remote sensing

MDPI

Letter

SPMF-Net: Weakly Supervised Building Segmentation by Combining Superpixel Pooling and Multi-Scale Feature Fusion

Jie Chen, Fen He, Yi Zhang, Geng Sun and Min Deng *

School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; cj2011@csu.edu.cn (J.C.); fennyh@csu.edu.cn (F.H.); zhangyi_csu@csu.edu.cn (Y.Z.); 0106170206@csu.edu.cn (G.S.)

* Correspondence: dengmin@csu.edu.cn; Tel.: +86-1350-746-7258

Received: 4 February 2020; Accepted: 23 March 2020; Published: 24 March 2020

Abstract: The lack of pixel-level labeling limits the practicality of deep learning-based building semantic segmentation. Weakly supervised semantic segmentation based on image-level labeling results in incomplete object regions and missing boundary information. This paper proposes a weakly supervised semantic segmentation method for building detection. The proposed method takes the image-level label as supervision information in a classification network that combines superpixel pooling and multi-scale feature fusion structures. The main advantage of the proposed strategy is its ability to improve the intactness and boundary accuracy of a detected building. Our method achieves impressive results on two 2D semantic labeling datasets, which outperform some competing weakly supervised methods and are close to the result of the fully supervised method.

Keywords: building detection; weakly supervised learning; superpixel; semantic segmentation; deep learning

1. Introduction

Building detection plays an important role in urban development planning, urban infrastructure planning, urban land use and management, land use change monitoring, digital cities, and real-time updates of urban traffic maps. Ensuring high efficiency and the precision of the automatic detection of buildings, based on massive high-resolution remote sensing data, is a challenging task in remote sensing.

Traditional algorithms for building detection based on remote sensing images are mainly driven by visual features via bottom-up approaches. These methods, such as geometric boundary-based [1], image segmentation-based [2], and building-specific auxiliary information (shadows, elevations, etc.) based [3], consider a building as a combination of low-level features that merge a building as a whole under some rules. These methods focus on the characteristic feature of buildings. However, the feature design requires experimental testing in decision making, thereby increasing the algorithm complexity.

In recent years, deep convolutional neural networks (DCNNs) [4] have been widely used in image classification [5], object detection [6] and semantic segmentation [7,8], because of their end-to-end learning mechanism and feature representation. In the remote sensing field, some DCNN-based segmentation approaches, such as U-Net [8], Deeplabv3+ [7], etc., are used to achieve excellent results in building detection [9–14]. All of these approaches are fully supervised, which means a pixel-level label benchmark is essential for training the semantic segmentation networks. However, pixel-level label datasets of high-resolution remote sensing images are scarce because of the time consuming and expensive work. The accurate extraction of buildings without a large-scale pixel-level dataset is a considerable problem to be solved in remote sensing.

In order to alleviate the problem of lacking pixel-level labels, in the field of image semantic segmentation, weaker labels instead of pixel-level labels are used, including bounding boxes [15], scribbles [16], and image-level labels [17]. It is because, although the weaker labels cannot indicate objects at the pixel-level, they potentially contain the high-level semantic and rough location information of objects. Among these weaker labels, image-level ones are the cheapest and are easier to obtain than others, and they obtain more and more attention in weakly supervised semantic segmentation [17–21]. In accordance with the different means of introducing weak supervision information, these methods can be divided into bottom-up and top-down approaches.

In the bottom-up approach, the saliency map of an image is generated by extracting its low-level features. Pseudo-labels are generated through the threshold-based segmentation of the saliency map and are used as the supervision information for the segmentation network. Wei et al. used a saliency map for supervision to train the initial segmentation model of simple images and train complex images to enhance segmentation performance [17]. This method can achieve good results on natural images containing a prominent object and simple background. However, this method misidentifies the buildings with other geo-objects when it is applied on remote sensing images with complex backgrounds and intra-class differences, because the saliency map only utilizes the low-level features of images.

In the top-down approach, classification networks, like VGG16 [5], and visualization methods, such as CAM(class activation map) [22] and Grad-CAM [23], are generally used to find the discriminative regions of objects. CAM obtains the weight of pixels in the feature maps through a global average pooling layer and generates a heatmap with different heat values in accordance with the different weights. The heatmap shows the approximate location and region of the object. Pseudo-labels can be generated through threshold-based segmentation of the objects. Wei et al. proposed an adversarial erasing method that uses a classification network to mine the object region [19]. It relies on a classification network to sequentially activate the most discriminative areas of an object. Durand et al. proposed a method for learning local visual features related to some classes, named WILDCAT(weakly supervised learning of deep convolutional neural networks), which can be used as weakly supervised semantic segmentation [21]. The object region located by a classification network can be considered as a segmentation seed that usually highlights a local discriminative area of an object and then propagates to the entire object region. The top-down approach can extract the high-level semantic information of images and can identify and locate the objects. However, the generated object region frequently contains the most discriminative area of the object, but cannot cover the entire object. Thus, this approach has poor performance in depicting the shapes and edges of objects.

