MSG-SR-Net: A Weakly Supervised Network Integrating Multiscale Generation and Superpixel Refinement for Building Extraction From High-Resolution Remotely Sensed Imageries

Xin Yan, Li Shen, Jicheng Wang, Xu Deng, and Zhilin Li

***Abstract*—Weakly supervised semantic segmentation (WSSS) methods based on image-level labels can relieve the tedious pixel-level annotation burden, and these methods are mainly based on class activation maps (CAMs). However, it is challenging to generate high-quality CAMs for high-resolution remotely sensed imagery (HRSI). In this article, we propose a WSSS method for building extraction from HRSI using image-level labels. The proposed method, termed as the MSG-SR-Net, integrates two novel modules, i.e., multiscale generation (MSG) and superpixel refinement (SR), to obtain high-quality CAMs so as to provide reliable pixel-level training samples for subsequent semantic segmentation steps. The MSG module is proposed to use global semantic information to guide the learning of multiple features across different levels, and then, respectively, to utilize multilevel features for generating multiscale CAMs. This component can effectively suppress the interference of the class-irrelevant noise and strengthen the use of profitable information in multilevel features. The SR module is designed to take advantage of superpixels to improve multiscale CAMs in target integrity and details preserving. Extensive experiments on two public building datasets demonstrated that the proposed modules made the MSG-SR-Net obtain more integral and accurate CAMs for building extraction. Moreover, experimental results also showed the proposed method achieved excellent performance with over 67% in F1-score, and outperformed other weakly supervised methods in effectiveness and generalization ability.**

***Index Terms*—Building extraction, class activation map, high-resolution remotely sensed imagery, superpixel refinement, weakly supervised deep learning.**

I. Introduction

BUILDING extraction from high-resolution remotely sensed imageries (HRSI) plays a vital role in many important applications, such as population estimation, urbanization evaluation, and urban planning [1]. This task aims to assign each pixel of the image as a building or nonbuilding label, which can be regarded as a binary-class semantic segmentation problem. However, due to the diversity of building objects and the confusion between man-made objects (e.g., building, roads), it is challenging to automatically extract building footprints from HRSI.

In recent years, extensive investigations have been presented to address this challenge, and among them, the fully convolutional neural networks (FCNs) [2]–[7] based methods have become the mainstream methods. As data-driven deep learning methods, the FCNs rely on to some extent the sufficient training on a large number of training images annotated with pixel-level labels. Nevertheless, collecting large-scale pixel-level annotations is very expensive and prohibitively time-consuming. On the other hand, weak annotations, in the form of scribble-level [8], point-level [9], bounding-box-level [10], [11], or image-level labels [12]–[27], however, can be obtained in a relatively cheap manner. Therefore, in this context, weakly supervised semantic segmentation (WSSS) methods have shown a growing potential in the domain of object extraction from HRSI. As stated above, although weak annotations can have different forms, in this article, we focus on pixel-level building extraction by adopting only image-level labels, which indicates the existence of object classes in images, and do not provide any information about their locations or boundaries.

The WSSS based on image-level labels is very difficult, because it needs to infer the precise spatial information only from object presence in the images. To this end, existing works usually rely on class activation maps (CAMs) for obtaining object masks, and then make them into pseudolabels to train a semantic segmentation network. Therefore, the quality of CAMs has a crucial impact on the performance of these methods. Nevertheless, existing methods cannot generate high-quality CAMs for building extraction from HRSI, as they are mainly designed for natural scene images (e.g., PASCAL VOC 2012 dataset [28]), and do not consider the characteristics of building objects in HRSI: 1) more scale variance of building objects in an image, 2) more complicated confusion between building objects and background areas, and 3) the need for more accurate boundaries of building objects.

Manuscript received November 2, 2021; revised November 30, 2021 and December 13, 2021; accepted December 14, 2021. Date of publication December 29, 2021; date of current version January 20, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071386 and in part by the Sichuan Science and Technology Program under Grant 2020JDTD0003. *(Corresponding author: Li Shen.)*

Xin Yan, Li Shen, Xu Deng, and Zhilin Li are with the State-Province Joint Engineering Laboratory of Spatial Information Technology for High-Speed Railway Safety, Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: yxecho.swjtu@gmail.com; rsshenli@outlook.com; rsxudeng@my.swjtu.edu.cn; dean.ge@swjtu.edu.cn).

Jicheng Wang is with the Key Laboratory of Ministry of Education on Land Resources Evaluation and Monitoring in Southwest China, Sichuan Normal University, Chengdu 610068, China (e-mail: wangjicheng123@sicnu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3137450

Considering the characteristics of building objects in HRSI, making full use of multiple features at different levels are key to generating high-quality CAMs for building extraction. Specifically, due to the existence of subsampling layers, multilevel features of CNNs can embed inherent multiscale information, which are beneficial to the identification of building objects with different sizes. Moreover, low-level features of CNNs also contain much low-level information (e.g., edge information), which can contribute to characterizing accurate boundaries of building objects. Therefore, many researchers have tried to generate CAMs utilizing multilevel features of CNNs [29], [30]. MS-CAM [29], first, adopted convolution and up-sample operations to extract available multilevel features, and then designed a fully connected layer with an attention mechanism to enhance the significant informative features and suppress less useful ones. WSF-Net [30] utilized the top-down architecture with skip connections to progressively merge different level features of CNNs. Benefiting from the use of multilevel features, these methods enhanced the quality of CAMs to some extent. However, there exist much complicated confusion between building objects and background areas in HRSI, which may result in accompanying class-irrelevant noises in low-level features of CNNs (e.g., too much noisy texture). And these class-irrelevant noises can impair the quality of CAMs. Yet, these methods ignored this issue.

With the aforementioned considerations in mind, in this article, we propose a weakly supervised method for building extraction from HRSI using image-level labels. The proposed method integrates two novel modules, i.e., multiscale generation (MSG) and superpixel refinement (SR), into a unified framework to obtain high-quality CAM for reliable pseudolabels generation. The MSG module is proposed to use global semantic information to guide the learning of multiple features across different levels, and then, respectively, to utilize multilevel features for generating multiscale CAMs. This component can effectively suppress the interference of the class-irrelevant noise and strengthen the use of profitable information in multilevel features. The SR module is designed to take advantage of superpixels to improve multiscale CAMs in target integrity and details preserving. A superpixel is defined as a group of similar neighboring pixels clustered based on low-level features, such as color histograms and texture features, so superpixels can effectively separate building objects from surrounding background areas, and can also preserve edge details of building objects. Therefore, the fusion of the two modules can ensure the generation of high-quality CAMs, and so as to provide reliable pixel-level training samples for subsequent semantic segmentation steps.

The rest of this article is organized as follows. Related work is reviewed in Section II. The introduction of the proposed method and its components are detailed in Section III. The performance of the proposed method is evaluated in Section IV. Finally, Section V concludes this article.

II. Related Works

A. Building Extraction With CNNs

Extensive investigations have been presented for building extraction based on convolutional neural networks (CNNs) owing to their capacity in hierarchical feature learning. Early studies [31], [32] have achieved pixel-level results through a classification network using the sliding window or superpixel. These methods determined a pixel's label by using CNNs to classify its corresponding sliding window or superpixel. However, these strategies are time-consuming, and ignore the relationship between different sliding windows or superpixels. Soon afterwards, fully convolutional network (FCN) [2] has been proposed, by extending the original CNN structure to enable dense prediction and efficiently generating pixel-level segmentation results. Since then, a variety of FCN-based networks have been developed, such as SegNet [3], U-Net [4], and DeepLab [5]–[7], [33], and also have been applied to building extraction from remote sensing images. However, all FCN-based networks need a large number of pixel-level labels, and collecting such a training dataset is time-consuming and expensive.

