and cars, and with decrease of the inter-class variance [5,28,29]. In other words, different objects may present the same spectral values within the remote sensing imagery, which make it more difficult to extract reasonable spatial features to resolve the classification of pixels in extracting the buildings.

In the last years, fully convolutional networks (FCNs) have shown a good performance of semantic segmentation [30–32]. Indeed, FCNs can not only learn how to classify pixels and determine what it is, but they can also predict the structures of the spatial objects [33]. The model is able to detect different classes of objects on the ground and predict their shapes, such as buildings, the curves of the roads, trees, and so on. However, it is a little short of being capable of detecting small objects or objects with many boundaries, because the boundaries of the objects are blurred and the results are visually degraded during classing when using FCNs [12].

There has been some research that tries to improve the performance of semantic segmentation and develop a deep neural network structure either by adding skip connections so as to reintroduce the high-frequency detailed information of an imagery after upsampling [34,35] or by using dilated convolution combined with CRFs [36]. The improved FCN model, which is designed as a multi-scale network architecture by adding a skip-layer structure, was trained to perform state-of-the-art natural image semantic segmentation [31]. A deep FCN with no downsampling was introduced to boost the effective training sample size and improve the classification accuracy [37].

The application of research into urban district classification using VHR remote sensing imagery ranges from urban management to flow monitoring. Recent research makes an effort to improve the accuracy in areas such as encoding of images, extraction of features from raw images [38,39], and the use of deep neural networks such as CNNs, FCNs, and so on, to label pixels, especially for the VHR remote sensing imagery [40,41]. However, pixel labelling of the VHR imagery in urban districts offers challenges relating to the varied semantic classes and geometry shapes. Because buildings and the other imperviousness objects in urban areas are very complicated with respect to both their spectral and spatial characteristics, it is inefficient and difficult to extract them. The VHR imagery is usually limited to three or four broad bands, and these spectral features alone may lack the ability to distinguish the objects because different objects have similar spectral values, for example, roads and roofs. Additionally, the same objects may have different spectral values, for example a roof that is divided into two parts by exposure to the sun and the shade. Therefore, discriminative appearance-based features are needed to improve the performance. Fortunately, most of the VHR remote sensing imageries usually have the corresponding overlapping image (or combined camera +

LiDAR systems) [12], and the digital surface model (DSM) is available, which can be regarded as an additional depth channel.

Previous researchers have provided useful insights into the various methods that can be used in pixel labelling. However, these methods cannot clearly detect the boundary of the objects, and lack the ability to remove the salt-and-pepper class noise; some pixels with similar spectral values are usually misclassify. To resolve these problems, this work attempts to take semantic labelling methods from computer vision and apply them to building extraction from VHR remote sensing imageries.

In this paper, we try to improve the classification accuracy by a new model based on deep residual networks (ResNet) [42]. At the same time, we introduce an object-oriented guided filter to improve the performance of classification. This method, on paper, involves three steps. First, imagery pre-processing is needed to prepare the dataset for deep learning. Second, a deep network is trained to segment VHR remote sensing imagery into two classes: buildings and clutter/unknown. Third, a guided filter is employed to optimize the extraction buildings and an ultimate spectral-spatial classification map of the urban district is achieved by fusing the object-oriented optimized results. All the challenges have resulted in improving the classification accuracy of complex urban area remote sensing imagery. The major contribution of this work is proposing a new model based on ResNet that we defined as Res-U-Net, and exploring a novel framework to perform classification of VHR remote sensing imagery. The experimental results show that the novel framework is more effective at extracting buildings.

The remainder of this paper is organized as follows: Section 2 presents the building extraction using VHR imagery in urban areas based on deep learning and guided filters; Section 3 describes the experimental results and how to set the parameters; Section 4 is a discussion of our method and Section 5 presents our concluding remarks.

2. Methods for Classification in Very High Resolution Remote Sensing Imagery

In this work a pixel classification method to extract buildings from urban districts within VHR remote sensing imageries based on deep learning and guided filters is proposed. First, the imageries are pre-processed and edge enhancing is used to emphasize the pixels which exist at the edges of the buildings. Some hand-crafted features including the normalized differential vegetation index (NDVI), the normalized digital surface model (NDSM), and the first component of the principal component analysis (PCA1) are extracted based on the color infrared (CIR) imagery, red green blue (RGB) satellite imagery as well as the corresponding digital surface model (DSM). Then, the proposed deep neural network Res-U-Net is introduced for pixel classification, where the hand-crafted features, the original bands, and the ground truth (labeled artificially) are treated as inputs to train the network. The output of the deep neural network is the segmentation map that represents the pixel labeling results. Finally, we briefly introduce the concept of a guided filter to fine-tune the pixel labeling results because the convolutional network tends to blur object boundaries and visually degrade the result when it is applied to remote sensing data [12]. An overview of the proposed pixel classification framework is illustrated in Figure 1.

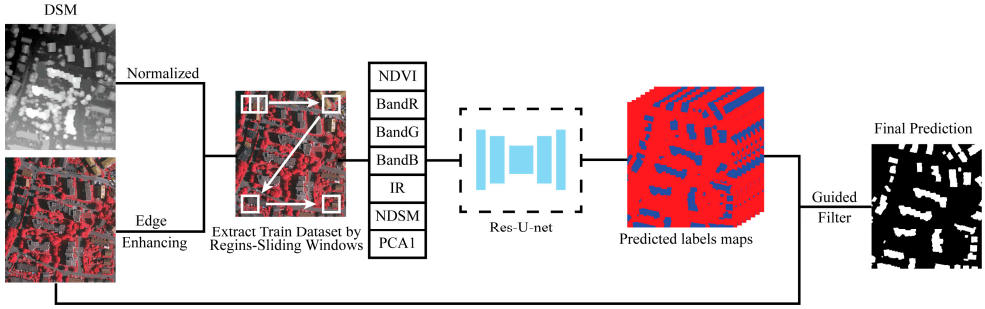

Figure 1. The framework of the pixel classification using deep learning and a guided filter.

2.1. Deep Learning for Remote Sensing Imagery Classification

Convolutional networks have been widely utilized in applications ranging from whole-image classification [43–45] to pixel classification as semantic segmentation in computer vision. Pixel classification includes automatically building maps of geo-localized semantic classes (for example: buildings, impervious surfer, vegetation, and so on) from the earth-observation data [46]. In recent years, deep learning has become a state-of-the-art tool for pixel classification in remote sensing, as well as other fields. Fully convolutional networks are adapted as effective tools for the semantic labelling of high-resolution remote sensing data. This paper uses the modified and extended architecture ResNet, named Res-U-Net (Figure 2).