Several studies have been conducted on weakly supervised semantic segmentation in remote sensing. Fu et al. proposed a weakly supervised semantic segmentation network with feature fusion and tested it on water and cloud datasets [24]. This network combines top-down and bottom-up approaches. However, this method is not flexible, because the a priori information extracted from the bottom-up approach is introduced in the postprocessing stage. For the segmentation of high-resolution synthetic aperture radar (SAR) images, Ma et al. subdivided images into superpixel maps and input them into a condition generative adversarial network for training [25]. However, this method relies excessively on the low-level feature information extracted from bottom-up approaches, thereby causing misclassification.

Kwak et al. [26] added a superpixel pooling layer to the classification network and used iterative training to improve segmentation performance. This method achieves good results on natural image datasets. Without decoding and multi-scale feature fusion, the network only uses an upsampling module to restore the feature map generated through convolution downsampling. The heatmap generated using this method cannot reflect the object details.

The building results with an accurate boundary and intact region are essential in the building detection using high-resolution remote sensing images. Therefore, this letter proposes a weakly supervised learning segmentation method to extract buildings, by fully exploiting the implicit semantic

information from image-level labels. In particular, the proposed method combines the advantages of top-down and bottom-up approaches to generate pixel-level weak supervision building labels. The proposed method makes the generated building labels have intact bodies and accurate boundaries by embedding the structures of superpixel pooling and multi-scale feature fusion. The proposed method is named as SPMF-Net (weakly supervised building segmentation method by combining the *S*uperpixel *P*ooling and *M*ulti-scale *F*eature fusion structures).

The remainder of this article is organized as follows. Section 2 introduces the methodology. Section 3 presents the experimental result, with quantitative and qualitative analyses. Section 4 provides the conclusions and future work directions.

2. Methodology

The proposed weakly supervised segmentation method consists of two parts: (1) giving the image-level class label of the building image and the corresponding superpixel map and generating pixel-level pseudo-label using a DCNN-based classification; (2) training a DCNN-based segmentation model with generated pixel-level pseudo-label to extract buildings. The entire framework includes two DCNNs, namely, a classification network that generates an object activation map and a segmentation network for building extraction. The former network is used to generate the CAM of the building region with image-level labels. On the basis of the threshold-based segmentation of the building region, pixel-level pseudo-labels are obtained as training labels for the latter network to extract buildings. The former network used to generate pixel-level labels is described in Section 2.1. Subsequently, the segmentation network is briefly introduced in Section 2.2.

2.1. Weak Supervision Label Generation

VGG16 [5] is a simple and excellent image classification network. We adopted it as the backbone of our classification network, where a superpixel pooling structure and a multi-scale upsampling feature fusion structure are combined, as shown in Figure 1.