B. Weakly Supervised Semantic Segmentation Based on Image-Level Labels

To relieve the cost of pixel-level labeling, weakly supervised semantic segmentation methods based on image-level labels have been studied in recent years. Early methods trained the segmentation network based on prior assumptions about the class distribution from image-level labels, such as multiple instance learning [13], [14] or expectation-maximization formulation [15], [16]. Recently, the methods [17]–[19] utilized pretrained classification networks to generate CAMs for obtaining object masks. Most of existing methods [17]–[28] are based on CAMs to obtain object masks as pseudo labels, and then train a semantic segmentation network on the pseudolabels. However, CAMs generated by the methods [17]–[19] only activate coarse object regions, which cannot be used for training an accurate segmentation network. Therefore, subsequent methods aimed at obtaining CAMs for covering more integral regions of objects.

Some researchers have tried to expand CAMs into integral and accurate regions. SEC [25] designed three losses, i.e., seed loss, expand loss, and constrain loss. DSRG [26] proposed to dynamically expand CAMs based on pseudolabels from CAMs by region growing. AffinityNet [34] used CAMs as pseudolabels and utilized pixel affinity to expand CAMs. IRNet [27] generated class boundary maps and displacement fields from CAMs and utilized them to expand CAMs. BENet [35], first, synthesized boundary annotations by exploiting CAMs, and then trained on the annotations for excavating more object boundaries to provide constraints for the segmentation model. However, these methods still start from initial CAMs, and they learn and expand based on initial CAMs. If initial CAMs only focus on the discriminative parts of buildings or even cover many nonbuilding regions, it is difficult to expand CAMs into integral and accurate regions of buildings. Therefore, the quality of the CAMs has a crucial impact on the performance of these methods.

Other researchers have made improvements in generating CAMs. AE-PSL [20] adopted an iterative erasing strategy to capture complementary regions. SPN [21] adopted superpixel pooling to generate more integral regions. MDC [22] used multiple convolutional layers with different dilation rates to

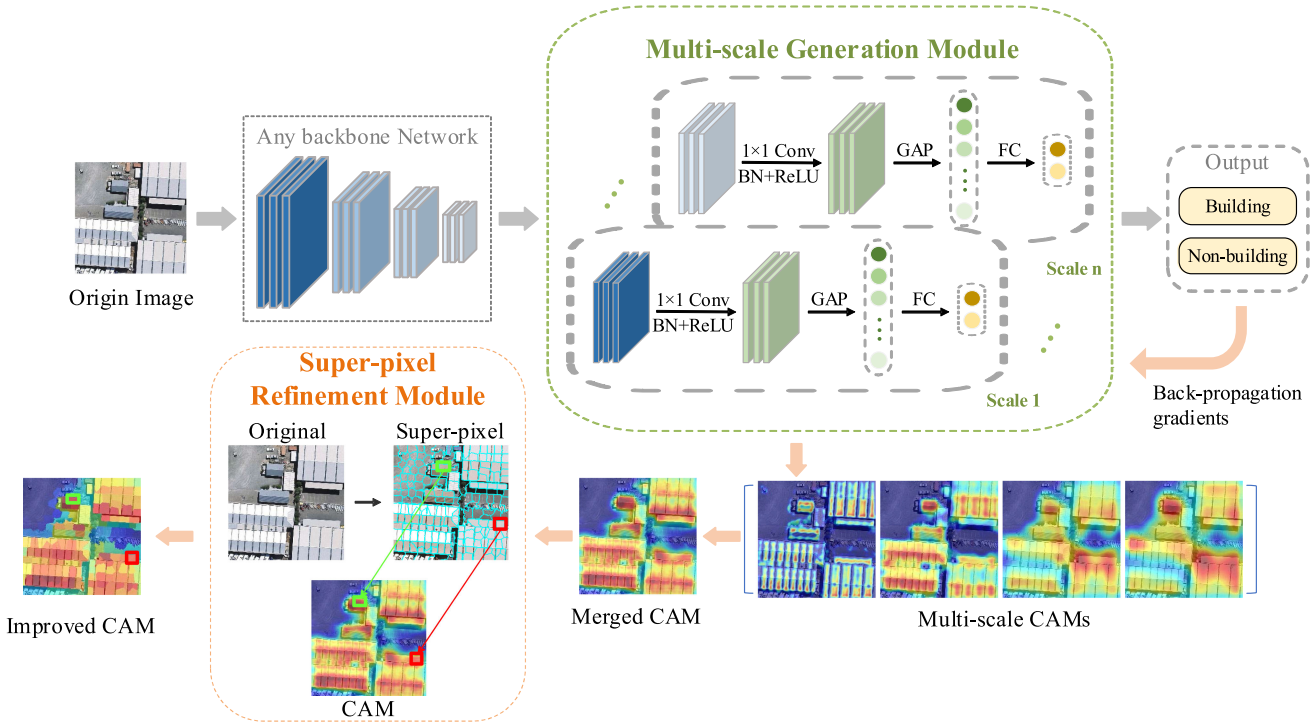


Fig. 1. Framework of the proposed MSG-SR-Net.

expand the activated regions. FickleNet [23] randomly dropped connections in each sliding window and then merged multiple inference results. SEAM [36] forced CAMs predicted from the same images after various transformations to be consistent, to generate more consistent and integral object regions. CIAN [37] exploited cross-image affinity containing same class objects to obtain more consistent object regions. Splitting versus Merging model [38] proposed two losses, i.e., discrepancy loss and Intersection loss, for optimizing the classification model to obtain integral object regions. Although these methods have made improvements in generating CAMs, most of them only focus on the highest feature maps of CNNs, ignoring low-level detailed information, and thus, often generate relatively coarse CAMs. Yet, the coarse CAMs could misclassify surrounding background regions as the object class, and are also unable to obtain integral regions and accurate boundaries of objects. Therefore, many researchers have tried to generate CAMs using multilevel features of CNNs [29], [30]. MS-CAM [29], first, adopted convolution and up-sample operations to extract available multilevel features and then designed a fully connected layer with an attention mechanism to enhance the significant informative features and suppress less useful ones. WSF-Net [30] utilized a top-down architecture with skip connections to progressively merge different level features of CNNs. However, when utilizing multilevel features, these methods ignored that low-level features of CNNs also contain much class-irrelevant noise (e.g., too much noisy texture), which would impair the quality of CAMs.

In this article, our main contribution is to improve the quality of generated CAMs for building extraction by integrating two novel modules, i.e., the MSG and the SR.

III. Proposed Method

In this section, we introduce the proposed weakly supervised building extraction method. It includes two sequential stages: obtaining CAMs via image-level labels, and training a building extraction model with CAMs. During the first stage, we first train a classification network based on image-level labels, then generate CAMs by using the trained classification network, and further improve CAMs. Then, in the second stage, we process improved CAMs into pseudolabels, and train a segmentation model based on the pseudolabels.

The proposed network, termed as the MSG-SR-Net, is comprised of two modules (i.e., the MSG and the SR). As shown in Fig. 1, it is a universal design framework, which carries out a straightforward extension of any classification-based network architecture. Besides, it allows for exploiting pretrained classification models for parameter preconditioning. The MSG-SR-Net is proposed to obtain integral and accurate CAMs. We will give detailed presentation about the MSG and the SR in Section III-A and III-B, respectively. The obtained high-quality CAMs are used for training a building extraction model. For obtaining better building extraction results, we adopt a reliable label selection strategy, which selects confident regions in CAMs for training and ignores uncertain regions. We will introduce this part in Section III-C.

