Improved Pseudomasks Generation for Weakly Supervised Building Extraction From High-Resolution Remote Sensing Imagery

Fang Fang, *Member, IEEE*, Daoyuan Zheng, Shengwen Li, *Member, IEEE*, Yuanyuan Liu, *Member, IEEE*, Linyun Zeng, Jiahui Zhang, and Bo Wan

***Abstract*—Benefiting from free labeling pixel-level samples, weakly supervised semantic segmentation (WSSS) is making progress in automatically extracting building from high-resolution (HR) remote sensing (RS) imagery. For WSSS methods, generating high-quality pseudomasks is crucial for accurate building extraction. To improve the performance of generating pseudomasks by using image-level labels, this article proposes a weakly supervised building extraction method by combining adversarial climbing and gated convolution. The proposed method optimizes class activation maps (CAMs) by using adversarial climbing strategy, generates accurate class boundary maps by introducing a gated convolution module, and further refines building pseudomasks by fusing pairing semantic affinities and CAMs with a random walk strategy. Experimental results on three datasets—two ISPRS datasets and a self-annotated dataset—demonstrate that the proposed approach outperformed SOTA WSSS methods, leading to improvement of building extraction from HR RS imager. This article provides a new approach for optimizing pseudomasks generation, and a methodological reference for the applications of weakly supervised on RS images.**

***Index Terms*—Adversarial climbing (AC), building extraction, gated convolution, high-resolution (HR) remote sensing (RS) imagery, weakly supervised semantic segmentation (WSSS).**

I. Introduction

Automatic building extraction plays an important role in urban planning [1], building change detection [2], [3], geographic data mapping, and updating [4]. With the increasing amount of high-resolution (HR) remote sensing (RS) imagery becoming an important and reliable data source, efficiently and accurately extracting buildings from RS images is significant and challenging.

Building extraction from RS imagery has been driven by advancing deep learning technology in recent years. Buildings can be extracted by classifying pixels of RS images as building or nonbuildings, which is regarded as a semantic segmentation task in computer vision [5]. Usually, semantic segmentation methods first train a model by using dense pixel-level samples through an end-to-end mechanism, and then classify each pixel of the unlabeled image by using the trained model. Represented by fully convolutional networks (FCNs) [6], an increasing body of supervised approaches have been developed to advance building extraction tasks, achieving significant performance improvements. These methods usually require a large number of labeled pixel-level samples to learn model parameters, especially given the great variation of buildings in HR imagery across regions. Although RS images are easily collected, labeling a large number of pixel-level samples is labor-intensive and time-consuming. In addition, due to the high complexity and diversity of building distribution scenes in RS images, greater difficulties are involved in efficiently obtaining high-quality pixel-level labels [7]. In general, developing new methods to effectively extract buildings from HR RS images by utilizing available labeled datasets or easily obtained labels is urgently needed [8].

Inspired by weakly supervised learning that constructs predictive models by learning with weak supervision [9], weakly supervised semantic segmentation (WSSS) methods are promoted to alleviate the issue of lacking pixel-level labels. They utilize easily obtained labels as week supervision to train models, such as bounding boxes, scribbles, points, and image-level labels. The image-level class label indicates the category of the object presented in an image, which is one of the easiest types to acquire among these weak labels. Nowadays, a two-step WSSS framework using image-level labels is widely used, which includes a classification network and a segmentation network. The framework generates pixel-level pseudomasks from given

Manuscript received November 15, 2021; revised December 30, 2021; accepted January 12, 2022. Date of publication January 19, 2022; date of current version February 11, 2022. This work was supported in part by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources under Grant KF-2021-06-088 and in part by the Wuhan Applied Fundamental Frontier Project under Grant 2020010601012166. *(Corresponding author: Bo Wan.)*

Fang Fang is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China, with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034, China, and also with the National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430074, China (e-mail: fangfang@cug.edu.cn).

Daoyuan Zheng is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China, and also with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034, China.

Shengwen Li, Yuanyuan Liu, and Bo Wan are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China, and also with the National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430074, China (e-mail: swli@cug.edu.cn; liuyy@cug.edu.cn; wanbo@cug.edu.cn).

Linyun Zeng is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: ly.zeng@cug.edu.cn).

Jiahui Zhang is with the National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430074, China (e-mail: zhangjiahui@cug.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3144176

image-level labels in its first step. Usually, the step obtains a class activation map (CAM) [10] from the classification network trained with image labels and generates pseudomasks of objects by expanding an initial CAM seed. In the second step of the framework, a segmentation network model is trained by taking the pseudomasks and the corresponding images as input, then predicting final masks of objects by using the trained networks [11].

Several WSSS methods have been developed to extract buildings from RS imagery and achieved promising results in the past two years. For example, SPMF-Net [12] boosted building segmentation by combining superpixel pooling and fused multiscale features. The approach generated CAMs that retain the shape and boundary information in superpixel and contributed to the accurate extraction of buildings. Li *et al.* [13] adopted conditional random field (CRF) to optimize CAMs obtained by a classification network and introduced CRF loss [14] and CRF postprocessing in the segmentation network for accurate building extraction. For WSSS methods, generating high-quality pseudomasks is crucial for accurate building extraction. There are two gaps in generating pseudomasks. First, CAMs focus on the most discriminative parts, thus is prone to partial activation of buildings regions, resulting in incomplete pseudomasks. In addition, previous models are not optimized for refining building boundaries in pseudomasks, failing to produce accurate buildings boundaries. There is still potential for improving the performance of building extraction.

To this end, this article proposes a WSSS building extraction method that integrates an adversarial climbing (AC) mechanism and gated convolution module (GCM) to accurately extract buildings from HR RS imagery. The proposed method utilizes an AC strategy to optimize CAMs, designs a GCM to generate class boundary maps (CBMs), and refines pseudomasks by fusing pairing semantic affinities and CAMs with a random walk strategy, contributing to enhancing the performance of building extraction from HR RS imagery. The original contributions of this article are as follows.

1) This article develops a new building pseudomasks generation framework based on image-level labels, which advances weakly supervised building extraction from HR RS imagery.
2) An antiadversarial attack mechanism named AC is introduced to the pseudomask's generation process, which generates CAMs through iterative manipulation, helping CAMs better cover buildings of various shapes.
3) The article introduces CBMs to present the latent boundaries of buildings and designs a GCM to generate accurate CBMs by keeping the low-level boundary-relevant features, enhancing the ability to extract building boundaries.
4) Experimental results on three building extraction datasets—two ISPRS standard datasets and a self-annotation dataset—show that the proposed method outperforms state-of-the-art weakly supervised methods.

The rest of this article is organized as follows. Section II reviews related works. Section III introduces the proposed framework, which is a weakly supervised building extraction method by combining AC and gated convolution. Section IV describes the details of experimental settings and results. Section V discusses factors that affect model performance, including network structure and some hyperparameters. Finally, Section VI concludes this article.

II. Related Works

Related works can be classified into the following three categories: building extraction based on handcrafted features, building extraction based on deep leaning, and WSSS methods on RS imagery. They will be discussed in the following three sections.

A. Building Extraction Based on Handcrafted Features

Traditionally, a large number of methods for building extraction from RS images are based on handcrafted features, such as color, spectrum, contextual, shadow, and geometry information [15]–[18]. For example, Sirmacek *et al.* [15] proposed an approach for building detection using multiple cues, which includes invariant color features and shadow information. Zhang [16] developed a texture filtering technique for detecting urban buildings. Awrangjeb *et al.* [17] proposed a novel method to extract buildings in HR RS images based on shadow detection. These features vary under different circumstances of light, atmospheric conditions, sensor quality, scale, surroundings, and building architectures. The performances of those methods are highly dependent on the selected features, which require strong domain-specific knowledge. Especially, the design of empirical features is varied with datasets sourced from different geographical regions or sensors. Developing data-adaptive automatic methods for building extraction from RS images is valuable and challenging.

B. Building Extraction Based on Deep Learning

Recently, the handcrafted features in an increasing body of applications are gradually replaced by data-driven deep learning technology, such as convolutional neural networks (CNN). Deep learning-based studies have achieved excellent performance in extracting building from RS images. Building extraction from HR RS imagery can be treated as a pixel-wise classification task of deep learning, initially known as semantic segmentation in the computer vision field. Semantic segmentation methods extract buildings by assigning each pixel to a class label that indicates whether the pixel is a building or not.

