On the Effectiveness of Weakly Supervised Semantic Segmentation for Building Extraction From High-Resolution Remote Sensing Imagery

Zhenshi Li, Xueliang Zhang, *Member, IEEE*, Pengfeng Xiao, *Senior Member, IEEE*, and Zixian Zheng

***Abstract*—A critical obstacle to achieve semantic segmentation of remote sensing images by the deep convolutional neural network is the requirement of huge pixel-level labels. Taking building extraction as an example, this study focuses on how to effectively apply weakly supervised semantic segmentation (WSSS) to high-resolution remote sensing (HR) images with image-level labels, which is a prominent solution for the huge labeling challenge. The widely used two-step WSSS framework is adopted, in which the pseudo-masks are first produced from image-level labels and followed by a segmentation network trained by the pseudo-masks. In addition, the fully connected conditional random field (CRF) is utilized to explore spatial context in both training and prediction stages. Detailed analyzes are implemented on applying WSSS on HR images in terms of producing pseudo-masks, training segmentation network, and optimizing predictions. We show that the trade-off between precision and recall of pseudo-masks, as well as the boundary accuracy and the background, needs to be carefully considered. The benefits of the segmentation network in the two-step framework are demonstrated in comparison to using classification network only for WSSS, and the effects of CRF-loss are identified to be powerful for improving the segmentation network while it is not appropriate for dense buildings. An overlapping strategy and CRF postprocessing are further demonstrated to be effective for optimizing the segmentation results during inferencing. Through deliberate settings, we can generate results comparable to fully supervised on the ISPRS Potsdam and Vaihingen dataset, which is meaningful for promoting WSSS applications for extracting geographic information from HR images.**

***Index Terms*—Building extraction, fully convolutional network, high-resolution remote sensing imagery, weakly supervised semantic segmentation (WSSS).**

I. Introduction

With the rapid progress of high-spatial resolution satellites, an increasing amount of high-resolution remote sensing (HR) images are getting available. It is crucial to automatically and accurately extract geographic information from HR images for applications. Semantic segmentation of HR images aims at assigning a geographic label to every pixel through an end-to-end mechanism, which was significantly promoted by deep convolutional neural networks (DCNNs) [1]–[3], especially by the progress of fully convolutional network [4]. It has been widely used for many geographic applications, e.g., cloud detection [5], [6], land cover mapping [7], [8], and urban target localization [9].

Under the supervision of rich pixel-level label dataset, e.g., the ISPRS 2-D benchmark [10], the Gaofen Image Dataset [11], and the Zurich Summer Dataset [12], fully convolutional networks were reported to be able to make use of the spatial context in images and to extract multiple-level features with escalating receptive field, which greatly pushed forward the performance of semantic segmentation for remote sensing images [13]–[15].

However, obtaining the huge amount of pixel-level labels that are required for training fully convolutional network is time-consuming, laborious, and expensive and even demands expertize and fieldwork. According to statistics, it takes 10.1 minutes on average to label a natural image in pixel-level, which is nearly 150 times than the time needed for labeling in image-level [16]. Additionally, it is more challenging to label remote sensing images, which possess a great variety of geographic objects, not only because of the large data size, but also due to the conceptual difficulty [17]. Consequently, it is urgent to develop methods which can perform impressively on remote sensing images with easily labeled datasets. To cope with the difficulty, several studies adopted transfer learning method [18], [19], which still needs a small quantity of pixel-level labeled data to fine-tune the network trained by other datasets. In the other way, semantic segmentation with weak supervision shed new light on overcoming the labeling difficulty for remote sensing images [20]–[22]. Instead of pixel-level label, segmentation network can be effectively trained by leveraging weak annotations such as image-level label, point supervision, scribble annotation, and bounding box, which are easily obtained because of the low annotation costs [23].

Among various types of weak annotations, the image-level annotation (image tag), which only indicates presence or absence of objects in an image, is the cheapest yet the most difficult type for weakly supervised semantic segmentation (WSSS) [23]. Compared with pixel-level label, image-level label does not

Manuscript received November 12, 2020; revised January 28, 2021; accepted March 1, 2021. Date of publication March 4, 2021; date of current version March 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297 and Grant 41871235, in part by the Fundamental Research Funds for the Central Universities under Grant 020914380080, and in part by the High-level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China. *(Corresponding author: Xueliang Zhang.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: lzhenshi@outlook.com; zxl@nju.edu.cn; xiaopf@nju.edu.cn; zhengzx95@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3063788

provide any spatial prior information of geographic objects that is necessary for training segmentation network, such as the location, extent, boundary, and shape of objects.

Semantic segmentation trained by image-level labels has achieved rapid progress in computer vision field, in which the two-step training strategy is widely exploited to solve the problem of lacking spatial prior information. Specifically, the pseudo-masks of segmentation are first produced by image tags, and the segmentation network, such as FCN [4], U-Net [24], DeepLab [25], is then trained with the pseudo-masks [26]–[29]. When producing pseudo-masks, the location cues (seeds) of objects are first obtained by image tags and then expanded to the whole object extents. With the benefit of indicating object location by classification network (trained from image tags), there are several visualization methods to obtain the location cues, including deconvolution [30], back-propagation [31], and global average pooling (GAP) [32]. The obtained seeds are always ineffective to be directly used for training a segmentation network because they only cover a small range of the whole object extents. Hence, much attention has been paid to further augmenting the seeds to the whole object extents, e.g., by adversarial erasing [29], seed region growing [26], and expanding the seeds themselves [28].

In remote sensing community, image-level label has attracted increasing attention for weakly supervised geographical information extraction. For example, several weakly supervised object detection (WSOD) methods for remote sensing images have been proposed, e.g., the progressive contextual instance refinement method [33], and the dynamic curriculum learning method [34]. As to image-level WSSS for remote sensing images, it needs more spatial prior information for training a segmentation model compared with WSOD, and several pioneering works have also been proposed [20]. On one hand, based on the two-step training strategy, Fu *et al.* [35] and Chen *et al.* [36] achieved binary segmentation of water, cloud, and building with fully convolutional network trained by image tags, where they focused on improving the quality of pseudo-masks, and thus, increasing segmentation accuracies. However, what a suitable pseudo-mask should be remains unclear when applying two-step WSSS on remote sensing images. It is always difficult to produce pseudo-masks with both high completeness and high accuracy. In this case, we need to understand what is the influence of the completeness and accuracy of pseudo-masks on WSSS for HR images and whether other impact factors exist, which can illustrate the direction of improving WSSS performance. On the other hand, the location cues from classification network trained with image tags were directly explored for geographic information extraction, in which case the advantage of training a powerful segmentation network was ignored. For example, Wang *et al.* [37] performed cropland segmentation with two types of weak annotation: single point label and image tag. Ali *et al.* [38] used an attention-based method to detect destruction regions. Li *et al.* [39] proposed a new global convolutional pooling operation and the local pooling pruning strategy to improve the quality of CAM for cloud detection. Accordingly, another question is raised—what is the benefit of the segmentation network within the two-step training framework? It will help demonstrate the effectiveness of the two-step training framework for weakly supervised segmentation of HR images by answering this question.

Fig. 1. Example of an HR image for illustrating the challenges of building extraction with fully convolutional network supervised by image-level labels.

Automatically extracting building information from HR images is of great significance to urban planning, population modeling, and environmental improvement, etc. Geographic object-based image analysis (GEOBIA) is the main method for extracting buildings from HR images [40], but it is difficult to determine an optimal image segmentation scale [41]–[43] and always requires a strong domain-specific knowledge for feature extraction [44]. Nowadays, fully convolutional network has received a lot of attention for this task because of its ability to extract features at multiple semantic levels from HR images [44]–[46]. However, the fully supervised model was mostly adopted, which relies on a wealth of pixel-level labels. When supervised by image-level labels, as shown in Fig. 1, the following challenges of extracting buildings from HR images need to be overcome by exploring fully convolutional network: 1) the large size HR images tend to cover multiple buildings, which makes it difficult to identify all the buildings by an image-level label; 2) buildings are distributed as multiple scales in HR images; 3) the high intraclass heterogeneity in HR images makes it difficult to extract the robust features; and 4) the low interclass heterogeneity will bring difficulty for discriminating buildings from backgrounds.

The focus of this study is on how to effectively achieve segmentation of buildings from HR images with the supervision of image-level labels, aiming at providing a technical reference for alleviating the difficulty of collecting pixel-level annotations for training the fully convolutional network. Elaborate analyzes are carried out in order to illustrate the key impactors of applying two-step WSSS to HR images. The main contributions of this study can be summarized as follows.

1) We demonstrate how the quality of pseudo-masks influences the successive training of segmentation network. To produce appropriate pseudo-masks, the contradiction between completeness and accuracy, as well as the boundary accuracy should be taken good care of.