A. Multiscale Generation Module

The MSG module is proposed to adequately utilize profitable information in multilevel features for generating multiscale CAMs. Aimed at eliminating class-irrelative noise in features and avoiding the overuse of high-level semantic information,

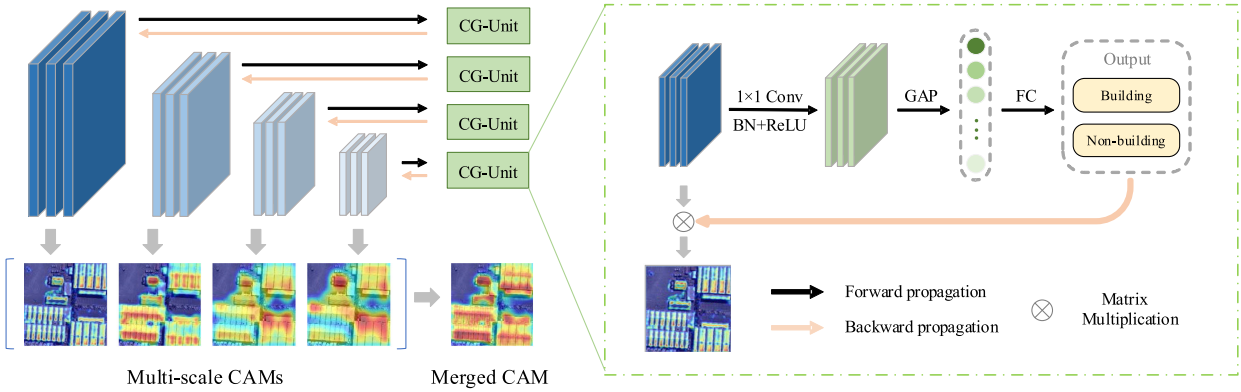


Fig. 2. Multiscale generation module.

the MSG encodes the class-specific semantic information into multilevel features, and then separately utilizes multilevel features to generate multiscale CAMs, as illustrated in Fig. 2.

The MSG comprises multiple CAMs generation units (CG-Unit), which are used to capture multilevel features, as shown in Fig. 2(a). Each CG-Unit consists of a 1×1 convolution layer, followed by a batch normalization layer with a rectified linear unit layer, and a general classification layer, as illustrated in Fig. 2(b). Specifically, a 1×1 convolution kernel is employed to integrate inputted feature maps into new embeddings for the benefit of image classification. The rectified linear unit layer is adopted because we only focus on features that have a positive influence on classification. Subsequently, the filtered features are inputted into a general classification layer, including a globe average pooling layer and a fully connected layer. Finally, the output of the CG-Unit is a tensor, which represents the predicted score of each class. In the training phase, the output is used to calculate a classification loss. In particular, the cross-entropy loss is adopted in this article. Minimizing the classification loss guides multiples features at lower levels to encode the global semantics, thus eliminating class-irrelative noise mixed in features.

Then, in the inference phase, we can exploit multilevel features without class-irrelative noise to generate CAMs. The CAM for each category is obtained by a set of selected feature maps and corresponding weights. We adopt the Grad-CAM++ technique [19] to calculate multiple CAMs with multilevel features, which are separately from different CG-Units. In detail, for each CG-Unit including a C categories classification layer, let a set of corresponding-level feature maps be expressed as $\Omega = \{F^1, F^2, \ldots, F^k, \ldots, F^n\}$, where n is the number of channels and $F^k \in R^{h \times w}$ corresponds to each feature map with $h \times w$ pixels. We represent the contribution score of the k_{th} feature map on the specific class c as w_k^c. So, the spatial location (i, j) in a class-specific CAM A^c is calculated as

$$A_{i,j}^c = \sum_k w_k^c \cdot F_{i,j}^k \,. \tag{1}$$

According to Grad-CAM++ technique, w_k^c is calculated by

$$w_k^c = \sum_i \sum_j \alpha_{i,j}^{k,c} \cdot relu\left(\frac{\partial Y^c}{\partial F_{i,j}^k}\right) \tag{2}$$

where Y^c represents the classification score for class c and $\alpha_{i,j}^{k,c}$ represents the gradient weights at the spatial location (i, j) for the specific class c on the feature map F^k, which can be formulated as

$$\alpha_{i,j}^{k,c} = \frac{\frac{\partial^2 Y^c}{\left(\partial F_{i,j}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial F_{i,j}^k\right)^2} + \sum_a \sum_b F_{a,b}^k \cdot \left\{\frac{\partial^3 Y^c}{\left(\partial F_{i,j}^k\right)^3}\right\}}. \tag{3}$$

Here, both (i, j) and (a, b) are iterators over the same class-specific CAM A^c and are used to avoid the confusion.

In practice, we adopt ResNet-50 [39] as the basic architecture, and we select multilevel features from stages 1–4. Correspondingly, the MSG consists of four CG-Units, which are, respectively, added at the end of each stage. So, we compute four losses in all, and the overall loss is computed as the sum of these losses. Through training with the MSG and the overall loss, we can obtain features of four levels without class-irrelative noise. And in the inference phase, we obtain multiscale CAMs separately generated by four CG-Units and corresponding-level feature maps.

After the abovementioned procedures, CAMs of four scales are calculated. CAMs from low-level features capture more detailed information, while CAMs from high-level features focus on rough building areas, as depicted in Fig. 2(a). Finally, adopting the fusion strategy proposed in [22], we merge multiscale CAMs into final CAMs by $A = \frac{1}{3}\Sigma_{i=1}^{3} A_i + A_4$, where A_i $(i = 12, 34)$ represents CAMs of four scales. In the merged CAMs, nonbuilding areas are suppressed, while building regions are sharply highlighted, as shown in Fig. 2(a).

B. Superpixel Refinement Module

When the MSG is devised to exploit multilevel features for generating CAMs, the SR module is proposed to improve CAMs for better ensuring accurate boundaries and local consistency, i.e., nearby pixels with similar appearance should share the same label.

A superpixel clusters a group of similar pixels in a neighborhood according to some low-level feature-based rules, so it can effectively separate building objects from surrounding

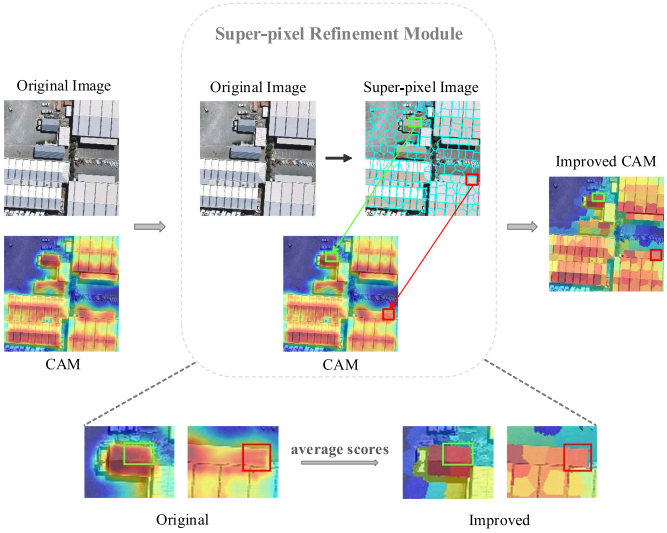


Fig. 3. Illustration of the SR module.

background areas, and can also provide edge information of building objects. To take advantage of this characteristic, the SR module is designed in this article, as illustrated in Fig. 3. The merged CAMs $A \in R^{W\times H}$ and their original image $I \in R^{W\times H\times C}$ are inputted into the SR, where W denotes width, H denotes height, and C denotes the number of channels. First, based on the original image, we adopt the SLIC algorithm [40] to generate its corresponding superpixel map $S \in M^{W\times H}$. $M = [1, N]$ indicates the sign of superpixels, and $S_{i,j} = n$ represents that the pixel at location (i, j) belongs to the n_{th} superpixel. Then, for each pixel in the same superpixel, the average of their building scores is assigned as the final score. In summary, the CAM improved by the SR is expressed as $O^n = \text{mean}\{A^n_{i,j}|(i, j) : S_{i,j} = n,\ n \in [1, N]\}$.

C. Building Extraction Model

Based on the procedures described previously, integral and accurate CAMs can be generated based on sample images with image-level labels. And then, we adopt a reliable label selection strategy to use the CAMs for training the building extraction network in a fully supervised manner.