Most image semantic segmentation methods based on deep learning are derived from FCNs [6], a pioneer in pixel-wise classification, and are continuously developing as a result of new emerging deep learning models. For example, DeConvNet [19], an encoder–decoder architecture, was developed with deconvolution and unpooling layers based on VGG [20]. U-Net [21], an end-to-end deep FCN, introduced skip connections between downsampling and upsampling layers. Different variants of U-Net have achieved superior performance in different image segmentation tasks. SegNet [22] is an encoder–decoder neural network for semantic segmentation, in which the decoder introduced pooling indices to perform nonlinear upsampling. DeepLab and its variants [23]–[25] use dilated convolution to address the decreasing resolution caused by downsampling. Recently, a spatial pyramid pooling strategy was introduced to capture image context to segment objects accurately at multiple scales. The pyramid scene parsing network [26] utilizes global contextual information through a pyramid pooling module.

These semantic segmentation models, which were initially developed for extracting basic objects from classic images, have also achieved significant performance when applied to RS image-related tasks. Building extraction from RS has advanced with numerous variants of these semantic segmentation models. For example, Siamese U-Net [27] tried improving segmentation accuracies with shared weights. BRRNet [28] was developed by combining a prediction module with a residual refinement module, improving the accuracy of building extraction. Manual characteristics and the guided filtering technique were applied to optimize building extraction with a novel ResUNet network [29]. MTPA-Net [30] is a scene-driven multitask parallel attention convolutional network for resolving the semantic gaps caused by the large intraclass variance among different kinds of buildings. Two UNet-based models, called MCG-UNet and BCL-UNet, are proposed for road and building segmentation from aerial imagery [31]. MFCNN [32] is a multifeature CNN for pixel-level segmentation of buildings and utilizes morphological filtering to regularize building boundaries. A novel FCN was proposed for accurately extracting buildings, in which a boundary learning task was embedded to help maintain the boundaries of buildings [33]. E-D-Net [34] was derived for building segmentation from visible aerial images, which preserves the edge information of the images and achieves prediction with higher detail quality. GRRNet [35] fuses HR aerial images and LiDAR point clouds for building extraction, which introduces a gated residual mechanism including a gated feature labeling unit to reduce unnecessary feature information. BR-Net [36] consists of a shared backend utilizing a modified U-Net and a multitask framework to generate superior building outlines by addressing restrictions and regulations of additional boundary information. SENet [37] integrated three individual segmentation models to obtain fine-scale spatial and spectral building information. However, these methods follow a supervised machine learning paradigm. Their model parameters need to be trained using a large number of samples with pixel-level labels.

C. WSSS Methods on RS Imagery

Several studies have documented the effectiveness of weakly supervised deep learning methods on semantic segmentation tasks. Several weak annotations, including bounding boxes [38]–[40], scribble [41], [42], and points [43], have been employed in recent WSSS models. Image-level labels are the most widely used type of weak annotation mainly because various image classification tasks have built a large number of image-level samples and accumulated many pretrained image classification models. The WSSS models usually generate pixel-level pseudomasks from weak annotations. Several studies [44]–[50] focused on generating better CAMs and improving the quality of seed area. Some research [51]–[53] aimed to expand the initial CAM seed to identify more regions of the target object. For example, AffinityNet [52] and IRNet [53] employed pairwise pixel semantic affinities to optimize the pseudomasks with CAMs.

Nowadays, WSSS models have been derived for extracting geographical entities from HR RS imagery. For example, U-CAM [54], a new image classification network combined with CAMs, was created for extracting cropland by considering the distribution of image-level labels. In addition, a masked U-Net was also proposed to obtain segmentation from sparse pixel labels. A global convolutional pooling operation and a local pooling pruning strategy were introduced into a WSSS framework to address cloud detection [55]. Hierarchical residual saliency maps combined with superpixel were generated to fulfill residential-area segmentation with a novel hierarchical weakly supervised learning method [56]. A weakly supervised feature-fusion network was proposed to accomplish water and cloud segmentation in RS images [57]. Adversarial learning and self-training strategies were combined to develop the framework for building segmentation in unsupervised domain adaptation [58].

For building extraction, several studies have been conducted with image-level labels. For example, SPMF-Net [12] generated pseudomasks by combining superpixel pooling and multiscale feature fusion with image-level labels. In SPMF-Net, a superpixel pooling layer was added to the classification network to improve the integrity and boundary accuracy of a detected building. A two-step training strategy approach was derived by Li *et al.* [13], in which the fully connected CRF was utilized to explore the spatial context in both training and prediction stages.

Although those methods have made significant improvements in building extraction, their performances are highly dependent on the quality of pseudomasks [13]. Due to the CAMs focusing on the most discriminative parts, the generated pseudomasks have a big gap with the ground-truth (GT) of buildings. It is promising to improve the accuracy of building extraction by optimizing the quality of the pseudomasks. To optimize, the generated pseudomasks is expected to serve for a series of applications on RS images, and is the research motivation of this article.

III. Method

The article presents a weakly supervised building extraction framework that combines AC and gated convolution (ACGC). The ACGC framework takes image-level labels as weak supervision, aiming to improve pseudomask generation for accurately extracting buildings from HR RS images. As shown in Fig. 1, the whole pipeline contains three key components. In the first component, the CAMs of buildings are generated and optimized with AC. Then, the CBMs of buildings are obtained with a GCM in the second component, and the pseudomasks of buildings are expanded by CAMs and pairwise affinities from CBMs in the third component. The three components will be discussed in the following three sections, respectively.

With the generated pseudomasks, a classic supervised segmentation model can be trained by taking pseudomasks as the

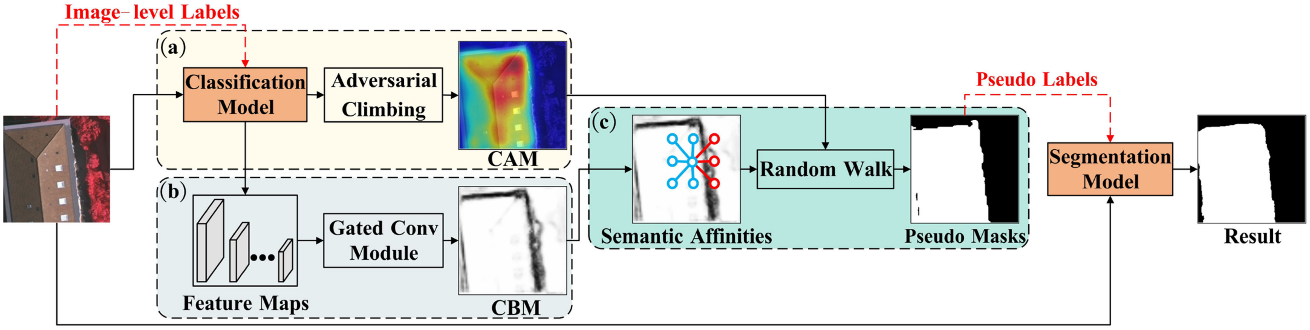

Fig. 1. Overview of the proposed framework for building extraction based on WSSS. (a) Optimizing CAMs with AC. (b) Generating CBMs based on gated convolution. (c) Producing pseudomasks with pairwise semantic affinities.

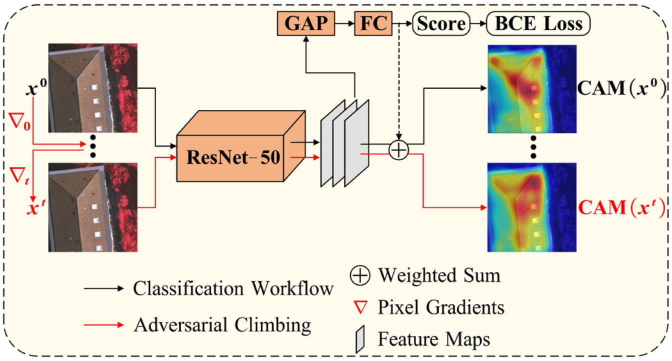

Fig. 2. Two workflows of CAMs optimization with AC.

GT of training samples. The model is expected to learn semantic information of HR images to identify buildings. Specifically, the learned supervised segmentation model can be fed with HR RS images and output the building masks that present extracted buildings.

A. *Optimizing CAMs With AC*

The classic WSSS framework generates CAMs mainly through a trained classification network, which suffers from the inaccuracy of obtained CAMs [10]. To improve the integrity of the generated CAMs, this component introduces an antiadversarial technique named AC [49] into the proposed framework to optimize generated CAMs.

The component that generates CAMs with AC is composed of two workflows, as shown in Fig. 2. The first workflow is classification workflow, which generates initial CAMs based on image-level labels. The second workflow is AC workflow, which iteratively optimizes CAMs.