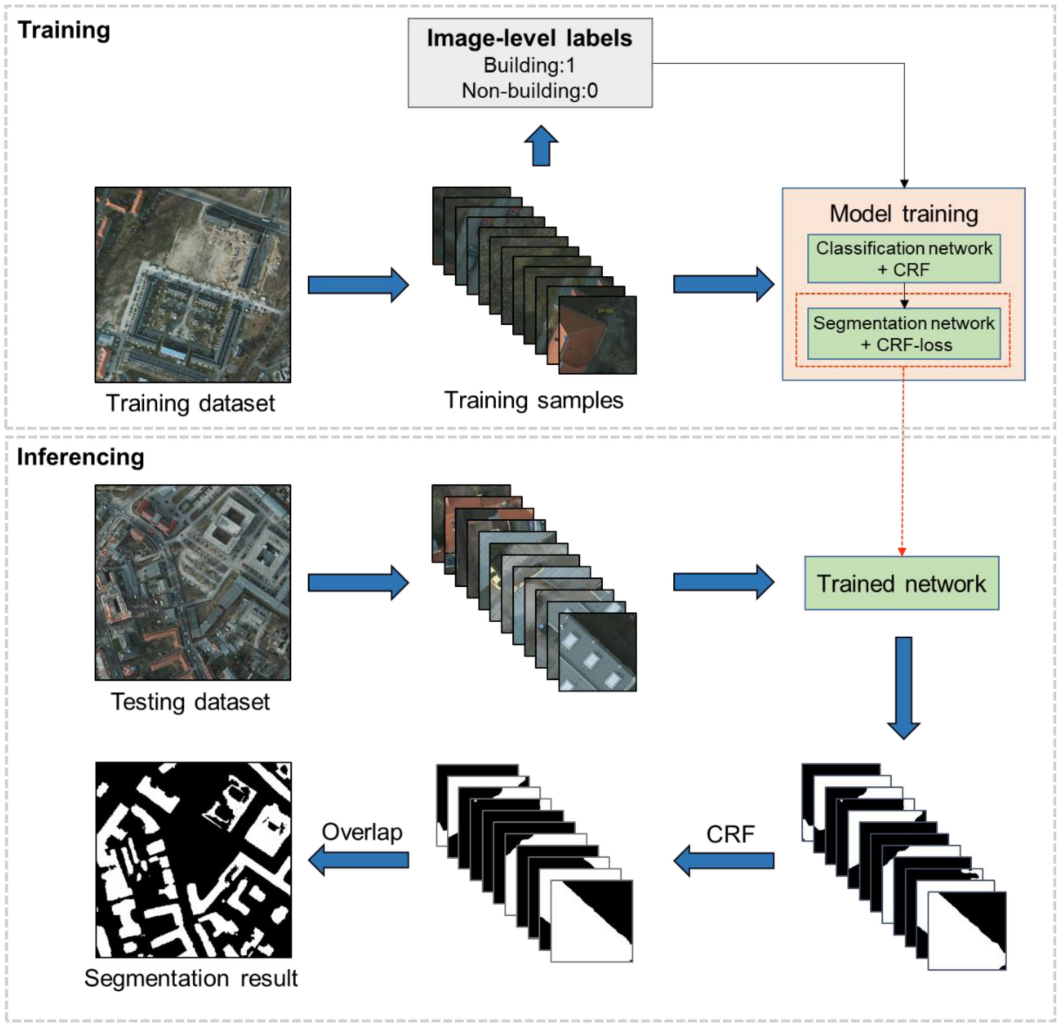


Fig. 2. Pipeline of the WSSS framework.

2) We show the benefits of training a segmentation network on improving WSSS performance for HR images by comparing with using only classification network.
3) We demonstrate that although the CRF-loss is able to taking advantage of the spatial context information in HR images for training the segmentation network, it may not be appropriate for images with dense buildings.
4) Through deliberate settings, outstanding results can be generated within our two-step WSSS framework whose *F*-score accounts for 95.3% and 91.3% performance of fully supervised model on the ISPRS Potsdam and Vaihingen dataset, respectively.

II. Methodology

A. Overview

In this study, we implement detailed analyzes on the procedure of two-step WSSS, including producing pseudo-masks, training segmentation network, and optimizing predictions. The workflow of our two-step image-level WSSS method is illustrated in Fig. 2. In terms of weakly supervised training, HR images are cropped into patches as training samples together with their image-level labels. A classification network is first trained by image tags to produce pseudo-masks, the precision and boundaries of which are improved by applying CRF. The pseudo-masks are then used to train a segmentation network with CRF-loss. When inferencing, the image patches are predicted by the trained segmentation network and the results are further optimized with CRF. An overlapping strategy is adopted to fuse result patches which can prevent context limitations and give full play to the network.

B. Training Procedure

The weakly supervised training procedure is shown in Fig. 3, which is roughly divided into two steps: generating pseudo-masks and training a segmentation network. Image tags are used to train a classification network to produce location cues of buildings, which are then optimized by CRF to generate pseudo-masks. The pseudo-masks consist of three forms of pixels: foreground (building), background (nonbuilding), and ignored pixels (unlabeled). The segmentation network is trained using the pseudo-masks, which is the final semantic segmentation model in the WSSS framework. The training procedure of

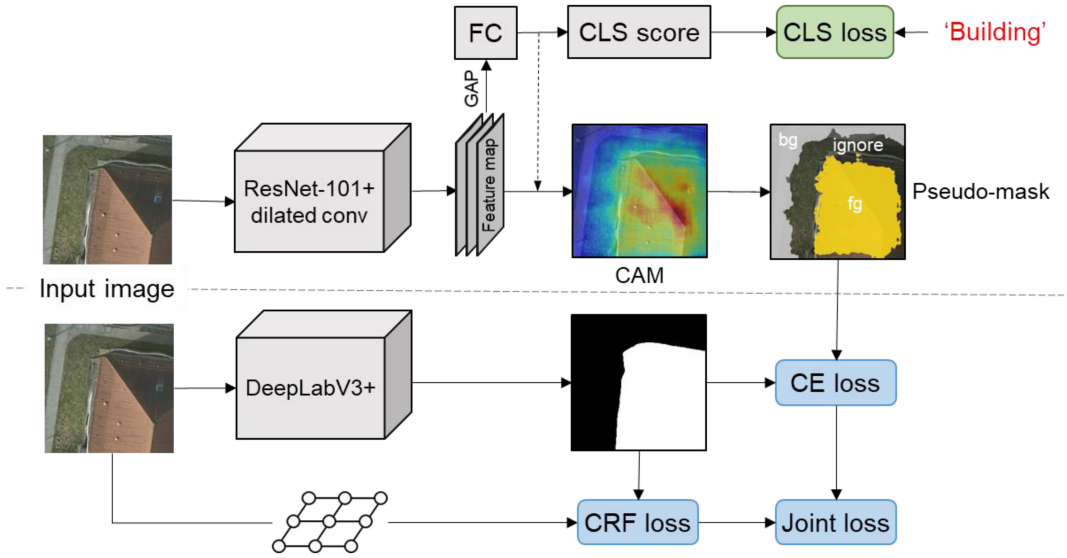


Fig. 3. Two-step weakly supervised training procedure of our method with image-level labels. The upper part is to produce pseudo-masks by training a classification network with image tags (Section II-B1). The lower part is to train a segmentation network with pseudo-masks under a joint loss (Section II-B2).

the segmentation network exploits a joint loss function, which combines the cross entropy loss on foreground and background pixels in the pseudo-mask and the CRF-loss considering all the pixels in the image patch to utilize spatial context information.

1) Generating Pseudo-Masks: Pseudo-masks can provide partial spatial prior information and thus be used to train a segmentation network. Hence, it is necessary to generate high-quality pseudo-masks [47] and to know how the quality of pseudo-masks affects the successive segmentation network training. Although image-level labels provide no spatial prior information, it has been proved that DCNN for classification trained from image tags is capable of retrieving object location cues according to the contribution of different positions to final classification score [31], [32].

In our framework, we adopt the method of class activation map (CAM) [32] to retrieve location cues of buildings, which denotes the probability of every pixel belonging to the target category of building. To improve the completeness of CAM, the classification network is replaced as a modified ResNet-101 [48] in which we import dilated convolution into the latter convolutional layers to perceive larger receptive field [49]. To be specific, dilated rate is set as two for conv4_x in ResNet-101 and four for conv5_x [25]. GAP is applied on the last feature map to connect the 3-D feature map and a fully-connected layer (FC). Then, the fully connected layer outputs one classification (CLS) score to compute CLS loss as follows:

$$l\ (y,\hat{y}) = -\left[y\log\sigma\left(\hat{y}\right) + (1-y)\log\left(1-\sigma\left(\hat{y}\right)\right)\right] \quad (1)$$

where σ is the sigmoid function, $\hat{y}$ is the CLS score, and y is the image tag.

After training the classification network, the fully connected layer weights are applied to the final feature map to generate a heatmap (CAM) for buildings.

To alleviate the problem caused by various scales of buildings in HR images, the multiscale aggregation strategy [50] is used to generate CAM with higher generality. Specifically, given an image and scale ratio, multiscale CAM is generated by taking average of CAMs from different scaled images as shown as follows:

$$M_{ms} = \sum_{i=0}^{n}\left(M_c\left(s_i\right)/\left(n+1\right)\right) \quad (2)$$

where $M_c(s_i)$ is the CAM of the scaled image with scale ratio s_i, $s_i \in \{s_0, s_1, \ldots, s_n\}$.

Usually, the foreground region is obtained by applying a hard foreground-threshold (fg-thre) on CAM to extract pixels with scores greater than the threshold [27]. To further improve the quality of the foreground region and explore the direction of optimizing pseudo-mask, we utilize CRF to realize optimization of CAM, which could result in two kinds of results: 1) producing foreground region that closely adheres to ground-truth for relatively regular buildings, as shown in Fig. 4(a); and 2) covering most or even the entire image patch for complex buildings or very large buildings, as shown in Fig. 4(b). For the first case, the foreground region can work similarly as fully supervised labels, which can help improve network capabilities. For the second case, the CRF result would bring additional errors, which need to be prevented. Hence, we make an intersection of the fg-thre result and the CRF result to obtain foreground region, which can balance the above two cases.

The CAM is a score map with single channel indicating the probabilities of pixels belonging to building. When applying CRF on CAM, it requires another score map that indicates the nonbuilding probabilities. We refer to the method [51] as shown as follows:

$$M_{nb}\ (x,y) = \left\{1 - M_b\left(x,y\right)\right\}^{\propto} \quad (3)$$

where M_b is the normalized result of CAM so that the maximum value equals to 1 and the minimum value equals to 0, and $\propto \geq 1$ represents the decay parameter.

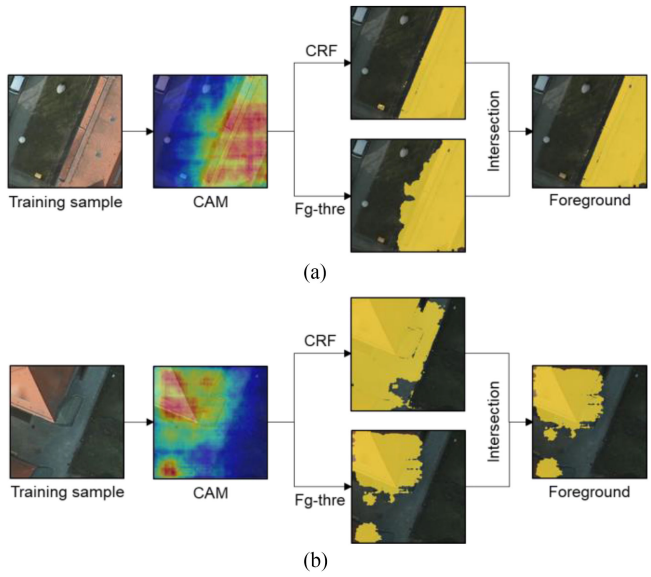


Fig. 4. Procedure of producing foreground region from CAM. (a) Case of obtaining accurate foreground region for regular building. (b) Case of producing numerous error pixels after applying CRF on CAM. The color ranging from blue to red in CAM indicates the increasing probability of foreground.

