FlipCAM: A Feature-Level Flipping Augmentation Method for Weakly Supervised Building Extraction From High-Resolution Remote Sensing Imagery

Xueliang Zhang, *Member, IEEE*, Qi Su, Pengfeng Xiao, *Senior Member, IEEE*, Wenye Wang, Zhenshi Li, and Guangjun He

***Abstract*— It is time-consuming to collect a huge number of pixel-level annotations for accurately extracting buildings by deep neural networks. Supported by class activation map (CAM), weakly supervised semantic segmentation (WSSS) methods with image-level annotations serve as an efficient solution for building extraction. However, it is a great challenge to generate high-quality CAM heatmaps for buildings from high-resolution remote sensing images. On one hand, image-level labels lack spatial information, resulting in partial integrity and hollow phenomenon for building extraction. On the other hand, complex backgrounds in remote sensing images can lead to inaccurate extraction of building boundaries. In this study, we propose a novel weakly supervised building extraction method called FlipCAM to deal with these challenges. The Flip module based on feature-level flipping augmentation is designed to improve the integrity of CAM heatmaps by fusing the original and flipped feature maps. In addition, by combining the Flip module with the slice and merge (SAM) module based on consistency architecture, FlipCAM is able to generate high-quality CAM heatmaps with both boundary fineness and internal integrity in an end-to-end manner, which also alleviates special difficulties for building extraction, including adhesions in dense buildings and confusions with background and shadows, providing reliable pixel-level pseudo masks for training segmentation network to extract buildings. Extensive experiments on three high-resolution datasets show that FlipCAM achieves excellent performance and outperforms other weakly supervised methods in terms of effectiveness and robustness capabilities. Our code is public at https://github.com/NJU-LHRS/FlipCAM-master.**

***Index Terms*— Building extraction, class activation map (CAM), feature-level flipping augmentation, high-resolution remote sensing imagery, weakly supervised deep learning.**

I. Introduction

With the rapid development of high-resolution satellites and remote sensing technology, building extraction from remote sensing imagery is of great significance for geographic applications, such as urban planning [1], [2], population estimation [3], and land cover mapping [4]. As a binary segmentation task, the main purpose of building extraction is to assign each pixel in a remote-sensing image as a building or nonbuilding label. With the increasing building extraction requirements and the growing number of high-resolution remote sensing images, it is crucial to find an efficient way to accurately and automatically extract buildings.

As data-driven methods, deep convolutional neural networks (DCNNs) have been widely utilized to extract buildings due to their powerful capability to handle abundant data [5], [6]. Under the supervision of pixel-level annotations, fully convolutional networks (FCNs) [7] can make full use of spatial context information among pixels and extract multilevel features by CNN receptive fields, which are extremely capable of building extraction tasks [8], [9]. However, collecting the huge amount of pixel-level labels for training FCNs is expensive and time-consuming [10], especially annotating remote sensing images with large spatial ranges and high professional requirements. Therefore, many incomplete annotation methods have been proposed for extracting buildings in remote sensing images, such as semisupervised method [11], self-supervised method [12], and weakly supervised method by scribble [13], bounding-box [14], point label [15], and image-level label [16].

Among various types of incomplete annotation methods, the image-level weakly supervised method is more practical and challenging because the image-level annotations are cost-optimal among them. Image-level labels only indicate the existence of buildings in images without any prior spatial information, such as spatial location or boundary information, which makes it difficult for weakly supervised methods to achieve the same performance as fully supervised semantic segmentation (FSSS) methods.

Manuscript received 10 November 2023; revised 8 January 2024; accepted 24 January 2024. Date of publication 30 January 2024; date of current version 8 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297, in part by the AI & AI for Science Project of Nanjing University under Grant 02091480605203, and in part by the Fundamental Research Funds for the Central Universities under Grant 020914380119. *(Corresponding author: Xueliang Zhang.)*

Xueliang Zhang is 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, Jiangsu 210023, China, and also with the Observation Research Station of Land Ecology and Land Use in the Yangtze River Delta, MNR, Nanjing, Jiangsu 210000, China (e-mail: zxl@nju.edu.cn).

Qi Su, Pengfeng Xiao, Wenye Wang, and Zhenshi Li 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, Jiangsu 210023, China.

Guangjun He is with the State Key Laboratory of Space-Ground Integrated Information Technology, Space Star Technology Company Ltd., Beijing 100095, China.

Digital Object Identifier 10.1109/TGRS.2024.3360276

In general, image-level weakly supervised semantic segmentation (WSSS) methods consist of two main steps. First, a classification network with image-level labels needs to be trained to obtain the class activation map (CAM) with the capability of object localization [17]. The pseudo masks are generated from CAM heatmaps as the coarse building extraction results from a series of postprocessing. Second, pseudo masks are utilized to train a conventional semantic segmentation network to extract fine-grained building regions [18], [19], [20], [21]. It can be inferred that the accuracy of pseudo masks has a direct impact on the building extraction performance in the second step.

Due to the serious lack of boundary information and spatial location information in image-level labels, pseudo masks usually face two deficiencies compared to fully supervised pixel-level labels, namely, insufficient boundary fineness and lack of internal integrity. Plenty of approaches have been proposed to alleviate the above-mentioned problems [22], [23]. As a traditional way, conditional random field (CRF) [24] and CRF-loss [25] were used to improve the boundary fineness of the pseudo masks by learning the spatial relationships between pixels [26]. Based on the principle of CRF, AffinityNet [27] and IRN [28] detected boundary information by predicting semantic affinity between adjacent pixels [27], [28], [29], [30], [31]. Inspired by CRF-loss, boundary loss [34], and boundary modules [35], [36], [37] were proposed to solve boundary problem in an end-to-end manner.

In addition, consistency regularization methods [38], [39], [40], [41] derived from self-supervised principles become popular due to better performance on mining boundary and multiscale information [42]. However, for weakly supervised building extraction tasks, it is crucial to create appropriate augmentation images in the consistency architecture, as these images can provide supervision beyond image-level labels, enabling the network to learn more about building boundary features. Ensuring that no misalignment occurs between the images is also important, and slice and merge (SAM) operations are an effective augmentation strategy in this regard. Otherwise, consistency regularization methods may fail to be effective.

To enhance the integrity, several methods tried to mine more category information in feature space [43], [44], [45] by multiple seeds [46], [47], [48], [49], [50], clustering [51], improved loss function [52], [53], and graph neural network [54], [55], [56], [57]. Combined with novel ideas in deep learning, the weakly supervised methods with uncertainty estimation [58], contrastive learning [59], prototype exploration [41], [60], and attention module [61], [62], [63], [64], [65], [66], [67], [68] have been proposed to more effectively address the lack of integrity. However, these weakly supervised methods in computer vision cannot be directly applied to remote sensing images due to different categories, different background complexity, and scale variation between natural images and remote sensing images. Therefore, it is crucial to design WSSS methods suitable for high-resolution remote sensing images.

Based on the urgent requirements for extracting geographic objects in a low-cost and high-efficiency way, existing studies have made targeted improvements for image-level weakly supervised extraction tasks of high-resolution remote sensing images [69], [70]. Specifically, some geographic objects that are relatively easy to separate from the background, such as cloud [71], water [72], green plastic cover [73], and solar panel [74], can be better extracted by customizing weakly supervised network architecture according to their own characteristics. For example, according to the characteristics of simple internal features and little heterogeneity, Li et al. [71] extracted cloud by pruning pooling layers in the classification network and greatly improved the fineness of cloud results. To improve the extraction performance of green plastic cover, a coarse-to-fine weakly supervised method was proposed to solve the special extracting problems, such as noisy labels, blurry boundaries, and imbalance categories [73].

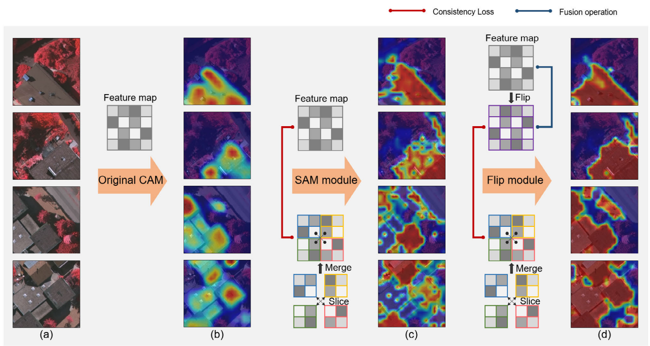


Fig. 1. Visualizations of CAM heatmaps in FlipCAM scheme. (a) Input image. (b) CAM heatmaps generated by the original CAM method. (c) CAM heatmaps generated by consistency architecture with SAM module. (d) CAM heatmaps generated by consistency architecture with SAM and Flip module.

Compared to the above-mentioned geographic objects, more studies focused on weakly supervised building extraction [75], [76], [77], [78], [79], [80]. Due to more complex representations and scenarios of buildings, such as dense buildings, and buildings confused with backgrounds and influenced by shadows, these studies preferred to separate the steps of boundary refinement and integrity enhancement without an end-to-end manner. To improve boundary fineness, adversarial climbing strategy [43], attention module, and multiscale fusion were adopted in adversarial climbing and gated convolution (ACGC) [77], weakly supervised attention network (WSAN) [75], and SPMF-Net [76], respectively. To improve internal integrity, superpixel refinement was adopted in MSG-SR-Net [79]. However, simultaneously optimizing both boundary fineness and internal integrity by iterative training is more efficient than separately optimizing one by one. Since many substeps already exist in weakly supervised methods, continuing to separate steps will greatly reduce extraction efficiency.

Given the above challenges, we propose a novel weakly supervised method for building extraction called FlipCAM (see Fig. 1). The proposed method integrates consistency architecture with two main novel modules, i.e., SAM module and Flip module, into a universal classification network to generate high-quality CAM heatmaps for obtaining high-confidence pseudo masks. In weakly supervised tasks, the scarcity of supervisory information leads to issues such as insufficient boundary fineness and lacking internal integrity. To tackle these issues, we proposed the FlipCAM method, comprised of the consistency architecture, SAM module, and Flip module, to introduce multiple branches in the classification network in a consistent and feature-enhanced manner, thereby increasing pixel correlations to achieve the goal of enriching supervisory information. This approach effectively enhances boundary fineness and internal integrity. Specifically, the consistency architecture is proposed to well fit the object boundary using consistency regularization method, and by incorporating appropriate modules in the branch, such as SAM module and Flip module, which can provide supplementary supervisory information for weakly supervised building extraction tasks. SAM module enhances boundary capability and multiscale feature extraction capability through SAM operations and multiscale image inputs. Flip module improves internal integrity by augmenting and integrating high-dimensional feature information at the feature level. It is worth noting that, unlike common augmentation methods that perform transformation on the original remote sensing image, the Flip module implements interaction between original and flipped feature maps in deep neural networks. This means that additional supervision is fused pixel-by-pixel in high-dimensional feature space, allowing the classification network to learn more pixel-level information. In addition, unlike the previously proposed weakly supervised building extraction approaches, both modules in FlipCAM improve building extraction performance simultaneously during iterative network training, thus it is an end-to-end training strategy during CAM heatmaps generation.