In the first workflow, a typical classification network is employed to generate the initial CAMs with image-level labels. The network consists of a ResNet [59] backbone followed by a global average pooling (GAP) layer and a fully connected layer. The training task is formulated as a binary classification problem and employs the binary cross-entropy as its loss function as follows:

$$-\frac{1}{n}\sum\left(y_B \log\sigma(\hat{y}_B) + (1-y_B)\log(1-\sigma(\hat{y}_B))\right) \quad (1)$$

where y_B denotes the class label, $\hat{y}_B$ denotes the classification score, and σ represents the sigmoid function. The CAMs are computed from the weighted sum of the feature maps of the last convolutional layer of the classification network. The value of each pixel in the CAM is normalized so that the maximum activation probability of each pixel is up to 1. Specifically, CAM(x) of an image x can be calculated as follows:

$$CAM(x) = \frac{\mathbf{W}^{\mathbf{T}} f(x)}{\max(\mathbf{W}^{\mathbf{T}} f(x))} \quad (2)$$

where $f(x)$ denotes the feature map of image x before the GAP layer, and $\mathbf{W}$ denotes the weight matrix of the fully connected layer of the classification network.

In the second workflow, AC is introduced to perturb HR RS images along pixel gradients, thus increasing the classification score of the building [60]. AC is iteratively executed *NI* times to generate manipulated images that are fed into the trained classifier to obtain CAMs, where *NI* is a hyperparameter. Given the initial image x^0, the AC process can be presented by

$$\begin{aligned}
x^t &= x^{t-1} + \varepsilon\nabla_{x^{t-1}} Y \\
Y &= \hat{y}_B^{t-1} - \lambda\left\|R_B^{t-1} \odot H_B^{t-1}\right\| \\
R_B^{t-1} &= \begin{cases} 1, & \text{if } \mathrm{CAM}(x^{t-1}) \geq \theta \\ 0, & \text{if } \mathrm{CAM}(x^{t-1}) < \theta \end{cases} \\
H_B^{t-1} &= \left|\mathrm{CAM}(x^{t-1}) - \mathrm{CAM}(x^0)\right| \quad (3)
\end{aligned}$$

where t is the number of iterations ranging from 1 to *NI*, ε is the size of the adversarial perturbation, and x^t denotes the manipulated image by changing the value of each pixel using gradient updates. $\nabla_{x^{t-1}}\ Y$ is the gradient of Y with respect to x^{t-1} computed using backpropagation. Y is the regularization formula of the classification score, $\mathrm{R}_{\mathrm{B}}^{t-1}$ is a restricting mask computed by thresholding $\mathrm{CAM}(x^{t-1})$ with θ, H_{B}^{t-1} is the differences of CAMs between x^{t-1} and x^0. $\odot$ is the element-wise product, $\|\cdot\|$ is the sum operation, and $\hat{\mathrm{y}}_{\mathrm{B}}^{t-1}$ is the classification score of x^{t-1} of buildings. λ is a hyperparameter of masking regularization. By regularization, scores of high-activated building areas remain unchanged, whereas scores of low-activated building areas can be iteratively increased. The nondiscriminative building regions in the manipulated images are gradually optimized. Thus, the CAMs can identify more regions of buildings. The optimized

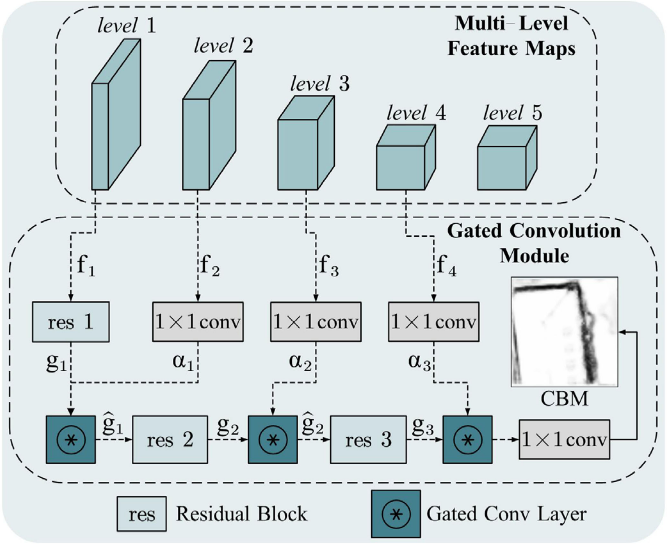

Fig. 3. Generating CBMs based on GCM. Pixels with higher values (blacker) in CBMs represent likely class boundaries.

CAMs, $\mathrm{M_B}$, can be normalized by aggregating the CAMs obtained from the manipulated image *x* at each iteration *t* as follows:

$$M_B = \frac{\sum_{t=0}^{NI} \mathrm{CAM}(x^t)}{\max \sum_{t=0}^{NI} \mathrm{CAM}(x^t)}. \tag{4}$$

B. Generating CBMs Based on Gated Convolution

Although the CAMs generated in the first component have been optimized by AC, the building boundaries have not been accurately refined. In this section, the proposed method further detects class boundaries to obtain pairwise semantic affinities [52] as additional information to generate pseudomasks. Here, a GCM is designed to generate CBMs. CBMs will present potential class boundaries between pixels of the confident building category and that of the nonbuilding category. The GCM will filter out the higher-level semantic information that may be inappropriate for class boundary detection and identify building boundaries with low-level features of HR images [61].

As shown in Fig. 3, feature maps of the four levels of the backbone network are fed to the GCM as the input. A set of residual blocks and gated convolution layers (GCLs) are employed to detect class boundaries. Specifically, f_n and g_n denote the nth-level feature maps of the backbone and the output of the nth residual block, respectively, where $n \in 1, 2, 3$. Then, an attention map α_n is calculated as the input of GCL

$$\alpha_n = \sigma(C_{1\times1}(f_{n+1} \parallel g_n)) \tag{5}$$

where σ is the sigmoid function, $\mathrm{C}_{1\times1}$ denotes the 1 × 1 convolution operation, and ‖ denotes the concatenation operation.

Afterward, the attention map α_n and feature maps f_n conduct element-wise product ⊙, following a residual connection, and channel-wise weighting with kernel w_n. GCL can be calculated as follows:

$$\hat{g}_n^{(i,j)} = (g_n * w_n)_{(i,j)} = ((g_{n_{(i,j)}} \odot \alpha_{n_{(i,j)}}) + g_{n_{(i,j)}})^T w_n \tag{6}$$

where $*$ denotes GCL operation, and $\hat{g}_n^{(i,j)}$ is the output of GCL. The kernel w_n with a size of 1 × 1 is applied for each pixel (i, j). g_{n+1} is derived by feeding $\hat{g}_n^{(i,j)}$ to the residual block as follows:

$$g_{n+1} = \mathrm{res}(\hat{g}_n^{(i,j)}). \tag{7}$$

g_{n+1} is fed into the next GCL layer in the GCM. The first residual block takes the feature map f_1 as input, and outputs g_1, and the last 1 × 1 convolution produces a CBM. When two pixels are separated by class boundaries in CBM, they are considered a pair with a low semantic affinity.

For a pair of pixels x_i and x_j, the semantic affinity a_{ij} between them is defined as follows:

$$a_{ij} = 1 - \max_{k \in \prod_{ij}} B(x_k) \tag{8}$$

where $B \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}}$ denotes the CBM, and $\prod_{\mathrm{ij}}$ represents a collection of pixels on the line between x_i and x_j. The maximum distance of a pair should be less than the radius γ. That is, the method ignores pairs whose distance is greater than γ. The GT binary edges are not given, which is why this article takes semantic affinity labels of pairwise pixels defined in IRNet [53] as the supervision for training our GCM. The semantic affinity labels derived from the confident region in CAMs have been proved to be effective in learning and generating class boundaries.

C. Refining Pseudomask With Pairwise Semantic Affinity

In this section, the pseudomasks of buildings are refined based on CAMs and CBMs generated in the two previous subsections. With the trained GCM, the semantic affinity matrix of each pairwise pixel can be computed from the predicted CBMs according to (8). Then, the component performs random walk propagation [62] on the CAMs with a transition probability matrix derived from the semantic affinity matrix. Specifically, the transition probability matrix T, in which diagonal elements are set to 1, is obtained as follows:

$$T = S^{-1} A^{\circ\beta}, \text{ where } S_{ii} = \sum_j a_{ij}^{\beta}, A = [a_{ij}] \in \mathbb{R}^{wh\times wh} \tag{9}$$

where A denotes the semantic affinity matrix, and $A^{\circ\beta}$ is A to the Hadamard power. The diagonal matrix S is computed for row-wise normalization of $A^{\circ\beta}$. The semantic propagation conducted by random walk is performed with T by

$$\mathrm{vec}(M_B^*) = T^p \cdot \mathrm{vec}(M_B \odot (1 - B)) \tag{10}$$

where ⊙ is the element-wise product, $\mathrm{vec}(\,\cdot\,)$ means vectorization of a matrix, and p denotes the number of iterations of random walk. Thereby, the activation value of each pixel is spread to neighboring regions with the same semantic information according to dense semantic affinities.