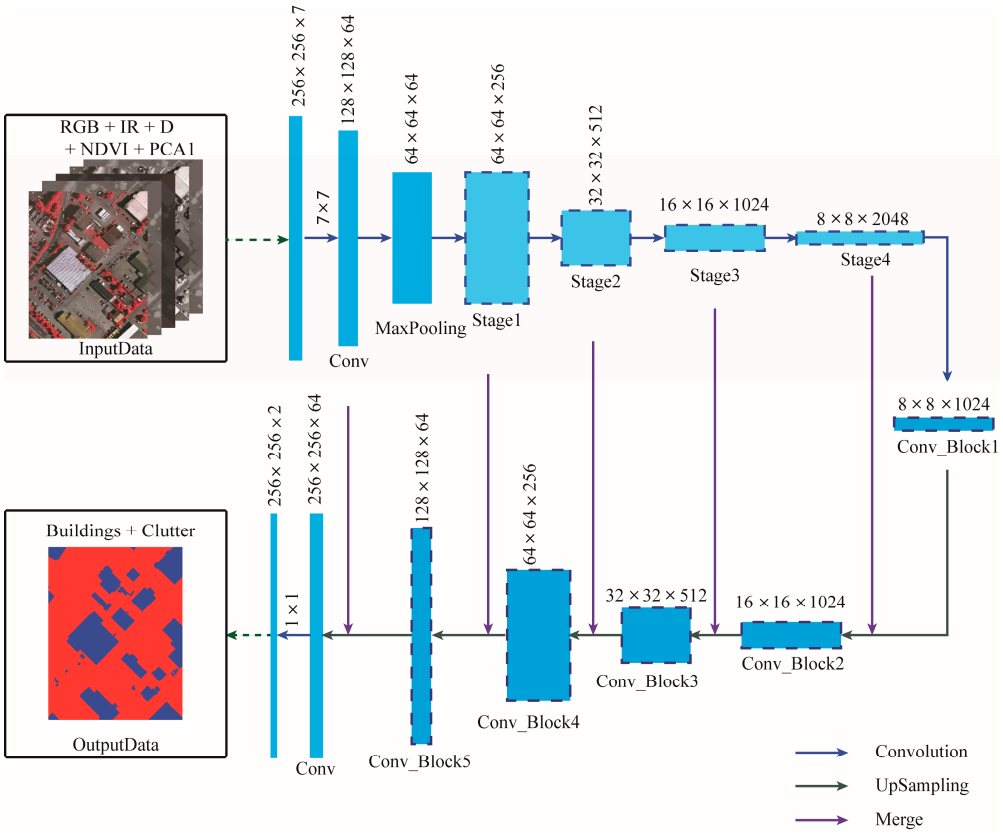

Figure 2. The architecture of the Res-U-Net used in this work. Each box represents the feature map and the x-y-size of the map is provided at the lower right edge of the box. The arrows denote the different operations which are explained at the lower right of the figure.

In this paper, we trained the Res-U-Net by adopting the approach of reference [47], which is famous for having the ability to work with very little training data but still obtain precise segmentation. The Res-U-Net network consists of two paths: contracting (left) and expansive (right). The left part is the ResNet, which is used to extract the features of input data, and we modified the input layer to adapt the seven elements of the input data. The input layer is followed by a normalization layer and a max pooling layer. The activation layer in the network contains a rectified linear unit (ReLU) and a 2 × 2 max-pooling operation for the subsampling, both of them improve the robustness of the network against distortions and small translations [44]. During the features extraction, there are four stages and every stage includes several residual blocks. The feature maps in the same block have the same size, and the feature maps in the following blocks are half that of the previous ones. The feature maps in different blocks have different scale features. The expansive part aims to extract the buildings using the feature maps. The number of stages in contracting and expansive is the same.

Inspired by the feature pyramid networks [48], to obtain the features in multiple scales, a concatenation with the corresponding stage from the contracting part is designed in the deep neural network. Every stage in the expansive part includes the upsampling of the feature map, a concatenation block and a convolution block, which consists of a 3×3 convolution layer, a normalization layer and a rectified linear unit. At the end of the network, a 1×1 convolutional layer is added to map the feature vectors to the two classes of buildings and clutter, the outputs of this layer indicate the class scores for the pixel. A softmax layer, used to calculate the classification results, is added at the end of the network. In this work, the deep convolutional network uses the ResNet as a feature extractor, which solves the degradation problem during the layer increases, and it is useful to extract the features in contracting. The concatenation in the expansive part is able to learn multiple scales and different level features, which increases the robustness of the network and improves the accuracy of the building extraction. The output of the softmax layer is a probability map with two channels. It presents the result of the classification between buildings and clutter in every pixel.

Within the remote sensing imagery and their corresponding normalized digital surface model, hand-crafted features such as NDVI, PCA1 as well as the classified segmentation maps are regarded as the inputs to train the network. The Res-U-Net builds higher level features by the grouping of mapping features of lower level features, and therefore, the results are located more accurately. It transmits the error from a high level to a low level and speeds up the training [47]. The size of the output of the network is the same as the input and it usesnd-to-end processing. At the beginning of the network, max-polling and convolution layers produce more abstract feature maps, which are beneficial for the up-convolution in order to calculate an accurate pixel classification result.

The building extraction problem can be regarded as a binary classification problem. During the training of the parameters, it can be solved by a logistic regression using the optimization of the energy function. As with other training methods [47], we train the network using the gradient descent to minimize the energy function. The energy function is calculated by the softmax as well as the cross entropy loss function. The softmax is used to calculate the probability map, defined as:

$$p_k\left(x^i\right) = \frac{\exp\left(\left(w_k\right)^T x^i\right)}{\sum_{j=1}^{K} \exp\left(\left(w_k\right)^T x^i\right)} \tag{1}$$

where $k \in \{1, 2\}$ which corresponds to the buildings and the clutter, and K represents the number of classes as two. $p_k\left(x^i\right)$ is the probability that sample x^i belongs to class k. The energy function is defined as follows:

$$E = \sum_{i=1}^{m} \sum_{k=1}^{K} w\left(x^i\right) \log p_k\left(x^i\right) \tag{2}$$

where $\left\{\left(x^i, y^i\right)\right\}_{i=1}^{m}$ is assumed to be the training data, x^i represents the vectored features, and y^i is the labeled data, m represents the number of samples, and w is a weight map in the network to be optimized.

2.2. Guided Filtering

To fine-tune the buildings extracted by deep learning, the guided filter, which was firstly proposed by He [49], is introduced in this work. Like the bilateral filter, it is an edge-preserving smoothing technique. Thanks to the guiding of the input image (guidance image), the filtering result is more structured and less smoothed. The guided filter is better than the bilateral filter in terms of detail and it is more effective [49], which makes it widely applicable in computer vision and graphics [50]. The guided filter assumes that the local linear model exists between the guidance image and the filtering result, so that it will benefit to optimize object classification like buildings.

The guided filter involves two input images including a guidance image I_c and a filtering image I_{in}. The filtering output O is assumed to be a linear transform of I_c in a window w_k:

$$O(i) = a_k I_c(i) + b_k \tag{3}$$

where a_k and b_k are the coefficients of the linear transform between the guidance image I_c and the filtering image O within window w_k (the size of window is $w \times w$). They can be calculated as follows:

$$a_k = \frac{\frac{1}{w^2}\sum_{i \in w_k} I_c(i) I_{in}(i) - u_k \overline{p}_k}{\sigma_k^2 + \varepsilon} \tag{4}$$

$$b_k = \overline{p}_k - a_k u_k \tag{5}$$

where, u_k and σ_k are the mean and variance of the guidance image I_c within the window w_k, and $\overline{p}_k$ is the mean of the filtering image I_{in} within the window w_k, and ε controls the blur degree of the guided filter. Because pixel i has a relationship with all the windows that cover it, the output of filtering O (i) is calculated as:

$$O(i) = \overline{a}_i I_c(i) + \overline{b}_i \tag{6}$$

where $\overline{a}_i$ and $\overline{b}_i$ are the mean of coefficients of all the windows that cover the pixel i. For simplicity, the equation can be rewritten as:

$$O = G(I_{in}, I_c, w, \varepsilon) \tag{7}$$

The original imageries are treated as guiders to optimize the boundaries in order to remove the salt-and-pepper class noise. The result, directly fine-tuned by the guided filter, will result in the over-smoothness of the extracted buildings in the output. However, the building maps should be binary and the pixels in the boundaries change gradient in reality. Therefore, we set a threshold during filtering. If the value is larger than the threshold it will be set to 255, which represents buildings, otherwise, it is equal to 0, which represents the clutter.