First, we make CAMs into pseudo pixel-level labels. The CAMs suggest that the higher the score value is, the higher the likelihood of the area belonging to building class will be, while the lower the value is, the higher the likelihood of the area belonging to nonbuilding class will be. Meanwhile, when the score value lies in the middle, the area could belong to building or nonbuilding class. Consequently, to utilize more reliable labels for training a segmentation model, we divide pixels into three groups: 1) building class, 2) nonbuilding class, and 3) uncertain class. We, first, normalize the values of score maps into the range of [0, 1]. Then, a high prior threshold of 0.5 is set, and pixels higher than 0.5 are regarded as building class, whereas pixels lower than a low prior threshold of 0.2 are regarded as nonbuilding class. Particularly, we divide pixels with the score value between 0.2 and 0.5 as the uncertain class, which will be ignored in the training stage. Until now, pseudolabels $Y \in (01, 2)^{W\times H}$ are generated for training the building extraction model, where 0 for nonbuilding class, 1 for building class, 2 for uncertain class, respectively.

We train our building segmentation model based on pseudolabels. DeepLabv3+ [7], one of the most popular fully supervised segmentation models, is selected as our building segmentation model, and the cross-entropy loss function is used as its objective function. For our pseudolabels, the loss is expressed as

$$L = -\sum_{(i,j)\in\Phi_b} \frac{\log\left(P^b_{i,j}\right)}{|\Phi_b|} - \sum_{(i,j)\in\Phi_n} \frac{\log\left(P^n_{i,j}\right)}{|\Phi_n|} \tag{4}$$

where $\Phi_b = \{(i, j)|\ Y_{i,j} = 1\}$, $\Phi_n = \{(i, j)|\ Y_{i,j} = 0\}$ are pixel sets of building and nonbuilding, respectively, $P^b_{i,j}, P^n_{i,j}$ are the building probability and nonbuilding probability predicted by the model for pixel (i, j). Especially, pixel sets of uncertain class are ignored in the training stage. Optimizing on the loss function means minimizing the difference between the real value and predicted value of the model, so that the model can classify building pixels and nonbuilding pixels, and even learn to identify whether pixels of uncertain class in pseudolabels belong to building class or not.

IV. Experiments

A. Experimental Setting

1) Datasets: We evaluate the proposed method on two public building datasets that are popular for evaluating building extraction methods, i.e., the WHU building dataset [41] and the InriaAID building dataset [42]. The two building datasets cover varied urban landscapes, where there are various and versatile architecture types of buildings with different colors, sizes, and usage. Therefore, they are ideal study data to evaluate the effectiveness and robustness of building extraction methods.

The WHU aerial imagery building dataset is a large, high-resolution, accurate, and open-source building dataset consisting of 8189 images with RGB bands. Each image has a size of 512 × 512 pixel and a 0.3-m spatial resolution. The dataset is divided into three parts: 1) a training set of 4736 images, 2) a validation set of 1036 images, and 3) a testing set of 2416 images.

Because the original WHU building dataset is created for fully supervised building extraction, we process it into a weakly supervised segmentation dataset. We keep the original division into training, validation, and testing dataset. With a sliding step size of 128, we crop the images into image blocks with a size of 256 × 256. For the training set, which is for training our weakly supervised building extraction method based on image-level labels, we label the blocks without any building pixels as negative samples and annotate the blocks, whose building coverage rate is over 15%, as positive samples for training stability. In total, 34 142 blocks and corresponding image-level labels are collected for training. For the validation and testing set, which is, respectively, to determine the hyperparameters of the method and to evaluate building extraction performance, the origin pixel-level labels are retained. In total, 9315 blocks

are collected for validation and 21717 blocks for testing, with corresponding pixel-level labels.

The InriaAID building dataset in Chicago consists of 36 aerial images with RGB bands. Each image has a size of 1500 $\times$ 1500 pixel and a 0.3-m spatial resolution. This dataset is labeled in pixel-level into two semantic classes: building class and nonbuilding class.

For the InriaAID building dataset, we first divide it into three parts: a training set of 24 images, a validation set of four images, and a testing set of eight images. Then, we process it into a weakly supervised learning dataset using the same strategies as the WHU dataset. With a sliding step size of 128, we crop the images into image blocks with a size of 256 $\times$ 256. Then, we label the blocks without any building pixel as negative samples and annotate the blocks, whose building coverage rate is over 15%, as positive samples. For the training set, 28925 blocks and corresponding image-level labels are collected for training. For the validation set and testing set, 6084 blocks and 12 168 blocks are, respectively, collected, with corresponding pixel-level labels.

2) Network Settings: The proposed MSG-SR-Net is implemented in the PyTorch [43] platform. We adopt ResNet-50 [39] as our backbone, which is pretrained by the ImageNet dataset [44], and modify it according to the design of the proposed network. We use the SGD optimizer with momentum 0.9 and weight decay 5e$-$4. The initial learning rate is 0.01 and is poly decayed by power 0.9 every epoch. We train the model with batch size 24 for 50 epochs. The training images are augmented by random horizontal flipping, color jittering, and random rotation between $-90°$ and $90°$.

For our final building extraction model, we adopt the DeepLabv3+ network, which adopts ResNet-101 as the backbone and is pretrained by the PASCAL VOC 2012 dataset [28]. For the building extraction model, we also use the SGD optimizer with momentum 0.9 and weight decay 5e$-$4. The initial learning rate is 0.01 and is poly decayed by power 0.9 every epoch. The batch size is set as 32, and the training time is 10 epochs. The final building extraction model is also implemented in the PyTorch platform.

3) Quantitative Metrics: We select several comprehensive metrics for evaluating the quality of our pixel-level building extraction including overall accuracy (OA), intersection-over-unions (IOU) score, and F1 score. In order to adhere to definitions used in the literature, we call building class as positive class and nonbuilding class as negative class. The metrics are calculated as

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{IOU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{6}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{Prec} \times \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}} \tag{7}$$

where TP, TN, FP, FN represent true positives, true negative, false positive, and false negative, respectively; Prec and Rec, respectively, represent Precision rate and Recall rate, calculated by

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{9}$$

TABLE I
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS FOR EACH MODULE OF THE PROPOSED METHOD

Methods	WHU dataset			InriaAID dataset		
	OA	IOU-score	F1-score	OA	IOU-score	F1-score
Baseline[19]	84.12	41.65	58.81	71.96	44.87	61.95
Baseline+SR	88.57	45.43	62.47	79.59	46.79	63.75
Baseline+MSG	91.37	51.85	68.29	80.5	49.97	66.64
Ours	**91.8**	**52.64**	**68.98**	**81.14**	**50.44**	**67.06**

B. Performance Evaluation

The pipeline of our weakly supervised building extraction method based on image-level labels includes: 1) obtain CAMs via image-level labels and 2) train building extraction model with CAMs in a fully supervised manner. Since the proposed network mainly improves on the first step, so to show the effectiveness of our proposed network for obtaining CAMs, model analysis with each proposed module and comparison with other weakly supervised methods are reported. Especially for quantitative analyses of CAMs, a threshold is set on CAMs to obtain segmentation results, and then quantitative analyses are obtained by comparing segmentation results with ground truth labels. Besides, we also compare our building extraction model with models obtained by other weakly supervised methods.

The proposed method mainly make improvement in generating CAMs, so five existing weakly supervised segmentation methods, which are similarly aimed at generating CAMs are adopted for comparison: 1) CAM method [17], 2) GradCAM++ method [19], 3) WILDCAT method [45], 4) superpixel pooling network (SPN) [21], and 5) SEAM method [36]. Note that for all the weakly supervised methods, we follow the same pipeline as our method.

C. CAM Results

1) Ablation Study: To illustrate the effectiveness of our proposed modules in the MSG-SR-Net for obtaining CAMs, we carry out ablation experiments on both the WHU building dataset and the InriaAID building dataset. First, we remove both the MSG and the SR from the MSG-SR-Net, and thus, obtain a simplified network as our baseline method, which is exactly the GradCAM++ method. Second, only the MSG is added into the baseline method, and the obtained network is termed as the baseline+MSG method, which is specifically designed to analyze the impact of the MSG. Third, we only add the SR into the baseline method, creating the baseline+SR method, which is specifically designed to assess the effectiveness of the SR. Finally, we add the SR into the baseline+MSG method, exactly obtained the proposed unified network.