The main contributions of this study can be summarized as follows:

1) Based on feature-level flipping augmentation and fusion strategy, an original module named Flip module is designed to improve the internal integrity for extracting buildings.
2) Based on the SAM strategy, an original module named SAM module is designed to abundant multiscale information for extracting building.
3) Combining the Flip module and SAM module by consistency architecture, a novel weakly supervised method named FlipCAM is proposed to greatly improve boundary fineness and internal integrity in an end-to-end manner for generating CAM heatmaps.

II. Methodology

As an image-level weakly supervised building extraction method, FlipCAM consists of two steps (see Fig. 2): 1) train classification network and generate pseudo mask and 2) train segmentation network and output result. In Section II-A, we briefly review how to generate CAM. In Section II-B, we describe the principle of consistency architecture and explain how it works in a theoretical way. In Section II-C, we describe the principle of SAM module and illustrate how it improves building extraction performance by multiscale information. In Section II-D, we describe in detail the principle of Flip module and illustrate its great improvement of the integrity for building extraction in a theoretical way. In Section II-E, we elaborate on how FlipCAM method converts CAM heatmaps into pseudo masks for segmentation network training.

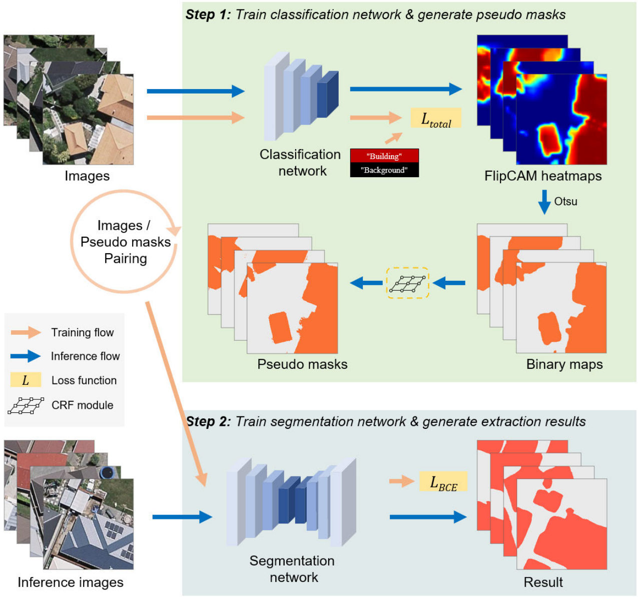


Fig. 2. Flowchart of the proposed FlipCAM method. In the first step, a classification network is trained for generating pseudo masks. Then in the second step, the pseudo masks are utilized as pixel-level labels to train a segmentation network and generate extraction results.

A. Preliminary

The core of the CAM method is converting image-level labels into coarse pixel-level labels. First, a classification network should be trained before generating CAM. To enhance the feature extraction capability of the classification network, we use binary cross-entropy loss function L_{BCE}

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \quad (1)$$

where N is the number of samples; y_i represents the category to which the ith sample belongs; and p_i is the predicted value of the ith sample. Due to the extreme versatility of the CAM method [78], there is no special restriction on the choice of classification network when generating CAM. However, considering that the generation of CAM requires the use of global average pooling (GAP) layer, ResNet-50, which comes with GAP layer, is chosen as the backbone in this study. Then we denote the feature maps on the last convolutional layer by $f \in R^{C\times HW}$, where C, H, and W denote the total number of channels, height, and width of the feature maps, respectively. The GAP layer compresses these feature maps for easy connection to the fully connected layer with parameters $w \in R^{2C}$. The fully connected layer is applied to calculate the building scores for determining the probability of containing buildings in an image. Specifically, the score s_b for building is obtained as follows:

$$s_b = \frac{1}{HW}\sum_{j=1}^{C} w_j \sum_{i} f_j^i. \quad (2)$$

Since this study focuses on building extraction, i.e., binary classification, s_b needs to be normalized and processed by the softmax activation function into the classification probability

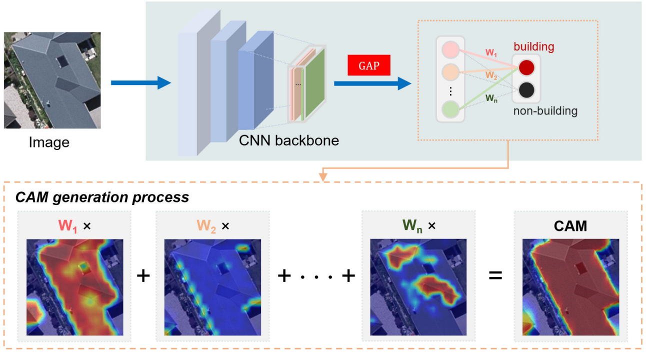


Fig. 3. Generation process of CAM. The predicted values of buildings are linearly combined with the feature maps generated by the last convolutional layer to generate CAM. CAM highlights the discriminative regions for building category.

of building p_b

$$p_b = \frac{\exp(s_b)}{\sum_{c=1}^{C} \exp(S_c)} \tag{3}$$

where $\exp(\bullet)$ is the exponential function; C contains two classes, i.e., building and background; and S_c represents the score calculated for each category after the fully connected layer. Then, as shown in Fig. 3, the original CAM M_b for extracting buildings is given by

$$M_b = \mathrm{ReLU}\left(\sum_{j=1}^{C} w_j f_j\right) \tag{4}$$

where $\mathrm{ReLU}(\cdot)$ is the rectified linear unit activation function. In short, the principle of generating CAM is a linear combination of the predicted values of buildings and the feature maps generated by the last convolutional layer. It is worth noting that the size of M_b is the same as the feature maps in the last convolutional layer, thus it is necessary to upsample M_b by bilinear interpolation to make it the same size as the input image to make subsequent processing easier.

As a pioneering work, the CAM method utilizes image-level labels to achieve coarse building extraction. However, there are still two challenges: 1) due to the influence of coarse-grained annotations, single loss function setting, convolutional process, and pooling process, building boundary information bears a serious loss when classification network is used to extract building features; and 2) classification networks tend to focus only on the most discriminative feature regions of the category to achieve accurate classification, leading to that CAM results keep an eye on local areas of the building. Hence, it is essential to improve the internal integrity of CAM results.

B. Consistency Architecture

The consistency architecture was proposed to well fit the object boundary in weakly supervised building extraction tasks. The core idea of the architecture is to enhance the model performance by incorporating additional supervision through consistency regularization [42]. Specifically, our model achieves consistency regularization by introducing an additional input branch, where the input images can undergo various data augmentation strategies such as SAM, flipping, rotation, rescaling, and so on. By minimizing the consistency loss during the training phase, the model continuously absorbs additional supervisory information, especially more fine-grained object boundary information, by narrowing the output variance between the two branches, ultimately improving the model's performance.

The essence of building extraction by the CAM method is to utilize a classification network for coarse semantic segmentation tasks. However, there is a huge difference in the way the parameters of classification networks and segmentation networks are optimized. In this case, the segmentation network performs pixel-level extraction of buildings, and when transforming and inverted transforming building images, ideally the segmentation network tends to be equivalent and can achieve the effect of the following:

$$N_{\mathrm{ip}}(I_b) = T^{-1}\big(N_{\mathrm{ip}}(T(I_b))\big) \tag{5}$$

where I_b is the image sample of building; T and T^{-1} respectively represents the transformation and inverted transformation; and N_{ip} represents the deep network with ideal parameters for building extraction. We refer to this phenomenon as equivalent segmentation. Compared with the segmentation network, the classification network is more focused on category invariance instead of equivalent segmentation, which makes the building boundary in CAM results not fine-grained enough. Therefore, we add consistency architecture to impose a consistency regularization on the classification network to achieve a better building boundary extraction effect, as shown in Fig. 4. The consistency architecture consists of two branches with shared network weights but outputs different feature maps. Based on the feature maps from the original images and the inverted transformation feature maps, we can establish the following loss function of consistency regularization:

$$L_{\mathrm{consistency}} = \|N(I_1) - T^{-1}(N(T(I_2)))\|_1 \tag{6}$$

where N, I_1, I_2, and $\|\cdot\|_1$ represent the classification network, input images from two branches, and $L1$-regularization, respectively. During the training of the classification network, $L_{\mathrm{consistency}}$ is continuously optimized to ensure that the output activation maps from two branches are constantly regularized, thereby enhancing the model's ability to extract building boundaries.

C. SAM Module

In order to improve the performance of the classification network on extracting multiscale buildings, SAM module is designed and integrated into consistency architecture as one branch.

Compared with the rescaled transformation in the self-supervised equivariant attention mechanism (SEAM) method [42], the SAM module [see Fig. 4(b)] puts remote sensing images with different scales into the network, which is more suitable for multiscale remote sensing scenarios. Specifically, each image for training is sliced into four pieces along two central axes and each piece utilizes the classification network to generate sliced feature maps. Notably, we have discovered that upsampling these slices to the original image size before inputting them into the classification network is unnecessary. Although the variation in image sizes

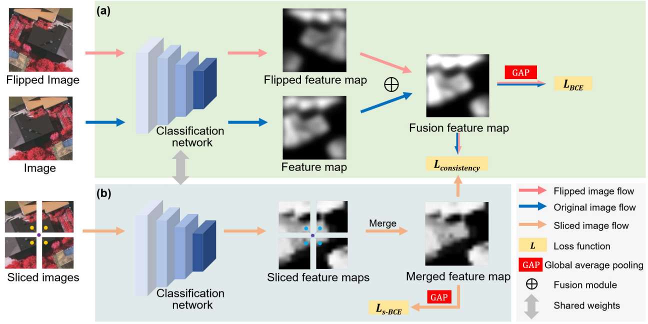


Fig. 4. Training process of FlipCAM. (a) Flip module. This module focuses on enhancing the integrity of building extraction. (b) SAM module. This module improves the performance of multiscale building extraction. These two modules, together with the consistency architecture ($L_{\text{consistency}}$), focus on improving the boundary fineness of building extraction.

result in altered feature map dimensions, causing a mismatch between the dimensions of flattened feature maps and fully connected layers of the classification network, SAM module has addressed this issue by performing a merge operation on the sliced feature maps before feature maps entering the fully connected layers. The merged feature maps have the same dimensions as the feature map produced by the original image. This approach without performing resampling has the advantage of significantly reducing memory consumption. Through the SAM operations, the classification network can learn multiscale information that is lacking in another branch. Connected by $L_{\text{consistency}}$ to the feature map in another branch, the merged feature map continuously contributes multiscale building information to the network.

D. Flip Module

The consistency architecture achieves to optimize the boundary of the building extraction result, but the integrity of the extraction result is still not assured, which would lead to the hollow phenomenon. Therefore, the Flip module [see Fig. 4(a)] is proposed to eliminate the incomplete extraction problem.