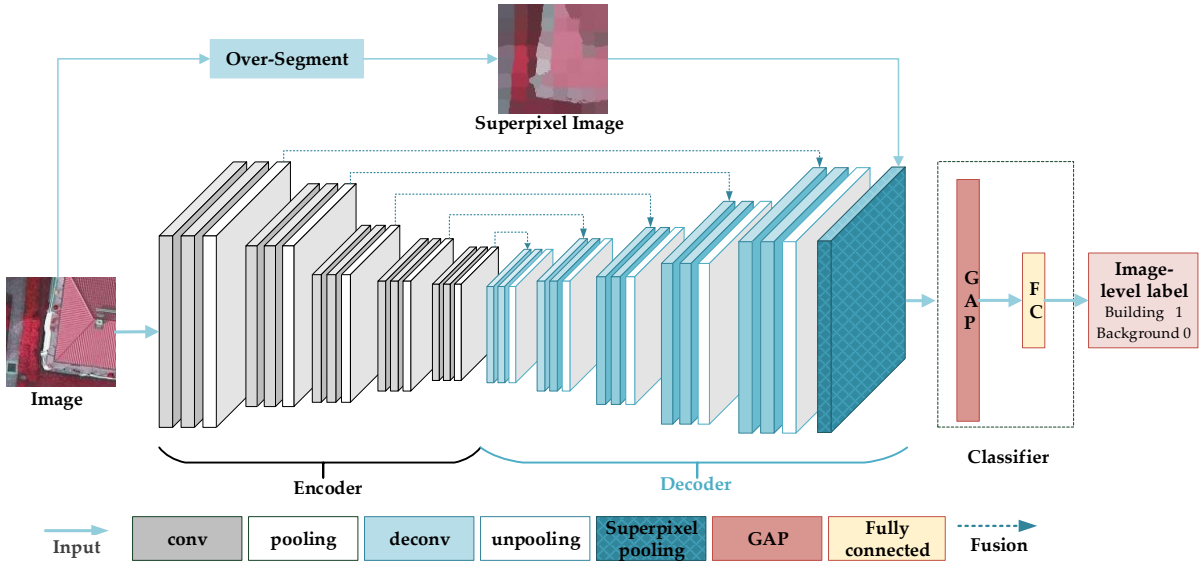

Figure 1. Weak supervision label generation network consists of a downsampling module, a multi-scale upsampling module, a superpixel pooling module, and a classifier.

2.1.1. Superpixel Pooling

The top-down weak supervision method transforms the feature map extracted by the last convolution layer into a vector, through global average pooling or global maximum pooling. Then, the vector is inputted into the fully connected layer to learn the weight in accordance with the image-level

label. The weight is visualized using CAM to obtain the object activation map. However, the object activation map obtained using this method cannot meet the requirements of semantic segmentation. On the one hand, the object region is constantly not intact, because only a partial region of the object with significant discriminant features will gain high weight. On the other hand, the resolution of the feature map is excessively low to recover the boundary and shape information of geo-objects after continuous downsampling, which is unsuitable for the pixel-level extraction of buildings. Some studies have shown that the use of prior shape information can significantly improve segmentation performance [27,28].

The superpixel can provide prior shape information of building objects, because it merges similar pixels into superpixels through some low-level feature-based rules. Therefore, a superpixel pooling layer is designed in this paper, as shown in Figure 2. The principle of the superpixel pooling layer is similar to the average pooling layer. The features of the local regions are aggregated through average pooling. Feature map $Z \in \mathbb{R}^{W \times H \times C}$ is inputted into the superpixel pooling layer, where W denotes the width, H denotes the height, and C denotes the number of channels. Superpixel map $P \in A^{W \times H}$ is a single channel map, with length H and width W. A= [1, M] indicates the label of each superpixel. P_i = M represents that pixel i belongs to the M_{th} superpixel. Output O of the superpixel pooling layer is a matrix with size $M \times C$, represented as $O \in R^{M \times C}$, where M represents the number of superpixels, and C is the number of channels of the feature map. The output of the superpixel pooling layer is expressed as

$$O_{\text{m,C}} = \text{average}\left\{Z_{\text{C},i} | i : P_i = M\right\} \tag{1}$$

where average {·} denotes the average pooling function.

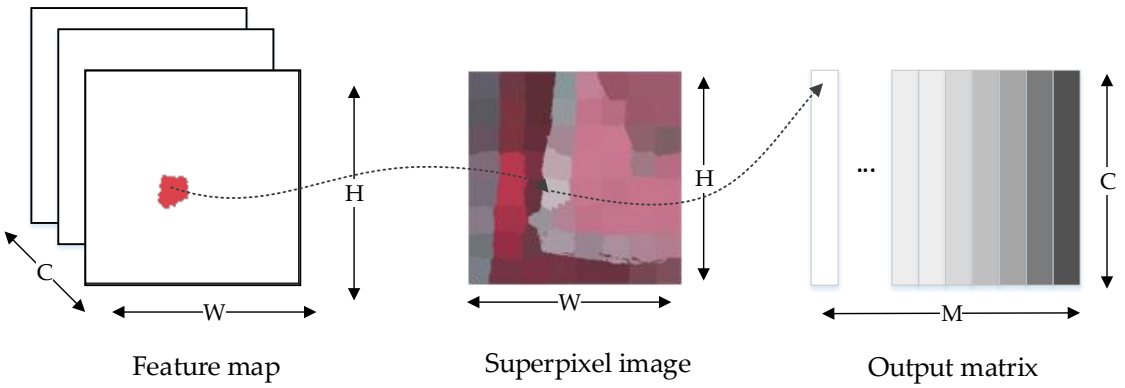

Figure 2. Superpixel pooling layer.