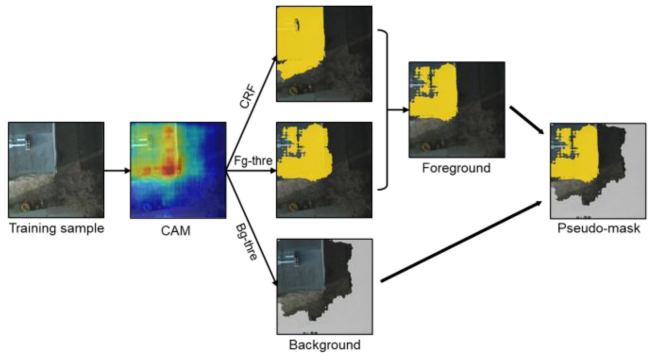


Fig. 5. Procedure of generating pseudo-mask for training segmentation network.

In addition to obtaining foreground region, we apply a background threshold (bg-thre) to CAM and extract the pixels with scores lower than the bg-thre. Finally, a pseudo-mask is generated by combining the foreground and the background regions. To sum up, the procedure of generating pseudo-mask is shown in Fig. 5. We will make a discussion about the quality of pseudo-masks by setting different thresholds as well as ingredient study, aiming at illustrating the influence of pseudo-masks on WSSS in terms of precision, completeness, background, and boundary accuracy.

2) Training Segmentation Network With CRF-Loss: For two-step weakly supervised segmentation, the segmentation network training procedure is of great importance because it is the final segmentation model. In this study, we set our sights on the benefits of the segmentation network itself and the effects of the loss function for training the segmentation network within the two-step workflow.

Since the generated pseudo-masks are inevitably incomplete and inaccurate, the segmentation network directly trained by this kind of labels would thus output errors in addition the network itself. Accordingly, we adopt CRF-loss [52], [53] into our loss function to prompt the network observing implicitly boundary and spectral consistency by taking advantage of the spatial context information in HR images during the training procedure.

Specifically, the standard Potts/CRF model could be expressed as follows [53]:

$$E\ (i,j) = \sum_{i,j\in\Omega} G\,(i,j)\,[S_i \neq S_j] \tag{4}$$

where Ω represents all the pixels in the image, $S_i, S_j \in \{0,1\}$ indicates the binary class label assigned to pixel i and j, [•] means the Iverson bracket, and G(i, j) is a matrix of pairwise discontinuity costs, for which we use the appearance kernel by [54]

$$G\ (i,j) = w \cdot \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) \tag{5}$$

where p is the pixel position and I is the pixel spectral value, θ_α and θ_β are hyper parameters controlling the "scale" of the Gaussian kernels, and w is the normalized term which can be found efficiently by a combination of expectation maximization and high-dimensional filtering [54]. This function penalizes the nearby similar pixels with different labels assigned.

In order to adapt the above function to the segmentation network, a quadratic relaxation is proposed in [53] as follows:

$$E\ (i,j) = \sum_{i,j\in\Omega} G\,(i,j)\,P_c\,(i)\,(1 - P_c\,(j)) \tag{6}$$

where P is the softmax output of the segmentation network and c∈{"building", "nonbuilding"}. Besides, to mitigate the negative effects caused by inaccurate pseudo-masks, we add the constraint term [50] to our binary CRF-loss as in (8). Note that the labeled pixels refer to the produced foreground and background pixels as illustrated in Section II-B1, and the unlabeled pixels refer to the ignored pixels which are neither foreground nor background. The final CRF-loss is calculated as follows, and the fast computation of the gradient for CRF-loss with dense Gaussian kernel is described in [54]:

$$L_{\mathrm{CRF}} = \sum_{\substack{i,\,j\,\in\,\Omega \\ i\,\neq\,j}} R\,(i)\,E\,(i,j) \tag{7}$$

$$R\ (i) = \begin{cases} 1 - \max\limits_{c\in C}\left(P_c\,(i)\right),\ i \in \text{labeled} \\ 1,\ i \in \text{unlabeled} \end{cases}. \tag{8}$$

The segmentation network is trained under a joint loss, as shown in (9), which consists of the CRF-loss in (8) calculated on all the pixels and the cross entropy loss (CE-loss) calculated only on labeled pixels.

$$L_{\mathrm{joint}} = L_{\mathrm{CRF}} + L_{CE}. \tag{9}$$

C. Inference Considering Spatial Context From Different Perspectives

The inference procedure is illustrated in the "inferencing" part of Fig. 2. The large HR images also need to be cropped into patches for inference. CRF is used for postprocessing the segmentation network output [54], [55].

It is noted that the cropped patches risk to split buildings, and thus, lead to incomplete buildings in a patch with inadequate spatial context information, which could lower the segmentation accuracy. This defect would limit the performance of a segmentation network, particularly for weakly supervised model whose identification ability is not so strong as that of fully supervised model. We adopt a simple yet effective overlapping strategy to alleviate this problem.

The overlapping strategy is used for concatenating segmentation results of cropped patches to obtain the full-size result. We crop a large image into patches with a certain overlapping degree (ratio of overlapping size to patch size), and input the patches into the trained segmentation network to obtain building segmentation outputs, which indicate the probability of pixels belonging to building or nonbuilding. After that, the average of overlapping areas is calculated as the final score. Finally, argmax is applied to the fused image to obtain semantic segmentation results. Though seemingly ordinary, this strategy can significantly improve the final accuracy caused by object incompleteness. Detailed validations will be given in Section IV-E.

III. Experiment Setup

A. Dataset

The Potsdam and Vaihingen benchmark dataset from ISPRS 2-D Semantic Labeling Challenge [10] are used to validate our framework. The Potsdam dataset is composed of 38 images with size of 6000×6000 pixels and spatial resolution of 5 cm. We extract the RGB bands from the original 4-band IRRGB (Infrared, Red, Green, and Blue). The Vaihingen dataset is composed of 33 images with approximately 2100×2100 pixels and spatial resolution of 9 cm. We remove the image 7_10 in Potsdam dataset because of its error annotations of buildings [15]. Among the remaining 37 images, we use 23 images for training and the other 14 images for testing. For Vaihingen dataset, we use 16 images for training and the other 17 images for testing.

The training images are cropped into 256×256 patches with a sliding stride of 128. Then we assign image-level label "building" to the images occupying building pixels more than 25% of the total pixels and "non-building" to images without building pixels. Finally, we obtain 36245 training patches from Potsdam dataset with 18416 positive samples ("building") and 17829 negative samples ("non-building"), and 2255 training patches from Vaihingen with 1841 positive samples and 414 negative samples. In addition, randomly horizontal flipping is used for data augmentation.

B. Implementation Details

Environment: The experiment is conducted on PyTorch1.4.0 and Python3.7. The whole model is trained on a computer with an Intel Core i7-9700KF CPU, one NVIDIA GeForce RTX 2080 Ti GPU, and 64 GB memory.

Network architecture: A modified ResNet-101 [48] imbedded with dilated convolution is used as the classification network for producing CAM, as described in Section II-B1. The DeepLabV3+ [56] with output stride 8 is used as the segmentation network. In order to enhance feature extraction ability of the networks, both classification and segmentation networks are initialized by the ResNet-101 pretrained on ImageNet [57].

Training: The training procedure is performed as follows: weak annotations (image tags) are used to train the classification network, which is used to produce CAMs and pseudo-masks of the whole training dataset. Then the pseudo-masks are used to train the DeepLabV3+ segmentation network based on the joint loss in (9).

When training the networks, we utilize the batched stochastic gradient descent (SGD) optimizer with momentum $= 0.9$ and weight decay $= 0.0005$. The initial learning rate is set as 0.001 and the learning rate decay strategy of "poly" is deployed. For both networks, we train 30 epochs with the batch size as 10. Besides, the scale ratio of multiscale CAM is set as $\{0.5; 1; 1.5; 2\}$ as in [46] and the decay parameter $\propto$ is set as 4. The parameters of CRF-loss follow the setting in [50] without further optimization. The thresholds for obtaining the foreground and background regions when generating pseudo-masks will be discussed in Section IV-A.

Testing: The trained segmentation network is applied on testing images to obtain building segmentation results. For convenience, we split the original large testing images into patches of size 250×250 pixels when inferencing. The overlapping strategy is then adopted to fuse segmentations of cropped patches, and its effectiveness will be analyzed in Section IV-E.

Evaluation metrics: Four accuracy metrics are used to evaluate the accuracies of results, including the intersection-of-union (*IoU*), *precision*, *recall*, and *F*-score, which are formulized as follows:

$$IoU = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \tag{10}$$

$$precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{11}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

$$F\text{-score} = 2 \cdot \frac{precision \times recall}{precision + recall} \tag{13}$$

where TP, FN, and FP represent to truly predict the "building" pixels as positive, to falsely predict the "building" pixels as negative, and to falsely predict the "nonbuilding" pixels as positive, respectively. *IoU* and *F*-score indicate the overall segmentation accuracy. *Precision* and *recall* indicate the omission and commission errors, respectively. It is noted that we calculate the accuracies based on the "no boundary" ground-truth provided by ISPRS.