The use of data augmentation techniques has already been successfully applied in weakly supervised segmentation to improve the accuracy of CAMs [89], [90]. However, these methods only enhance the original images through random masks [89] or heatmap-based masks [90], lacking further integration of high-dimensional feature information extracted by the network. In contrast to traditional weakly supervised image augmentation methods, the Flip module integrates high-dimensional feature information by fusing feature maps at the feature level. This allows the network to have a deeper understanding of the overall representation of buildings, thereby enhancing the integrity of buildings. Unlike common data augmentation in deep learning preprocessing, the Flip module aims at achieving feature-level augmentation. Specifically, it places the flipped image into the weight-shared classification network and generates the flipped feature maps. It enhances the integrity of the original CAM by fusing the original and flipped feature maps pixel-by-pixel

$$M_{\text{fu}}^{i,j} = \left(M_{\text{o}}^{i,j} + M_{\text{f}}^{i,W-j}\right), 0 \leq i \leq H, 0 \leq j \leq W \qquad (7)$$

Algorithm 1 Training phase of FlipCAM for WSSS.

```
Input: $\mathbf{C}$(classification label), $\mathbf{I} = \{(\mathbf{I_p},\mathbf{I_s^{\mathit{i}}})\}$
  ($\mathbf{I_p} \rightarrow$ image for generating pseudo mask, $\mathbf{I_s^{\mathit{i}}} \rightarrow$ sub-image)
Output: $\mathbf{L_{total}}$ (the final loss)
//step1: generate flipped image and sliced image
$\mathbf{I_f}$ = Horizontal Flip ($\mathbf{I_p}$)
$\mathbf{I_s^1}$,$\mathbf{I_s^2}$,$\mathbf{I_s^3}$,$\mathbf{I_s^4}$ = Split ($\mathbf{I_p}$)
// step2: generate feature maps by a CNN backbone
$\mathbf{X_p}$ = CNN_Backbone ($\mathbf{I_p}$)
$\mathbf{X_f}$ = Horizontal_Flip$^{-1}$ (CNN_Backbone($\mathbf{I_f}$))
$\mathbf{X_{fu}}$ = Fusion ($\mathbf{X_p}$,$\mathbf{X_f}$)
for $i$ in {1,2,3,4} do
  | $\mathbf{X_s^{\mathit{i}}}$= CNN_Backbone($\mathbf{I_s^{\mathit{i}}}$)
end
$\mathbf{X_s}$ = Merge ($\mathbf{X_s^1}$,$\mathbf{X_s^2}$,$\mathbf{X_s^3}$,$\mathbf{X_s^4}$)
// step3: calculate losses and backpropagation
$\mathbf{L_{BCE}}$= BCE_Loss (GAP ($\mathbf{X_{fu}}$), $\mathbf{C}$)
$\mathbf{L_{s-BCE}}$ = BCE_Loss (GAP ($\mathbf{X_s}$), $\mathbf{C}$)
$\mathbf{L_{consistency}}$= L1_Loss ($\mathbf{X_{fu}}$, $\mathbf{X_s}$)
$\mathbf{L_{total}}$= $\mathbf{L_{BCE}}$+ $\mathbf{L_{s-BCE}}$ +$\alpha\times$ $\mathbf{L_{consistency}}$
//$\mathbf{L_{total}}$ back propagation
```

where M_{fu}, M_{f}, and M_{o} represent the fused feature map, flipped feature map, and original feature map, respectively. Then, we perform GAP operations on the fused feature map and calculate binary cross-entropy loss $L_{s-\text{BCE}}$. Meanwhile, $L_{\text{consistency}}$ is calculated by the fused feature maps and the above-mentioned merged feature maps.

Instead of utilizing other common augmentation methods to enhance the feature map, flip transformation is selected to improve the integrity of CAM for the following reasons. First, some image augmentation methods such as image rotation, image rescaling, or image translation are unable to provide enough supervision because of little changes before and after data augmentation. Specifically, most of the image augmentation methods do not change the relative position of the pixels, i.e., geographic object B is to the left of geographic object A and remains so after data augmentation. Therefore, from the principle of deep learning architecture, CNN backbone considers that the features of the two images before and after data augmentation are almost the same. Second, as a widely used augmentation method, image cropping is not applicable to remote sensing images where geographic objects may appear at any location, and it is also difficult to fuse the feature maps generated from the cropped images with the original feature maps. Third, flip transformation neither provides limited supervision nor loses spatial and location information compared to the original image. As a kind of mirror transformation, flip transformation can provide enough supervision. In addition, since it has been proved that flip transformation is effective in both image augmentation [42] and label augmentation [81], [82], we assume that it would be reasonable to have an effect on feature-level augmentation.

In the training process (see Algorithm 1), the consistency scaling coefficient α is set to maintain the balance of three loss functions

$$L_{\text{total}} = L_{\text{BCE}} + L_{s-\text{BCE}} + \alpha \times L_{\text{consistency}}. \tag{8}$$

With L_{total} decreasing by the deep learning optimizer, the classification network iterates over the parameters to improve the performance of building extraction.

E. Building Extraction Model

In the training phase, two branches of FlipCAM share the weights of the CNN backbone, and the trained network parameters are unique and consistent across both branches. After training the classification network, we directly input original remote sensing images into the network to generate FlipCAM heatmaps without flip and feature fusion operations, and threshold heatmaps by the Otsu algorithm [83] to distinguish the buildings from the background. CRF postprocessing [24] is utilized to convert coarse heatmaps into refined pseudo masks, which integrates contextual information by considering the spatial relationships between pixels and the similarity of pixel values to improve the fineness of the building boundary.

It is evident that the combination of image-level labels and CAM method aligns well with the structure based on the classification network. Moreover, utilizing the Otsu algorithm for threshold processing ensures the generated pseudo-labels are well-matched with the segmentation network (step 2 in Fig. 2).

DeepLabv3+ [86] is selected as the segmentation network in this study because of its outstanding performance in geographic object extraction and popularity. Specifically, the main innovations of DeepLabv3+ are the integration of atrous spatial pyramid pooling (ASPP) and encoder-decoder architecture. ASPP allows the model to capture multiscale contextual information by utilizing multiple parallel atrous convolutional layers with different rates. This enables the model to effectively capture both local and global image features, resulting in more precise segmentation results. The encoder-decoder architecture in DeepLabv3+ further enhances the segmentation performance by combining the high-level semantic information from the encoder with the detailed spatial information from the decoder. This fusion of information helps refine the segmentation boundaries and improve the overall segmentation accuracy.

As with the classification network, the binary cross-entropy loss function is also utilized in our segmentation network. After training the segmentation network (DeepLabv3+) by L_{BCE}, we put the testing images without any screening process into the trained segmentation network to generate the building extraction results. In order to further analyze the results in large-scale remote sensing images, all the results are stitched back together in the same way as the original images were cut (a 50% repetition rate between adjacent results).

III. Experimental Results

A. Experimental Setup

1) Datasets: In order to involve as many different types of buildings, different building alignment densities, different background situations, etc., as possible to verify the effectiveness and robustness of the proposed FlipCAM, three datasets containing various types of buildings were selected, including the ISPRS 2-D semantic segmentation benchmark dataset (Potsdam dataset and Vaihingen dataset) [82], and Aerial Imagery for Roof Segmentation (AIRS) dataset [83] (see Table I and Fig. 5). Specifically, Potsdam dataset consists of four bands (IRRGB) of TIFF files. AIRS dataset and Vaihingen dataset consist of RGB and IRRG bands, respectively. The Potsdam dataset comprises 38 raw aerial images with 6000 × 6000 pixel and a ground resolution of 5 cm/pixel, with 24 images for training and 14 for testing. An image named "top_potsdam_7_10_RGB" in Potsdam dataset was removed because of the annotation error. The Vaihingen dataset contains 33 aerial images of different sizes and a ground resolution of 9 cm/pixel, with 16 images for training and 17 for testing. The AIRS dataset comprises 24 raw aerial images with 10 000 × 10 000 pixel and a ground resolution of 7.5 cm/pixel, with 20 images for training and 4 for testing. The original images in each dataset are cropped into patches with 256 × 256 pixel with a 50% repetition rate between adjacent samples. In detail, classification samples for training are annotated as image-level annotation "building" when its pixel ratio is greater than 25% and "nonbuilding" when no building pixels in the image, which are divided into training and validation sets approximately in the ratio of 8:1. For the image patches with a pixel ratio of building between 0% and 25%, including them as positive samples in the training phase may adversely affect the performance of building feature extraction, as the majority of their regions still belong to the background. Therefore, we discard images with a pixel ratio of the building between 0% and 25% for training the classification network. For testing the performance of methods in real scenarios, segmentation samples with pixel-level annotations were derived

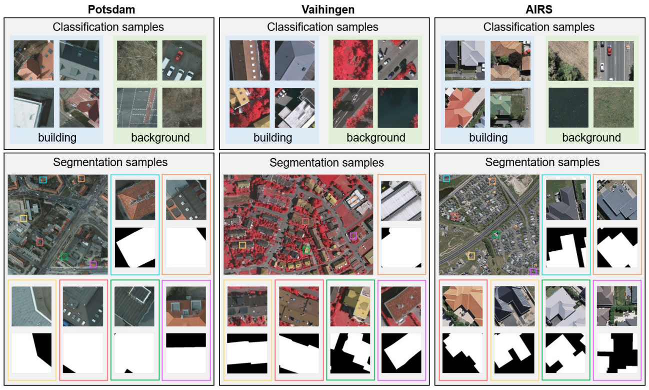


Fig. 5. Visual example of the classification samples and segmentation samples from the Potsdam dataset, Vaihingen dataset, and AIRS dataset. Specifically, classification samples are image-level annotations for training and validation of the classification network, and segmentation samples are pixel-level annotations for testing the segmentation network.

TABLE I
Composition of Potsdam, Vaihingen, and AIRS Datasets

Datasets	Training	Validation	Test	Resolution (m)
Potsdam	32000	4244	28350 (14 images)	0.05
Vaihingen	8000	1019	4498 (17 images)	0.09
AIRS	39000	4971	23716 (4 images)	0.075

from specific testing images without any screening process. The detailed datasets composition is in Table I.

2) Network Settings: As an effective convolutional neural network backbone in both classification and segmentation tasks, the ResNet [85] series is adopted as our CNN backbone. In step 1 (see Fig. 2), ResNet-50 is adopted as a classification network for generating FlipCAM, which is pretrained by the ImageNet dataset [10]. As for hyperparameter settings, PolyOptimizer is utilized as the optimizer in the training phase with momentum 0.9 and weight decay 0.0005. The learning rate is equal to 0.01 at the beginning of network training and is polydecayed by power 0.9 for each epoch. In addition, other training configurations such as batch size, the number of epochs, and seed are set as 32, 100, and 0, respectively. In step 2 (see Fig. 2), ResNet-101 and DeepLabv3+ are adopted as our segmentation network backbone and architecture. Most of the training settings in the segmentation network are the same as the classification network except the batch size and the number of epochs are 64 and 50, respectively.

The experiments are conducted on PyTorch 1.10.1 and Python 3.9.7. Both the classification network and segmentation network are trained on a computer with an Intel Core i7-11700K CPU, one NVIDIA GeForce RTX 3080 GPU and 64 GB memory.