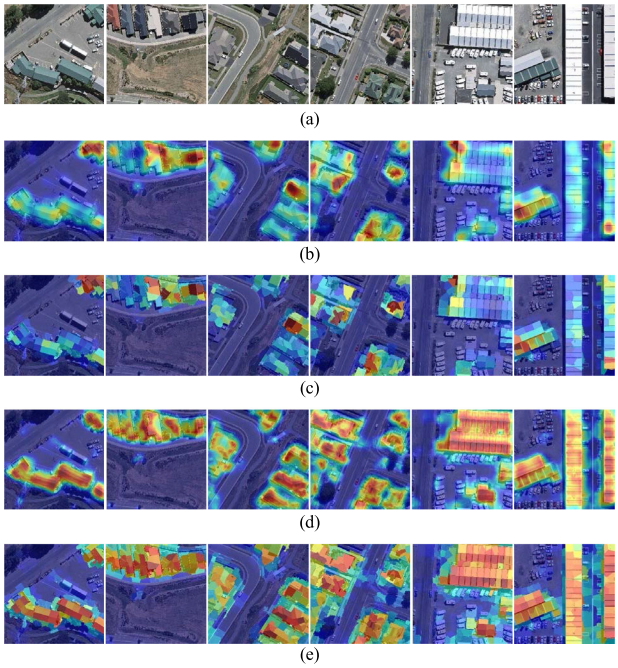


Fig. 4. Qualitative results of ablation experiments for each module of the proposed method.

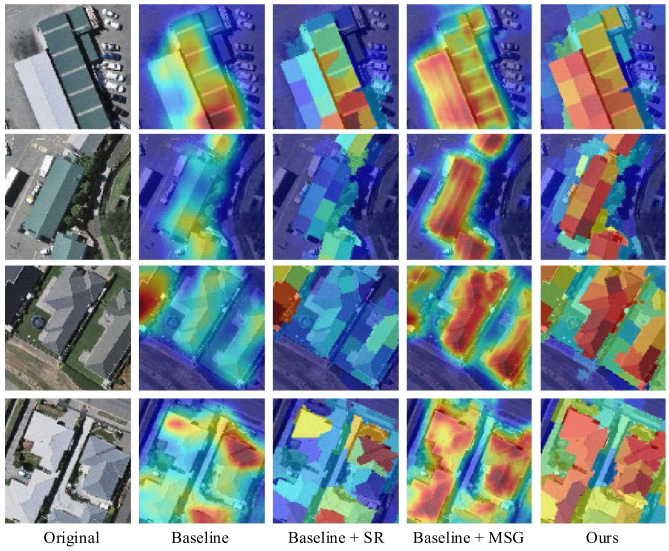


Fig. 5. More detailed qualitative results of ablation experiments for each module of the proposed method.

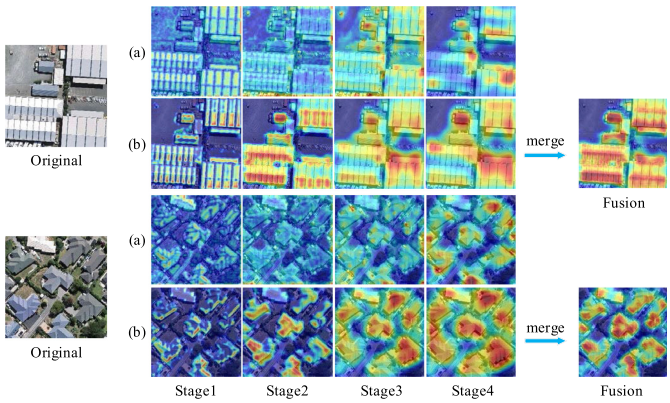


Fig. 6. Qualitative results of multiscale CAMs from stages1–4 (a) without the MSG or (b) with the MSG, and fusion CAMs.

The quantitative results listed in Table I demonstrate that the proposed method achieved the best performance, and both of the proposed two modules make improvements in generating CAMs. It can be seen that the incorporation of the MSG alone or the SR alone can gain a quite considerable improvement. With the help of the SR, the baseline+SR method outperforms the baseline method in all the metrics on both building datasets. It is not surprising because the SR can improve CAMs on accurate boundaries and local consistency. Comparing the baseline+MSG method with the baseline method, the MSG yields the improvement of 7.25 points in overall accuracy, 10.2 points in IOU-score, 9.48 points in F1-score on WHU building dataset, and 5.1 points in IOU-score, 4.69 points in F1-score on InriaAID building dataset. We argue that this is because the MSG can eliminate class-irrelative noise in features and, thus, take advantage of multilevel features to generate multiscale CAMs. And multilevel features, especially low-level features, can contribute to generating high-quality CAMs. Moreover, through the comparison between the baseline+MSG method and the proposed method, we also can see that the addition of the SR further improves performance on two building datasets. Finally, the integration of two novel modules makes the proposed method outperform the baseline method by an extremely significant margin on both two datasets, in detail, 7.68 points in overall accuracy, 10.99 points in IOU-score, 10.17 points in F1-score on WHU building dataset, and 5.57 points in IOU-score, 5.11 points in F1-score on InriaAID building dataset.

For an all-around comparison, we show the benefit of the MSG and the SR qualitatively in Figs. 4 and 5. We can see that CAMs generated by the baseline method focus on the most discriminative parts of buildings, shown in Fig. 4(b); whereas benefiting from both modules, the proposed method can obtain more integral and accurate building regions, as shown in Fig. 4(e). And the more detailed comparison, as shown in Fig. 5, shows that by introducing the MSG, the baseline+MSG method can perform better in obtaining integral regions of buildings and identifying nonbuilding areas around buildings. The representative examples can also be found in the second and fourth rows of Fig. 5. Besides, comparing between Fig. 4(b) and (c), or (d) and (e), we can observe that due to the SR, the CAMs in Fig. 4(c) or (e) can obtain more accurate boundaries of buildings and suppress nonbuilding disturbance. The more representative examples can also be found in the first and third rows of Fig. 5. This means that the SR can further improve CAMs in building boundaries, regardless of the improvement of the MSG.

2) Effect of MSG Module in Utilizing Multilevel Features: In terms of both visual and quantitative results, the MSG gains a quite significant improvement in obtaining CAMs. To better understand the effectiveness of the MSG, we carry out further experiments. In Fig. 6, we, respectively, show a) multiscale

TABLE II
QUANTITATIVE RESULTS OF CAMS WITH DIFFERENT MULTISCALE METHODS

Methods	WHU dataset			InriaAID dataset		
	OA	IOU-score	F1-score	OA	IOU-score	F1-score
Baseline [19]	84.12	41.65	58.81	71.96	44.87	61.95
MS-CAM [29]	89.80	42.60	59.75	64.73	14.74	25.69
WSF-Net [30]	88.99	45.18	62.24	79.52	45.04	62.11
Ours (Baseline + MSG)	**91.37**	**51.85**	**68.29**	**80.5**	**49.97**	**66.64**

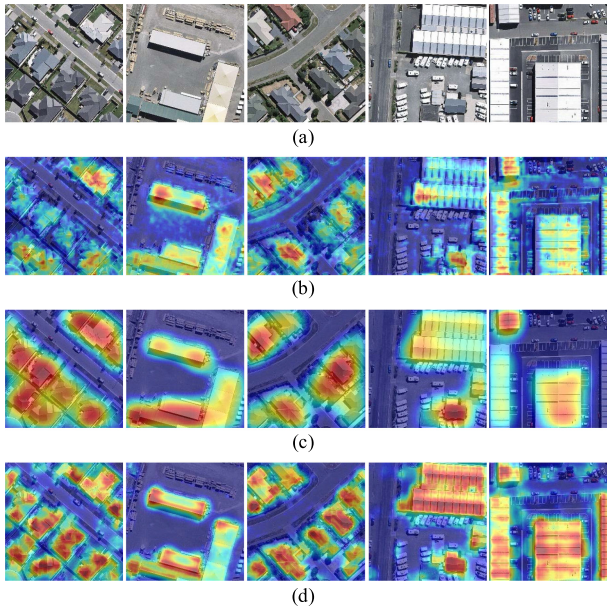


Fig. 7. Qualitative results of CAMs obtained by different multiscale methods.