The backward gradient calculation formula for superpixel pooling is expressed as

$$\frac{\delta O_{C,M}}{\delta Z_{C',i}} = \begin{cases} \frac{1}{N(P_i)}, \; if \; P_i = M \; and \; C' = C \\ 0 \; , otherwise \end{cases} \tag{2}$$

where $N(P_i)$ is the number of pixels in the superpixel with label M.

We can aggregate the feature vectors aligned in the superpixel space using the superpixel pooling layer. Different from the traditional pooling structure, superpixel pooling does not have a rectangular layout determined by predefined raster pixels (such as 2 × 2 maximum pooling), but an irregular region determined by the superpixel shape. In other words, superpixel pooling is a pooling operation under the constraint of a single superpixel boundary region. The superpixel map includes shape prior information, which is good for detailing object boundaries. In this paper, simple linear iterative clustering (SLIC) [29] is used to generate superpixel maps of building images. It first converts the image from RGB color space to CIELAB color space at first and then generates homogeneous regions of different sizes through SLIC.

2.1.2. Multi-Scale Feature Fusion

CAM uses the weight of the last convolutional layer to generate the object activation map. The direct method is to add a superpixel pooling structure into a classification network to restore the last feature map to a superpixel image size at first and connect it with the superpixel pooling layer. Accordingly, we need to restore the feature map size to the input image size and retain the semantic detail of the object. In [26], only one upsampling module (including two deconvolution layers and one unpooling layer) is used to restore the feature map size to the input image size, which is connected to the superpixel pooling layer. The generated activation region cannot recover the main part of the building when only one upsampling module is used, and the features from convolution downsampling are not fused, because the resolution of the feature map generated after continuous downsampling is excessively low. Although the object activation map contains a highly abstract semantic feature, a superpixel inevitably corresponds to a large number of pixels in the feature map when the superpixel pooling layer is used. This condition will result in superpixel feature vector smoothing and the loss of many details with high probability.

The function of the upsampling module is to restore the size of the convolutional feature map and match its size with the input superpixel map of the superpixel pooling layer for calculation. Different from [26], we use five consecutive upsampling modules to form an encoder–decoder structure corresponding to downsampling and restore the size of the feature map step by step. We merge the output features at different scales and inputted them into the superpixel pooling layer. Assume that $V_1 \in \mathbb{R}^a$, $V_2 \in \mathbb{R}^b$, where V_1 represents the feature map of the output of the upsampling module, and V_2 represents the feature map of the output of the downsampling module. Fused feature V is expressed as

$$\mathrm{V} = [V_1, V_2] \in \mathbb{R}^{a+b} \tag{3}$$

A corresponding multi-scale upsampling feature fusion module is adopted after the downsampling module in VGG16, as shown in Figure 1. The module consists of five upsampling modules, corresponding to five downsampling modules. Each upsampling module consists of two deconvolution layers and an unpooling layer, each of which is accompanied by a batch normalization layer and a rectified linear unit layer. Each upsampling module is added with features from the corresponding downsampling module.

2.1.3. CAM Calculation

The superpixel pooling layer is followed by a classifier. The classifier consists of a global average pooling layer and a fully connected layer. The classifier transforms the output M × C matrix of the superpixel pooling into a 1 × C vector, through global average pooling, to average each superpixel. Each superpixel is assigned a feature vector through the superpixel pooling layer. The vector is connected to the fully connected layer in the classifier for classification. As only image-level building and background labels are available in our method, SPMF-Net is optimized with a binary classification loss. The classification loss is calculated and backpropagated according to the formula (2).

Training the classification network with image-level labels aims to generate pixel-level labels using CAM, rather than classifying the images. CAM assigns an activation score to each individual pixel. Unlike the original CAM, the proposed strategy assigns an activation score to a single superpixel. In this way, the object activation heatmap can be generated with the shape and boundary of the preserved building. Many accurate building regions can be achieved, because the structure combines the detailed information under the multiresolution feature map and retains the boundary information of the building with high confidence.

The generated heatmap suggests that the higher the heat value is, the higher the likelihood of the building area will be. We normalize the pixel value of heatmaps. The pixels with a heat value greater than 0.5 are regarded as building, and the pixels below this threshold are regarded as other classes.

Thus, the pseudo-labels of the building can be obtained as training data for network segmentation to extract buildings.

2.2. Building Extraction

Based on the procedures described in Section 2.1, the weak supervision labels can be generated. These weakly supervision labels have reached the pixel-wise supervision, which can be employed for training the semantic segmentation network in a supervised way.