3) Evaluation Metrics: Four accuracy metrics are selected to evaluate the performance of the proposed FlipCAM method on building extraction, including precision, recall, $F1$-score, and intersection over union (IOU), which are formulized as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$F1 - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{12}$$

where TP (true positive), FP (false positive), and TN (false negative) represent the number of pixels correctly predicted as building, the number of pixels incorrectly predicted as building, and the number of pixels correctly predicted as nonbuilding, respectively. Precision and recall represent the commission and omission errors, respectively. $F1$-score is the harmonic average of precision and recall, which can better reflect the extraction performance when the number of buildings and background pixels is imbalanced. IoU describes the extent of overlap between prediction and ground truth, which integrally represents the quality of building extraction.

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

In this section, to demonstrate the superiority of the proposed method, we compare FlipCAM with nine state-of-the-art WSSS methods, namely CAM [17], IRNet [28], PuzzleCAM [39], SEAM [42], AdvCAM [43], ReCAM [52], CONTA [60], AMR [67], and ACGC [77]. Specifically, CAM is the pioneer and basic framework of WSSS. IRNet optimizes extraction boundary by training with interpixel relations on the attention maps. The core idea of SEAM and PuzzleCAM is creating additional supervision by consistency regularization. AdvCAM forces regions initially considered not to be discriminative to become involved in subsequent classifications in an anti-adversarial manner. AMR leverages a spotlight branch and a compensation branch to obtain weighted CAMs that can provide recalibration supervision and task-specific concepts. ReCAM reactivates the converged CAM with BCE by using softmax cross-entropy loss (SCE). As a novel method in weakly supervised building extraction, ACGC improves building extraction performance by combining ACGC. In a word, these methods are typical in WSSS with their diverse innovation and good performance. It is noted that since there is no open-source code for ACGC, we reproduced all open-source codes except for ACGC on the three datasets to obtain the experimental results. For all the comparison experiments we conduct, remote sensing images are subjected to the same data augmentation methods such as random horizontal flipping, random cropping, and random rotation in image preprocessing phase. Furthermore, we employed DeepLabv3+ as a benchmark FSSS method for comparison, since the performance gap between WSSS and FSSS is a critical indicator for the effectiveness of WSSS.

1) Performance Evaluation on Pseudo Masks: As the pixel-level labels in step 2 (see Fig. 2), pseudo masks directly affect the performance of the segmentation network for building extraction. The comparison results of pseudo masks on three datasets are presented in Table II. It is demonstrated that the pseudo masks extracted by the FlipCAM method perform the best among the weakly supervised methods and achieve a balance between accuracy and integrity. Specifically, some weakly supervised methods like SEAM place greater emphasis on boosting the foreground confidence of pseudo masks, resulting in lower integrity of the pseudo masks and increased uncertainty in background information. To ensure effective segmentation network training, it is necessary to additionally adopt the strategy of ignoring pixel values and background thresholds [16]. As shown in Fig. 6, with the excellent performance of consistency architecture and two submodules, FlipCAM effectively mitigates the problem of image noise and background misclassification in pseudo masks.

2) Performance Evaluation on Building Extraction Results: The comparison results for building extraction on three datasets are shown in Table III. It is demonstrated that the proposed FlipCAM also performs the best. Specifically, from the precision and recall metrics of each method, it can be observed that the methods with balanced accuracy and completeness yield superior extraction results, e.g., the proposed FlipCAM, while methods that prefer a particular capability are not robust enough due to insufficient feature information extraction of building category, e.g., IRNet, CONTA, and ReCAM.

It is worth noting that some weakly supervised methods with consistency principles, such as SEAM and PuzzleCAM, show low integrity performance in both Potsdam and AIRS datasets. However, with the addition of Flip module, FlipCAM solves the problem of low integrity and thus maintaining high accuracy (see Table III). Since Vaihingen dataset contains

TABLE II

COMPARISON OF PSEUDO MASKS ON THREE DATASETS

	Potsdam				Vaihingen				AIRS			
	precision	/ *recall*	/ *F1-score* /	*IoU*	*precision*	/ *recall*	/ *F1-score* /	*IoU*	*precision*	/ *recall*	/ *F1-score* /	*IoU*
CAM	0.752	0.888	0.814	0.687	**0.961**	0.537	0.689	0.526	0.922	0.612	0.736	0.582
IRNet	0.915	0.796	0.852	0.742	0.761	0.796	0.778	0.637	0.894	0.664	0.762	0.616
CONTA	0.870	0.816	0.842	0.727	0.902	0.738	0.812	0.684	0.888	0.778	0.829	0.708
SEAM	**0.926**	0.858	0.891	0.803	0.822	0.876	0.848	0.737	**0.934**	0.724	0.816	0.689
AdvCAM	0.910	0.858	0.883	0.791	0.886	0.780	0.830	0.709	0.861	0.833	0.847	0.734
PuzzleCAM	0.867	0.920	0.893	0.806	0.909	0.735	0.813	0.685	0.855	0.740	0.847	0.735
AMR	0.860	0.928	0.893	0.806	0.917	0.737	0.818	0.691	0.752	**0.888**	0.814	0.686
ReCAM	0.820	**0.941**	0.877	0.781	0.765	0.886	0.821	0.697	0.870	0.870	0.870	0.770
FlipCAM (ours)	0.902	0.917	**0.909**	**0.834**	0.887	**0.888**	**0.887**	**0.797**	0.919	0.870	**0.894**	**0.808**

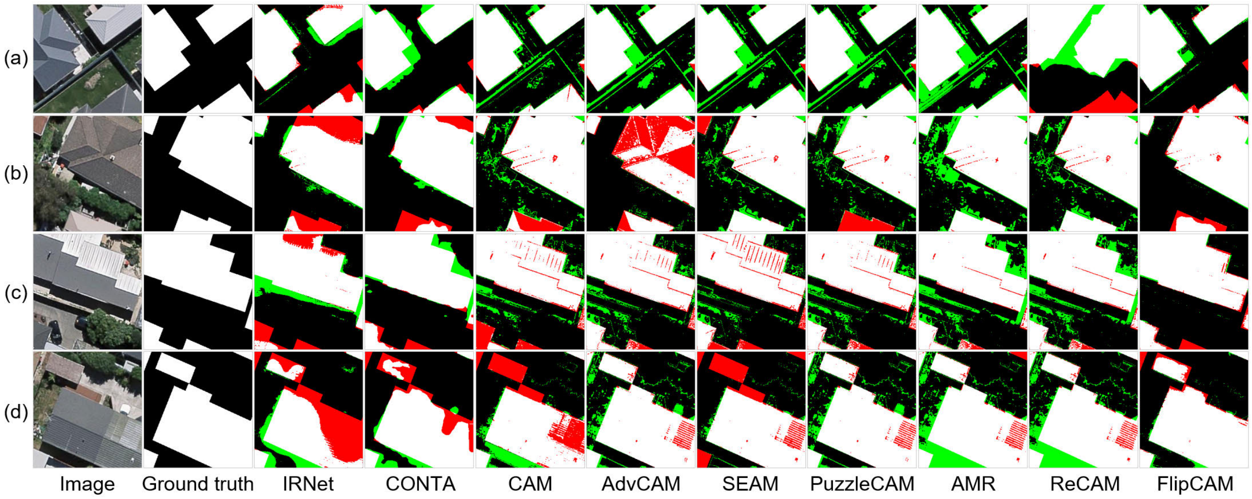


Fig. 6. Visualization of pseudo masks results of different weakly supervised methods. Different colors are used to distinguish the commission and omission situations, i.e., white for TP, black for TN, red for FP, and green for FN. (a)–(d) Pseudo masks of different samples produced by the compared methods.

TABLE III

COMPARISON OF BUILDING EXTRACTION RESULTS ON THREE DATASETS

	Potsdam				Vaihingen				AIRS			
	precision	/ *recall*	/ *F1-score* /	*IoU*	*precision*	/ *recall*	/ *F1-score* /	*IoU*	*precision*	/ *recall*	/ *F1-score* /	*IoU*
CAM	0.852	0.888	0.870	0.769	**0.968**	0.725	0.829	0.707	0.857	0.782	0.818	0.692
IRNet	0.918	0.889	0.903	0.823	0.825	0.838	0.831	0.711	0.932	0.741	0.825	0.703
CONTA	0.907	0.837	0.871	0.771	0.912	0.836	0.873	0.774	0.873	0.826	0.849	0.737
SEAM	0.943	0.884	0.913	0.839	0.882	0.865	0.873	0.774	**0.952**	0.817	0.880	0.785
AdvCAM	0.910	0.863	0.886	0.795	0.876	0.836	0.856	0.748	0.893	0.843	0.868	0.766
PuzzleCAM	0.914	0.899	0.906	0.829	0.899	0.851	0.874	0.777	0.890	0.843	0.866	0.764
AMR	0.918	0.875	0.896	0.812	0.905	0.815	0.857	0.750	0.848	0.881	0.864	0.761
ReCAM	0.865	0.922	0.893	0.806	0.893	0.840	0.866	0.763	0.927	0.844	0.883	0.791
ACGC	0.920	0.912	0.916	0.845	0.928	0.834	0.879	0.784	/	/	/	/
FlipCAM (ours)	**0.945**	**0.926**	**0.935**	**0.878**	0.926	**0.878**	**0.902**	**0.821**	0.913	**0.883**	**0.898**	**0.815**
FSSS	0.959	0.937	0.948	0.901	0.955	0.921	0.938	0.882	0.957	0.941	0.949	0.903

more dense buildings, the accuracy of SEAM and PuzzleCAM decreases significantly. In this case, the Flip module and SAM module provide more feature information of buildings to the network, which enables FlipCAM to maintain excellent building extraction performance in complex scenarios. To further understand the building extraction capability of FlipCAM,