TABLE III
QUANTITATIVE RESULTS OF CAMS REFINED BY THE SR AND DENSE CRF (CAM* REPRESENTS CAMS OBTAINED BY THE BASELINE+MSG METHOD)

Methods	WHU dataset			InriaAID dataset		
	OA	IOU-score	F1-score	OA	IOU-score	F1-score
CAM*	91.37	51.85	68.29	80.5	49.97	66.64
CAM*+ CRF [46]	**92.16**	45.49	62.53	76.14	48.04	64.91
Ours (CAM*+ SR)	91.8	**52.64**	**68.98**	**81.14**	**50.44**	**67.06**

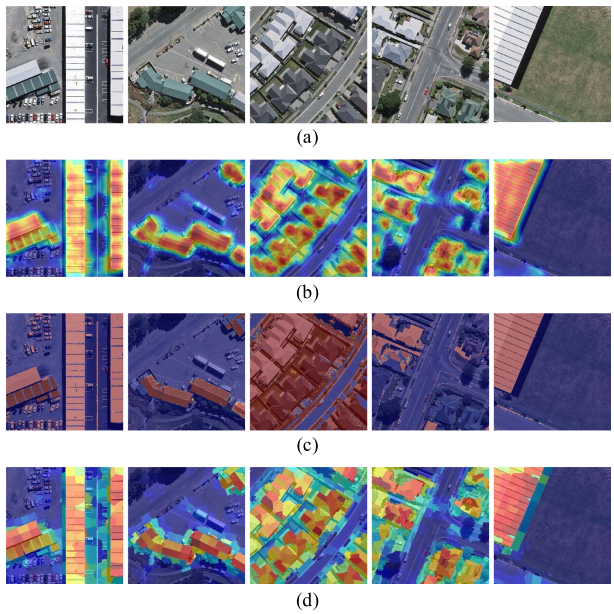


Fig. 8. Qualitative results of CAMs refined by the SR and dense CRF.

CAMs without the MSG and b) multiscale CAMs with the MSG and their fusion CAMs. Multilevel features used for two kinds of CAMs are the same, all from stages 1–4 in ResNet-50. As depicted in Fig. 6, it is obvious that low-level features reveal more spatial details such as edge and texture information. Especially, as shown in Fig. 6(a), in the origin CAMs from low-level features without the MSG, such as the CAMs from stages 1 and 2, there exists much class-irrelevant noise, which can interfere with building extraction. Compared with Fig. 6(a), Fig. 6(b) manifests the effectiveness of the MSG in eliminating class-irrelevant noise in CAMs, which makes CAMs focus on building regions. Moreover, in the fusion CAMs, misclassified nonbuilding areas are further suppressed, while building regions are sharply highlighted.

Besides, Table II and Fig. 7 compare the performance of different multiscale methods for generating CAMs. Obviously, the MSG obtains the best performance of CAMs in terms of both quantitative assessment and qualitative inspection. As seen in Fig. 7, MS-CAM confuses some roads with building regions, and WSF-Net cannot separate different building objects with surrounding background areas. Instead, the MSG module is able to obtain accurate building regions with only a small amount of background noise involved. Besides, the MSG can better characterize accurate boundary details, which is different from other methods. It also should be noted that although all these multiscale methods can improve the performance of CAMs on the WHU dataset, the MS-CAM performs quite badly on the InriaAID dataset, indicating its weak generalization ability. By comparison, the MSG and the WSF-Net show better generalization.

3) Effect of the SR in Refining CAMs: The results in Table III and Fig. 5 show the performance of the SR in refining CAMs. To better understand the effectiveness of the SR, we compare it with the state-of-the-art postprocessing method, dense CRF [46]. The experiments, as illustrated in Fig. 8, demonstrate that dense CRF can refine CAMs to some extent, but its performance is not stable enough. To be specific, for the scenes where building region are significantly different from background areas, CAMs can be refined quite well by dense CRF into integral building regions with accurate boundaries, as shown in the second column and fifth column of Fig. 8. But the dense CRF can also confuse building objects with other surrounding objects (e.g., cars and roads). For example, as shown in the first column and third column of Fig. 8, many cars and roads are misclassified as building class. Besides, some building regions identified in CAMs can be filtered out by dense CRF refinement, as shown in the fourth column of Fig. 8. Therefore, to sum up, the dense CRF postprocessing method may result in a significant decreased accuracy for building extraction. As shown in Table III, the

TABLE IV
QUANTITATIVE COMPARISON OF CAMS WITH THE PROPOSED METHOD AND OTHER METHODS

Methods	WHU dataset			InriaAID dataset		
	OA	IOU-score	F1-score	OA	IOU-score	F1-score
CAM[17]	84.67	42.5	59.66	73.4	39.65	56.78
GradCAM++[19]	84.12	41.65	58.81	71.96	44.87	61.95
WILDCAT[45]	85.51	38.06	55.13	78.64	47.94	64.81
SPN[21]	85.6	36.26	53.23	75.45	44.76	61.84
SEAM[36]	89.74	52.47	68.82	**81.38**	47.28	64.2
Ours	**91.8**	**52.64**	**68.98**	81.14	**50.44**	**67.06**

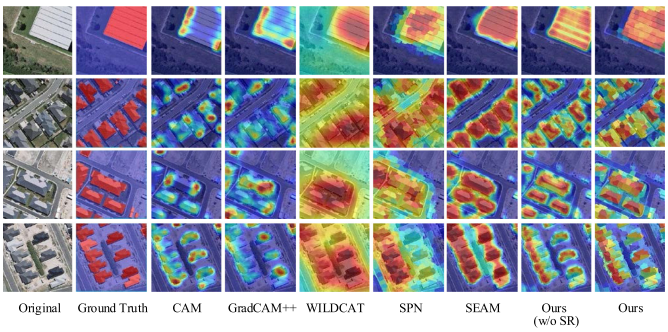


Fig. 9. Qualitative results of CAMs with the proposed method and other methods.

TABLE V
COMPARISON OF QUANTITATIVE RESULTS ON THE WHU DATASET

Methods	Validation dataset			Test dataset		
	OA	IOU-score	F1-score	OA	IOU-score	F1-score
CAM[17]	85.69	45.03	62.1	84.58	40.36	57.51
GradCAM++[19]	82.32	39.98	57.12	81.64	36.13	53.09
WILDCAT[45]	86.25	27.49	43.12	84.97	25.38	40.49
SPN[21]	87.5	42.22	59.38	86.8	38.22	55.31
SEAM[36]	91.81	**56.69**	**72.36**	91.72	**53.73**	**69.90**
Ours	**92.18**	**56.69**	**72.36**	**91.81**	53.66	69.84

TABLE VI
COMPARISON OF QUANTITATIVE RESULTS ON THE INRIAAID DATASET

Methods	Validation dataset			Test dataset		
	OA	IOU-score	F1-score	OA	IOU-score	F1-score
CAM[17]	76.72	43.61	60.73	73.35	42.44	59.59
GradCAM++[19]	78.66	51.9	68.34	75.39	48.38	65.21
WILDCAT[45]	75.08	37.36	54.39	73.0	33.09	49.73
SPN[21]	71.89	37.58	54.63	68.28	34.36	51.15
SEAM[36]	82.46	49.62	66.32	81.61	51.36	67.86
Ours	**85.25**	**56.07**	**71.85**	**85.98**	**58.87**	**74.11**

IOU-score drops after dense CRF refinement from 51.85% to 45.49% on the WHU dataset. Different from dense CRF, the SR module can stably improve CAMs for diverse building scenes. As shown in the Fig. 8, it can refine CAMs to better ensure accurate boundaries and local consistency for different scenes. And the results in Table III also demonstrate that it can consistently improve CAMs quantitatively in the extraction accuracy. Therefore, the SR is more suitable to refine CAMs for building extraction in high-resolution remotely sensed imageries.