The Deeplabv3+ [7], as one of the well-known fully supervised semantic segmentation networks, uses an encoder–decoder multi-scale network structure and has achieved advanced performance on many datasets. In this letter, we adopt Deeplabv3+ as the building segmentation network with Xception as its backbone, and the cross-entropy loss function as its objective function. The loss function L is used in network training to quantitatively evaluate the difference between real value $y^{(i)}$ and predicted value $y^{(i')}$. $i = 1, \ldots, n$, where n represents the number of training data. The loss function of network training is expressed as

$$L = -\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} log\, y^{(i')} + \left(1 - y^{(i)}\right) log(1 - y^{(i')})). \tag{4}$$

The loss function is minimized, which can be expressed as

$$\underset{\theta}{Min} \sum_{w \in I_w} L_W(f(_w; \theta)), \tag{5}$$

where I_w represents the images from the weakly supervised training dataset.

3. Results and Analysis

In this section, we will introduce the datasets, comparisons and evaluation metrics, and experimental details.

3.1. Dataset

The public 2D semantic labeling contest Potsdam dataset and Vaihingen dataset are elaborately labeled, which are provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) Commission II/4. These datasets have been used to verify the effectiveness of many methods as test datasets, and they are widely recognized in the field of remote sensing [30,31]. The two datasets consist of four-band image data (near-infrared, red, green, blue) and corresponding digital surface model (DSM) data. Potsdam contains 38 patches with the same size of 6000 × 6000 pixels. Vaihingen contains 33 patches of different sizes. The ground sampling distance is 5 cm. Each patch is labeled into six categories, namely, impervious surfaces, building, low vegetation, tree, car, and clutter/background.

3.2. Data Processing

We select 20 patches as training data, 4 patches as verification data, and 14 patches as test data from the Potsdam dataset. From the Vaihingen dataset, we select 17 patches as training data, 4 patches as verification data, and 12 patches as test data. With a sliding step size of 128, we crop the patches in the training data into image blocks with a size of 256 × 256. All image blocks are rotated by 90°, 180°, and 270°. All the rotated image blocks are horizontally mirrored. Thus, the processed Potsdam dataset for fully supervised building segmentation contains 62,437 blocks for training and 12,488 blocks for verification, while the Vaihingen dataset contains 24,793 blocks for training and 8811 blocks for verification.

In particular, the blocks with a building pixel ratio of more than 50% from the training data are collected as building class, and the blocks without building pixels are collected as another class, to

establish the dataset for weakly supervised classification. Considering that accurate building region detection relies on good classification performance, the rest of the blocks that contain building pixels, but have a ratio which is less than 50%, are excluded from the training data to ensure the performance of the classification network. Consequently, the processed Potsdam dataset for classification contains 17,863 building blocks and 31,271 other class blocks, while the Vaihingen dataset contains 5056 building blocks and 3944 other class blocks. All the weak supervision methods for comparison use this same dataset for training. Each block was segmented into 64 superpixels using SLIC.

3.3. Parameter Settings

All experiments were conducted on the Pytorch framework version 0.4.1. The training of the proposed weak supervision information generation network was iterated 10 times. The batch size was set to 10, and the learning rate was set to 0.001. The network was optimized through stochastic gradient descent, with a momentum of 0.9 and a weight decay rate of 0.0005. The training of adopted segmentation network was iterated 10 times. The batch size was set to 8, and Adam was used as the optimizer. The initial learning rate was 0.001 and the learning rate was reduced to 1/10 at every epoch. The other methods used for comparison keep the experimental setup in their original paper. We used the Deeplabv3+ model as the segmentation network. All experiments were performed on a computer with NVIDIA 1080Ti GPU, i7-9700k CPU, and 16 GB memory.

3.4. Evaluation Metrics and Comparisons

In order to verify the effectiveness of the proposed method quantificationally, OA(Overall Accuracy), DA(Detection Accuracy), FAR(False Alarm Rate) and mIOU (mean Intersection Over-Unions) were used as evaluation metrics [18].

In our experiments, three weakly supervised semantic segmentation models are used for comparison with the proposed method: (1) CAM-based weakly supervised method [22], which is a simple model of the top-down approach, has been used as the baseline in many tasks; (2) WILDCAT method [21], which is an efficient model in the top-down approach, improves the attention to the feature of different objects based on the CAM method; (3) superpixel pooling network (SPN) [26], which adds a superpixel pooling layer based on the CAM method.

Besides, in order to verify the effectiveness of the multi-scale feature fusion integrated in SPMF-Net, an ablation experiment is conducted on SPMF-Net_v1, which is only embedded in the encoder–decoder upsampling module on the basis of SPN. Moreover, the fully supervised approach Deeplabv3+ [7] is adopted to verify the potential of the proposed method.