TABLE I
ACCURACIES OF THE FOREGROUND REGIONS IN PSEUDO-MASKS WITH DIFFERENT FG-THRE AND THE CORRESPONDING SEGMENTATION NETWORK OUTPUTS (BG-THRE = 0.2) ON THE ISPRS POTSDAM DATASET, WHERE THE ACCURACIES OF PSEUDO-MASKS ARE CALCULATED ON TRAINING SAMPLES AND THOSE OF SEGMENTATION NETWORK OUTPUTS ON TESTING SAMPLES

Fg-thre	Foreground in pseudo-mask				Segmentation network output			
	IoU	*Precision*	*Recall*	*F*-score	*IoU*	*Precision*	*Recall*	*F*-score
0.7	0.207	0.930	0.210	0.343	0.779	0.860	0.894	0.877
0.6	0.413	0.920	0.428	0.584	0.789	0.875	0.891	0.883
0.5	0.609	0.901	0.653	0.757	0.806	0.872	0.914	0.893
0.4	0.728	0.870	0.818	0.843	0.781	0.833	0.926	0.877
0.3	0.769	0.831	0.913	0.870	0.777	0.839	0.914	0.875

C. Experiment Design for Analyzing the Effectiveness of WSSS

The process of our analyzes is organized as follows. Effects of the pseudo-mask quality on the segmentation network are evaluated in Section IV-A. Section IV-B shows the benefits of training a segmentation network by comparing with only using a classification network for WSSS of HR images. Section IV-C presents the advantages and disadvantages of the CRF-loss when training segmentation network, as well as the optimization effects of the CRF postprocessing. Our WSSS results are compared with fully supervised case in Section IV-D, and the gap can be further narrowed by the overlapping strategy as described in Section IV-E. Finally, the comparison results with other state-of-the-art two-step WSSS methods are presented in Section IV-F.

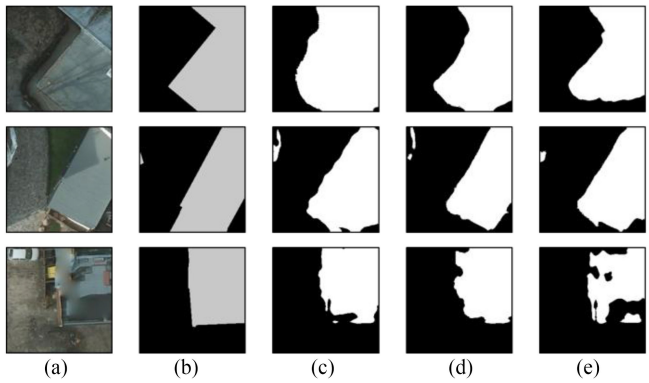


Fig. 6. Examples of segmentation network outputs with different fg-thre. The images are taken from the ISPRS Potsdam dataset. (a) Image. (b) Ground-truth. (c) Fg-thre = 0.4. (d) Fg-thre = 0.5. (e) Fg-thre = 0.6.

IV. RESULTS

A. Influence of Pseudo-Mask Quality

In this section, we analyze how the quality of pseudo-mask affects the successive training of the segmentation network. The quality of the pseudo-masks is directly influenced by the thresholds of both foreground (fg-thre) and background (bg-thre) regions applied on CAM, as well as by the CRF optimization. Therefore, the accuracies of segmentation network outputs are evaluated and compared, where the segmentation networks are trained by pseudo-masks with different qualities, aiming at revealing the relative importance of the accuracy (indicated by *precision*) and the completeness (indicated by *recall*) of the pseudo-mask for training segmentation network in WSSS of HR images.

Influence of fg-thre: The accuracies of the foreground regions in pseudo-masks by setting different fg-thre and the corresponding segmentation network outputs on the Potsdam dataset are presented in Table I, given bg-thre as 0.2. As fg-thre changes from 0.3 to 0.7, the *precision* and *recall* of pseudo-masks gradually increases and decreases, respectively, because fewer foreground pixels with higher probability of building are left in pseudo-masks. The *IoU* and *F*-score of pseudo-mask achieves the highest value when fg-thre is set as 0.3. However, the accuracy change trend of the segmentation network outputs with increasing fg-thre is different with that of the pseudo-masks. When fg-thre is 0.5, the *precision* and *recall* of the segmentation network output achieve the best tradeoff, and thus, the highest *IoU* of 0.806.

The comparison of the two change trends indicates that the highest overall accuracy of pseudo-mask (indicated by *IoU* and *F*-score) may not achieve the best training performance of segmentation network under weakly supervised condition, which means that the best *precision-recall* tradeoff for pseudo-mask cannot assure the best *precision-recall* tradeoff for segmentation network. It is also shown that too many or too few foreground pixels in pseudo-mask would do harm to the performance of segmentation network, as the cases of setting fg-thre as 0.3 and 0.7, respectively. In the case of too few foreground pixels by setting fg-thre as 0.7, the spatial prior information is not enough for training the segmentation network, while in the case of too many foreground pixels by setting fg-thre as 0.3, the pseudo-mask contains many error labels. Furthermore, the same experiments are conducted on the Vaihingen dataset, as shown in Table II, whose accuracy changes similarly with Potsdam.

To further illustrate the influence of the pseudo-mask quality on segmentation network when the foreground pixels are neither too few nor too many, the example segmentation outputs on Potsdam by setting fg-thre as 0.4, 0.5, and 0.6 are presented in Fig. 6. Combining Fig. 6 and the accuracies in Table I, it shows that more foreground pixels (higher completeness but lower accuracy) in pseudo-mask would make the segmentation network produce more target pixels. By contrary, fewer foreground pixels in pseudo-mask would result in fewer target pixels. In this range, a suitable fg-thre needs to be cautiously set with the consideration of the successive segmentation performance, rather than only for achieving the best tradeoff between the completeness and the accuracy of pseudo-masks. It is worth

TABLE II
ACCURACIES OF THE FOREGROUND REGIONS IN PSEUDO-MASKS WITH DIFFERENT FG-THRE AND THE CORRESPONDING SEGMENTATION NETWORK OUTPUTS (BG-THRE = 0.3) ON THE ISPRS VAIHINGEN DATASET

Fg-thre	Foreground in pseudo-mask				Segmentation network output			
	IoU	*Precision*	*Recall*	*F*-score	*IoU*	*Precision*	*Recall*	*F*-score
0.9	0.037	0.991	0.037	0.071	0.163	0.953	0.165	0.281
0.8	0.199	0.981	0.200	0.332	0.599	0.958	0.615	0.749
0.7	0.427	0.953	0.437	0.599	0.721	0.834	0.842	0.838
0.6	0.612	0.895	0.659	0.759	0.718	0.803	0.872	0.836
0.5	0.684	0.806	0.819	0.812	0.695	0.768	0.881	0.821

TABLE III
ACCURACIES OF BACKGROUND REGIONS IN PSEUDO-MASKS WITH DIFFERENT BG-THRE AND THE CORRESPONDING SEGMENTATION NETWORK OUTPUTS (FG-THRE = 0.5) ON THE ISPRS POTSDAM DATASET

Bg-thre	Background in pseudo-mask				Segmentation network output			
	IoU	*Precision*	*Recall*	*F*-score	*IoU*	*Precision*	*Recall*	*F*-score
0.1	0.235	0.955	0.238	0.381	0.776	0.844	0.906	0.874
0.2	0.419	0.906	0.438	0.591	0.806	0.872	0.914	0.893
0.3	0.552	0.828	0.623	0.711	0.792	0.896	0.873	0.884

TABLE IV
ACCURACIES OF BACKGROUND REGIONS IN PSEUDO-MASKS WITH DIFFERENT BG-THRE AND THE CORRESPONDING SEGMENTATION NETWORK OUTPUTS (FG-THRE = 0.7) ON THE ISPRS VAIHINGEN DATASET

Bg-thre	Background in pseudo-mask				Segmentation network output			
	IoU	*Precision*	*Recall*	*F*-score	*IoU*	*Precision*	*Recall*	*F*-score
0.2	0.296	0.960	0.300	0.457	0.696	0.783	0.863	0.821
0.3	0.488	0.941	0.503	0.656	0.721	0.834	0.842	0.838
0.4	0.646	0.908	0.691	0.785	0.671	0.920	0.712	0.803

mentioning that the bg-thre is set as a unified value here (0.2 for Potsdam and 0.3 for Vaihingen) to analyze the influence of fg-thre. The change trend also presents the same pattern as demonstrated above when setting different unified bg-thre, which we do not show due to space constraints.

Influence of bg-thre: The background label is also crucial for binary segmentation since it indirectly imposes constraint to foreground. Tables III and IV show the accuracies of the background regions in pseudo-masks and the corresponding segmentation network outputs on the Potsdam and Vaihingen dataset, respectively, when setting different bg-thre.

Even though different bg-thre values would not lead to a change of the foreground region, the accuracies of segmentation network outputs are truly changed with different bg-thre values. As presented in Tables III and IV, a too small bg-thre value (0.1 for Potsdam or 0.2 for Vaihingen) would produce background region in pseudo-mask with high *precision* but low *recall*. In this case, the trained segmentation network tends to produce output with low *precision* because they may cover more commission building pixels due to the lack of sufficient supervision on nonbuilding pixels, as shown in Fig. 7(c). By contrary, a too large bg-thre value (0.3 for Potsdam or 0.4 for Vaihingen) would produce background region in pseudo-mask with low *precision* and high *recall*, which indicates that building pixels are mistakenly labeled as background in the pseudo-mask. This makes a trained segmentation network sustain too much constraint when predicting foreground, and thereby output results with excessively low *recall*, as shown in Fig. 7(e). In summary, inadequate background constraint may introduce more false predictions while too much constraint may lead to incomplete results.

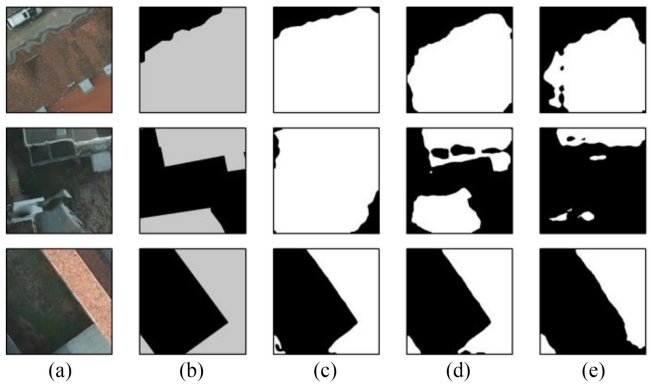


Fig. 7. Examples of segmentation network outputs with different bg-thre. (a) Image. (b) Ground-truth. (c) Bg-thre = 0.1. (d) Bg-thre = 0.2. (e) Bg-thre = 0.3.