4) Comparison With Other Weakly Supervised Methods: In Table IV and Fig. 9, we show the performance of our proposed method on generating CAMs quantitatively and visually, compared with other weakly supervised methods. From Table IV, we can see that our proposed method achieves over 50% in IOU-score, 67% in F1-score on both WHU dataset and InriaAID dataset, and outperforms most of other weakly supervised methods by an obvious margin. Especially, the SEAM method achieves similar performance to our proposed method in IOU-score and F1-score on WHU dataset and in overall accuracy on InriaAID dataset, but considering all the metrics, the proposed method performs better. As shown in the visualization results of Fig. 9, the proposed method can obtain more integral regions of buildings than the CAM method and GradCAM++ method. Particularly, as shown in the second and fourth rows of Fig. 9, it is obvious that the proposed method successfully separates adjacent buildings, whereas other methods including WILDCAT, SPN, SEAM, misclassify many background regions around buildings. This happened because both proposed modules make our method more effectively exploit multilevel features, particularly low-level features, to generate CAMs, and multilevel features (e.g., texture) can help to classify diverse building objects and distinguish between nonbuilding and building regions. Moreover, from the first row of Fig. 9, it can be seen that the proposed method also obtains more accurate boundaries of buildings. The reason is that the MSG and the SR can effectively utilize rich detailed information of multilevel features and the characteristic of superpixels, which both contribute to obtaining accurate edge information of buildings.

D. Building Extraction Results

In this section, we verify the effectiveness of our building extraction model, compared with models obtained by other weakly supervised methods using image-level labels. For further illustrating the robustness of our building extraction model, we evaluate it on two public building datasets, which contain various building objects with different colors, sizes, and usage.

1) Results on the WHU Building Dataset: The comparison results of our proposed method and other methods on the WHU building dataset are provided in Table V. Our building extraction model obtains an excellent performance of 92.18% in overall accuracy, 56.69% in IOU-score, 72.36% in F1-score on the validation dataset, and 91.81% in overall accuracy, 53.66% in IOU-score, 69.84% in F1-score on the test dataset. The comparison results in Table V also indicate that our model can outperform most compared models with an obvious margin, and the SEAM model can have a similar performance to ours. SEAM model performs slightly better on the test dataset than ours in the metrics of IOU-score and F1-score, but our model performs better in the OA metric.

In Fig. 10, the segmentation results on the WHU dataset for different methods are visualized for a better inspection.

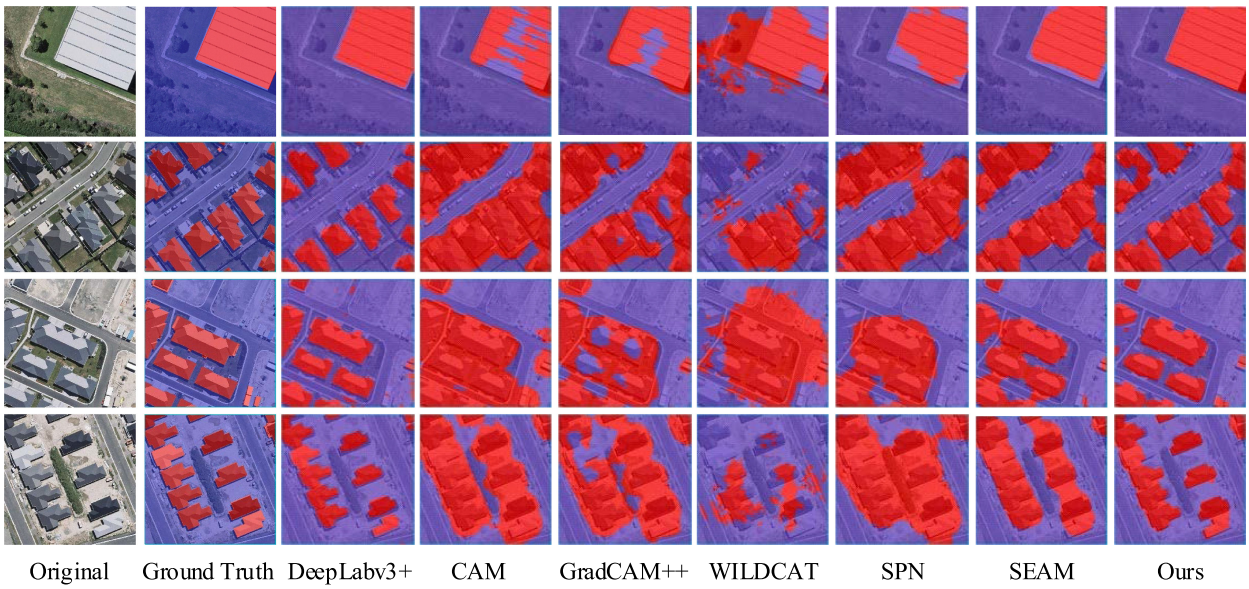


Fig. 10. Comparison of qualitative results on the WHU dataset.

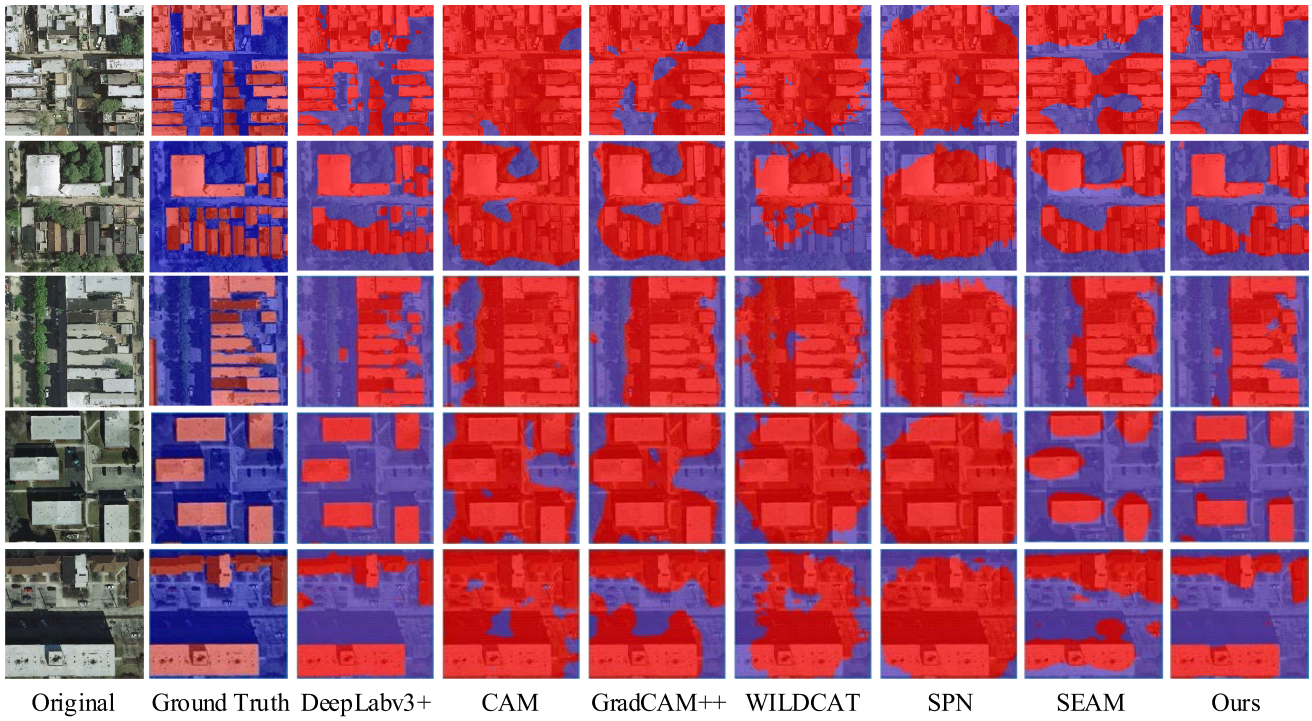


Fig. 11. Comparison of qualitative results on the InriaAID dataset.