3.5. Quantitative Analysis

The quantitative results of SPMF-Net are compared with other state-of-the-art approaches, as shown in Tables 1 and 2 respectively. Compared with CAM and WILDCAT, the accuracy of SPMF-Net with the superpixel pooling structure and multi-scale upsampling feature fusion module is significantly improved. On the Potsdam dataset, compared with CAM, OA of SPMF-Net increases by 4.01%, DA increases by 17.9%, FAR score increases by 9.84%, and the mIOU score increases by 8.58%. Compared with WILDCAT, OA increases by 2.28%, DA increases by 25.26%, FAR score increases by 14.15%, and the mIOU score increases by 9.64%. Compared with SPN, OA increases by 1.34%, DA increases by 3.09%, FAR score increases by 1.67%, and the mIOU score increases by 1.92%. As shown in Table 2, the performance of SPMF-Net is also impressive on the Vaihingen dataset. Compared with CAM, the OA of SPMF-Net increases by 2.24%, DA increases by 19.90%, FAR score increases by 9.83%, and the mIOU score increases by 7.84%. Compared with WILDCAT, OA increases by 3.95%, DA increases by 27.13%, FAR score increases by 17.47%, and the mIOU score increases by 13.10%. Compared with SPN, OA increases by 3.35%, DA increases by 4.05%, FAR score increases by 3.32%, and the mIOU score increases by 4.33%.

Table 1. Comparison of quantitative results on the Potsdam dataset.

Method	OA	DA	FAR	mIOU
CAM [22]	87.36%	63.61%	71.00%	69.86%
WILDCAT [21]	89.09%	55.95%	66.70%	68.81%
SPN [26]	90.03%	78.42%	79.17%	76.52%
SPMF-Net_v1	90.71%	80.37%	80.17%	77.60%
SPMF-Net	91.37%	81.51%	80.84%	78.45%
Deeplabv3+ [7]	96.08%	87.33%	91.02%	89.39%

Table 2. Comparison of quantitative results on the Vaihingen dataset.

Method	OA	DA	FAR	mIOU
CAM [22]	90.53%	58.91%	73.48%	73.58%
WILDCAT [21]	88.83%	51.67%	65.84%	68.31%
SPN [26]	89.42%	74.76%	79.99%	77.08%
SPMF-Net_v1	91.75%	76.76%	81.82%	79.77%
SPMF-Net	92.77%	78.81%	83.31%	81.42%
Deeplabv3+ [7]	95.54%	90.42%	91.70%	89.19%

The results of SPMF-Net are also promising compared with the results of the fully supervised Deeplabv3+ model. On the Potsdam dataset, the OA of SPMF-Net reaches 95%, and DA reaches 93%, FAR score reaches 89%, and the mIOU score reaches 88% compared with Deeplabv3+'s result. On the Vaihingen dataset, the OA of SPMF-Net reaches 97%, and DA reaches 87%, and both FAR score and mIOU score reach 91% compared with Deeplabv3+'s result. These results demonstrate that the addition of the superpixel pooling structure and multi-scale fusion module has a great advantage for improving the accuracy and integrity of building location. The low-level features are introduced and the boundary information is preserved because of superpixel pooling. The features from different convolutional layers are fused because of the multi-scale upsampling module.

We present the results of the ablation analysis for SPMF-Net to demonstrate the effectiveness of the proposed module, as listed in Tables 1 and 2. The comparison of SPN with CAM and WILDCAT shows that the addition of superpixel pooling modules can effectively improve the performance of building extraction. After adding the superpixel pooling layer, the FAR score improves by approximately 8% compared with CAM, and the mIOU score improves by approximately 7% on the Potsdam dataset, as shown in Table 1. On the Vaihingen dataset, the FAR score improves by approximately 6% compared with CAM, and the mIOU score improves by approximately 4%. The comparison of SPMF-Net_v1 with SPN illustrates the effectiveness of the continuous upsampling module with an encoder–decoder structure added to the network structure. On the Potsdam dataset, compared with SPN, the DA of SPMF-Net_v1 improves by 2%, and FAR and mIOU score improve by 1%. As shown in Table 2, the DA and FAR of SPMF-Net_v1 improves by 2%, and the mIOU score improves by approximately 3% compared with SPN on the Vaihingen dataset. This finding shows that the detailed information of images is retained, using continuous upsampling modules to restore the size of the feature map. Compared with SPMF-Net_v1, the FAR and mIOU scores of SPMF-Net approximately improve by 1% on the Potsdam dataset, while the FAR and mIOU scores of SPMF-Net approximately improve by 1.5% on the Vaihingen dataset. This finding indicates that the fusion of the corresponding downsampled features in the upsampling module is helpful to completely and accurately extract buildings.