3. Results

3.1. Datasets

The ISPRS 2D semantic labelling VHR remote sensing imageries of urban districts are used in the experiments, including the Vaihingen (Germany) and Potsdam (Germany) datasets, as these are open asset datasets provided online. Both of them consist of the near infra-red, red, and green ortho-rectified imagery (or color infra-red, CIR). The corresponding digital surface models (DSMs) generated by dense image matching and ground truth labels are annotated manually. Additionally, the Potsdam dataset has a blue channel, containing 38 ortho-rectified aerial IRRGB images of $\approx 6000 \times 6000$ (in total, over 1,368,000,000 pixels) at 5 cm spatial resolution, where 24 tiles are labelled with pixel-level ground truth. The Vaihingen dataset comprised of 33 large image patches of $\approx 2500 \times 2500$, extracted from a larger orthophoto imagery captured over Vaihingen. Overall, there are about 168,287,871 pixels, and the imageries have a ground sample distance (GSD) of 9 cm, where 16 tiles are labelled with pixel-level ground truth. Each of the ground truth labels are made up of building and unknown (clutter). The DSM is a value array which has the same size as the input image and the labelled ground truth. At the same time, the normalized DSMs [51] are available for us, where the height is computed using the off-ground pixels. The imageries with ground truth are divided into two parts, where 80% are used to train the Res-U-Net and 20% are used to validate the trained model.

3.2. Preprocessing the Data for Deep Learning

Although the urban remote sensing imagery used in this work is in high resolution, some object edges are still fuzzy, which result in the object being unrecognizable from the background. Therefore,

this work introduces the edge enhancement effect to the remote sensing imagery processing. The edge enhancement is an image-filter that reduces the effect of noise. It can also decrease the complexity of the image computation. Edge enhancement is widely used in fields such as pattern recognition, image semantic segmentation, and so on. This work enhances the edge of the imageries using the python imaging library (PIL). It is a kind of convolutional filter, where a $n \times n$ matrix is defined to operate with the digital imagery. Every pixel of the edge enhancement result is a sum-weighted value of the convolution region. The size of the convolution kernel used in the experiment is 5×5.

The size of the total from dataset is approximately 6000×6000. If the whole dataset is used as an input for the deep network, millions of paragraphs must be learned, which would lead to a lack of memory. Therefore, we processed the imageries using a 256×256 sliding window with a stride of 64 px to produce the samples. Every eight samples were regarded as a batch to train the network.

3.3. Experimental Setup and Results

To improve the accuracy of the vegetation in this experiment, we computed the NDVI from the near-infrared and the red channels, and it was used as an indicator for the vegetation ($NDVI = (NIR - R)/(NIR + R)$). A PCA transformation was introduced to extract the first component comprising of brightness, which will be beneficial to classify some special building roofs. The bands of R, G, B (there is no blue band in the Vaihingen data), and CIR, as well as the hand-crafted features including NDVI, NDSM, the first component of PCA, and the corresponding ground truth (Figure 1) are used as inputs to train the Res-U-Net. The architecture, as well as the parameters used in this work, is shown in Figure 3.

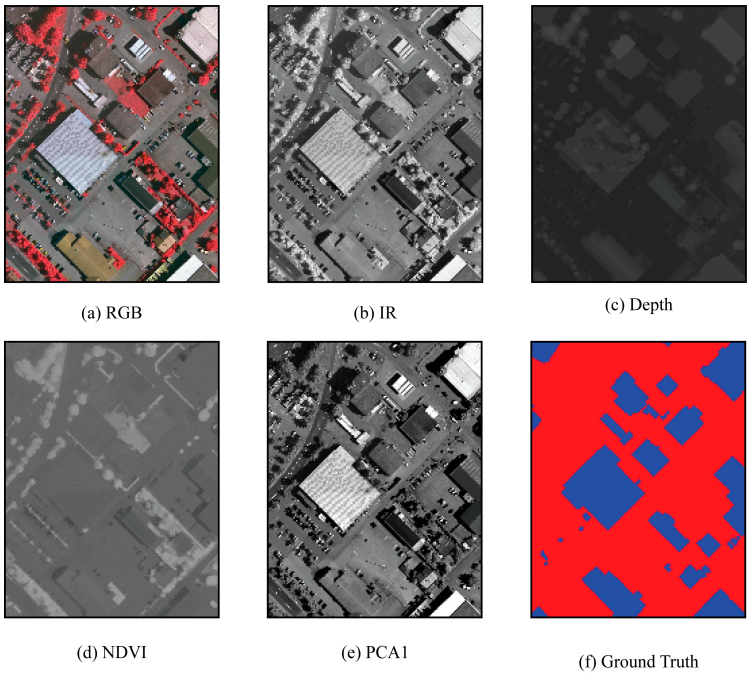

Figure 3. Samples of the urban remote sensing imageries used in the experiments.

For an individual network, we trained the network with a learning rate of 0.001. To ensure an outstanding learning result, we divided the learning rate by ten every ten epochs. There are 100 epochs during the training and each epoch has 2048 samples. We use the Adam as the optimizer to optimize the network when adjusting parameters like weights, biases, and so on. In case most of the evaluation data have targets, we set the size of evaluation data as 2000×2000.

The provided metrics of F_1 score and the global pixel-wise accuracy of each class are used to assess the quantitative performance. F_1 score is a representation of the harmonic mean of precision and recall, and it can be calculated as follows:

$$F_1{}^i = 2 \times \frac{precision_i \times recall_i}{precision_i + recall_i} \tag{8}$$

where

$$precision_i = \frac{TP_i}{TP_i + FP_i}, recall_i = \frac{TP_i}{TP_i + FN_i} \tag{9}$$

Here, TP_i is the number of true positives for class i, FP_i and FN_i represent false positive and false negative, respectively. These metrics are computed using the pixel-based confusion matrices per tile or by an accumulated confusion matrix. At the same time, the overall accuracy (OA) can be obtained by normalizing the trace from the confusion matrix [52].

The proposed deep learning of the Res-U-Net is implemented using Tensorflow and Keras in the Linux platform with a TITAN GPU (12 GB RAM). After 204,800 iterations, our best model achieves state-of-the-art results on the datasets (Table 1). The changing accuracies and losses of the Potsdam and Vaihingen datasets with the increasing epochs are shown in Figure 4.