Obviously, compared to other weakly supervised methods, our model performs better in integrity and accurate boundaries of buildings, and a representative example can be found in the first row of Fig. 10. Moreover, our model also can accurately distinguish different building objects. For example, as shown in the fourth row of Fig. 10, our model successfully separates adjacent buildings, where the background areas between the two buildings severely interfere with the predictions from other methods. Compared with ground truth labels, there are still some misclassified pixels in the results of our proposed method, however, which are less than other weakly supervised methods.

2) Results on the InriaAID Building Dataset: We also carry out experiments on the InriaAID dataset, to further evaluate the effectiveness and the generalization ability of our proposed weakly supervised method on building extraction. The quantitative comparison of the InriaAID dataset is listed in Table VI, and the visualization results are also shown in Fig. 11. From the visualization results in Fig. 11, it is obvious that our model

performs better in integral and accurate regions of buildings. We argue that the reason is that the performance of the final building extraction model is strongly related to the quality of CAMs. According to the analysis about the performance of CAMs, benefit from the MSG and the SR, the proposed method is capable of generating more accurate and integral CAMs. Therefore, our building extraction model trained on these high-quality CAMs can obtain more excellent extraction results.

According to Table VI, our building extraction model achieves the top performance of over 85% in overall accuracy on both validation and test dataset. As for the metrics of IOU-score and F1-score, our model reaches over 55% and 70% on both the validation dataset and the test dataset. Slightly different from the results on the WHU dataset, the proposed method can perform favorably against all the other compared methods on the InriaAID dataset, including the SEAM model, which has a similar performance to ours on the WHU dataset. It is not surprising because the InriaAID dataset contains more diverse building objects and more adjacent buildings, as shown in the first column of Fig. 11. And due to lack of multiscale information, most compared methods have an unsatisfactory performance on this kind of buildings, while benefit from the MSG and the SR, the proposed method can take advantage of multilevel features and superpixels, which contribute to separating adjacent buildings, and classifying diverse building objects with different sizes and types. Therefore, the proposed method can have more excellent and more robust performance.

V. Conclusion

In this article, we proposed the MSG-SR-Net to generate high-quality CAMs for weakly supervised building extraction based on image-level labels, which integrates multiscale CAMs and SR. Extensive experiments on two building datasets, i.e., the WHU building dataset and the InriaAID building dataset, show that the proposed MSG-SR-Net can identify accurate building regions and achieve excellent building extraction performance. Moreover, qualitative and quantitative analysis results verified that the proposed two novel modules, i.e., the MSG and the SR, can effectively utilize multilevel features of CNNs and the characteristic of superpixels, and thus, enable more precise weakly supervised building extraction. Ablation studies for the two modules convincingly demonstrated that the MSG can eliminate class-irrelative noise in features and adequately utilize multilevel features for generating multiscale CAMs, and the SR can further improve CAMs in target integrity and details preserving. Besides, through performance evaluation on two datasets, we demonstrate that our building extraction model obtained by the proposed MSG-SR-Net can achieve excellent building extraction performance and outperform other weakly supervised methods in the effectiveness and the generalization ability on building extraction.

References

[1] J. R. Jensen and D. C. Cowen, "Remote sensing of urban/suburban infrastructure and socio-economic attributes," *Photogramm. Eng. Remote Sens.*, vol. 65, pp. 611–622, 1999.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFS," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for semantic image segmentation," Dec. 2017. Accessed: Jul. 27, 2021. [Online]. Available: http://arxiv.org/abs/1706.05587

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[8] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 3159–3167.

[9] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[10] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1635–1643.

[11] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 876–885.

[12] J. Chen, F. He, Y. Zhang, G. Sun, and M. Deng, "SPMF-Net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion," *Remote Sens*, vol. 12, no. 6, 2020, Art. no. 1049.

[13] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," 2014, *arXiv:14127144*.

[14] F. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 413–432.

[15] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1742–1750.

[16] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1796–1804.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 2921–2929.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[19] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

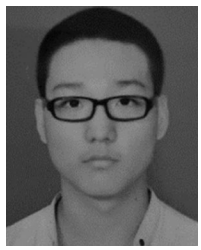
[20] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1568–1576.

[21] S. Kwak, S. Hong, and B. Han, "Weakly supervised semantic segmentation using superpixel pooling network," in *Proc. AAAI Conf. Artif. Intell.*, pp. 4111–4117, Feb. 2017.

[22] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7268–7277.

[23] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5267–5276.

[24] Y. Zhu, Y. Zhou, H. Xu, Q. Ye, D. Doermann, and J. Jiao, "Learning instance activation maps for weakly supervised instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3116–3125.
[25] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 695–711.
[26] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7014–7023.
[27] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2209–2218.
[28] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[29] X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin, and Q. Chen, "MS-CAM: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in SD-OCT images," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3443–3455, Dec. 2020.
[30] K. Fu *et al.*, "WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1970.
[31] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, 2017, doi: 10.1016/j.isprsjprs.2017.05.002.
[32] Z. Guo, Q. Chen, G. Wu, Y. Xu, R. Shibasaki, and X. Shao, "Village building identification based on ensemble convolutional neural networks," *Sensors*, vol. 17, no. 11, Nov. 2017, Art. no. 11, doi: 10.3390/s17112487ovb.
[33] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10548–10559, Oct. 2021, doi: 10.1109/JSTARS.2021.3119286.
[34] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4981–4990.
[35] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 347–362.
[36] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12275–12284.
[37] J. Fan, Z. Zhang, T. Tan, C. Song, and J. Xiao, "Cian: Cross-image affinity net for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10762–10769.
[38] T. Zhang, G. Lin, W. Liu, J. Cai, and A. Kot, "Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 663–679.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[41] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[42] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.
[43] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 8026–8037, 2019.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[45] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 642–651.
[46] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2011, pp. 109–117.



Xin Yan received the B.S. degree in remote sensing science and technology in 2018 from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, where he is currently working toward the Ph.D. degree in photogrammetry and remote science.

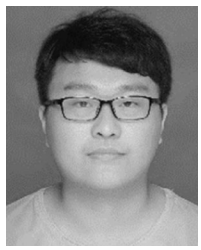
His research interests include computer vision and remote sensing image processing.

Li Shen received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008, and the Ph.D. degree in photogrammetry and remote science from the College of Resources Science and Technology, Beijing Normal University, Beijing, China, in 2013.

He also spent one year with the University of Waterloo as a Visiting Student from September 2011 to September 2012. He is currently an Associate Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China. His current research interests include computer vision, and machine learning for remote sensing image understanding and geospatial data analysis, with applications to disaster management and environmental monitoring.

Jicheng Wang received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2012 and 2020, respectively.

He is currently an Assistant Professor with the Key Laboratory of Ministry of Education on Land Resources Evaluation and Monitoring in Southwest China, Sichuan Normal University, Chengdu. His research interests include computer vision and remote sensing image processing.



Xu Deng received the B.S. degree in remote sensing science and technology in 2019 from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, where he is currently working toward the M.S. degree in photogrammetry and remote science.

His research interests include computer vision and remote sensing image processing.

Zhilin Li received the B.Eng. degree in photogrammetry and remote sensing from Southwest Jiaotong University, Chengdu, China, in 1982, and the Ph.D. degree in photogrammetry and remote sensing from the University of Glasgow, Glasgow, U.K., in 1990.

He is currently a Professor with the Faculty of Geosciences and Environmental Engineering, and the Director with the State-Province Joint Engineering Laboratory of Spatial Information Technology for High-Speed Railway Safety, Southwest Jiaotong University. He was a Research Associate/Fellow with the University of Newcastle upon Tyne, University of Southampton, and the Technical University of Berlin. He was with the Curtin University of Technology, Bentley, Australia, as a Lecturer for two years. He was with the Hong Kong Polytechnic University from 1996 to 2020 as Assistant/Associate Professor, Professor, and Chair Professor. He has authored more than 200 journal papers and authored three books in cartography, GIS and related areas. His current research interests include multiscale modeling and representation of spatial data, feature extraction from remote sensing imagery, and application of remote sensing in high-speed railway safety.