Effectiveness of optimizing pseudo-mask with CRF: The foreground in pseudo-mask is improved by intersecting the foreground region from CAM threshold with the result of applying CRF on CAM (CAM-CRF), as presented in Fig. 4. It should be noted that the CRF operation can hardly work on the CAMs of the Vaihingen dataset, because the building pixels produced by CAM-CRF are always covering nearly the entire image due to the densely distributed buildings in the Vaihingen dataset. Hence, we only conduct the CRF optimization for pseudo-mask

TABLE V
ACCURACIES OF THREE TYPES OF FOREGROUND REGION IN PSEUDO-MASK AND THE CORRESPONDING SEGMENTATION NETWORK OUTPUTS (FG-THRE = 0.5, BG-THRE = 0.2) ON THE ISPRS POTSDAM DATASET. "CAM-CRF" REPRESENTS THE METHOD APPLYING CRF ON CAM TO OBTAIN FOREGROUND IN PSEUDO-MASK, "FG-THRE = 0.5" ONLY APPLYING A HARD FOREGROUND THRESHOLD TO CAM, AND "INTERSECTION" INTERSECTING THE ABOVE TWO FOREGROUND REGIONS

Method	Foreground in pseudo-mask				Segmentation network output			
	IoU	*Precision*	*Recall*	*F*-score	*IoU*	*Precision*	*Recall*	*F*-score
CAM-CRF	0.767	0.797	0.954	0.868	0.771	0.830	0.916	0.871
Fg-thre=0.5	0.608	0.894	0.655	0.756	0.783	0.851	0.908	0.879
Intersection	0.609	0.901	0.653	0.757	0.806	0.872	0.914	0.893

on the Potsdam dataset here since this optimizing operation is not effective for the Vaihingen dataset. Table V shows the accuracies of the three types of foreground regions before and after intersection on the Potsdam dataset, where fg-thre and bg-thre are set as 0.5 and 0.2, respectively.

Compared with the foreground by threshold, CAM-CRF achieves apparently lower *precision* and extremely high *recall*, which leads to a 0.16 higher *IoU* and 0.11 higher *F*-score. As demonstrated above, a too large *recall* together with low *precision* indicates that CAM-CRF tends to cover too much false foreground, which is harmful for segmentation network training, and thus, leads to the worst segmentation network output as shown in Table V. In addition, this comparison further demonstrates a high *IoU* (or *F*-score) of pseudo-mask would not assure a high *IoU* (or *F*-score) of segmentation network output.

Though the performance of segmentation network directly trained on CAM-CRF is even negatively affected, the accuracy of segmentation network output is apparently improved by training on pseudo-mask with intersected foreground. However, the accuracy of the intersected foreground is very similar with that of the foreground from threshold, where the *IoU* and the *precision* of the intersected foreground are only 0.001 and 0.007 higher, as shown in Table V.

This comparison at first demonstrates the effectiveness of CAM-CRF on improving the performance of training segmentation network through intersection. However, it also raises another question—why such a small accuracy improvement of pseudo-mask could result in an apparently large accuracy improvement of segmentation network output? Fig. 5 has demonstrated that CRF can help improve the foreground boundary for regular buildings. To further prove that, quantitative boundary accuracies of the two kinds of pseudo-masks are calculated using the measure of edge location error [58]. The edge location error of the pseudo-mask before intersection (with only fg-thre) is 0.673 and that after intersection is 0.524 (a larger error value indicates more mismatching boundaries). The quantitative results demonstrate that foreground boundaries in pseudo-mask are improved, which would thus promote the segmentation network to capture more precise boundaries. Several example segmentation outputs are presented in Fig. 8 to show the improvement of building boundaries by intersected foreground. In summary, this comparison indicates that the boundary accuracy is another important character of the quality of pseudo-mask.

How to produce good pseudo-masks: As to the foreground region in pseudo-mask, it could be safe to summarize that both high *precision* and high *recall* would promote the training performance of segmentation network. However, the *precision* and *recall* are always needed to be tradeoff. When too few foreground pixels are obtained, which is indicated as high *precision* with low *recall*, the spatial context prior information is not enough to support the high-quality training of segmentation network. By contrary, when too many foreground pixels are obtained, which is indicated as low *precision* with high *recall*, the error foreground pixels outside the true objects would be harmful to segmentation network training. Hence, a suitable solution of obtaining foreground region should be cautiously settled within the range of neither too few nor too many foreground pixels. It is also demonstrated that the best *precision-recall* tradeoff case for foreground region would not assure the best final segmentation accuracy. Accordingly, we cannot simply determine the best solution of generating foreground region by the largest *IoU* or *F*-score value. Beyond that, improving the boundary accuracy of foreground region is another important factor for training segmentation network to improve segmentation accuracy.

As to the background region, which imposes constraints to foreground objects, it is also crucial to insure sufficient and accurate background pixels. A suitable tradeoff of background region needs to be found out for training an effective segmentation network.

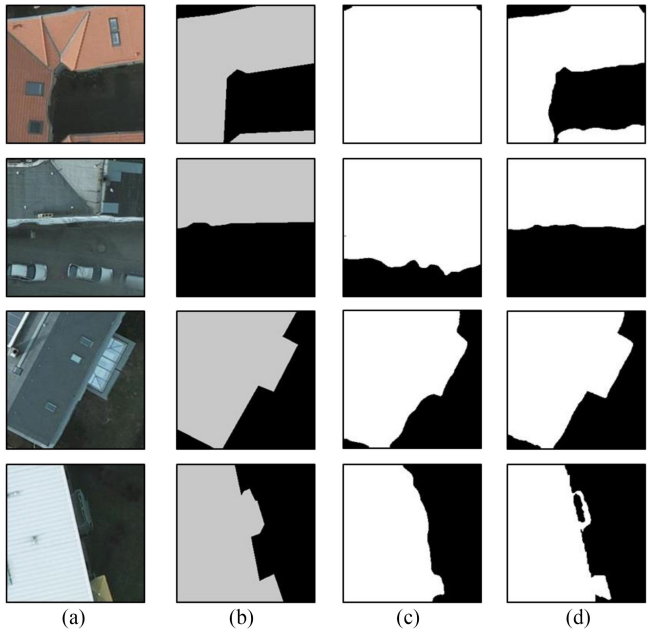


Fig. 8. Examples of segmentation network outputs trained by pseudo-masks containing different types of foreground regions on the Potsdam dataset. (a) Image. (b) Ground-truth. (c) Segmentation using foreground by applying hard threshold on CAM. (d) Segmentation using foreground by intersecting the threshold CAM and the CAM-CRF.

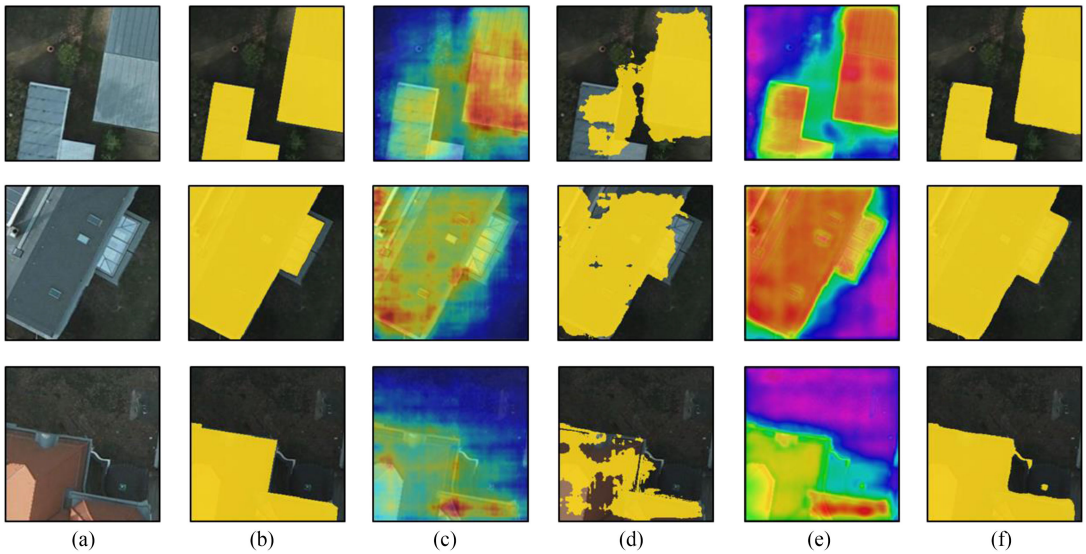


Fig. 9. Examples on the Potsdam dataset. (a) Image. (b) Ground-truth. (c) CAM. (d) Extracted foreground regions from the trained classification network. (e) Increasing probability of foreground in segmentation network output indicated by color range from purple to red. (f) Extracted foreground regions from the trained segmentation network.

TABLE VI
ACCURACIES OF EXTRACTED FOREGROUND REGIONS FROM THE TRAINED CLASSIFICATION NETWORK AND THE TRAINED SEGMENTATION NETWORK

Dataset	Method	*IoU*	*Precision*	*Recall*	*F-score*
Potsdam	Classification network	0.563	0.908	0.598	0.721
	Segmentation network	0.806	0.872	0.914	0.893
Vaihingen	Classification network	0.315	0.952	0.320	0.479
	Segmentation network	0.721	0.834	0.842	0.838

B. Benefits of the Segmentation Network in Two-Step WSSS Framework

A classification network holds inherent abilities to perceive the discriminative regions in an image by various visualization methods [30]–[32] as described in Introduction, which has been directly adopted for extracting geographic information from remote sensing images. For example, CAM has been applied to WSOD from HR images [59], which does not need dense prediction. In addition, the discriminative regions obtained from classification network were also directly used for extracting geographic information that needs dense prediction [37], [38]. However, in the two-step WSSS framework, the discriminative regions are used as pseudo-masks for training a segmentation network and the final inferencing procedure (dense prediction) is fulfilled by the segmentation network rather than the classification network. In this subsection, we aim at clarifying the benefits of segmentation network in the two-step WSSS framework and thus verifying the suitability of the two-step WSSS framework for extracting geographic information from HR images.

To illustrate the benefits of segmentation network, the trained classification network and the trained segmentation network in our framework are respectively applied on testing images to produce CAM and segmentation result for comparison. The parameters of producing foreground region from CAM in testing images are set same as that for training segmentation network. Accordingly, the differences between the extracted foregrounds by the two networks can clearly reveal the benefits of the segmentation network. As shown in Table VI, the accuracy of extracted foreground by segmentation network is apparently improved compared with that by inferencing CAM directly, which indicates that the segmentation network could learn to explore more spatial context information during training by itself and thus result in higher accuracy. Examples of inferencing results from CAM and segmentation network are presented in Fig. 9, showing that both the completeness and the boundary accuracy by segmentation network are apparently improved compared with those by CAM.