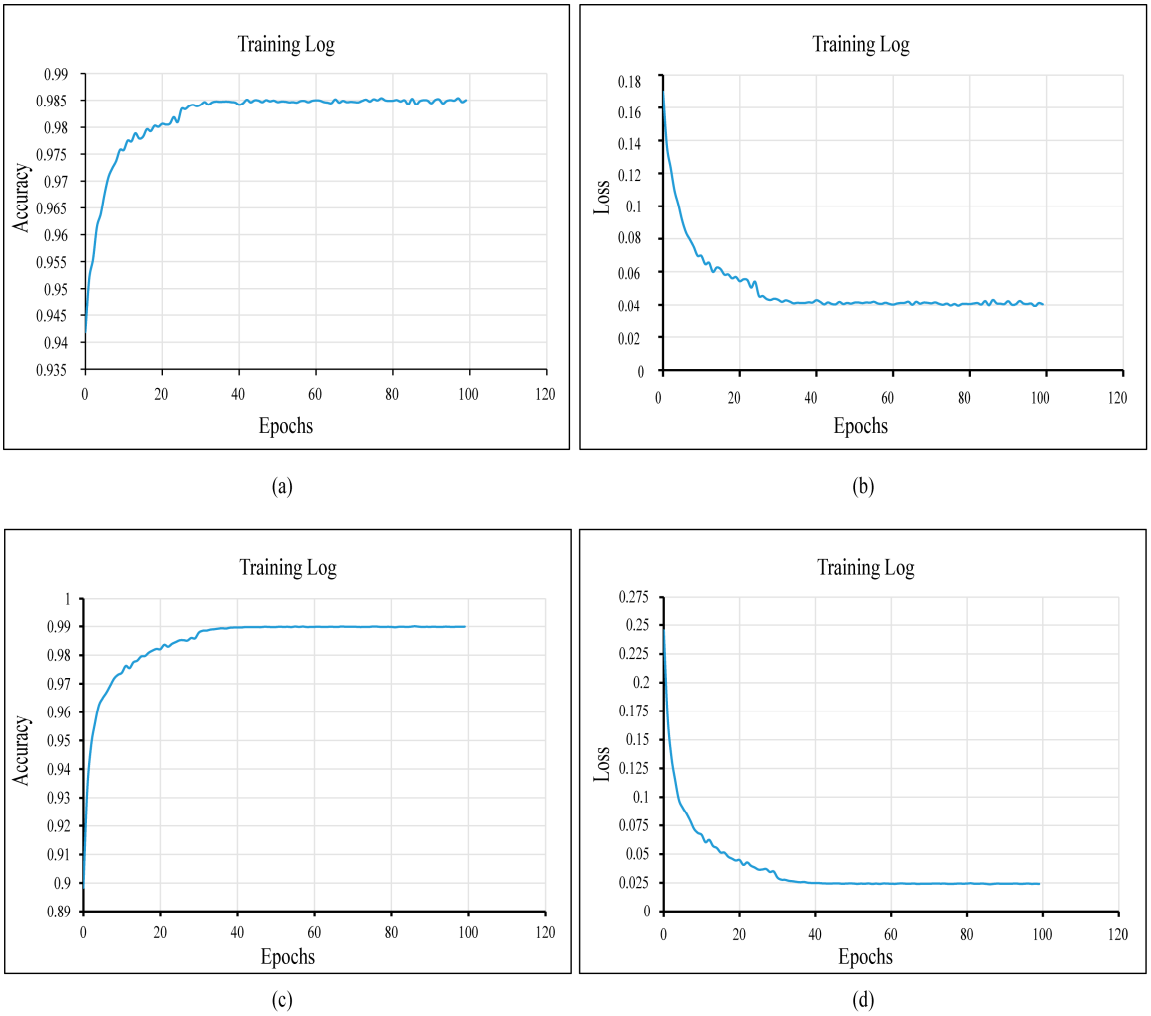

Figure 4. Plots showing the accuracy and loss of the Res-U-Net network for training the datasets. The training accuracy (**a**) and the loss (**b**) change with the epochs increasing in Potsdam. The training accuracy (**c**) and the loss (**d**) change with the epochs in Vaihingen.

The architecture reaches 96.91% overall accuracy over the Potsdam and 97.71% overall accuracy over Vaihingen, respectively. The deep learning frame performs particularly well on impervious ground and the buildings (Figure 5).

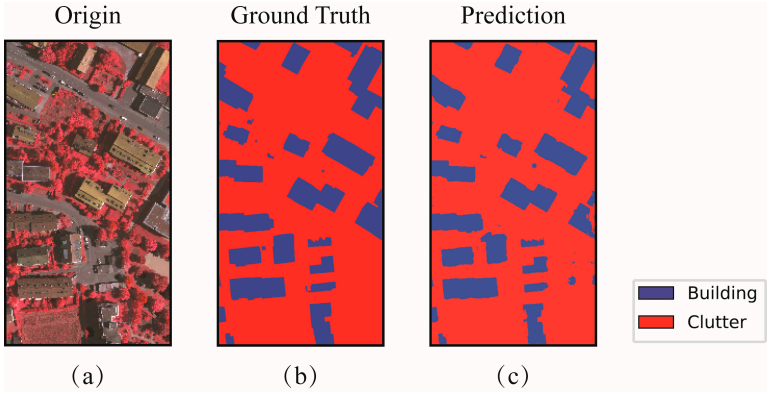

Figure 5. Results of the buildings extraction of the Vaihingen dataset using deep learning, the imagery (**a**) as well as the corresponding ground truth (**b**) and the prediction (**c**).

Although the accuracy of the pixel labelling improved by using edge enhancement and deep neural networks, the boundaries of the buildings were still blurry and some pixels belonging to the buildings were misclassified (Figure 6b,e). To improve the performance, a guided filter was introduced. During the optimization by the guided filter, we set values larger than the threshold (t = 90) to 255, which is mentioned in Section 2.2. Otherwise, the values are set to 0. The original imageries as well as the prediction results produced by deep learning are used as the input for the guided filter. From the results (Figure 6), it is clear that the performance in both of the datasets improved.

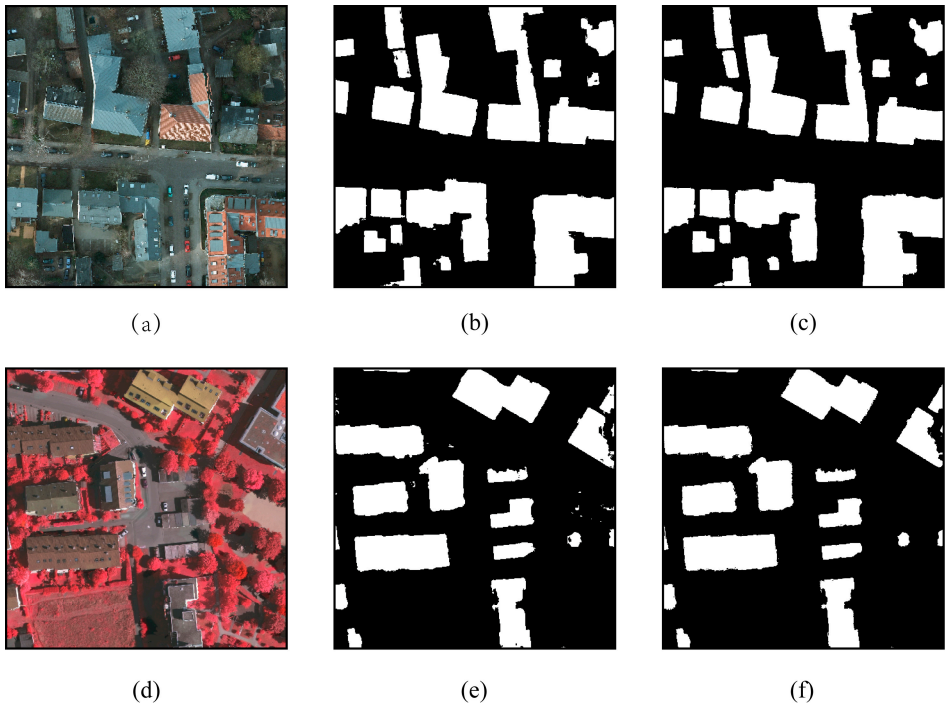

Figure 6. Results of the guided filter. (**a**) represent the original imageries in the Potsdam and Vaihingen datasets, respectively. (**b**) represent the corresponding prediction from deep learning. (**c**) represent the results of guided filter. (**d**–**f**) are the original imageries, prediction from deep learning and the results of guided filter in Vaihingen, respectively.

Table 1. Measures the average accuracy for the classification, precision, recall as well as F_1 score for buildings and clutter in Potsdam and Vaihingen, respectively.

Dataset	OA	Precision (B)	F_1 (B)	Recall (B)	Precision (C)	F_1 (C)	Recall (C)
Postdam	0.9691	0.9634	0.9390	0.9158	0.9709	0.9793	0.9878
Vaihingen	0.9771	0.9621	0.9515	0.9412	0.9816	0.9850	0.9883

Where B stand for buildings, and C represent clutter, and OA represents overall accuracy.

4. Discussion

4.1. Some Effects to the Result of Deep Learning

Although VHR remote sensing imagery is easily applied to distinguish objects on the ground, some edges are not obvious between objects with similar spectral values, so it is difficult to classify the pixels, especially in the urban districts. This work introduces edge enhancing to increase the differences among objects which leads to better performance during classification. We compared the overall accuracy for buildings and clutter classification, as well as precision, recall and F_1 (mentioned above) by both using and not using the preprocessing (Figure 7), respectively.