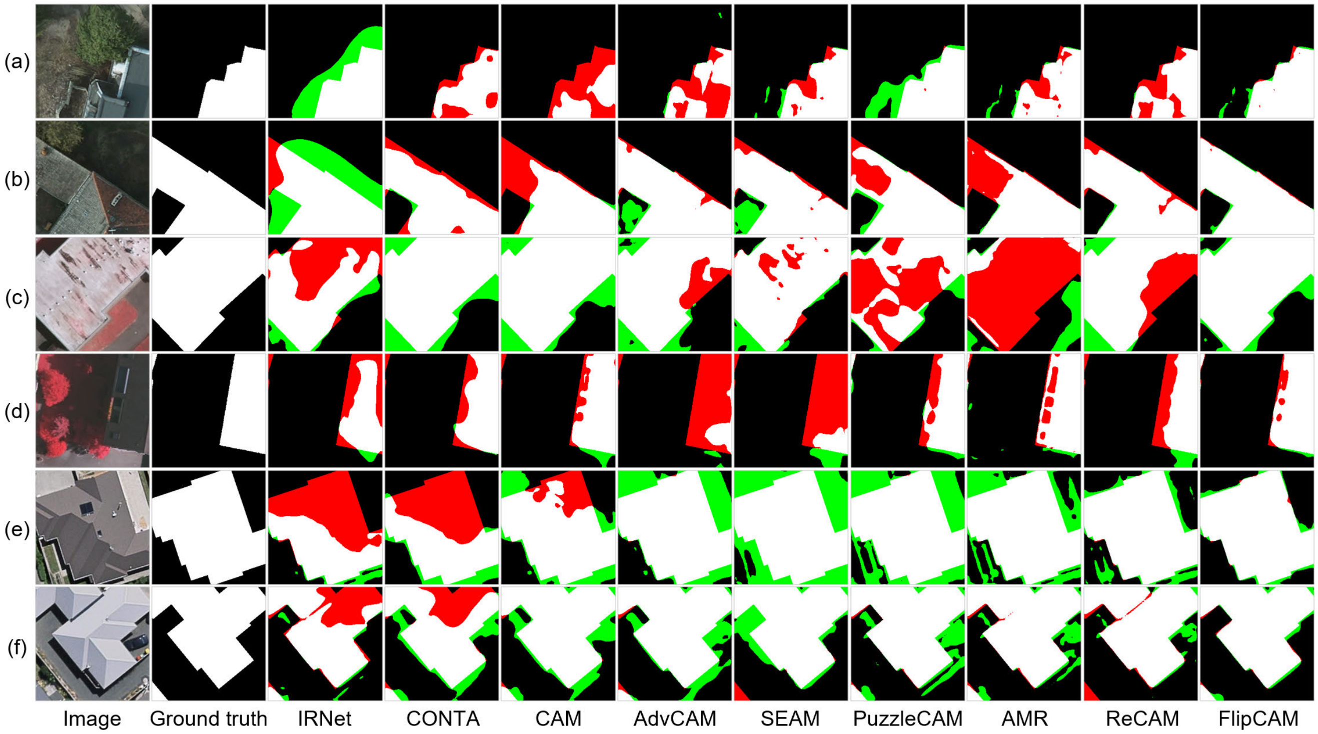


Fig. 7. Visualization of single building extraction results by different weakly supervised methods. Different colors are used to clearly distinguish the commission and omission situations, i.e., white for TP, black for TN, red for FP, and green for FN. (a)–(f) Extraction results of different samples by all the compared methods.

we analyze and compare different weakly supervised methods in various scenarios.

a) Single-building extraction performance: Single-building extraction refers to the precise segmentation and outlining of the internal shapes and external contours of individual structures. This task mainly faces the challenges of fine-grained building shapes and contour extraction (see Fig. 7). As shown in Fig. 7(a), due to the complex composition of the building, most methods cannot completely extract the building shape. A few methods like PuzzleCAM can extract the full shape, but at the same time, many background pixels are confused with the foreground, resulting in rough building contours. As a weakly supervised method that balances extraction accuracy and integrity, FlipCAM makes excellent performance when describing the building shapes and contours.

Commissioning of building attachments is another challenge in single-building extraction. Unlike FSSS with pixel-level annotations, weakly supervised methods only have image-level annotations, thus building attachments like impervious surfaces or cars [see Fig. 7(e) and (f)] that frequently co-occur with buildings are easily misclassified with buildings. Except for CAM, IRNet, and CONTA which cause the omission phenomenon, other methods extract building attachments as buildings to a certain extent, but thanks to the excellent coordination of the consistency architecture and Flip module, FlipCAM alleviates this problem and performs the best in this scenario.

b) Dense-building extraction performance: Different from single-building extraction, it is more challenging to deal with dense buildings by weakly supervised methods. First, weakly supervised methods tend to focus more on the most discriminative regions in remote sensing images, making it particularly easy to overlook entire individual buildings, especially small-scale ones, when extracting dense buildings with high intraclass heterogeneity. Consequently, this can result in an inaccurate count of the buildings extracted. As shown in Fig. 8(e), in addition to the severe over-segmentation phenomenon in AMR, most of the weakly supervised methods are attracted to the large building in the upper left corner when extracting features, ignoring the dense buildings in the lower half of the image, thus causing errors in the number of buildings extracted. On the contrary, the proposed FlipCAM accurately extracts both large buildings and dense buildings, indicating that building features extracted by FlipCAM are more comprehensive. Second, building adhesion caused by network upsampling is also a common phenomenon in multi-scale dense building scenes. As shown by the yellow circles in Fig. 8(b), (c), and (e), thanks to the Flip module and SAM module enriching the supervised information, although sacrifices a little boundary information, FlipCAM method still effectively avoids the adhesion phenomenon and the omission of small buildings.

c) Extraction performance of buildings confused with background: Due to the intricate background information around buildings, accurately extracting buildings from complicated remote sensing scenarios is extremely difficult. Moreover, due to the lack of pixel-level supervision, it is more difficult for weakly supervised methods to distinguish between buildings and background pixels with low interclass heterogeneity, as these pixels have similar spectral and textural information. With similar spectral characteristics to buildings,

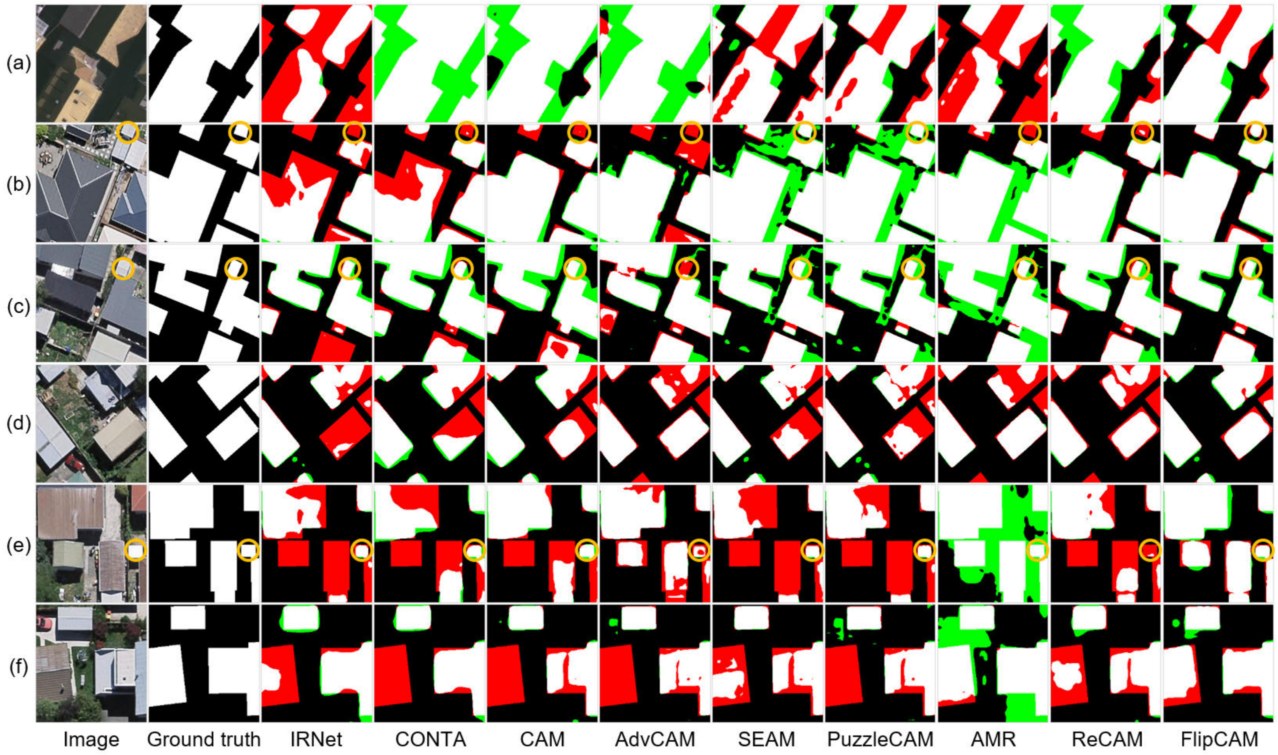


Fig. 8. Visualization of dense building extraction results by different weakly supervised methods. Different colors are used to clearly distinguish the commission and omission situations, i.e., white for TP, black for TN, red for FP, and green for FN. (a)–(f) Extraction results of different samples by all the compared methods.

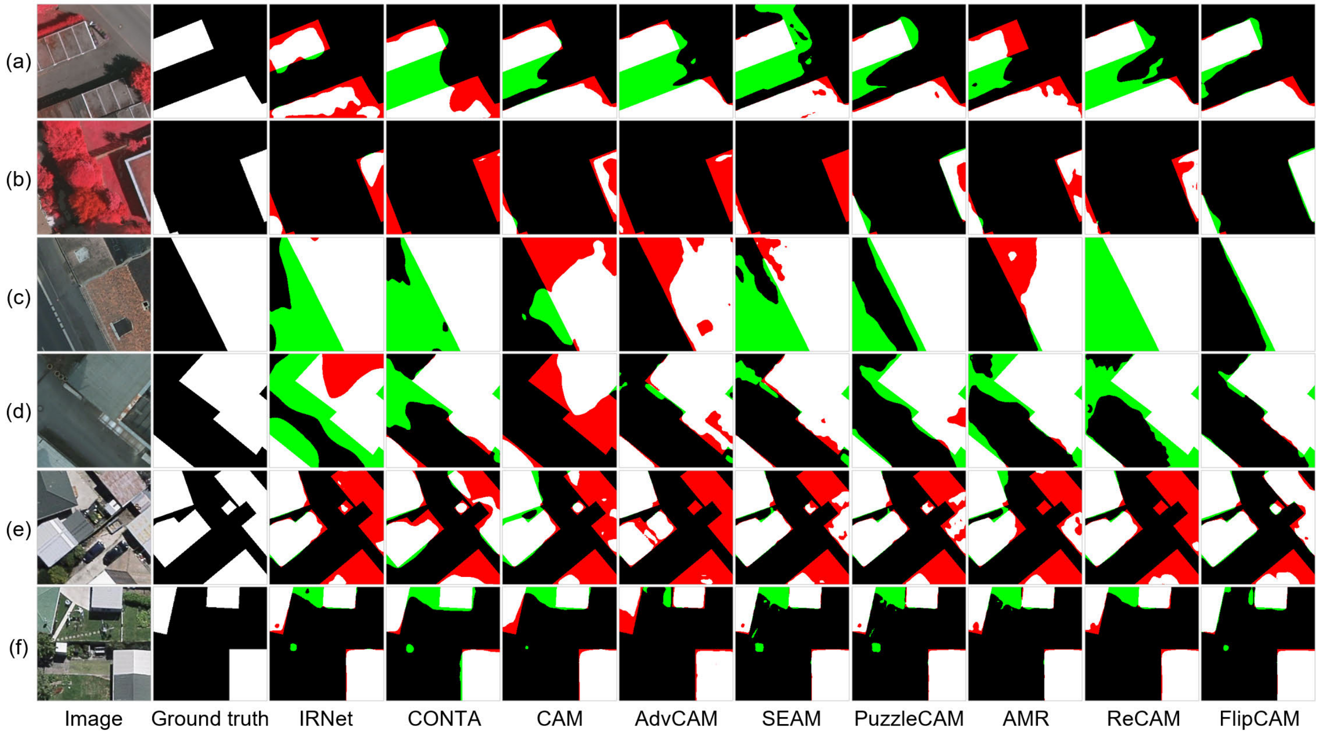


Fig. 9. Visualized extraction results of buildings confused with background by different weakly supervised methods. Different colors are used to clearly distinguish the commission and omission situations, i.e., white for TP, black for TN, red for FP, and green for FN. (a)–(f) Extraction results of different samples by all the compared methods.