3.6. Qualitative Analysis

CAM, WILDCAT, and SPMF-Net for the pseudo-labels of buildings are shown in Figure 3. CAM has a larger located region of buildings, whereas WILDCAT prefers to locate many discriminating regions of buildings. Compared with the ground truth, some nonbuilding regions in the heatmap generated by SPMF-Net are included in the building regions (as shown in the red rectangles of the

second line in Figure 3d,h). This condition is because the feature of these regions is similar to the feature of the building. However, the pseudo-labels generated by SPMF-Net can completely and accurately mark the building region. Compared with CAM and WILDCAT, the results of SPMF-Net are more accurate in building regions and contain less noise regions. This finding is because SPMF-Net combines the high-level semantic features using a top-down approach and the low-level structural features extracted using a bottom-up approach, which are suitable for generating a complete and accurate heatmap. The value of pixels in the same superpixel in the heatmap generated by SPMF-Net is limited to the same value, because of the superpixel pooling module. On the contrary, every pixel of the heatmap in CAM and WILDCAT is different in the local area, because of the different weights of pixels.

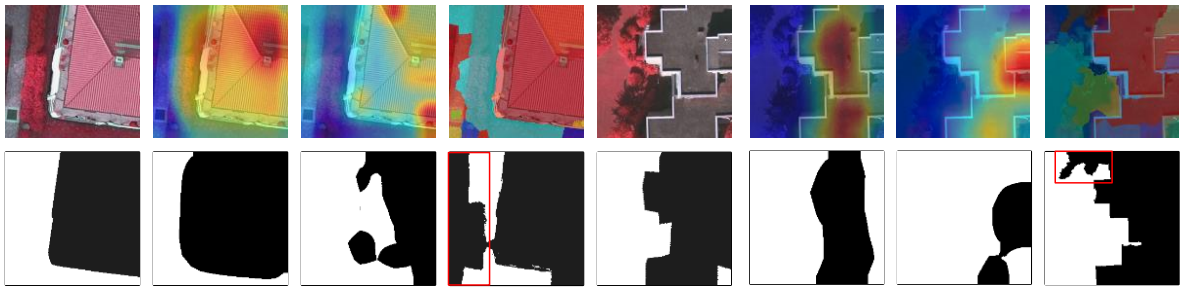

Figure 3. Pseudo-labels of building. (**a**) and (**e**) are the input image and ground truth respectively; (**b**) and (**f**) are the heatmaps of CAM and its generated pseudo-label; (**c**) and (**g**) are the heatmaps of WILDCAT and its generated pseudo-label; and (**d**) and (**h**) are the heatmaps of SPMF-Net and its generated pseudo-label. (**a–d**) are from the Potsdam dataset, (**e–f**) are from the Vaihingen dataset. Buildings are represented by black color, and other classes are represented by white color.

The results of all the methods used in the experiments are shown in Figure 4. The building detection results generated by CAM and WILDCAT cannot completely preserve the boundaries of the building. This condition is because the unit of building region generated by CAM and WILDCAT is at the pixel level rather than a local superpixel. The value of heatmap varies for pixels because the weight of each pixel is different. The shape of the object region is lumpy in the heatmap, indicating that the two methods are insensitive to the boundary information of the building. Compared with the SPN, the noise in the result of SPMF-Net is significantly reduced, the boundary is clearer, and the possibility of misclassification decreases.

Figure 4. Qualitative results comparison between our method and other methods on the Potsdam dataset (the first and second columns) and Vaihingen dataset (the third and fourth columns). (**a**) is the original remote sensing image; (**b**) is result of CAM, (**c**) is result of WILDCAT; (**d**) is result of SPN; (**e**) is result of SPMF-Net_v1; (**f**) is result of SPMF-Net; (**g**) is result of Deeplabv3 +; and (**h**) is the ground truth. Buildings are displayed by black color, and other classes are displayed by white color.

4. Conclusions

This paper proposes a weakly supervised network for pixel-level building label generation. It uses image-level label as supervision to locate the building region using a top-down approach. The low-level detail features extracted using a bottom-up approach can be used to exhibit the boundary of the building, by combining the superpixel pooling layer module, in accordance with the characteristics of the remote sensing image. The validity of the proposed method is verified by comparing it with several related methods on the ISPRS Potsdam dataset and Vaihingen dataset. SPMF-Net can achieve the boundary of buildings and improve the accuracy and integrity of building results.