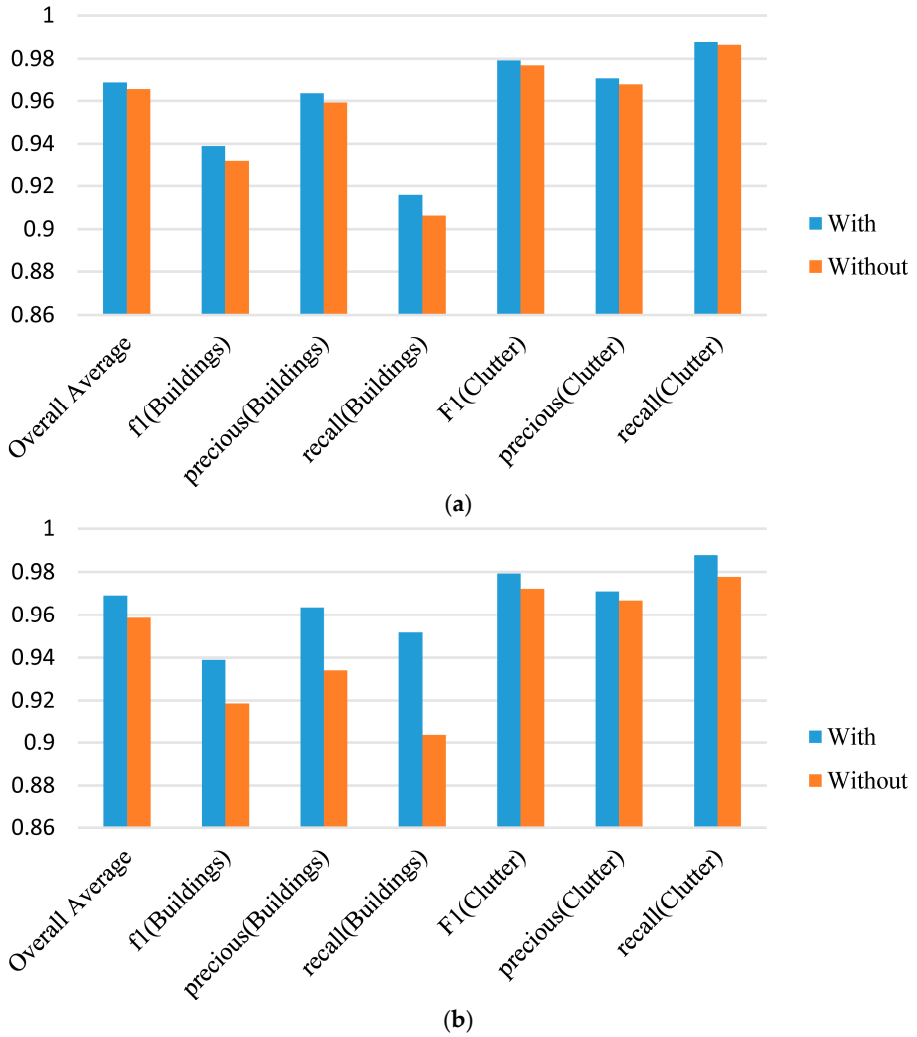

Figure 7. The results of the building extraction from the urban districts of the Potsdam (**a**) and Vaihingen (**b**) datasets with and without edge enhancing in the preprocessing stage.

As we can see, the overall accuracy of Potsdam has improved by 0.43% and the overall accuracy of Vaihingen has improved by 2.94%. At the same time, the precision and recall for buildings has improved compared to the results computed using the inputs without edge enhancing. Edge enhancement is able to emphasize the indistinct pixels at the edges of the buildings so that they can be classified more precisely, as shown in Figure 8. It can be easily observed that the performance is poor in some parts like A, B without the edge enhancing preprocessing.

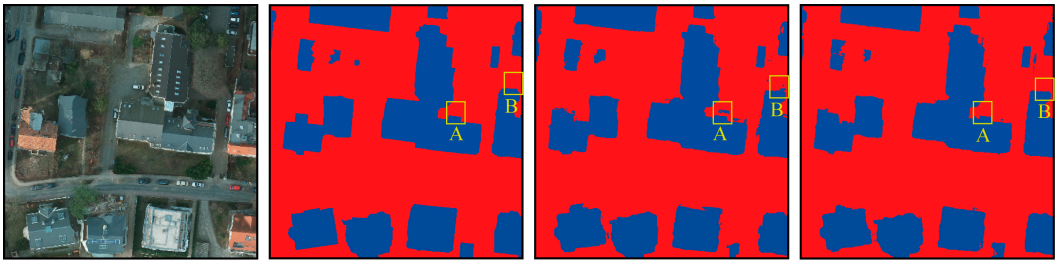

Figure 8. Results of the building extraction in the urban district of the Potsdam dataset with and without edge enhancing preprocessing. (**a**) The original imagery of this urban district. (**b**) The ground truth of this region. (**c**) The prediction results using the Res-U-Net without enhanced preprocessing. (**d**) The prediction results with enhanced preprocessing.

Precision, recall and the F_1 scores have significantly improved thanks to the discriminative power of the digital surface model (DSM) and NDVI. To illustrate some differences between the results achieved by the DSM and NDVI, the controlling variable method was adopted to analysis of the effects of the elements. We compared the performance of deep learning whilst exclude either the DSM or the NDVI and the performance of deep learning only treat the RBG images as input. Table 2 compares the results on the Vaihingen and Potsdam datasets. It can be clearly observed that the results support the idea that it is beneficial to use the DSM and the NDVI, and that they improve the overall accuracy by 1.64% and 0.39% for the Potsdam dataset and 1.45% and 2.19% for the Vaihingen dataset. They also improve the F_1 by 3.92% and 0.83% for Potsdam and 2.42% and 3.89% for Vaihingen. Compared with the results, it is clear that the limitation of the input only with RBG images, the overall accuracy of deep learning decreased by 2.66% and 2.7%, respectively; and F_1 for building decreased by 5.71% and 4.13%, respectively, whilst exclude both the DSM and the NDVI.

Table 2. (**a**) Compared with the results whilst exclude either the DSM or the NDVI for Potsdam dataset. (**b**) Compared with the results whilst exclude either the DSM or the NDVI for Vaihingen dataset.

Elements	OA	Precision (B)	F_1 (B)	Recall (B)	Precision (C)	F_1 (C)	Recall (C)
all	0.9691	0.9634	0.9390	0.9158	0.9709	0.9793	0.9878
Without DSM	0.9527	0.9606	0.8998	0.8462	0.9483	0.9688	0.9901
Without NDVI	0.9652	0.9644	0.9307	0.8993	0.9655	0.9768	0.9883
Only RGB	0.9425	0.9471	0.8819	0.8251	0.9412	0.9621	0.9838
Elements	**OA**	**Precision (B)**	**F_1 (B)**	**Recall (B)**	**Precision (C)**	**F_1 (C)**	**Recall (C)**
all	0.9771	0.9621	0.9515	0.9412	0.9816	0.9850	0.9883
Without DSM	0.9626	0.9341	0,9273	0.9207	0.9725	0.9749	0.9773
Without NDVI	0.9552	0.9228	0.9126	0.9026	0.9663	0.9699	0.9737
Only IRRG	0.9501	0.9181	0.9102	0.9025	0.9618	0.9677	0.9736

B stand for buildings and C represent clutter, and OA means overall accuracy.

By analysis, it is clear that the performance using the DSM as a channel of input has improved when compared to the case without the DSM. The recall for buildings in the two datasets decreased by 6.96% and 2.05%, respectively. That is to say, the nature of some pixels that are buildings are

misclassified as clutter. Although the pixels that belong to a roof exposed to the sun and a roof out of the sun are different, they have the same DSM value, so it will perform well when extracting all kinds of building roofs. Some road pixels are very similar to the roof of the building in terms of spectral characteristics, but they have a large difference in DSM. As a result, DSM improves the capability of the model to extract buildings and the classification precision of OA, buildings and clutter. The results can be observed in Figure 9.