Finally, this component produces pseudomasks by assigning "Building" to a pixel with an activation score above the segmentation threshold ST and "Nonbuilding" to other pixels.

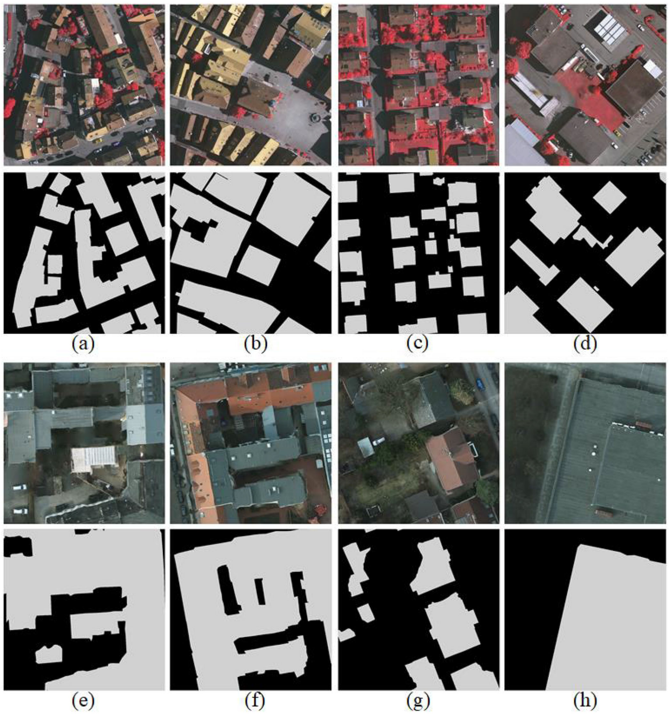

Fig. 4. Samples of the ISPRS datasets and corresponding GT masks. (a)–(d) Vaihingen. (e)–(f) Potsdam.

IV. Experiments and Results

A. Data

1) ISPRS Potsdam and Vaihingen Dataset: The Potsdam and Vaihingen datasets, which are two public semantic labeling contest datasets provided by the ISPRS II/4 committee, are used to evaluate the proposed method. The two datasets consist of three bands of TIFF files and their corresponding digital surface model data. The Potsdam dataset contains 38 raw large aerial images, with 24 images in its training set and 14 images in its test set. All the images are 6000 × 6000 pixels at 5 cm spatial resolution. The Vaihingen dataset comprises 33 raw large image patches of different sizes with a ground resolution of 9 cm/pixel, with 16 images in its training set and 17 images in its test set. The two datasets can be downloaded from the ISPRS official website [63].

In this article, raw images with three channels of red, green, and blue were selected from the Potsdam dataset, and images with three channels of near-infrared, red, and green were chosen from the Vaihingen dataset. For the Potsdam dataset, the image named “top_potsdam_7_10_RGB” that has an error in its annotations [64] was removed from its test dataset. Fig. 4 shows some samples and corresponding building GT masks.

Following the data preprocessing in previous works [13], the raw images in the training dataset are cropped into 256 × 256 patches with a sliding step size of 128. The image-level labels in this article are similarly determined by the pixel ratio of buildings in an image patch. Specifically, a patch is labeled as building when its pixel ratio is greater than 0.25 and is labeled as nonbuilding when the image does not contain any building pixels. The remaining unlabeled images with a pixel ratio between 0 and 0.25 are removed. The raw images in the test dataset are also cropped into 256 × 256 patches for adopting overlapping strategy when inferencing. Finally, the processed Potsdam dataset contains 38 281 image patches from 23 raw images for training, with 18 243 building patches and 20 038 nonbuilding patches, and 32 256 image patches from 14 raw images for testing. The processed Vaihingen dataset contains 2588 image patches from 16 raw images for training, with 2092 building patches and 496 nonbuilding patches, and 5530 image patches from 17 raw images for testing.

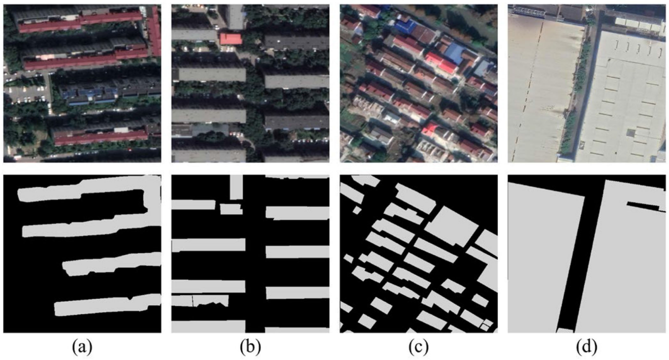

Fig. 5. Samples of the self-annotated building segmentation dataset and corresponding GT masks.

2) Self-Annotated Building Segmentation Dataset: The self-annotated building segmentation dataset is a manually labeled pixel-level HR RS imagery dataset. Its raw images are sampled from Google Maps. This dataset contains 7260 patches located in four cities (Beijing, Wuhan, Shanghai, and Shenzhen) in China. Each patch has a resolution of 500 × 500 pixels with a ground resolution of 0.29 m.

The processed self-annotated dataset contains 2352 training patches, with 2009 building patches, 343 nonbuilding patches, and 1275 testing patches. Four samples and corresponding GT masks are shown in Fig. 5.

B. Experimental Settings

The training procedure of this article consists of three steps: training a classification network, training GCM, and training a segmentation network. It is performed in the following manner: image labels are first employed to train a classification network for producing CAMs. The semantic affinity labels from CAMs are then utilized to train GCM, which generates CBMs for refining pseudomasks of all images in the training dataset. Finally, the pseudomasks are used to tune a segmentation network, a DeepLabv3+ model pretrained on ImageNet. The training of the segmentation network is supervised by cross-entropy loss on both foreground and background pixels in the pseudomask.

For training the classification network, the backbone of the classification network in the first component is a pretrained ResNet-50 [59] model, in which the stride of the last downsampling layer is adjusted to 2. This article uses the stochastic gradient descent optimizer with a momentum of 0.9 and weight decay of 1×10^{-4} for training classification networks. The initial learning rate is set to 0.001 and the learning rate decay strategy of "poly" is adopted. The batch size is set to 16. The randomly horizontal flipping and multiscale training were performed for 40 epochs. For training GCM, the GCM from the ResNet-50 backbone network was trained for ten epochs. In this step, the initial learning rate was set to 0.01, the batch size was set to 32, and the backbone was frozen during GCM training. The radius γ was set to 5 for training and 3 for testing. For training the segmentation network, the value of batch size was set to 8, and the number of classes was set to 2. In its inference process, the testing images were evaluated by an overlapping strategy [13].

The three key components of ACGC were implemented with PyTorch, and the segmentation network of ACGC was implemented with PaddleSeg [65]. All the experiments were conducted on a 2080 super GPU.

C. Evaluation Metrics

Considering the building extraction from HR RS images used evaluation metrics in semantic segmentation, tasks were employed for examining model performances, including the intersection over union (IoU), precision, recall, and F1-score, which are formulated as follows:

$$\begin{aligned}\mathrm{IoU} &= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})\\ \mathrm{Precision} &= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})\\ \mathrm{Recall} &= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\\ \mathrm{F1} &= 2\times(\mathrm{Precision}\times\mathrm{Recall})/(\mathrm{Precision}+\mathrm{Recall})\end{aligned} \tag{11}$$

where TP means the true positive rates, FP means the false positive rates, and FN means the false negative rates. Precision represents the percentage of TP in the ground truth, recall indicates the percentage of TP in the segmentation result, the F1-score is the weighted average of precision and recall, and IoU is the average value of the intersection of the prediction and ground truth over their union.

D. Baselines

In this article, the proposed method was compared with three WSSS models: CAM [10], IRNet [53], and AdvCAM [49]. IRNet is designed for weakly supervised instance segmentation with image-level labels. IRNet consists of two branches: one is to predict displacement field and generates instance-wise CAMs, and the other is to detect CBMs and predict pairwise semantic affinities between pixels. AdvCAM utilizes adversarial image manipulation to increase classification scores and produce CAMs, which helps identify regions occupied by objects more accurately. In [53], IRNet was employed on the CAMs to obtain pseudomasks. For the three models, their best pseudomask results in several experiments are chosen to train their segmentation network [48].

Moreover, a widely used fully supervised segmentation model (Fully) called DeepLabv3+ [25] is used to examine the effectiveness of the proposed method. DeepLabv3+ employs GT masks instead of pseudomasks to train the model, which means that it does not indicate that the model is superior even if it has higher accuracy than ACGC or the WSSS baseline models.