In future studies, we will focus on making the building region generated in weakly supervision to be intact and accurate. Auxiliary data, such as DSM and normalized difference vegetation index, will be used to enhance the performance.

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

Funding: The work was supported by the National Natural Science Foundation of China, grant number No.41671357; the Scientific Research Fund of Hunan Provincial Education Department, grant number No.16K093, and the Fundamental Research Funds for the Central Universities of Central South University, grant number No. 1053320183827.

Acknowledgments: The authors would like to thank all their colleagues in the lab and the anonymous reviewers for their very competent comments and helpful suggestions.

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

References

1. Cui, S.; Yan, Q.; Reinartz, P. Complex building description and extraction based on Hough transformation and cycle detection. *Remote Sens. Lett.* **2012**, *3*, 151–159. [CrossRef]
2. Tian, J.; Chen, D.M. Optimization in multi-scale segmentation of high-resolution satellite images for artificial feature recognition. *Int. J. Remote Sens.* **2007**, *28*, 4625–4644. [CrossRef]
3. Brunn, A.; Weidner, U. Hierarchical Bayesian nets for building extraction using dense digital surface models. *ISPRS J. Photogramm. Remote Sens.* **1998**, *53*, 296–307. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3 December 2012; pp. 1097–1105.
5. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
6. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 July 2017; pp. 7263–7271.
7. Chen, L.-C.; Zhu, Y.; 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, 9 September 2018; pp. 801–818.
8. 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 October 2015; pp. 234–241.
9. Yuan, J. Learning Building Extraction in Aerial Scenes with Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2793–2798. [CrossRef] [PubMed]
10. Yang, H.L.; Jiangye, Y.; Dalton, L.; Melanie, L.; Amy, R.; Budhendra, B. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [CrossRef]

11. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building Extraction From Satellite Images Using Mask R-CNN With Building Boundary Regularization. In Proceedings of the CVPR Workshops, Salt Lake City, UT, USA, 18 June 2018; pp. 247–251.
12. Zuo, T.; Feng, J.; Chen, X. HF-FCN: Hierarchically fused fully convolutional network for robust building extraction. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 24 November 2016; pp. 291–302.
13. Papadopoulos, G.; Vassilas, N.; Kesidis, A. Convolutional Neural Network for Detection of Building Contours Using Multisource Spatial Data. In Proceedings of the International Conference on Engineering Applications of Neural Networks, Crete, Greece, 24 May 2019; pp. 335–346.
14. Lin, J.; Jing, W.; Song, H.; Chen, G. ESFNet: Efficient Network for Building Extraction From High-Resolution Aerial Images. *IEEE Access* **2019**, *7*, 54285–54294. [CrossRef]
15. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13 December 2015; pp. 1635–1643.
16. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June 2016; pp. 3159–3167.
17. Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.M.; Feng, J.; Zhao, Y.; Yan, S. STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 2314–2320. [CrossRef] [PubMed]
18. Zhang, T.; Lin, G.; Cai, J.; Shen, T.; Shen, C.; Kot, A.C. Decoupled spatial neural attention for weakly supervised semantic segmentation. *IEEE Trans. Multimed.* **2019**, *21*, 2930–2941. [CrossRef]
19. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; Yan, S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 July 2017; pp. 1568–1576.
20. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18 June 2018; pp. 7268–7277.
21. Durand, T.; Mordan, T.; Thome, N.; Cord, M. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 July 2017; pp. 642–651.
22. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
23. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21 July 2017; pp. 618–626.
24. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
25. Ma, F.; Gao, F.; Sun, J.; Zhou, H.; Hussain, A. Weakly supervised segmentation of SAR imagery using superpixel and hierarchically adversarial CRF. *Remote Sens.* **2019**, *11*, 512. [CrossRef]
26. Kwak, S.; Hong, S.; Han, B. Weakly supervised semantic segmentation using superpixel pooling network. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4 February 2017; pp. 4111–4117.
27. Pathak, D.; Krahenbuhl, P.; Darrell, T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1796–1804.
28. Pinheiro, P.O.; Collobert, R. From image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1713–1721.
29. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; SüSstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]

30. Sun, Y.; Zhang, X.; Xin, Q.; Huang, J. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 3–14. [CrossRef]
31. Fu, Z.; Sun, Y.; Fan, L.; Han, Y. Multiscale and multifeature segmentation of high-spatial resolution remote sensing images using superpixels with mutual optimal strategy. *Remote Sens.* **2018**, *10*, 1289. [CrossRef]