impervious surfaces cause great disturbance for building extraction. However, to ensure the building extraction integrity, most of the weakly supervised methods prefer to sacrifice the precision of the building extraction results and consider impervious surfaces with low heterogeneity as building [see Fig. 9(a), (c), and (d)], which indicates that the building feature space of these weakly supervised methods contains a lot of background noise that extremely inhibits network

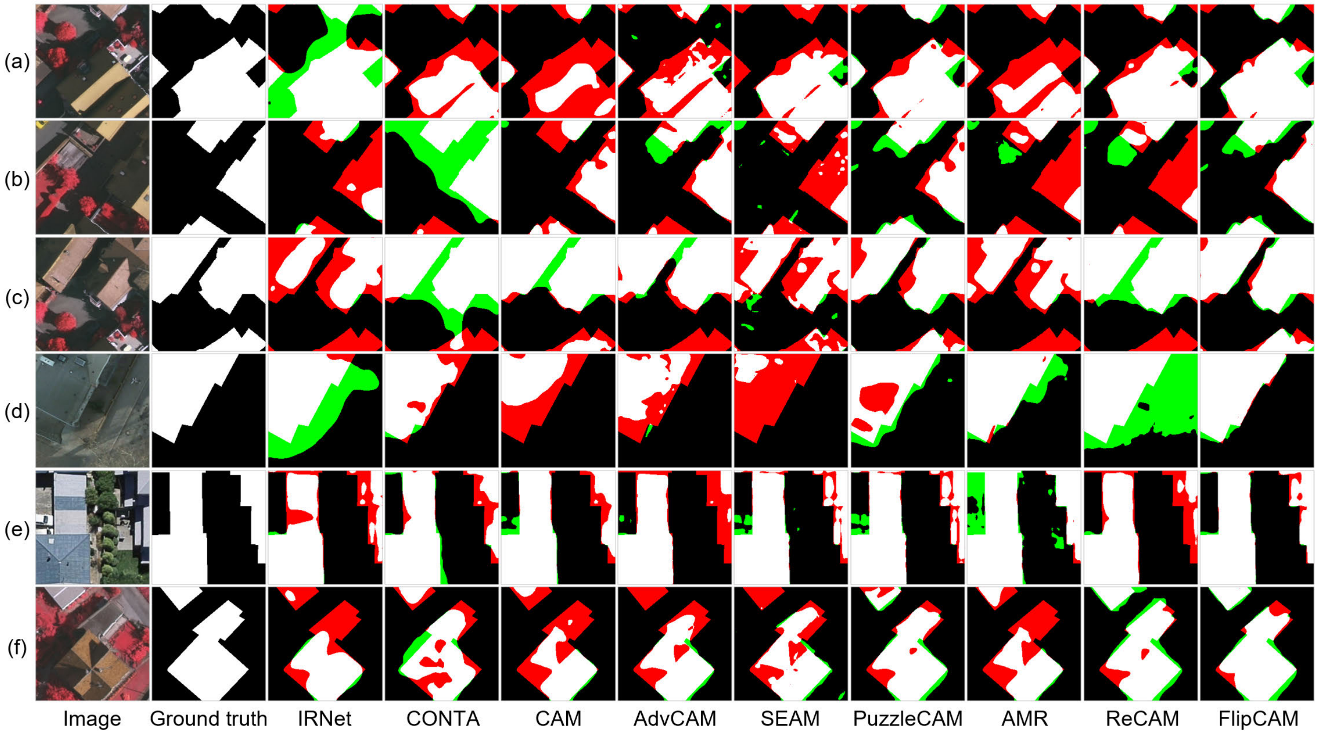


Fig. 10. Visualized extraction results of buildings influenced by shadows from different weakly supervised methods. Different colors are used to clearly distinguish the commission and omission situations, i.e., white for TP, black for TN, red for FP, and green for FN. (a)–(f) Extraction results of different samples by all the compared methods.

performance. Similarly, buildings are considered as impervious surfaces in some scenarios [see Fig. 9(e)]. However, FlipCAM incorporates comprehensive building information in feature space and thus eliminates as much noise as possible from the background, achieving excellent building extraction performance on complicated backgrounds.

d) Extraction performance of buildings influenced by shadows: The shadows produced by buildings or trees affect building extraction. Owing to the presence of shadows, weakly supervised methods need strong robustness in distinguishing whether it is a building or not in the shadow region based on the category and location information of geographic objects. As shown in Fig. 10, most weakly supervised methods treat shadows as background because the spectral and textural characteristics of shadows are not similar to those of buildings. However, due to the specificity of shadows generated by buildings, in addition to referencing spectral and texture features, the capability of a network to accurately extract buildings in shadow regions mainly depends on a deep understanding of the building macro information and shadow location information. Specifically, a network with strong robustness can make judgments based on comprehensive features even if part of the buildings is obscured by shadows. FlipCAM eliminates shadow interference in a variety of scenes, whether it is filtering interference in a variety of scenes, whether it is filtering the background [see Fig. 10(c) and (d)], or extracting the buildings [see Fig. 10(a), (b), and (e)] under shadows.

C. Component Analysis

1) Ablation Study for FlipCAM: In order to illustrate the effectiveness of two major modules in the proposed FlipCAM method, the ablation study is designed on three datasets. First, it is essential to test the performance of the baseline, i.e., the original CAM method, in generating CAM heatmaps. Second, only the SAM module and consistency architecture are joined to the baseline to illustrate the effectiveness. Finally, based on the previous design, the Flip module is added to the baseline, forming the proposed FlipCAM method. The quantitative results of the ablation study are presented in Table IV. It is demonstrated that each module in the FlipCAM method improves the quality of generated heatmaps. Furthermore, the proposed method with two effective modules achieves the best performance among them. Specifically, with the addition of the SAM module, the accuracy of CAM results improves by 6.0%, 7.0%, and 6.1% on Potsdam, Vaihingen, and AIRS datasets. As can be seen from Fig. 11(a)–(c), the CAM heatmaps extracted by the baseline have almost no local response to building boundary information, and the response

TABLE IV

Ablation Study for Each Module of FlipCAM Method. The Right Three Columns Indicate the Accuracy of Different Ablation Experiments on Different Datasets, and the Evaluation Metric is IoU

Baseline	SAM module	Flip module	Potsdam	Vaihingen	AIRS
✓			0.769	0.707	0.703
✓		✓	0.764	0.703	0.705
✓	✓		0.829 (+ 0.060)	0.777 (+ 0.070)	0.764 (+ 0.061)
✓	✓	✓	**0.878** (+ 0.109)	**0.821** (+ 0.104)	**0.815** (+ 0.112)

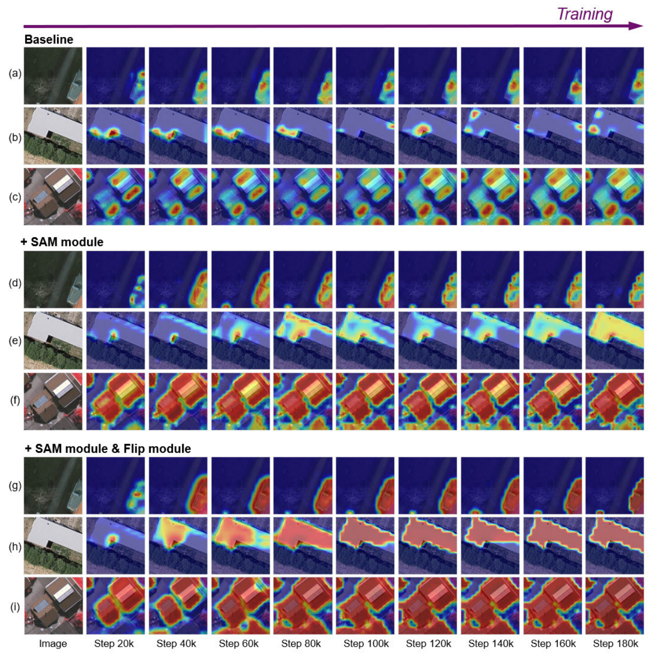


Fig. 11. Visualized training phase of three experiments in ablation study. (a)–(c), (d)–(f), and (g)–(i) Three lines of images separately show the CAM heatmaps by the baseline method, the addition of the SAM module, and the addition of the Flip module when training, respectively.

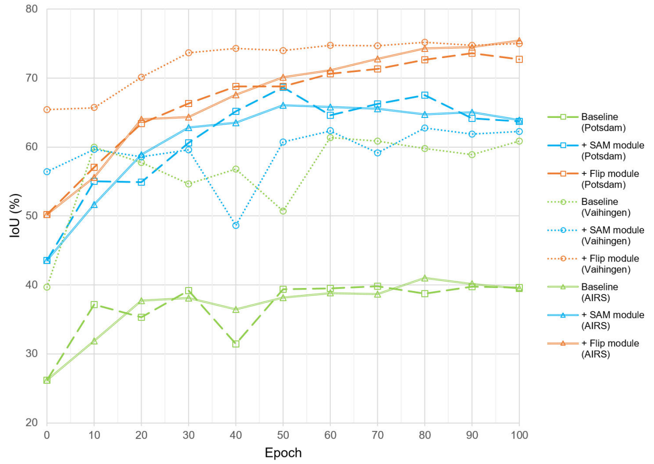


Fig. 12. Accuracy line chart of the baseline method, the addition of SAM module, and the addition of Flip module on three datasets. During the training of the classification network, the addition of the Flip module performs best and demonstrates strong stability and robustness.

to internal parts of the building is also very discrete. With the constraint of the consistency loss function, the fineness of the building boundary is greatly improved and a certain degree of integration of building internal information is achieved, which means the response is no longer randomly extracted in the building region compared with the baseline. The SAM module performs better on Vaihingen dataset where there are more dense buildings compared to the other datasets. Therefore, it can be inferred that the SAM module can effectively handle the case of dense buildings in the image. For example, the small house in Fig. 11(d) is extracted by SAM module, but not by the baseline.