TABLE I
PSEUDOMASK RESULTS ON THE POTSDAM AND VAIHINGEN DATASET

Dataset	Mothod	IoU		F1	
		BP	NBP	BP	NBP
Potsdam	CAM [10]	0.783	0.636	0.879	0.777
	IRNet [53]	0.818	0.726	0.900	0.841
	AdvCAM [49]	0.830	0.736	0.907	0.847
	ACGC	**0.831**	**0.744**	**0.908**	**0.853**
Vaihingen	CAM [10]	0.655	0.741	0.791	0.851
	IRNet [53]	0.771	0.808	0.871	0.894
	AdvCAM [49]	0.762	0.814	0.865	0.897
	ACGC	**0.775**	**0.822**	**0.873**	**0.902**
Self-annotated	CAM [10]	0.388	0.560	0.560	0.718
	IRNet [53]	0.412	0.595	0.584	0.746
	AdvCAM [49]	0.405	0.612	0.576	0.759
	ACGC	**0.431**	**0.612**	**0.603**	**0.759**

The metrics used in the table are the accuracies of pseudomasks on training patches. BP presents building regions in pseudomasks and NBP presents nonbuilding regions in pseudomasks. The bold values denote the best results.

E. Pseudomask Results

Three group experiments were conducted on three datasets to examine the performances of generating pseudomasks. As shown in Table I, ACGC achieves the best performance among the four WSSS methods in terms of the IoU values and F1 values of the building (BP) and nonbuilding (NBP) pixels. Compared to pseudomask from the CAM method, the BP IoU increases by 0.048, 0.120, and 0.043 and the NBP IoU rises by 0.138, 0.081, and 0.052 in these three datasets. Furthermore, the results from two state-of-the-art methods, i.e., IRNet and AdvCAM, are also inferior to those of our method. The results suggest that introducing an antiadversarial attack mechanism and GCM helps to improve the generation of pseudomasks and further obtain better building extraction results. Since the segmentation model is trained with the pseudomasks, the improvement of pseudomasks will help to extract buildings more accurately.

F. Building Extraction Results

1) Results of ISPRS Potsdam and Vaihingen Dataset: Table II lists the comparison building extraction results on the Potsdam dataset and Vaihingen dataset. As shown in Table II, the proposed ACGC framework achieves the best building extraction performance among the four WSSS methods. Specifically, it obtains IoU and F1-score values of 0.784 and 0.879 on the Potsdam dataset and 0.845 and 0.916 on the Vaihingen dataset, respectively, making a significant improvement compared with the three baseline models. Compared with AdvCAM, ACGC produces output with lower accuracy in terms of precision and recall on the two datasets. The proposed method still achieves the best tradeoff in terms of IoU and F1-score. Therefore, ACGC

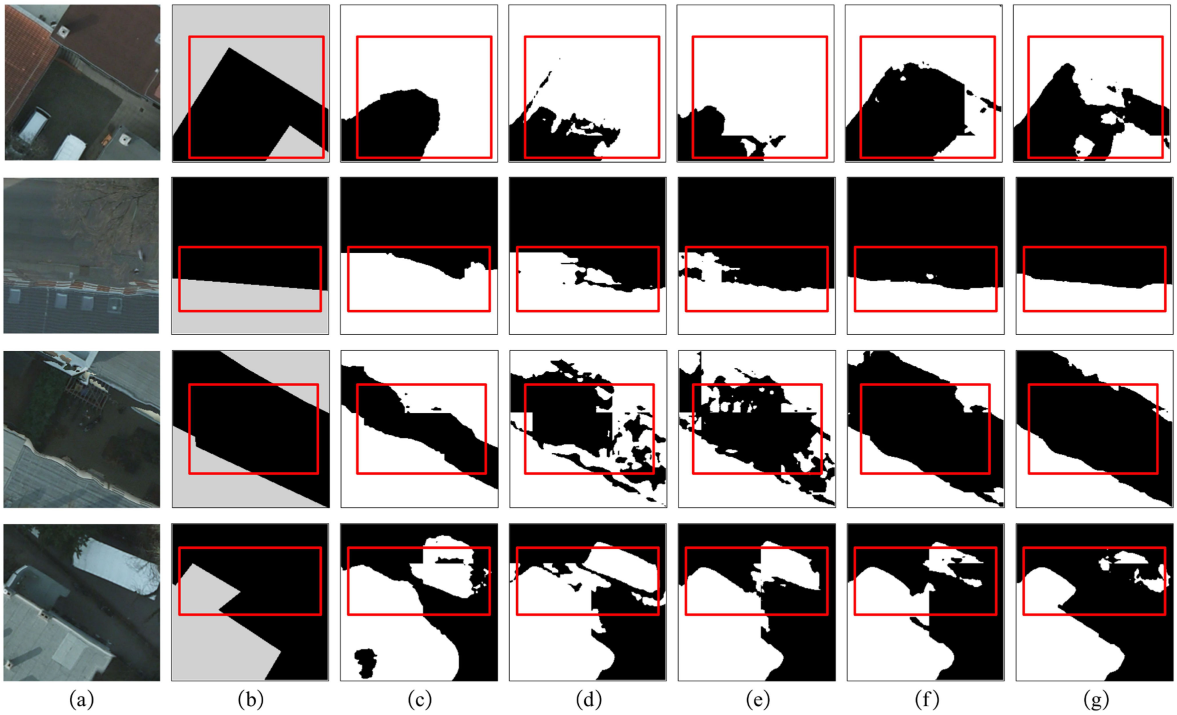

Fig. 6. Visualized results on the Potsdam dataset. (a) Original RS images. (b) GT. (c) CAM. (d) IRNet. (e) AdvCAM. (f) ACGC. (g) Fully.

TABLE II
Building Extraction Results on the Potsdam and Vaihingen Dataset

Dataset	Method	IoU	Precision	Recall	F1
Potsdam	CAM [10]	0.764	0.828	0.908	0.866
	IRNet [53]	0.827	0.908	0.903	0.905
	AdvCAM [49]	0.832	**0.931**	0.887	0.908
	ACGC	**0.845**	0.920	**0.912**	**0.916**
Vaihingen	CAM [10]	0.656	0.838	0.751	0.792
	IRNet [53]	0.782	0.872	**0.884**	0.878
	AdvCAM [49]	0.781	0.896	0.859	0.877
	ACGC	**0.784**	**0.928**	0.834	**0.879**

The presented segmentation accuracies are calculated on testing images with an overlapping evaluation. The optimal adversarial iterations on the two datasets are set as 2 and 3, respectively. The bold values denote the best results.

TABLE III
Building Extraction Results on the Self-Annotated Building Segmentation Dataset

Dataset	Method	IoU	Precision	Recall	F1
Self-annotated	CAM [10]	0.383	0.520	0.593	0.554
	IRNet [53]	0.386	0.513	0.609	0.557
	AdvCAM [49]	0.436	0.546	0.684	0.607
	ACGC	**0.457**	**0.558**	**0.717**	**0.627**

The presented segmentation accuracies are calculated on testing images with an overlapping evaluation. The optimal adversarial iterations on the dataset are set as 3. The bold values denote the best results.

shows better building extraction performance, which is mainly attributed to the quality of the pseudomasks.

To further observe model performances, the results of some samples in the Potsdam dataset are illustrated in Fig. 6. In some areas with narrow streets, the proposed method extracted more BP and misclassified fewer NBP, which can be seen in the first and third rows of the figure. In the complexity scene, even in the presence of similar building spectral with other features, the proposed method generates better boundaries than the baseline models, which can be seen in the second and fourth rows of the figure.

Some samples in the Vaihingen dataset are selected to further illustrate the results of the building extraction, as shown in Fig. 7. Compared with other WSSS models, ACGC captures more accurate boundaries and outputs fewer incorrect BP, leading to building boundaries closer to GTs. The results can be explained by the fact that ACGC captures boundary features during pseudomask generation and further improves the segmentation results.

Moreover, the proposed method also shows great performance for dense buildings. Fig. 8 shows the segmentation results of two local dense building areas. The proposed method can better extract BP from the background in those areas.