C. Effectiveness of CRF-Loss and CRF Postprocessing

CRF-loss allows the segmentation network to be trained with spatial context information from images in addition to that from pseudo-masks. CRF postprocessing is applied to improve segmentation network outputs by further exploring spatial context. For the Potsdam dataset, as shown in Table VII , the segmentation network trained by CE-loss purely on pseudo-masks could achieves *IoU* 0.749, and the *IoU* is improved to 0.806 by the extra CRF-loss supervision and improved to 0.774 by CRF postprocessing. The accuracy improvements demonstrate the effectiveness of both CRF-loss and CRF postprocessing, but the different ranges of improvements reveal that CRF-loss and CRF postprocessing work differently. In addition, the CRF postprocessing could even improve the segmentation accuracy of the network trained by joint loss.

Several examples on the Potsdam dataset are presented in Fig. 10 to illustrate the different effectiveness of CRF-loss and CRF postprocessing. As shown in Fig. 10(c) and (d), due to

TABLE VII
ACCURACIES OF SEGMENTATION NETWORK OUTPUTS WITH DIFFERENT LOSS FUNCTIONS AND WITH/WITHOUT CRF POSTPROCESSING (CRF-POST) ON THE ISPRS POTSDAM AND VAIHINGEN DATASET

Dataset	Method	*IoU*	*Precision*	*Recall*	*F*-score
Potsdam	CE-loss	0.749	0.816	0.901	0.857
	CE-loss+CRF-post	0.774	0.832	0.917	0.873
	CE-loss+CRF-loss	0.806	0.872	0.914	0.893
	CE-loss+CRF-loss+CRF-post	0.810	0.876	0.916	0.895
Vaihingen	CE-loss	0.701	0.833	0.815	0.824
	CE-loss+CRF-post	0.721	0.845	0.832	0.838
	CE-loss+CRF-loss	0.721	0.834	0.842	0.838
	CE-loss+CRF-loss+CRF-post	0.725	0.836	0.845	0.840

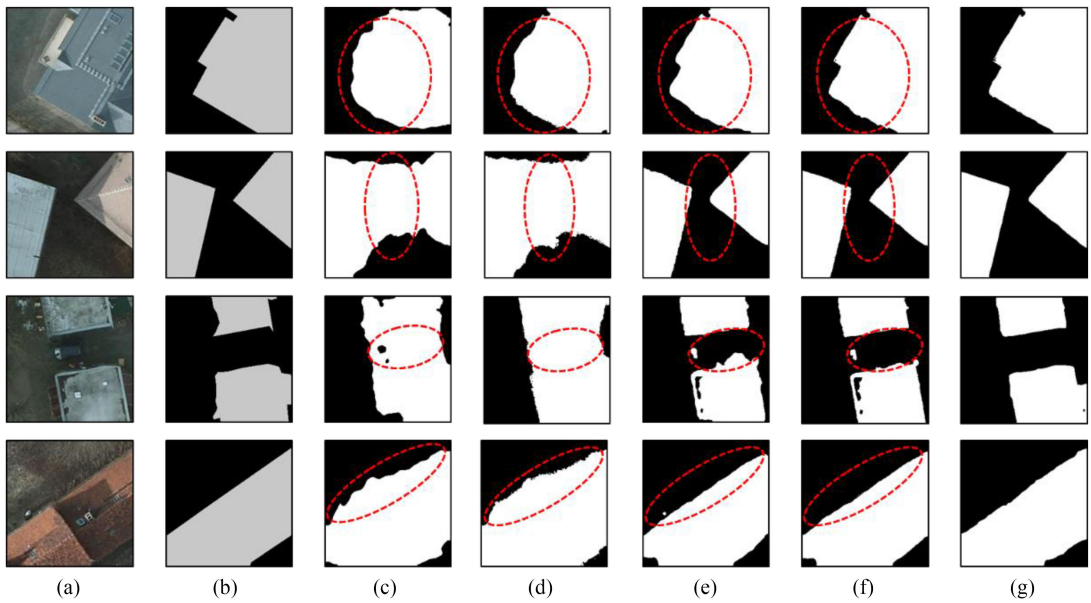


Fig. 10. Examples of segmentation network outputs trained by different loss functions and with/without postprocessing on the Potsdam dataset. "CE-loss" represents network trained with only CE-loss. "CRF" means postprocessing by CRF. "CRF-loss" represents network trained with the joint CE-loss and CRF-loss. "Fully" represents network trained with full supervision of pixel-level labels. (a) Image. (b) Ground-truth. (c) CE-loss. (d) CE-loss+CRF. (e) CRF-loss. (f) CRF-loss+CRF. (g) Fully.

TABLE VIII
COMPARISON OF THE ACCURACIES BETWEEN THE WEAKLY AND FULLY SUPERVISED METHOD

Dataset	Method	*IoU*	*Precision*	*Recall*	*F*-score
Potsdam	Weakly	0.810	0.876	0.916	0.895
	Fully	0.910	0.959	0.947	0.953
Vaihingen	Weakly	0.725	0.836	0.845	0.841
	Fully	0.882	0.946	0.929	0.937

the inevitable mistakes in pseudo-masks, the trained network may produce errors which are hardly corrected by CRF post-processing. However, the segmentation network trained with extra CRF-loss is capable of exploring the spatial context from images implicitly and thereby correcting such errors, as shown in Fig. 10(e). Beyond that, for the results from the network trained with extra CRF-loss, contextual information can be again used in the postprocessing by CRF to further improve the accuracies, through which several detailed mistakes can be corrected, as shown in Fig. 10(f).

As to the Vaihingen dataset, it is shown in Table VII that the results from combining CRF-loss and CRF postprocessing also achieve the best accuracy. However, the CRF-loss does not work as effectively as on the Potsdam dataset. This is because the segmentation network trained with CRF-loss can hardly distinguish the buildings that are densely distributed, as shown in Fig. 11(a). For the images with buildings that are not very crowded, the segmentation network trained with CRF-loss is able to output results with higher *precision* and boundary accuracy, as shown in Fig. 11(b).

D. Comparison With Fully Supervised Semantic Segmentation

The performance gap between weakly and fully supervised segmentations is a critical indicator for the effectiveness of WSSS. The accuracy of our WSSS method is compared with that of fully supervised method on the ISPRS Potsdam and Vaihingen dataset in Table VIII, where the fully supervised segmentation uses the same network DeepLabV3+ and CE-loss. It is shown that on the Potsdam (Vaihingen) dataset, the overall accuracy gap is only 0.1 (0.157) and 0.058 (0.096) in terms of *IoU* and *F*-score, respectively, which means that we can achieve 89.0% (82.2%) and 93.9% (89.8%) WSSS performance of fully supervised network indicated by *IoU* and *F*-score, respectively.

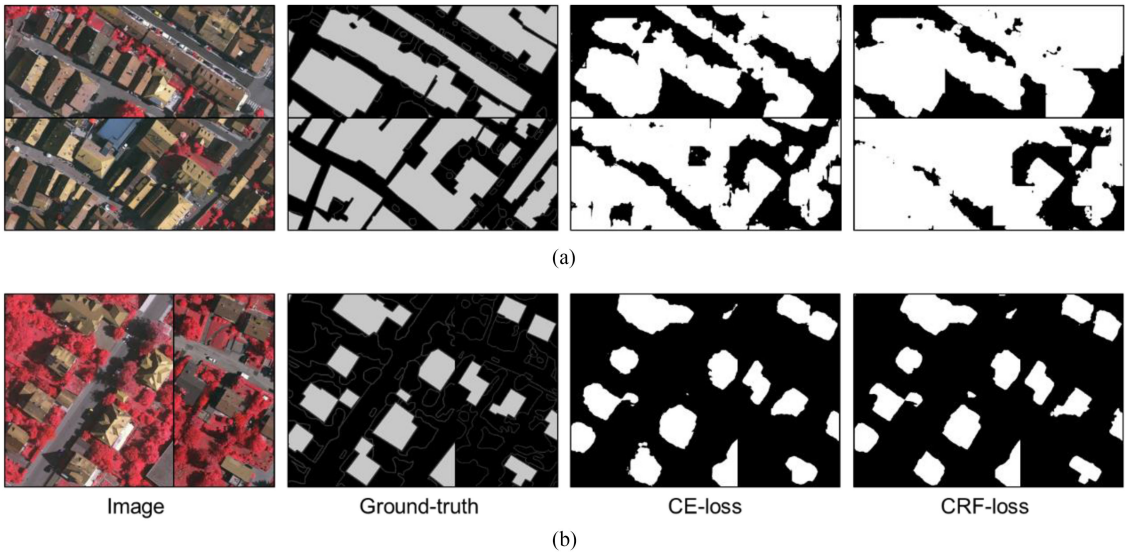


Fig. 11. Examples of segmentation network outputs trained by different loss functions on the Vaihingen dataset. "CE-loss" represents network trained with only CE-loss, and "CRF-loss" represents network trained with the joint CE-loss and CRF-loss. Segmentation network trained with CRF-loss performs unsatisfactorily on the images with too crowded buildings (a) and while works well otherwise (b).

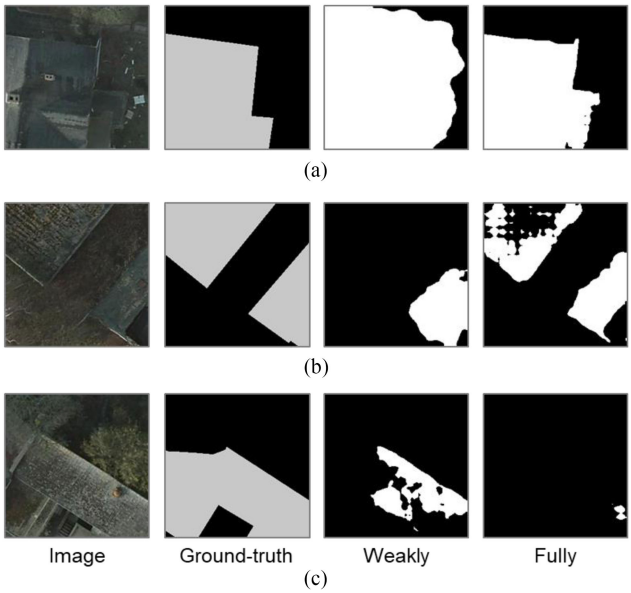


Fig. 12. Failure segmentation examples on the Potsdam dataset with comparison to fully supervised method.