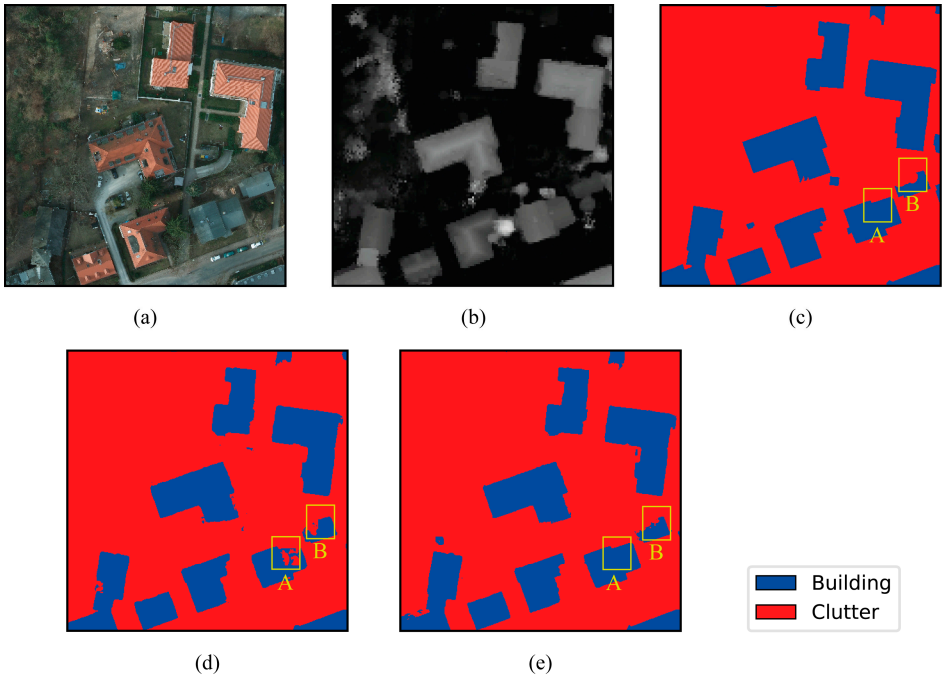

Figure 9. The results of the building extraction in the urban district of the Potsdam dataset with and without the DSM as a channel of the input. (**a**) The original imagery of the urban district. (**b**) The DSM of this district. (**c**) The ground of this region. (**d**) The prediction results using the Res-U-Net without the DSM as a channel of input. (**e**) The prediction results corresponding to the input with the DSM.

The NDVI can show the impact of the underlying background of buildings and the vegetation canopy structure to some degree. In urban areas, some low buildings are always covered with tress, which make it difficult to classify, like part A and B in Figure 10. When training the network without the NDVI, the overall accuracy and F_1 for both buildings and clutter in both Potsdam and Vaihingen datasets decreased. The recall for buildings in the two datasets decreased by 1.65% and 3.86%, respectively. The results (Figure 10) show that the NDVI as a channel of input to train the model is beneficial to solve the problem.

Compared with other methods using the same datasets (that is, the training and validation datasets), the results are reported in Table 3. The Res-U-Net proposed in this work shows improvements on building extraction in both datasets. The network extracts features using the ResNet, which works well in contracting, and it benefits a lot from solving the degradation problem during the increase of layers. The expansive concatenated with multiple scales in different blocks and is helpful in classifying the buildings of different sizes.

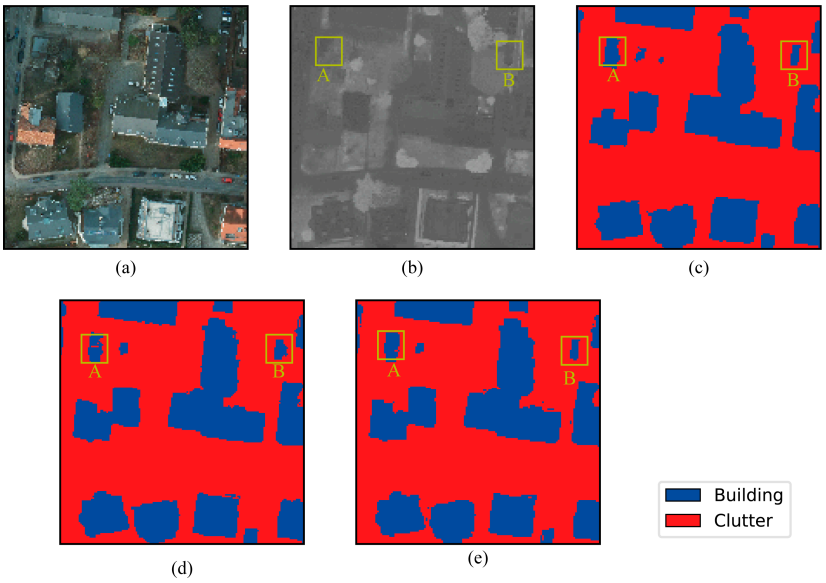

Figure 10. The results of the building extraction in the urban district of the Potsdam dataset with and without the NDVI as a channel of the input. (**a**) The original imagery of the urban district. (**b**) The NDVI of this district. (**c**) The ground of this region. (**d**) The prediction results using the Res-U-Net without the NDVI as a channel of input. (**e**) The prediction results corresponding to the input with the NDVI.

Table 3. Compared with the results of proposed method with others methods on Vaihingen and Potsdam datasets.

Dataset	SegNet	FCN	CNN + RF	Mulit-Scale Deep Network [33]	CNN + RF + CRF	Ours
Vaihingen	0.9078	0.9279	0.9423	0.945	0.943	0.9771
Potsdam	0.9174	0.9127	0.9303	0.9406	0.9392	0.9691

4.2. Influence of the Guided Filter

The threshold used in the optimization by the guided filter is important. Since some pixels near the building edge and the spectrum are similar to the buildings if the threshold is smaller, more pixels will be extracted as buildings and lead to the extracted building area being larger than the real building area. On the other hand, if the threshold is larger, some unclear edges will be excluded and the extracted building area will be smaller than the real area of the buildings (Figure 11).

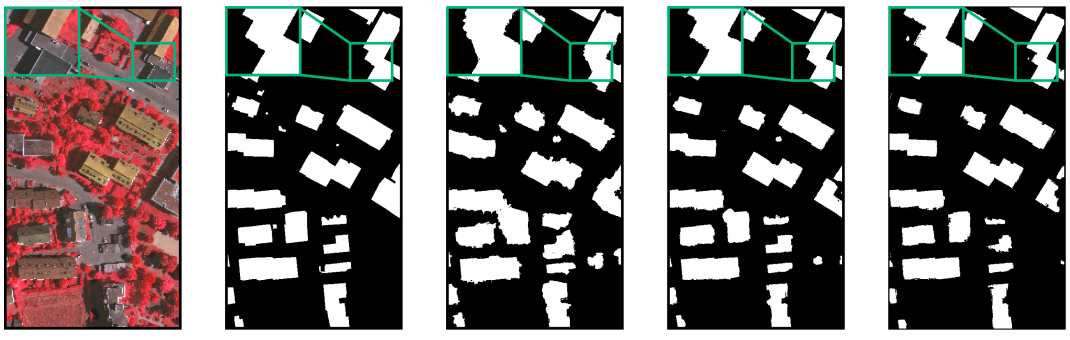

Figure 11. The results of the guided filter. (**a**) The imagery of the urban district. (**b**) The ground truth corresponding with the imagery. (**c**) The optimization result with a guided filter with a threshold $t = 40$. (**d**) The optimization result with a guided filter with a threshold of $t = 90$. (**e**) The optimization result with a guided filter with a threshold of $t = 165$.

To get the optimal threshold, we compared the overall accuracy and F_1 of the results using different thresholds. The guided filter with different thresholds was then used by the same predicted results of the Res-U-Net. The thresholds range was between 40 and 175 and the threshold value increased by every five steps. From the result (Figure 12) we can see that the accuracy increases as the threshold grows until it reaches a threshold of t = 90. After that, the overall accuracy and F_1 decreases with the growing threshold. In this way, the threshold in this work was set to 90 while optimizing using the guided filter.