2) Results of Self-Annotated Building Segmentation Dataset: A series of experiments is conducted on a self-annotated building segmentation dataset to further examine the performance of ACGC. The segmentation model trained by the high-quality pseudomasks presents the best advantages in extracting buildings, as shown in Table III. Although the metric values of all models on this dataset are lower than those of the above datasets, ACGC achieves a significant improvement of 0.071 and 0.021 in terms of building IoU compared with the

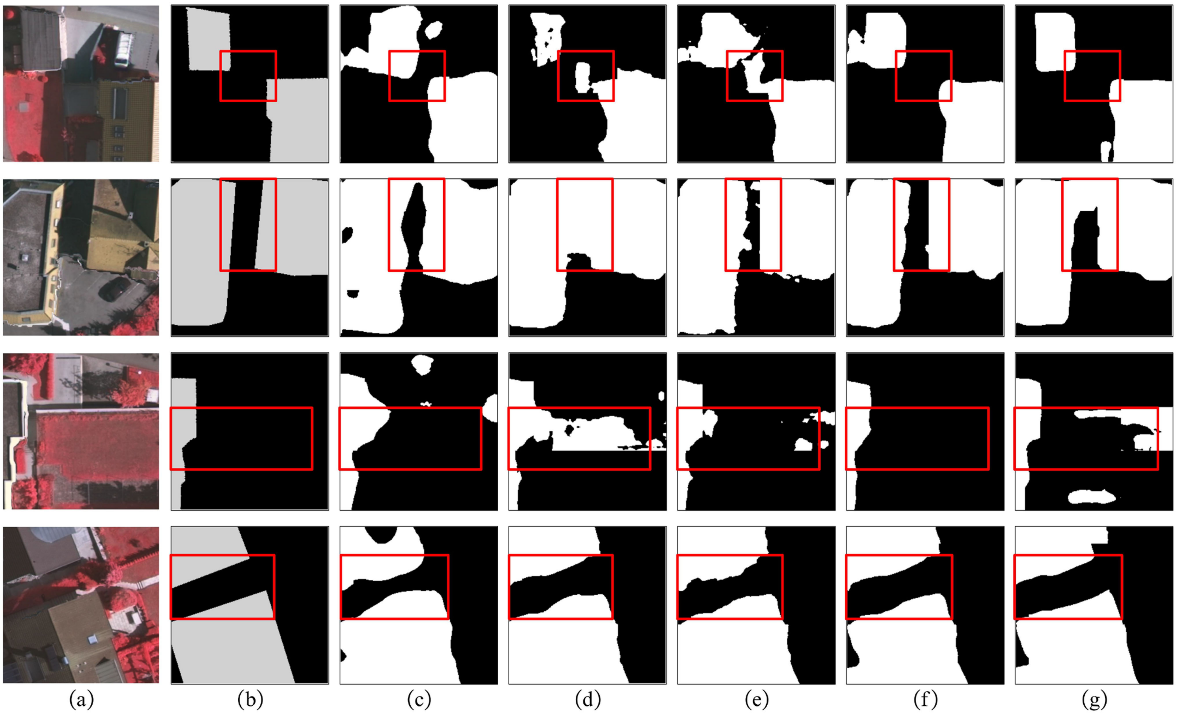

Fig. 7. Visualized results on the Vaihingen dataset. (a) Original RS images. (b) GT. (c) CAM. (d) IRNet. (e) AdvCAM. (f) ACGC. (g) Fully.

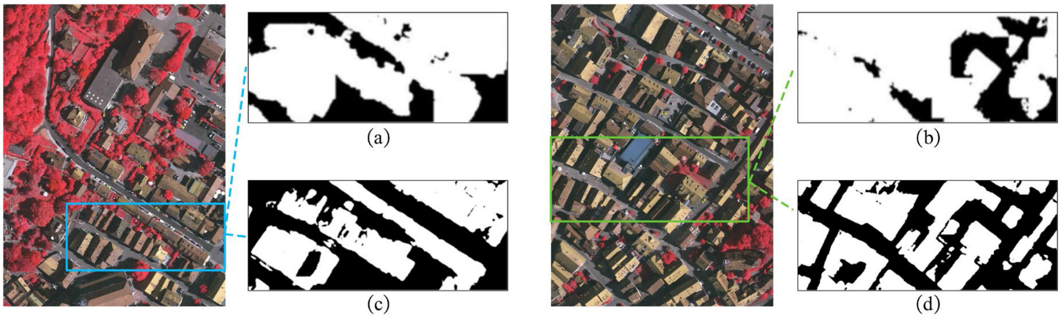

Fig. 8. Building extraction results in local dense building areas of two samples in the Vaihingen dataset. (a) and (b) Outputs of [13]. (c) and (d) Outputs of the proposed method.

IRNet and AdvCAM. This finding suggests that the proposed framework is effective.

Fig. 9 visualizes some comparison results on the self-annotated building segmentation dataset. For some challenging scenes, the shadow interference (as shown in the second row), the regular building shapes (as shown in the first and fourth rows), and irregular building shapes (as shown in the third row), the proposed method can reserve more integral building regions and clearer building boundaries. These results suggest that ACGC is promising.

3) Results Compared With Fully Supervised Semantic Segmentation: The segmentation results from the fully supervised method are also shown in our experiments. ACGC can outperform the supervised network in some cases, for example, the samples in the third row of Fig. 7. As shown in Table IV, the proposed method can achieve comparable results to the fully supervised method (Fully). In the table, ACGC achieves 93.4% and 96.4% performance of the fully supervised method in terms of IoU and F1-score on the Postsdam dataset and achieves 89.6% and 94.2% performance on the Vaihingen dataset. ACGC improves segmentation accuracy and narrows the performance gap between weakly and fully supervised building extraction models.

V. Discussion

A. Ablation Study

In this section, ablation experiments are conducted on the Vaihingen dataset to further evaluate the effectiveness of key components in ACGC. Specifically, we examine model performances after removing the AC and GCM. The w/o GCM method removes GCM as well as the random walk policy that depends on

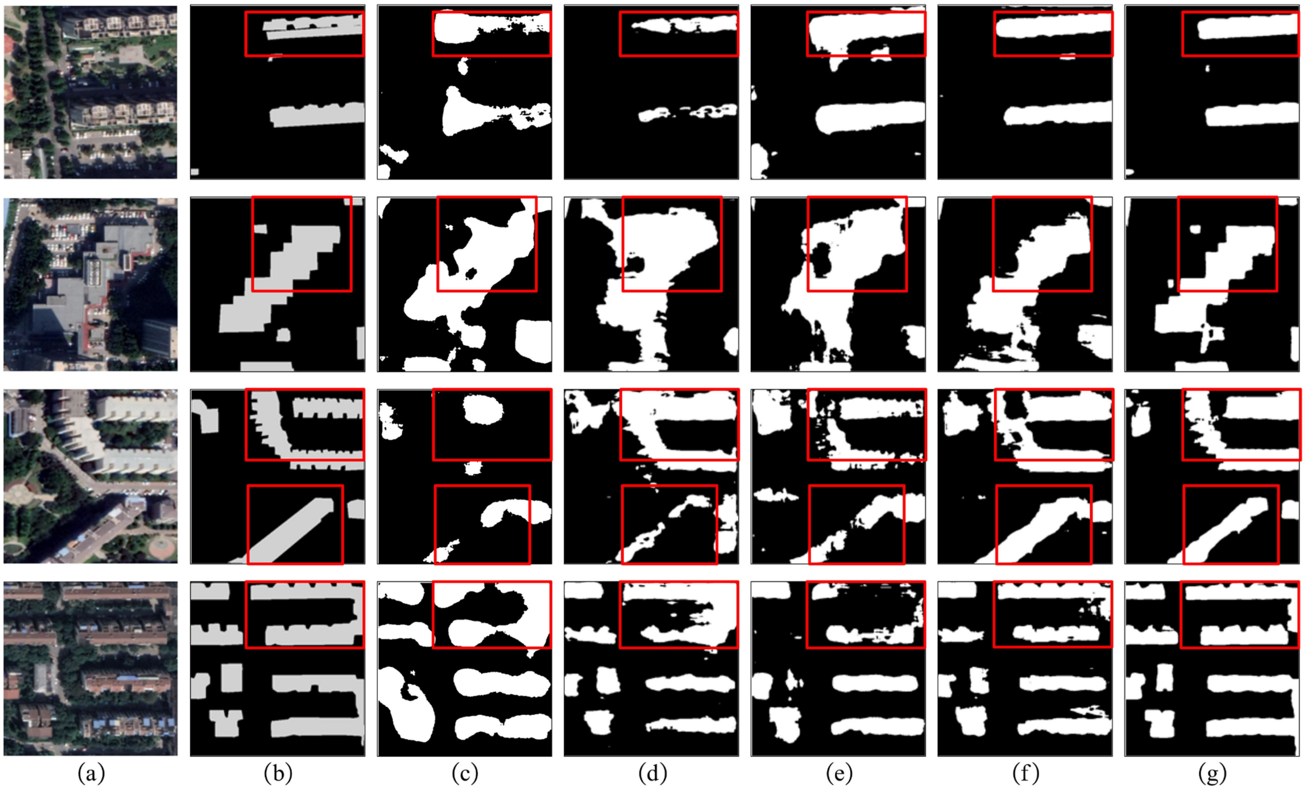

Fig. 9. Visualized results of building extraction on the self-annotated building segmentation dataset. (a) Original RS images. (b) GT. (c) CAM. (d) IRNet. (e) AdvCAM. (f) ACGC. (g) Fully.