Although the improvement of the classification network by the SAM module leads to a significant increase in extraction accuracy, the heatmaps still have some drawbacks. First, the SAM module focuses more on the response of the building boundary, which leads to the insufficient internal integrity of buildings [see Fig. 11(f)], and even the hollow phenomenon [see Fig. 11(d)]. Second, even without the hollow phenomenon [see Fig. 11(e)], it is obvious that the response values of the building edges extracted by SAM module are larger than those of building interiors. Therefore, building interior is easier to be classified as background than boundary in the binarization process. Furthermore, the hollow phenomenon leads to the worse omission of building interiors after CRF postprocessing and ultimately causes a significant decrease in the accuracy of pseudo masks. Based on SAM module, the accuracy of CAM results improves 4.9%, 3.4%, and 5.1% on Potsdam, Vaihingen, and AIRS datasets when the Flip module was added. This is because Flip module utilizes the feature-level augmentation strategy and pixel-level fusion to enhance the integrity of building extraction. Specifically, the feature maps generated from the upper branch [see Fig. 4(a)] are replaced with the fusion feature maps fused by original feature maps and flipped feature maps. Then, the classification network can greatly improve the integrity of building extraction while maintaining the accuracy of boundary extraction [see Fig. 11(g)–(i)] trained by the consistency regularization strategy. As expected, the hollow phenomenon is also eliminated. Moreover, after adding the Flip module, the CAM heatmaps show that the response value of the interior area of the building is greater than that of the boundary area, and the response value of the mixed pixels containing buildings and background is lower, which is more in line with the response logic of CAM heatmaps. In this way, even if some building boundary pixels in images are misclassified after binarization process, these pixels can be easily recompensed by CRF postprocessing, thus further improving the accuracy of pseudo masks.

Based on the consistency architecture, the Flip module can effectively improve the performance of the model in extracting buildings. However, the Flip module without the consistency architecture may not always be able to play its role. In order to verify the performance of the Flip module under Baseline conditions, we conducted an additional set of ablation experiments with the Baseline + Flip module, and the results are shown in Table IV. From the results, it can be observed that the accuracy of the three datasets did not improve with the addition of the Flip module alone. This is because the Flip module, as a standalone module, can only enhance the completeness of the original CAM heatmaps. However, without the edge constraints provided by consistency architecture and SAM module, the precision and overall accuracy of building extraction do not improve. Hence, the combined effect of the SAM module and Flip module is necessary to achieve optimal accuracy in building extraction.

Except for the accuracy advantages, FlipCAM also performs best in stability and robustness of classification network training (see Fig. 12). Classification network with only the SAM module has the common situation that the accuracy of the validation set increases and then decreases during the training process, and even a precipitous drop in accuracy occurs around the 40th epoch on Vaihingen dataset, which produces serious

TABLE V

Performance of FlipCAM Method Under Different α Parameter on Three Datasets. Different Lines Represent the Building Extraction Result Under Different α. In This Table, the Evaluation Metric is IoU

Parameter (α)	Potsdam	Vaihingen	AIRS
0	0.531	0.623	0.624
0.25	**0.829**	**0.759**	**0.781**
0.5	0.820	0.744	0.771
0.75	0.818	0.713	0.767
1	0.818	0.732	0.778
1.25	0.812	0.741	0.757
1.5	0.822	0.734	0.760
1.75	0.818	0.714	0.763
2	0.813	0.691	0.743

TABLE VI

Performance of FlipCAM Method on Different Feature-Level Augmentation Strategy on Three Datasets. Different Lines Represent the Building Extraction Result on Different Feature-Level Augmentation Strategy, Such as Flip, Rotation, and Rescale, Respectively. In This Table, the Evaluation Metric is IoU

Augmentation strategy		Potsdam	Vaihingen	AIRS
Flipping (direction)	Horizontal	**0.829**	**0.759**	**0.781**
	Vertical	0.824	0.758	0.775
Rotation (degree)	90°	0.773	0.694	0.691
	180°	0.753	0.705	0.619
	270°	0.776	0.705	0.732
Rescale (ratio)	0.25	0.796	0.692	0.727
	0.5	0.747	0.711	0.749
	0.75	0.808	0.679	0.751

consequences in real remote sensing scenarios without the validation set.

With the addition of SAM and Flip modules, the running time per iteration during the training process gradually increases. However, it does not impact the model's inference time. This is attributed to the weight-sharing mechanism employed in our FlipCAM method. While these modules incur extra computation time during training, the weight-sharing mechanism ensures that only the main branch is utilized for predicting input images during the inference phase. This results in no additional inference time consumption, allowing for superior accuracy within the same inference time frame.

2) Evaluation of Critical Parameter (α): As a critical parameter in the loss function, the α parameter determines the strength of the consistency regularization and has a significant impact on the performance of FlipCAM in extracting buildings. The performance of the FlipCAM method under different α values on three datasets is shown in Table V. It is illustrated that the performance is worst when α is 0 and optimal when it is 0.25. Then, the accuracy tends to decrease as the value of α increases.

In general, to reduce the workloads of parameter tuning, α defaults to 1. However, as a specific downstream task, it is necessary to balance multiple loss functions by tuning α. By adjusting the weights of each loss function, FlipCAM pays more attention to the optimization of the internal integrity of the building and the multiscale information brought by SAM module and Flip module. On the contrary, too much attention to the consistency regularization may lead to boundary information suppressing other feature information, which will bring down the accuracy instead.

3) Evaluation of Different Feature-Level Augmentation Strategies: In order to validate the effectiveness of the feature-level flipping strategy, it is essential to evaluate the performance of different feature-level augmentation strategies on three datasets. As presented in Table VI, it can be observed that using both horizontal flip and vertical flip as augmentation strategies resulted in similar experimental accuracy on three datasets. Since we already utilized random flipping as an image augmentation strategy during training, the network has learned robust visual knowledge of both vertical and horizontal flipped remote sensing images. Therefore, for feature-level augmentation strategy, whether we use vertical or horizontal flipping, the effect is essentially the same. Additionally, it is illustrated that the accuracy of feature-level flipping strategy is far ahead of feature-level rotation and rescale strategy. Furthermore, the rescale strategy is better than the rotation strategy to a certain extent. The reason is that the feature-level flipping strategy neither provides limited supervision nor loses spatial and location information compared to other feature-level augmentation strategies. Although the rescale strategy sacrifices a part of spatial information, it still provides effective supervision because of the different resolution images in the two branches. Therefore, we can conclude that the more supervision information and spatial information between different branches, the better the building extraction performance could be achieved.

4) Evaluation on w/o Segmentation Network Step: As a general process for weakly supervised methods, we first use the classification network to generate pseudo-masks for training samples, and then use the segmentation network to generate the building extraction results. However, some works directly use classification networks to generate geographic object results [69], which is more efficient. Therefore, we compared the building extraction performance with or without the segmentation network step, which is listed in Table VII. It is shown that the performance of the FlipCAM method with the segmentation step is significantly better than that without the step. As for weakly supervised building extraction, we indeed need the segmentation network to further refine extraction results.

IV. Discussion

Four perspectives are discussed from designing a weakly supervised network architecture for extracting buildings from

TABLE VII

Performance of FlipCAM Method Whether Containing Segmentation Network Step on Three Datasets. First Line Represents Building Extraction Result Without Segmentation Network Step. Second Line Represents Building Extraction Result With Segmentation Network Step. The Evaluation Metric is IoU

Segmentation network step	Potsdam	Vaihingen	AIRS
w/o	0.845	0.746	0.699
w	**0.878**	**0.821**	**0.815**

TABLE VIII

Performance of FlipCAM Method With Different Slices Number on Three Datasets. Different Lines Represent the Building Extraction Result Under Different Slices Number. The Evaluation Metric is IoU

Slices number	Potsdam	Vaihingen	AIRS
2×2 = 4	**0.829**	**0.759**	**0.781**
4×4 = 16	0.790	0.712	0.728
8×8 = 64	0.733	0.714	0.633

high-resolution remote sensing images based on the advantages and limitations of FlipCAM.

In the proposed FlipCAM method, generating the CAM heatmaps is an end-to-end process with simultaneous optimization. A weakly supervised building extraction method in an end-to-end manner can optimize boundary fineness and internal integrity simultaneously without any damage among optimization steps. As a multitask learning method with multiple loss functions, the classification network gradually improves the performance of both boundary and internal extraction simultaneously with smooth gradient descent [19]. On the contrary, we tried to consider combining multiple individual methods such as superpixel segmentation [79] and adversarial climbing [41] as distributed optimization. The first-step optimization can be partly damaged by the second-step optimization because distributed optimization has no way to consider the performance of both metrics at the same time.

In the context of weakly supervised building extraction tasks, the fundamental reasons behind phenomena such as adjacent buildings, buildings covered by shadows, low interclass heterogeneity, and high intraclass heterogeneity can be attributed to insufficient supervision information. Our FlipCAM method draws inspiration from the concept of consistency regularization in the self-supervised learning domain, enabling the network to learn not only image-level label information but also additional constraint information inherent in the images themselves. This allows the network to continuously improve its ability to extract fine-grained building boundaries in the training phase. For the SAM module, incorporating inputs of different image sizes allows the weakly supervised network to learn multiscale information. As for the Flip module, the feature-level augmentation strategy enables the weakly supervised network to learn more internal integrity information. Consequently, building boundary becomes more refined, with stronger internal integrity, and exhibit greater robustness in handling different scale scenes. This ultimately improves extraction capabilities in scenarios involving dense buildings, buildings covered by shadows, etc.

How to design the new branch brought by consistency architecture has a great impact on the improvement of building extraction. In the FlipCAM method, we utilize the SAM module, where the number of sliced subimages is 4. This slicing way balances the abundance of multiscale information with the GPU computing power, so that the batch size of the classification network can still reach 32 with the addition of one branch. In the case of higher GPU computing power, the remote sensing images can be divided into 16 or even 64 subimages. As is shown in Table VIII, we have conducted experiments to investigate the impact of different numbers of slices on the performance of the SAM module. The experimental results indicate that a small number ($2 \times 2 = 4$ slices) of sliced subimages tend to yield better performance compared to a bigger number of slices. Therefore, excessive slices may not be suitable for segmentation tasks as it would result in a significant loss of contextual information, which hampers building feature extraction in remote sensing images.

Furthermore, compared to other weakly supervised building extraction methods, FlipCAM enhances the mapping with the feature map of the original CAM at the end of the training process. Specifically, regardless of how to improve CAM by existing weakly supervised methods, they use the original CAM as one of the final mappings for computing the loss function. The reason is that the other mappings exist as complementary to the original CAM. However, in the building extraction task, the original CAM possesses the huge drawback of not being able to fully extract the internal features of the building as well as the boundary features. In FlipCAM, the Flip module fuses the original CAM with the flipped CAM pixel by pixel before using it as the final mapping, instead of directly using the original CAM as a separate mapping. In this way, the fused feature maps already contain both the representations of the original CAM and the additional information of the flipped feature maps. Therefore, there is no need to worry about the building extraction performance to return to the original CAM level if consistency architecture does not work.

Indeed, our method still has some shortcomings. First, we utilized three building datasets with a resolution in the centimeter range (0.05–0.09 m). It would be more challenging to extract buildings from lower-resolution remote sensing data, to which more attention is suggested to pay in the future. Second, while our method demonstrates excellent performance in extracting buildings, its effectiveness in extracting other small-scale artificial geographic objects, such as cars, or natural geographic objects, such as vegetation, remains unknown. Further studies are necessary to validate the applicability of our method to these additional scenarios.