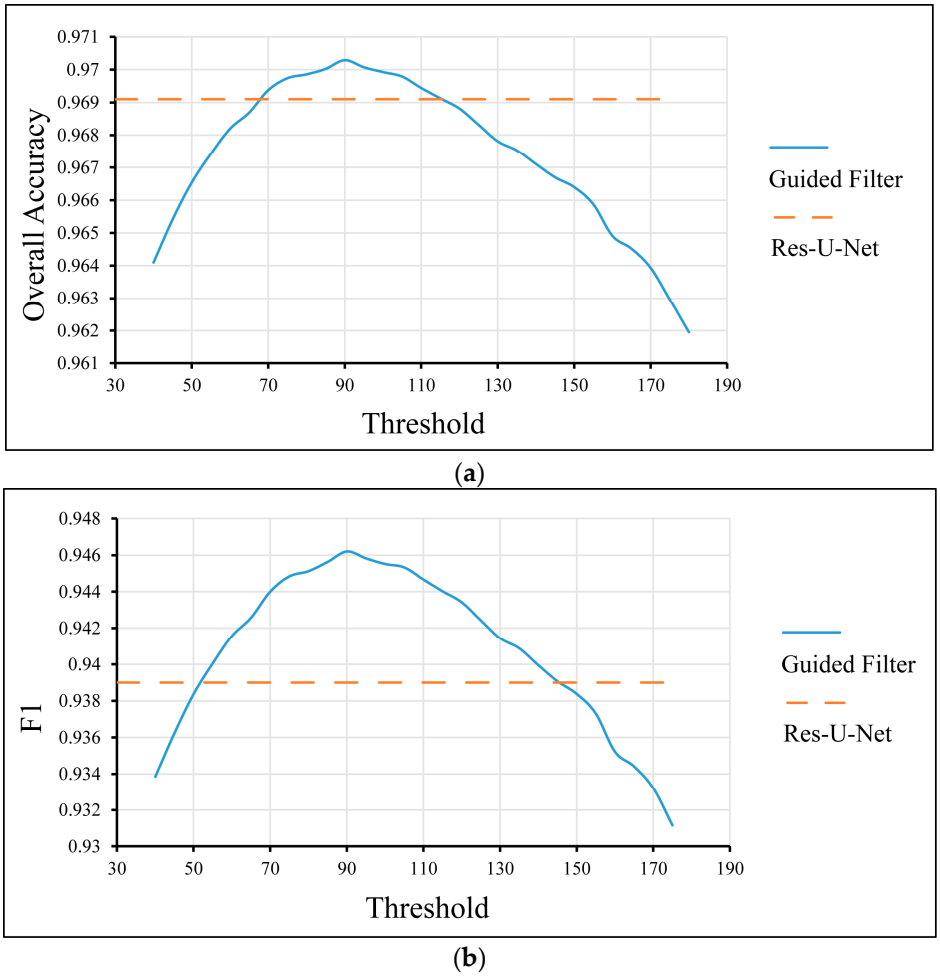

Figure 12. The overall accuracy (**a**) and F_1 (**b**) changing as the threshold increases while optimization by the guided filter.

The size of the window in the guided filter also affects the accuracy during optimization. If the size of window is too small, there will be less information in view to be used to guide the optimization and the filtered result will not be able to obtain enough surrounding information during optimization. On the contrary, if the size of window is too large, the information in the window will be mixed, which will mislead the filter optimization. To get the optimal window size in the guided filter, we compared the overall accuracy and F_1 of the results using different window sizes from two to 15. From the results (Figure 13) we can see that the overall accuracy and F_1 increased as the window size increased until it reached *size* = 5. After that, the overall accuracy and F_1 decreased with the growing window size. Therefore, the size of the window in the guided filter was set as five while optimizing the Res-U-Net results in the experiments.

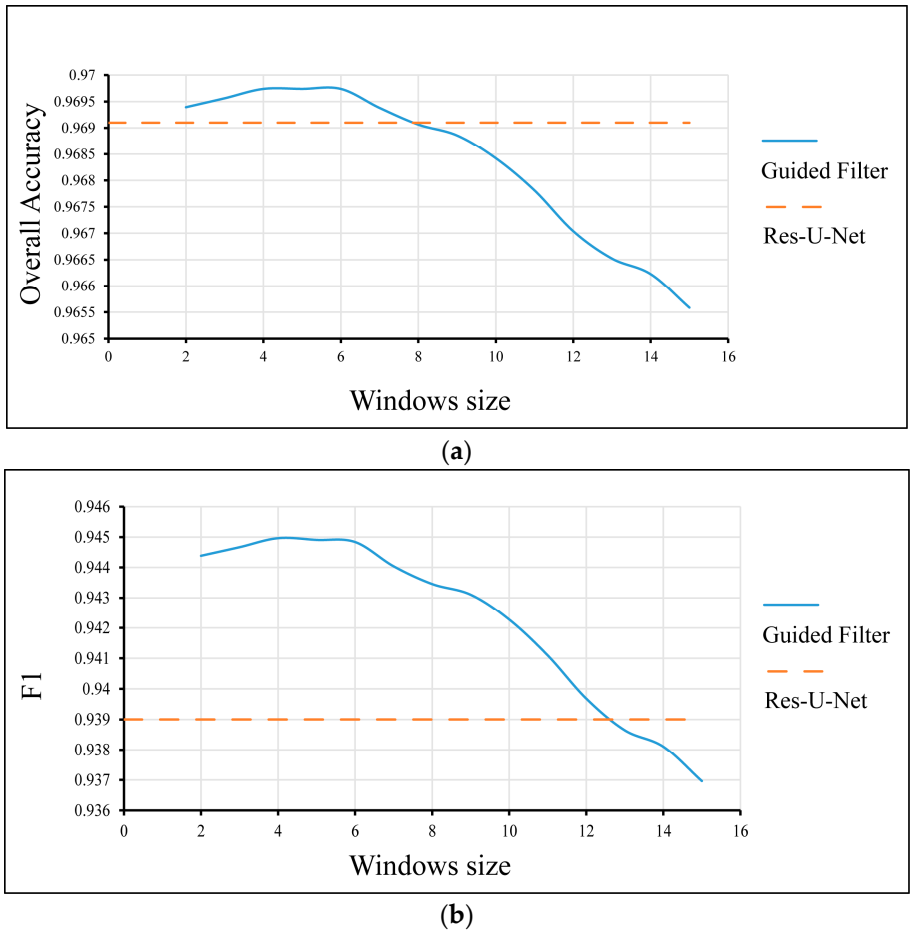

Figure 13. The overall accuracy (**a**) and F_1 (**b**) changing as the size of window increases while optimization by the guided filter.

5. Conclusions

In this paper, a novel framework to perform building extraction in urban districts with very high resolution (VHR) remote sensing imagery is presented. The major contribution of this work is to explore an alternative technique for labeling objects in urban districts, which combined deep learning and guided filtering. This project aimed to design a network which improved the accuracy of building extraction and introduced a guided filter into the post-processing of the results. In our work, during the preprocessing of the date, we used edge enhancing and it is helpful in improving the performance of the segmentation process. As the deep neural network, Res-U-Net did well in labeling different scales buildings; guided filtering was introduced after the Res-U-Net neural network stage, which optimized the classification results and removed the salt-and-pepper class noise. At the same time, it preserved the boundaries of the objects within the imagery effectively. Experiments were carried out on two VHR remote sensing imagery datasets. Every desirable object was extracted successfully using the method mentioned in this work and the results showed the effectiveness and feasibility of the proposed framework in improving the performance of the urban district remote sensing imagery classification. The method was compared with some classical VHR remote sensing classification such as the fully convolutional network (FCN) as well as the method that combined the convolutional neural network (CNN) and random forest (RF). Experimental results demonstrated that our methods were better than the other methods. The proposed method in this work can obtain improvements in terms of overall accuracy, precision and F_1 over the classical classification systems.