TABLE IV
BUILDING EXTRACTION RESULTS BETWEEN WSSS AND FULLY WSSS ON THE POTSDAM, VAIHINGEN, AND SELF-ANNOTATED BUILDING SEGMENTATION DATASETS

Dataset	Method	IoU	Precision	Recall	F1
Potsdam	ACGC	0.845	0.920	0.912	0.916
	Fully	0.905	0.940	0.960	0.950
Vaihingen	ACGC	0.784	0.928	0.834	0.879
	Fully	0.875	0.953	0.914	0.933
Self-annotated	ACGC	0.457	0.558	0.717	0.627
	Fully	0.743	0.849	0.857	0.852

GCM. To clearly elaborate the findings, the values of evaluation metrics for both the building pixels (BP) and nonbuilding pixels (NBP) are recorded and listed in Table IV.

As shown in Table V, the full framework (ACGC) exhibits the best performance, achieving the best tradeoff according to IoU and F1-score on pseudomasks. ACGC has a lower NBP precision and a lower BP recall compared with w/o Adv. We argue that AC is employed to generate more integral CAMs and consequently tend to a higher threshold to reach the best IoU of pseudomasks, resulting in fewer pixels being labeled as building. In addition, ACGC employs the GCM and learns pairwise semantic affinities to further expand the building regions. After the GCM is removed, all the performance metrics in Table V are reduced. This result suggests that the GCM helps correct some errors and learns reliable affinities from confident regions in CAMs, thus better expanding the building regions.

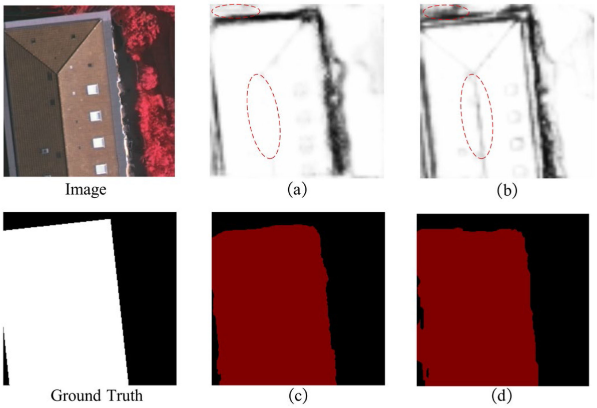

Fig. 10. Comparison results of the CBMs and pseudomasks generated by different methods. (a) and (b) CBMs generated by ACGC and IRNet. (c) and (d) Pseudomasks obtained by ACGC and IRNet.

B. Improvement of CBM From GCL

The GCL is introduced to optimize CBMs and thus improves the quality of the generated pseudomasks. To observe the improvement of CBM from introducing GCL, the results of a sample are visualized in Fig. 10.

TABLE V
ABLATION RESULTS WITH DIFFERENT COMPONENTS INTEGRATION IN ACGC ON THE VAIHINGEN DATASET

Model	IoU		Precision		Recall		F1	
	BP	NBP	BP	NBP	BP	NBP	BP	NBP
CAM	0.655	0.741	0.856	0.808	0.736	0.900	0.791	0.851
w/o Adv	0.772	0.812	0.873	**0. 895**	**0.870**	0.898	0.872	0.896
w/o GCM	0.723	0.768	0.814	0.822	0.866	0.922	0.838	0.869
Ours	**0.775**	**0.822**	**0.898**	0.884	0.850	**0.922**	**0.873**	**0.902**

The metrics used in the table are the accuracies of pseudomasks on training patches. BP presents building regions in pseudomasks and NBP presents nonbuilding regions in pseudomasks. The bold values denote the best results.

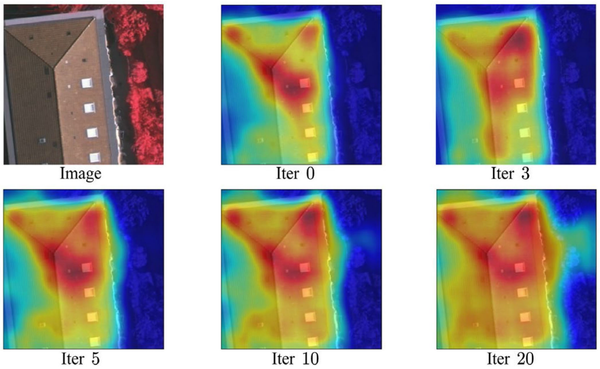

Fig. 11. Effect of adversarial iterations for CAMs in the Vaihingen dataset.

As shown in Fig. 10, the GCL contributes to detecting more precise class boundaries. In the figure, the closer the class boundaries in CBMs are to the boundaries of buildings, the more accurate the pseudomasks are. This result can be explained by the fact that the inappropriate information for class boundary identification is gradually filtered out with the help of the gating mechanism, and better CBMs are generated.

C. Effect of the Number of AC

Once the number of iterations reaches *NI*, the iterative process of the AC process will be finished. To examine the effect of *NI* values, experiments are conducted on the Vaihingen dataset in this section.

As shown in Fig. 11, when *NI* is set to 3, the CAMs fit building regions best. This finding suggests that an excessive number of iterations will raise the problem of overactivation. In addition, more nonbuilding regions are wrongly activated if *NI* is larger than 3. The *NI* value in this article is smaller than that in the classical application [49], in which *NI* is set to 27. We argue that the reason for this result is that building extraction is the process of identifying a single category of pixels, namely the building category pixels. ACGC extracts fewer categories than in the PASCAL VOC 2012 dataset task and should employ few adversarial iterations.

D. Improved Pseudomasks on Various Quality CAMs

ACGC generates optimized pseudomasks based on CAM, which is why its performance may vary somewhat when various-quality CAMs are inputted. This section will examine the effect of CAMs quality on the generated pseudomasks.

With the given CAMs, we conducted experiments by taking AdvCAM as a baseline and examined their BP and NBP IoU of pseudomasks. Three groups of CAMs, each with IoUs of 0.714, 0.677, and 0.653, respectively, are selected for our experiments.

Table VI shows that the IoU of CAMs and the IoU of pseudomasks are positively correlated. The results are consistent with our knowledge that better CAMs contribute to better quality pseudomasks in terms of BP and NBP.

Moreover, the experimental results show that a noisier CAM (smaller IoU values) corresponds to a greater advantage of our model over AdvCAM. This result occurred because the GCLs in the GCM retain the boundary features and simultaneously filter out irrelevant and incorrect features learned by poor confident semantic labels. Our GCM generates better CBMs and further corrects wrongly classified pixels by random walk propagation. Consequently, ACGC obtained better IoU of pseudomasks when CAMs have worse quality.

E. Experiments With More Segmentation Model

To further examine the quality of the pseudomasks generated by the proposed method, three SOTA segmentation networks are selected for the following experiments on the Vaihingen dataset. These segmentation networks are trained by feeding the pseudomasks from ACGC and AdvCAM, respectively. Their IoU, precision, recall, and F1 on the test dataset are reported in Table VII. As shown in the table, the three networks using the pseudomask generated by ACGC have higher IoU, precision, recall, and F1 compared to those using the pseudomask generated by AdvCAM. The results show that the improved pseudomasks generated by ACGC are robust, which can be combined with a series of semantic segmentation methods for improving the performance of weakly supervised building extraction.

F. Experiment on Model Efficiency

To examine the efficiency of ACGC, the comparison experiments were conducted with two WSSS baseline methods on the Vaihingen dataset. Both training time and inference time of pseudomasks are reported in Table VIII. As listed in this

TABLE VI
INFLUENCE OF QUALITY OF CAM ON THE PSEUDOMASKS BETWEEN ADVCAM AND ACGC ON THE VAIHINGEN DATASET

CAM	Method	IoU of CAM	IoU of Pseudo-masks		Gap	
			BP	NBP	BP	NBP
Group 1	AdvCAM [49]	0.714	0.773	0.818	0.002	0.004
	ACGC		**0.775**	**0.822**		
Group 2	AdvCAM [49]	0.677	0.721	0.758	0.012	0.014
	ACGC		**0.733**	**0.772**		
Group 3	AdvCAM [49]	0.653	0.699	0.739	0.015	0.017
	ACGC		**0.714**	**0.756**		

The metrics used in the table are the accuracies on training patches. BP represents building regions of pseudomasks and NBP represents nonbuilding regions of pseudomasks. The bold values denote the best results.