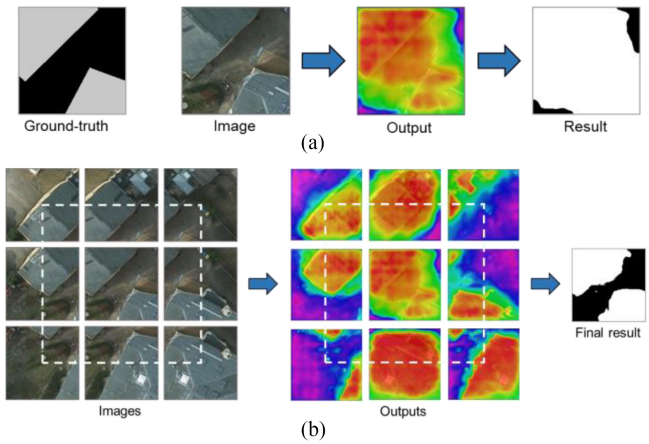


Fig. 13. Effectiveness of the overlapping strategy on correcting segmentation errors. The color range from purple to red in (b) indicates the increasing probability of foreground in segmentation network output. (a) Segmentation error without overlapping. (b) Correcting segmentation error by 1/2 overlapping degree.

As presented in Fig. 10(f) and (g), our WSSS method can produce segmentation results comparable to fully supervised network.

According to the gaps indicated by *precision* and *recall*, it shows that the difference mainly comes from the gap of *precision*, as shown in Fig. 12. The common failure case is shown in Fig. 12(a), where the segmentation network can hardly identify accurate boundaries of complex buildings and results in lower *precision* value compared with fully supervised network. Fig. 12(b) and (c) shows the cases that both the weakly and fully supervised networks perform poorly when dealing with complex buildings. A building may be missed and confused with background as shown in Fig. 12(b), while in a few cases, it can even perceive buildings that the fully supervised network cannot, as shown in Fig. 12(c).

E. Effects of the Overlapping Strategy for Inferencing

In this section, we conduct detailed analysis on the effects of the overlapping strategy during inferencing procedure, through which we can prevent context limitations and the quality of segmentation can be significantly improved for WSSS.

Usually, HR images are cropped into patches when inferencing due to their large size, which would split buildings near the edge of patches and thus lead to object incompleteness. It is easy to know that the overlapping strategy can eliminate the stitching seams from concatenating. Furthermore, the incomplete buildings in cropped patches would limit the spatial

TABLE IX
SEGMENTATION ACCURACIES WITH/WITHOUT USING OVERLAPPING STRATEGY FOR BOTH WEAKLY AND FULLY SUPERVISED NETWORK

Dataset	Supervision	Without overlapping				With overlapping			
		IoU	*Precision*	*Recall*	*F*-score	*IoU*	*Precision*	*Recall*	*F*-score
Potsdam	Weakly	0.810	0.876	0.916	0.896	0.843	0.901	0.930	0.915
	Fully	0.910	0.959	0.947	0.953	0.924	0.968	0.953	0.960
Vaihingen	Weakly	0.725	0.836	0.845	0.841	0.758	0.858	0.868	0.863
	Fully	0.882	0.946	0.929	0.937	0.896	0.958	0.933	0.945

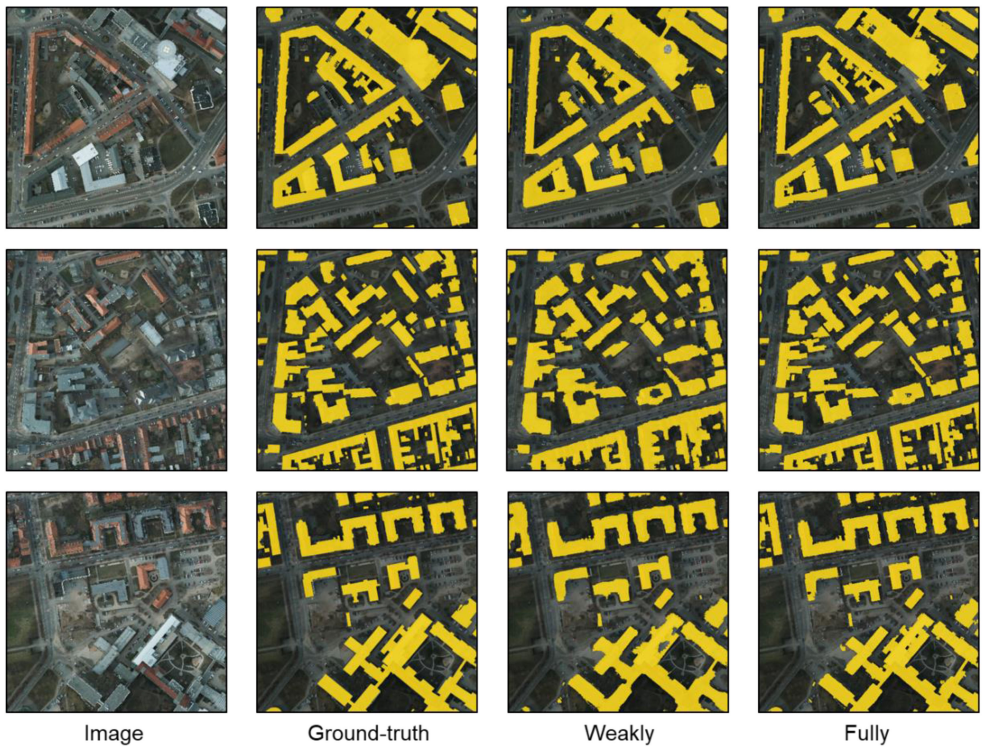


Fig. 14. Segmentation results of our weakly and fully supervised network on original Potsdam images.

context information to be used for inferencing, and thus, lead to mistakes, as can be seen in Fig. 13(a). The overlapping strategy helps to utilize spatial context information from neighboring patches and is thus capable of correcting such errors caused by incompleteness, as demonstrated by Fig. 13(b), where accurate buildings can be segmented with a larger field of view through the overlapping strategy. As shown in Table IX, the overlapping strategy is effective for both two datasets in terms of improving WSSS accuracy greatly.

As for fully supervised case, the segmentation accuracies are also improved by the overlapping strategy, as shown in Table IX. Besides, compared with the fully supervised case, the accuracy improvement range of the weakly supervised case is apparently larger. This is because the capacity of the weakly supervised network is not as strong as that of the fully supervised network, and thus suffers more from inadequate spatial context information. In this case, the performance gap between the weakly and fully supervised segmentation is further shortened thanks to the overlapping strategy. Two groups of final building segmentation results on two datasets are shown in Figs. 14 and 15.

F. Comparison With the State-of-the-Art Methods

Here, we compare the performance of our method with other state-of-the-art image-level WSSS methods. Considering the difficulty of method reproduction and the fairness of comparison, we choose two two-step image-level WSSS models, i.e., SEC [27] and DSRG [26], whose codes are available online.

Both SEC and DSRG generated foreground of pseudo-mask by applying a threshold to CAM [32], and generated background from the saliency map [31] by setting a customized threshold. Then, SEC trained a segmentation network based on the "SEC loss," and DSRG trained a segmentation network using the "deep seeded region growing" method.

The method as described in Section II-B1 is used to generate the pseudo-mask of SEC and DSRG in our comparison experiment. As to the threshold settings, two kinds of fg-thre are set, i.e., the optimum threshold as discussed in Section IV-A and the default threshold in SEC and DSRG where the fg-thre equals to 0.8. The results of the different segmentation networks using two kinds of pseudo-masks are shown in Table X. It is shown that the results of our segmentation network achieve the best accuracy on both datasets. Besides, for both SEC and DSRG, the results

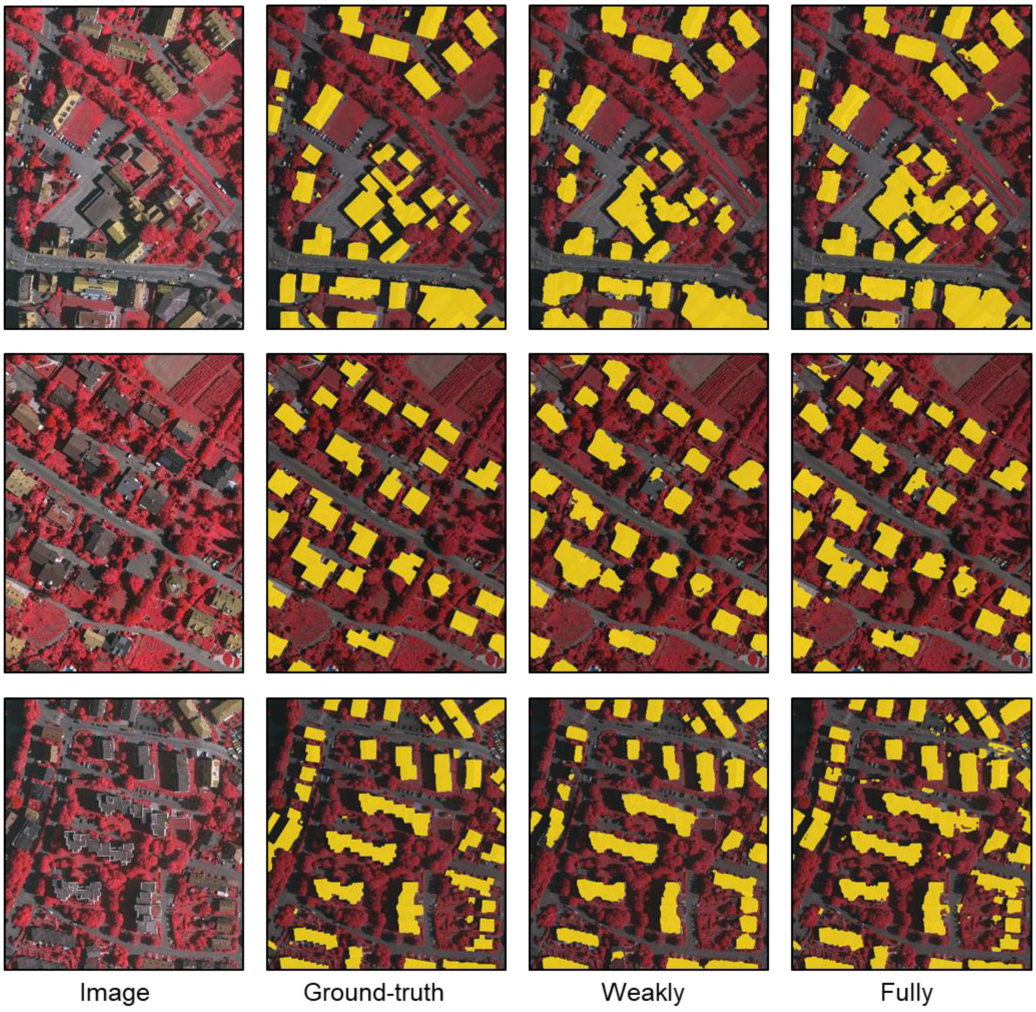


Fig. 15. Segmentation results of our weakly and fully supervised network on original Vaihingen images.