With the development of remote sensing technology, more and more VHR images can be accessed conveniently, and the classification of the urban district plays an important role in practical applications such as urban infrastructure, management, and so on. This work has provided an effective method to improve VHR image classification performance. However, the shape of some buildings that are covered by trees cannot be detected precisely, and some blurry and irregular boundaries are hardly classified. In the future, a more optimized deep neural network is required to improve efficiency and accuracy. At the same time, further improvement may be achieved by combining the deep neural network and the guided filter in an end-to-end model, which would combine the advantage of a guided filter that preserves boundaries and decreases the salt-and-pepper class noise whilst also being convenient to train like the FCN. Instead of treating non-building as a background class, we will take the scene semantic into account and extract the roads and trees as well as the cars and so on in future studies.

Acknowledgments: This study was financially supported by the National Natural Science Foundation of China (41671400, 41701446, 41401443), National key R & D program of China (No. 2017YFC0602204) and Hubei Natural Science Foundation of China (2015CFA012).

Author Contributions: Yongyang Xu, Zhong Xie proposed the network architecture design and the framework of extracting buildings. Yongyang Xu performed the experiments and analyzed the data. Yongyang Xu, Liang Wu wrote the paper. Zhanlong Chen revised the paper and provided valuable advices for the experiments.

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

References

1. Moser, G.; Serpico, S.B.; Benediktsson, J.A. Land-Cover Mapping by Markov Modeling of Spatial-Contextual Information in Very-High-Resolution Remote Sensing Images. *Proc. IEEE.* **2013**, *101*, 631–651. [CrossRef]
2. Longbotham, N.; Chaapel, C.; Bleiler, L.; Padwick, C.; Emery, W.J.; Pacifici, F. Very High Resolution Multiangle Urban Classification Analysis. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1155–1170. [CrossRef]
3. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In Proceedings of the Computer Vision—ECCV 2010—European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 210–223.
4. Dalponte, M.; Bruzzone, L.; Gianelle, D. Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data. *Remote Sens. Environ.* **2012**, *123*, 258–270. [CrossRef]
5. Lv, Z.Y.; He, H.; Benediktsson, J.A.; Huang, H. A Generalized Image Scene Decomposition-Based System for Supervised Classification of Very High Resolution Remote Sensing Imagery. *Remote Sens.* **2016**, *8*, 814. [CrossRef]
6. Tian, S.; Zhang, X.; Tian, J.; Sun, Q. Random Forest Classification of Wetland Landcovers from Multi-Sensor Data in the Arid Region of Xinjiang, China. *Remote Sens.* **2016**, *8*, 954. [CrossRef]
7. Kumar, A.L.; Sinha, P.; Taylor, S. Improving image classification in a complex wetland ecosystem through image fusion techniques. *J. Appl. Remote Sens.* **2014**, *8*, 083616. [CrossRef]
8. Wang, J.; Song, J.; Chen, M.; Yang, Z. Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine. *Int. J. Remote Sens.* **2015**, *36*, 3144–3169. [CrossRef]
9. Konstantinidis, D.; Stathaki, T.; Argyriou, V.; Grammalidis, N. Building Detection Using Enhanced HOG–LBP Features and Region Refinement Processes. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 888–905. [CrossRef]
10. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688.
11. Liu, F.; Jiao, L.; Hou, B.; Yang, S. POL-SAR Image Classification Based on Wishart DBN and Local Spatial Information. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3292–3308. [CrossRef]
12. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection. *arXiv* **2016**, arXiv:1612.01337.

13. Zhang, Y. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS J. Photogramm. Remote Sens.* **1999**, *54*, 50–60. [CrossRef]
14. Huang, X.; Zhang, L. A comparative study of spatial approaches for urban mapping using hyperspectral ROSIS images over Pavia City, northern Italy. *Int. J. Remote Sens.* **2009**, *30*, 3205–3221. [CrossRef]
15. Lv, Z.Y.; Zhang, P.; Benediktsson, J.A.; Shi, W.Z. Morphological Profiles Based on Differently Shaped Structuring Elements for Classification of Images With Very High Spatial Resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *7*, 4644–4652. [CrossRef]
16. Wang, Y.; Song, H.; Zhang, Y. Spectral-spatial classification of hyperspectral images using joint bilateral filter and graph cut based model. *Remote Sens.* **2016**, *8*, 748. [CrossRef]
17. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 512–519.
18. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *arXiv* **2014**, arXiv:1311.2524.
19. 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.
20. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv* **2016**, arXiv:1606.00915.
21. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H. Conditional random fields as recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1529–1537.
22. Audebert, N.; Boulch, A.; Lagrange, A.; Le Saux, B.; Lefevre, S. *Deep Learning for Remote Sensing*; Technical Report; ONERA The French Aerospace Lab, DTIM & Univ. Bretagne-Sud & ENSTA ParisTech: Palaiseau, France, 2016.
23. Penatti, O.A.B.; Nogueira, K.; Santos, J.A.D. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 44–51.
24. Audebert, N.; Boulch, A.; Randrianarivo, H.; Le Saux, B.; Ferecatu, M.; Lefèvre, S.; Marlet, R. Deep learning for urban remote sensing. In Proceedings of the Urban Remote Sensing Event (JURSE), Dubai, UAE, 6–8 March 2017; pp. 1–4.
25. Vakalopoulou, M.; Karantzalos, K.; Komodakis, N.; Paragios, N. Building detection in very high resolution multispectral data with deep learning features. In Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1873–1876.
26. Farabet, C.; Couprie, C.; Najman, L.; Lecun, Y. Learning Hierarchical Features for Scene Labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef] [PubMed]
27. Wilkinson, G.G. Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 433–440. [CrossRef]
28. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, V.-D. Effective semantic pixel labelling with convolutional networks and conditional random fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 36–43.
29. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, A.V.D. Semantic Labeling of Aerial and Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2868–2881. [CrossRef]
30. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
31. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]
32. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [CrossRef]
33. Audebert, N.; Saux, B.L.; Lefèvre, S. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. In Proceedings of the Computer Vision—ACCV 2016, Taipei, Taiwan, 20–24 November 2016; pp. 180–196.

34. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
35. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic Segmentation of Aerial Images with an Ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
36. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
37. Sherrah, J. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. *arXiv* **2016**, arXiv:1606.02585.
38. Dalla Mura, M.; Benediktsson, J.A.; Waske, B.; Bruzzone, L. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3747–3762. [CrossRef]
39. Tokarczyk, P.; Wegner, J.D.; Walk, S.; Schindler, K. Features, Color Spaces, and Boosting: New Insights on Semantic Classification of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 280–295. [CrossRef]
40. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
41. Volpi, M.; Tuia, D. Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 881–893. [CrossRef]
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
43. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
45. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
46. Lagrange, A.; Saux, B.L.; Beaupère, A.; Boulch, A.; Chan-Hon-Tong, A.; Herbin, S.; Randrianarivo, H.; Ferecatu, M. Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks. In Proceedings of the Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 4173–4176.
47. 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; pp. 234–241.
48. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, arXiv:1612.03144.
49. He, K.; Sun, J.; Tang, X. Guided Image Filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1397–1409. [CrossRef] [PubMed]
50. Pritika; Budhiraja, S. Multimodal medical image fusion using modified fusion rules and guided filter. In Proceedings of the International Conference on Computing, Communication & Automation, Noida, India, 15–16 May 2015; pp. 1067–1072.
51. Gerke, M. *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*; Technical Report; University of Twente: Enschede, The Netherlands, 2015.
52. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated Convolutional Neural Network for Semantic Segmentation in High-Resolution Images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]