TABLE VII
COMPARISON RESULTS ON THE VAIHINGEN DATASET OF THREE DIFFERENT SEGMENTATION NETWORKS USING PSEUDOMASKS FROM ADVCAM AND ACGC, RESPECTIVELY

Segmentation Network	Method	IoU	Precision	Recall	F1
UNet [21]	AdvCAM [49]	0.723	0.901	0.785	0.839
	ACGC	0.758	0.912	0.818	0.863
PSPNet [26]	AdvCAM [49]	0.769	0.902	0.839	0.869
	ACGC	0.780	0.901	0.853	0.876
HRNet [66]	AdvCAM [49]	0.774	0.900	0.847	0.872
	ACGC	0.792	0.921	0.849	0.884

TABLE VIII
TIME COST COMPARISON WITH TWO BASELINE METHODS

Method	Training time/img(ms)	Inference time/img(s)
IRNet [53]	14.727	0.175
AdvCAM [49]	14.727	1.235
ACGC	15.631	1.249

The pseudomasks inference time is the sum of time from AD, GCM, and random walk.

table, the training time cost of ACGC is slightly higher than that of the two baseline methods, and the inference time of ACGC is approximately the same as that of AdvCAM. The experimental results show that the introduction of GCM in ACGC does not significantly increase time consumption. We argue that the proposed ACGC offers a better tradeoff between accuracy and efficiency.

VI. CONCLUSION

This article proposed an approach to improve weak supervision-based building extraction from HR RS imagery. The approach uses antiadversarial manipulation to generate more integral CAMs as building seed regions and optimizes the boundaries in pseudomasks by introducing a GCM. Experimental results on three datasets show that the proposed method outperforms the state-of-the-art methods and achieves higher IoU and clearer boundaries. This article provides a new method for generating better pseudomasks of buildings and offers a methodological reference for the applications of weakly supervised RS images.

The main limitation of this article is that only one type of annotation was utilized to generate pseudomasks. The new model integrating more kinds of weakly supervised annotations is expected to enhance the proposed model. Moreover, the approach of expanding CAMs can be explored in greater detail in the future work.

ACKNOWLEDGMENT

The authors would like to thank Dr. Zhuo Zheng, who is studying in the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China, and the editors and reviewers, who provided valuable suggestions that refined this article effectively.

REFERENCES

[1] G. Sohn and I. Dowman, “Data fusion of high-resolution satellite imagery and lidar data for automatic building extraction,” *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 1, pp. 43–63, 2007.

[2] L. Khelifi and M. Mignotte, “Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis,” *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[3] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, “Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 8000605.

[4] A. E. Maxwell *et al.*, “Semantic segmentation deep learning for extracting surface mine extents from historic topographic maps,” *Remote Sens.*, vol. 12, no. 24, 2020, Art. no. 4145.

[5] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using deep learning: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3059968 Feb. 2021.

[6] 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.

[7] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, “Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 20–33, 2021.

[8] 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.*, vol. 129, no. 2, pp. 361–384, 2021.

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

[10] 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.

[11] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, “Causal intervention for weakly-supervised semantic segmentation,” in *Proc. 34th Conf. Neural Inf. Process. Syst.*, 2020, pp. 655–666.

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

[13] Z. Li, X. Zhang, P. Xiao, and Z. Zheng, "On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3266–3281, 2021.

[14] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated CRF loss for weakly supervised semantic image segmentation," 2019, *arXiv:1906.04651*.

[15] B. Sirmacek and C. Unsalan, "Building detection from aerial images using invariant color features and shadow information," in *Proc. 23rd Int. Symp. Comput. Inf. Sci.*, Dec. 2008, pp. 1–5.

[16] Y. Zhang, "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering," *ISPRS J. Photogramm. Remote Sens.*, vol. 54, no. 1, pp. 50–60, 1999.

[17] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Improved building detection using texture information," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 143–148, 2011.

[18] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image," *J. Multimedia*, vol. 9, no. 1, pp. 181–188, 2014.

[19] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Feb. 2015, pp. 1520–1528.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

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

[22] 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.

[23] L. 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. 2017.

[24] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[25] L. 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.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[27] 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. 2018.

[28] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 1050.

[29] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high-resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 144.

[30] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2020.

[31] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Multi-object segmentation in complex urban scenes from high-resolution remote sensing data," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3710.

[32] Y. Xie *et al.*, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, 2020.

[33] S. He and W. Jiang, "Boundary-assisted learning for building extraction from optical remote sensing imagery," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 760.

[34] Y. Zhu, Z. Liang, J. Yan, G. Chen, and X. Wang, "E-D-Net: Automatic building extraction from high-resolution aerial images with boundary information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4595–4606, 2021.

[35] 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. Photogramm. Remote Sens.*, vol. 151, pp. 91–105, 2019.

[36] G. Wu *et al.*, "A boundary regulated network for accurate roof segmentation and outline extraction," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1195.

[37] D. Cao, H. Xing, M. S. Wong, M.-P. Kwan, H. Xing, and Y. Meng, "A stacking ensemble deep learning model for building extraction from remote sensing images," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3898.

[38] 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.*, Dec. 2015, pp. 1635–1643.

[39] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3136–3145.

[40] J. Lee, J. Yi, C. Shin, and S. Yoon, "BBAM: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2643–2652.

[41] 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.*, Dec. 2016, pp. 3159–3167.

[42] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2017, pp. 7158–7166.

[43] 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.

[44] 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.

[45] 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.*, Dec. 2017, pp. 618–626.

[46] 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.*, Jun. 2018, pp. 7268–7277.

[47] 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.*, Jun. 2019, pp. 5267–5276.

[48] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2020, pp. 12275–12284.

[49] J. Lee, E. Kim, and S. Yoon, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nov. 2021, pp. 4071–4080.

[50] G. Sun, W. Wang, J. Dai, and L. Van Gool, "Mining cross-image semantics for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 347–365.

[51] 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.*, Jun. 2018, pp. 7014–7023.

[52] 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.*, Dec. 2018, pp. 4981–4990.

[53] 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.*, Jan. 2019, pp. 2209–2218.

[54] 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, 2020, Art. no. 207.

[55] 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, 2020, Art. no. 112045.

[56] L. Zhang, J. Ma, X. Lv, and D. Chen, "Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 117–121, Jan. 2019.

[57] 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.

[58] X. Yao, Y. Wang, Y. Wu, and Z. Liang, "Weakly-supervised domain adaptation with adversarial entropy for building segmentation in cross-domain aerial imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8407–8418, 2021.

[59] 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.

[60] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2018, pp. 1325–1334.

[61] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Nov. 2019, pp. 5229–5238.

[62] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 858–866.

[63] ISPRS, "International society for photogrammetry and remote sensing: 2D semantic labeling challenge," 2016. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[64] 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. Photogramm. Remote Sens.*, vol. 156, pp. 1–13, 2019.

[65] Y. Liu *et al.*, "Paddleseg: A high-efficient development toolkit for image segmentation," 2021, *arXiv:2101.06175*.

[66] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

Fang Fang (Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in management science and engineering from China University of Geosciences, Wuhan, China, in 1998 and 2012, respectively.

She is currently an Associate Professor with School of Computer Science, China University of Geosciences. Her research interests include spatial data mining, machine learning, and GIS application.

Daoyuan Zheng received the B.S. degree in software engineering from Wuhan Textile University, Wuhan, China, in 2020. He is currently working toward the M.S. degree in software engineering from China University of Geosciences, Wuhan, China.

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

Shengwen Li (Member, IEEE) received the B.S. degree in computer science and the Ph.D. degree in cartography and geographic information engineering from China University of Geosciences, Wuhan, China, in 2000 and 2010, respectively.

He is currently an Associate Professor with School of Computer Science, China University of Geosciences. His research interests include deep learning for remote sensing, spatial-temporal data mining, and knowledge graph.

Yuanyuan Liu (Member, IEEE) received the B.E. degree in communication engineering from Nanchang University, Nanchang, China, in 2005 and the Ph.D. degree in management science and engineering from Central China Normal University, Wuhan, China in 2015.

She is currently an Associate Professor with School of Computer Science, China University of Geosciences. Her research interests include image processing, computer vision, and pattern recognition.

Linyun Zeng received the B.E. degree in internet of things from Jilin University, Jilin, China, in 2019. She is currently working toward the M.S. degree in software engineering from China University of Geosciences, Wuhan, China.

Her research interest includes deep learning for urban function recognition.

Jiahui Zhang received the B.S. degree in computer science and technology from Huazhong Agricultural University, Wuhan, China, in 2020. She is currently working toward the M.S. degree in software engineering from China University of Geosciences, Wuhan, China.

Her research interests include multi-modality image fusion for urban analysis.

Bo Wan received the B.S. degree in computer science and technology and the Ph.D. degree in cartography and geographic information engineering from the China University of Geosciences, Wuhan, China, in 1998 and 2007, respectively.

He is currently a Professor with the School of Computer Science, China University of Geosciences. His research interests include remote sensing, GIS modeling, spatial-temporal information mining and analysis.