V. Conclusion

In this study, a novel weakly supervised building extraction method from high-resolution remote sensing imagery, named

FlipCAM, was proposed. To improve the ability of boundary fineness, consistency architecture is designed by extending a new branch in the classification network, which continually refines the boundary information in the training phase by the consistency regularization principle. As one of the branches in consistency architecture, the SAM module is utilized to provide abundant multiscale information. Furthermore, Flip module with feature-level flipping augmentation strategy is designed to improve the capability of internal integrity, which enhances the integrity of CAM heatmaps. Combining SAM module with Flip module, FlipCAM simultaneously improves boundary fineness and internal integrity in an end-to-end manner for generating CAM heatmaps, which is different from other weakly supervised methods for building extraction with two substeps. Under the image-level weak supervision on three representative high-resolution datasets, the proposed FlipCAM method achieves promising pseudo masks and building extraction results and outperforms the state-of-the-art approaches. In addition, FlipCAM also alleviates special difficulties in building extraction. In future work, we will explore the potential of building extraction methods with weak annotations in different types of remote sensing images in terms of spatial and spectral resolutions.

Acknowledgment

The authors are grateful to High Performance Computing Center, Nanjing University, Nanjing, China, for their help on GPU resources.

References

[1] R. Fan, F. Li, W. Han, J. Yan, J. Li, and L. Wang, "Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630316.

[2] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1835–1838.

[3] Y. Xie, A. Weng, and Q. Weng, "Population estimation of urban residential communities using remotely sensed morphologic data," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1111–1115, May 2015.

[4] 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. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[5] 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. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[6] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.

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

[8] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[9] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[10] 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.*, Jun. 2009, pp. 248–255.

[11] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589.

[12] D. Muhtar, X. Zhang, and P. Xiao, "Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 4411511.

[13] H. Chen et al., "Structure-aware weakly supervised network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412712.

[14] Y. Sun, L. Mou, Y. Wang, and X. X. Zhu, "Bounding box regression network for building height retrieval using a single SAR image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 56–59.

[15] X. Zhang, Z. Zheng, P. Xiao, Z. Li, and G. He, "Patch-based training of fully convolutional network for hyperspectral image classification with sparse point labels," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8884–8897, 2022.

[16] 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.

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

[18] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 695–711.

[19] 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.*, Jun. 2018, pp. 7014–7023.

[20] S. Lee, M. Lee, J. Lee, and H. Shim, "Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5491–5501.

[21] 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, Feb. 2021.

[22] D. He and Y. Zhong, "Deep hierarchical pyramid network with high-frequency-aware differential architecture for super-resolution mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503815.

[23] D. He, Q. Shi, X. Liu, Y. Zhong, G. Xia, and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using Landsat imagery," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 2036–2067, Dec. 2022.

[24] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

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

[26] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12765–12772.

[27] 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.

[28] 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. (CVPR)*, Jun. 2019, pp. 2204–2213.

[29] B. Zhang, J. Xiao, J. Jiao, Y. Wei, and Y. Zhao, "Affinity attention graph neural network for weakly supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8082–8096, Nov. 2022.

[30] J. Fan, Z. Zhang, T. Tan, C. Song, and J. Xiao, "CIAN: Cross-image affinity net for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10762–10769.

[31] L. Ru, Y. Zhan, B. Yu, and B. Du, "Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16825–16834.

[32] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, F. Sohel, and D. Xu, "Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6964–6973.

[33] X. Zhang et al., "Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5463–5472.

[34] J. Chen, S. Fang, H. Xie, Z.-J. Zha, Y. Hu, and J. Tan, "End-to-end boundary exploration for weakly-supervised semantic segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2381–2390.

[35] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 347–362.

[36] J. Liu, J. Zhang, Y. Hong, and N. Barnes, "Learning structure-aware semantic segmentation with image-level supervision," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[37] J. Wei, Q. Wang, Z. Li, S. Wang, S. K. Zhou, and S. Cui, "Shallow feature matters for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5989–5997.

[38] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5207–5216.

[39] S. Jo and I.-J. Yu, "Puzzle-CAM: Improved localization via matching partial and full features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 639–643.

[40] S.-H. Yoon, H. Kweon, J. Jeong, H. Kim, S. Kim, and K.-J. Yoon, "Exploring pixel-level self-supervision for weakly supervised semantic segmentation," 2021, *arXiv:2112.05351*.

[41] Q. Chen, L. Yang, J. Lai, and X. Xie, "Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4278–4288.

[42] 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. (CVPR)*, Jun. 2020, pp. 12272–12281.

[43] 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. (CVPR)*, Jun. 2021, pp. 4070–4078.

[44] J. Lee, J. Choi, J. Mok, and S. Yoon, "Reducing information bottleneck for weakly supervised semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 27408–27421.

[45] J. Fan, Z. Zhang, C. Song, and T. Tan, "Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4282–4291.

[46] J. Fan, Z. Zhang, and T. Tan, "Employing multi-estimations for weakly-supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 12362, 2020, pp. 332–348.

[47] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye, "DANet: Divergent activation for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6588–6597.

[48] P.-T. Jiang, L.-H. Han, Q. Hou, M.-M. Cheng, and Y. Wei, "Online attention accumulation for weakly supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7062–7077, Oct. 2022.

[49] 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. (CVPR)*, Jun. 2019, pp. 5262–5271.

[50] K. Sun, H. Shi, Z. Zhang, and Y. Huang, "ECS-Net: Improving weakly supervised semantic segmentation by using connections between class activation maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7263–7272.

[51] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Weakly-supervised semantic segmentation via subcategory exploration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8988–8997.

[52] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 959–968.

[53] P. Wu, W. Zhai, and Y. Cao, "Background activation suppression for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14228–14237.

[54] Y. Yao et al., "Non-salient region object mining for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2623–2632.

[55] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1984–1992.

[56] S.-Y. Pan, C.-Y. Lu, S.-P. Lee, and W.-H. Peng, "Weakly-supervised image semantic segmentation using graph convolutional networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[57] K. Zhang, S. Chen, Q. Ju, Y. Jiang, Y. Li, and X. He, "Maximize the exploration of congeneric semantics for weakly supervised semantic segmentation," 2021, *arXiv:2110.03982*.

[58] Y. Li, Y. Duan, Z. Kuang, Y. Chen, W. Zhang, and X. Li, "Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1447–1455.

[59] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen, "C^2AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 979–988.

[60] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 655–666.

[61] T. Wu et al., "Embedded discriminative attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16760–16769.

[62] R. Li, Z. Mai, Z. Zhang, J. Jang, and S. Sanner, "TransCAM: Transformer attention-based CAM refinement for weakly supervised semantic segmentation," *J. Vis. Commun. Image Represent.*, vol. 92, Apr. 2023, Art. no. 103800.

[63] W. Gao et al., "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2866–2875.

[64] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2214–2223.

[65] C. Liu, E. Xie, W. Wang, W. Wang, G. Li, and P. Luo, "WegFormer: Transformers for weakly supervised semantic segmentation," 2022, *arXiv:2203.08421*.

[66] B. Kim, S. Han, and J. Kim, "Discriminative region suppression for weakly-supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1754–1761.

[67] J. Qin, J. Wu, X. Xiao, L. Li, and X. Wang, "Activation modulation and recalibration scheme for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2117–2125.

[68] W. Sun, J. Zhang, Z. Liu, Y. Zhong, and N. Barnes, "GETAM: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation," 2021, *arXiv:2112.02841*.

[69] 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, p. 207, Jan. 2020.

[70] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.

[71] 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.

[72] K. Fu et al., "WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image," *Remote Sens.*, vol. 10, no. 12, p. 1970, Dec. 2018.

[73] Y. Cao and X. Huang, "A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 157–176, Jun. 2022.

[74] J. Zhang, X. Jia, and J. Hu, "SP-RAN: Self-paced residual aggregated network for solar panel mapping in weakly labeled aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612715.

[75] M. U. Ali, W. Sultani, and M. Ali, "Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 115–124, Apr. 2020.

[76] 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, p. 1049, Mar. 2020.

[77] F. Fang et al., “Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1629–1642, 2022.

[78] Q. Su, X. Zhang, P. Xiao, Z. Li, and W. Wang, “Which CAM is better for extracting geographic objects? A perspective from principles and experiments,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5623–5635, 2022.

[79] X. Yan, L. Shen, J. Wang, X. Deng, and Z. Li, “MSG-SR-Net: A weakly supervised network integrating multiscale generation and superpixel refinement for building extraction from high-resolution remotely sensed imageries,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1012–1023, 2022.

[80] 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. 2020.

[81] J. Zhou, Y. Zheng, J. Tang, L. Jian, and Z. Yang, “FlipDA: Effective and robust data augmentation for few-shot learning,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 8646–8665.

[82] M. Gardner et al., “Evaluating models’ local decision boundaries via contrast sets,” in *Proc. ACL*, 2020, pp. 1307–1323.

[83] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[84] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, “International society for photogrammetry and remote sensing: 2D semantic labeling challenge. Working group III/4—3D scene analysis,” ISPRS, Leopoldshöhe, Germany, 2014, vol. 1, no. 4.

[85] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, “TEMPORARY REMOVAL: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings,” *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 42–55, Jan. 2019.

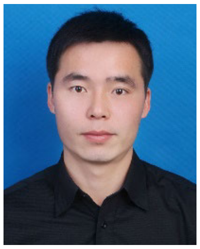
[86] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2015, pp. 770–778.

[87] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[88] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with Atrous separable convolution for semantic image segmentation,” in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.

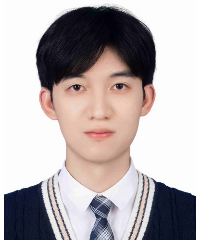
[89] K. K. Singh and Y. J. Lee, “Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[90] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.



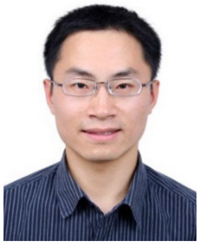
Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information systems and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a visiting student with the Informatics Institute, University of Missouri, Columbia, MO, 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. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.



Qi Su received the B.S. degree in surveying engineering from Hohai University, Nanjing, China, in 2021. He is currently pursuing the M.S. degree in remote sensing of resources and environment with Nanjing University, Nanjing.

His research interests include semantic segmentation and weakly supervised 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. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Hesse, Germany, from 2011 to 2012, and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. Since 2019, he has been a Professor with Nanjing University. He has authored four books and over 160 articles. His research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.

Wenye Wang received the B.S. degree in geographic information science from Nanjing University, Nanjing, China, in 2021, where he is currently pursuing the M.S. degree in cartography and geographical information system.

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

Zhenshi Li received the B.S. degree in geographic information science from Hohai University, Nanjing, China, in 2019, and the M.S. degree in cartography and geographic information system from Nanjing University, Nanjing, in 2022, where he is currently pursuing the Ph.D. degree in cartography and geographic information system.

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

Guangjun He was born in Hubei, China, in 1987. He received the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2015.

From 2015 to 2017, he was an Assistant Researcher Fellow with the State Key Laboratory of Space-Ground Integrated Information Technology, CAST, Beijing, China, where he has been an Associate Researcher Fellow since 2018. He is the author of more than 40 articles. His research interests include remote sensing image processing, multisensor information fusion, and machine learning.