TABLE X
COMPARISON OF THE ACCURACIES AMONG DIFFERENT WSSS METHODS ON THE TWO DATASETS. "D" REFERS TO GENERATING PSEUDO-MASKS WITH THE DEFAULT FG-THRE (0.8) OF SEC [27] AND DSRG [26], AND "O" THE OPTIMUM THRESHOLDS AS DISCUSSED IN SECTION IV-A

Dataset	Method	IoU	Precision	Recall	F-score
Potsdam	SEC - d	0.605	0.946	0.626	0.753
	DSRG - d	0.695	0.896	0.756	0.820
	SEC - o	0.710	0.904	0.767	0.830
	DSRG - o	0.739	0.849	0.851	0.850
	Ours	0.810	0.876	0.916	0.896
Vaihingen	SEC - d	0.697	0.930	0.738	0.823
	DSRG - d	0.681	0.763	0.868	0.812
	SEC - o	0.705	0.923	0.751	0.828
	DSRG - o	0.698	0.815	0.832	0.823
	Ours	0.725	0.836	0.845	0.841

by using their default thresholds to generate pseudo-masks are always inferior to that of using the optimum thresholds, where the gap is particularly obvious on the Potsdam dataset (*IoU* gap 0.105 for SEC and 0.044 for DSRG). This gap further proves the necessity to generate suitable pseudo-masks.

Furthermore, we compute the training and inference time of the three segmentation networks. As shown in Table XI, the training and inference time of our segmentation network are both the lowest among the three methods.

TABLE XI
TIME COST COMPARISON FOR THE THREE METHODS. THE INPUT IMAGE SIZE DURING TRAINING AND INFERENCE ARE 256×256 AND 250×250, RESPECTIVELY

Method	Training time / image (s)	Inference time / image (s)
SEC	0.893	0.125
DSRG	3.816	0.220
Ours	0.101	0.030

V. CONCLUSION

In this study, we investigated how to achieve accurate building extraction by fully convolutional network from HR images with no requirement of pixel-level labeled data. The image-level two-step WSSS framework is adopted, in which pseudo-masks are first generated by image tags and then a segmentation network is trained by pseudo-masks. Through detailed analyzes, we illustrate several key impact factors in the two-step weakly supervised training procedure. Specifically, the influence of precision,

completeness, boundary accuracy, and background should be carefully considered for producing suitable pseudo-mask. The segmentation network is able to inherently explore spatial context information during training procedure, and thus, improve the performance of the two-step WSSS. We also demonstrate the effects of CRF-loss on improving semantic segmentation accuracy and its limitations. Combining the two-step training framework and deliberate settings, we can generate results with *F*-score 0.915 and 0.863 on the ISPRS Potsdam and Vaihingen datasets, respectively, which account for 95.3% and 91.3% performance of the fully supervised model. The findings in this study should be meaningful for promoting WSSS applications for extracting geographic information from HR images.

Although the image-level weakly supervised semantic segmentation results for HR images by our method achieve a small gap to fully supervised case, more in-depth studies are still needed to improve the application potential of WSSS for remote sensing. First, this study is performed based on very-high resolution remote sensing images with centimeter-level resolution. The potential of WSSS needs to be further validated with different types of remote sensing images in terms of both spatial and spectral resolutions. Second, we only focus on a single category of building with relatively regular geometric features. More land cover categories with abundant spectral and geometric features remain to be explored by WSSS. Third, it remains to be discussed in the future about the influence of both the quantity and quality of weak annotations on the performance of WSSS.

REFERENCES

[1] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[2] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[3] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

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

[5] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, pp. 307–316, May 2019.

[6] M. Wieland, Y. Li, and S. Martinis, "Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network," *Remote Sens. Environ.*, vol. 230, Sep. 2019, Art. no. 111203.

[7] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[8] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 223–236, May 2019.

[9] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[10] ISPRS. "International Society for Photogrammetry and Remote Sensing: 2D semantic labeling challenge," 2016. Accessed: Jul. 30, 2020. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[11] X. Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[12] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 1–9.

[13] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[14] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[15] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 156, pp. 1–13, Oct. 2019.

[16] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Comput. Vis.*, 2014, vol. 8693, pp. 740–755.

[17] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019.

[18] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[19] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, Apr. 2019.

[20] M. Schmitt, J. Prexl, P. Ebel, L. Liebel, and X. X. Zhu, "Weakly supervised semantic segmentation of satellite images for land cover mapping–Challenges and opportunities," 2020, *arXiv:2002.08254*.

[21] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[22] L. Chan, M. S. Hosseini, and K. N. Plataniotis, "A comprehensive analysis of weakly-supervised semantic segmentation in different image domains," *Int. J. Comput. Vis.*, Sep. 2020.

[23] S. Hong, S. Kwak, and B. Han, "Weakly supervised learning with deep convolutional neural networks for semantic segmentation: Understanding semantic layout of images with minimum human supervision," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 39–49, Nov. 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, Oct. 2015, pp. 234–241.

[25] 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," 2017, *arXiv:1606.00915*.

[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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7014–7023.

[27] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," 2016, *arXiv:1603.06098*.

[28] 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. 5262–5271.

[29] 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. 6488–6496.

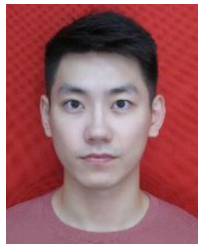
[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, vol. 86–89, 2014, pp. 818–833.

[31] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014, *arXiv:1312.6034*.

[32] 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.*, Jun. 2016, pp. 2921–2929.

[33] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8002–8012, Nov. 2020.

[34] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675–685, Jan. 2021.
[35] K. Fu *et al.*, "WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image," *Remote Sens.*, vol. 10, no. 12, Dec. 2018, Art. no. 1970.
[36] 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, Mar. 2020, Art. no. 1049.
[37] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sens.*, vol. 12, no. 2, Jan. 2020, Art. no. 207.
[38] M. U. Ali, W. Sultani, and M. Ali, "Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 115–124, Apr. 2020.
[39] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.
[40] P. Aravena Pelizari, K. Spröhnle, C. Geiß, E. Schoepfer, S. Plank, and H. Taubenböck, "Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements," *Remote Sens. Environ.*, vol. 209, pp. 793–807, May 2018.
[41] T. Blaschke *et al.*, "Geographic object-based image analysis – Towards a new paradigm," *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 180–191, Jan. 2014.
[42] G. Chen, Q. Weng, G. J. Hay, and Y. He, "Geographic object-based image analysis (GEOBIA): Emerging trends and future opportunities," *GISci. Remote Sens.*, vol. 55, no. 2, pp. 159–182, Mar. 2018.
[43] M. Kucharczyk, G. J. Hay, S. Ghaffarian, and C. H. Hugenholtz, "Geographic object-based image analysis: A primer and future directions," *Remote Sens.*, vol. 12, no. 12, Jun. 2020, Art. no. 2012.
[44] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020.
[45] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, May 2019.
[46] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.
[47] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[49] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.
[50] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An End-to-End weakly supervised semantic segmentation approach," *AAAI*, vol. 34, no. 7, pp. 12765–12772, Apr. 2020.
[51] J. Ahn and S. Kwak, "Learning Pixel-level semantic affinity with Image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.
[52] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated CRF loss for weakly supervised semantic image segmentation," 2019, *arXiv:1906.04651*.
[53] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *ECCV*, vol. 11220, 2018, pp. 524–540.
[54] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems* vol. 24, 2011, pp. 109–117.
[55] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.
[56] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vis.*, 2018, pp. 833–851.
[57] 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.
[58] C. Persello and L. Bruzzone, "A novel protocol for accuracy assessment in classification of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1232–1244, Mar. 2010.
[59] K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun, "MultiCAM: Multiple class activation mapping for aircraft recognition in remote sensing images," *Remote Sens.*, vol. 11, no. 5, Mar. 2019, Art. no. 544.

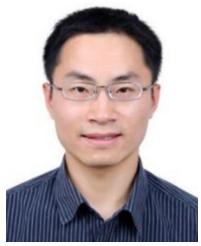


Zhenshi Li received the B.S. degree in geographic information science from Hohai University, Nanjing, China, in 2019. He is currently working toward the M.S. degree in cartography and geographical information system from Nanjing University, Nanjing, China.

His research interests include semantic segmentation and weakly supervised deep learning for remote sensing.

Xueliang Zhang received the B.S. degree in geographical information system and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, China, in 2010 and 2015.

From 2014 to 2015, he was a Visiting Student with Informatics Institute, University of Missouri, Columbia, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University, where he is currently an Associate Professor with the Department of Geographic Information Science. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.



Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002 and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. From 2011 to 2012, he was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Germany, and from 2014 to 2015, with the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA. He has authored four books and more than 60 articles. His current research interests include high.]resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.

Zixian Zheng received the B.S. degree in geographic information science from Sun Yat.]sen University, Guangzhou, China, in 2019. She is currently working toward the M.S. degree in cartography and geographical information system from Nanjing University, Nanjing, China.

Her research interests include semantic segmentation and deep learning for remote sensing.