One Model Is Enough: Toward Multiclass Weakly Supervised Remote Sensing Image Semantic Segmentation

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

***Abstract*— Semantic segmentation of remote sensing images (RSIs) is effective for large-scale land cover mapping, which heavily relies on a large amount of training data with laborious pixel-level labeling. Due to the easy availability of image-level labels, weakly supervised semantic segmentation (WSSS) based on them has attracted intensive attention. However, existing image-level WSSS methods for RSIs mainly focus on binary segmentation, which are difficult to apply to multiclass scenarios. This study proposes a comprehensive framework for image-level multiclass WSSS of RSIs, consisting of appropriate image-level label generation, high-quality pixel-level pseudo mask generation, and segmentation network iterative training. Specifically, a training sample filtering method, as well as a dataset co-occurrence evaluation metric, is proposed to demonstrate proper image-level training samples. Leveraging multiclass class activation maps (CAMs), an uncertainty-driven pixel-level weighted mask is proposed to relieve the overfitting of labeling noise in pseudo masks when training the segmentation network. Extensive experiments demonstrate that the proposed framework can achieve high-quality multiclass WSSS performance with image-level labels, which can attain 94.23% and 90.77% of the mean intersection over union (mIoU) from pixel-level labels for the ISPRS Potsdam and Vaihingen datasets, respectively. Beyond that, for the DeepGlobe dataset with more complex landscapes, the WSSS framework can achieve an accuracy close to 99% of the fully supervised case. In addition, we further demonstrate that compared to adopting multiple binary WSSS models, directly training a multiclass WSSS model can achieve better results, which can provide new thoughts to achieve WSSS of RSIs for multiclass application scenarios. Our code is publically available at https://github.com/NJU-LHRS/OME.**

***Index Terms*— Class activation map (CAM), image-level label, multiclass, pixel-level uncertainty, remote sensing image (RSI), weakly supervised semantic segmentation (WSSS).**

I. Introduction

With the vigorous development of Earth observation technology, massive remote sensing images (RSIs) at high spatial resolution are becoming available, which satisfies the fundamental requirement for timely and fine-grained land cover mapping. Semantic segmentation, aiming at assigning the land cover type to every pixel in an RSI, is a widely used and powerful technology for land cover mapping. Driven by big data and supported by deep learning methods, semantic segmentation has achieved unprecedented accuracy and efficiency [1], [2]. However, deep learning-based semantic segmentation approaches rely on an extensive amount of manually labeled annotations at pixel level, whose collection is time-consuming and labor-intensive. It is reported that labeling a single 1024 × 1024 RSI requires 2.5 h [1], which is significantly longer than the time required for labeling Cityscapes [3], a dataset for natural image semantic segmentation, indicating the difficulty of labeling RSIs.

To solve the abovementioned problem, semantic segmentation of RSIs under weak supervision has been explored [4], [5]. Compared with semantic segmentation using pixel-level labels, termed fully supervised semantic segmentation (FSSS), weakly supervised semantic segmentation (WSSS) only needs weak labels that require much lower labeling costs due to the coarse-grained form of image labeling [6]. There are several kinds of weak labeling types, including image-level annotations, points, scribbles, and bounding boxes [7]. Driven by these weak labels, WSSS has been successfully used for cloud detection [8], [9], water delineation [8], [10], building extraction [11], [12], [13], [14], [15], road extraction [16], [17], [18], and so on.

Image-level annotation only needs to point out whether a certain type of geographical objects exists in the image, without location, extent, or shape information. The lowest labeling cost attracts the most research attention among different weak labeling types for WSSS of RSIs. To achieve pixel-level semantic segmentation with only image-level annotations, image classification networks [19] are trained to produce class activation maps (CAMs) [20], which are used directly for semantic segmentation by thresholding [9], [21] or as pseudo masks to train a fully convolutional network (FCN)-based [22] segmentation network [8], [13], [15]. However, studies on image-level WSSS of RSIs primarily concern only a single category, i.e., a binary task to segment the specific target from the background, which are hardly used for tasks that need to extract multicategory information. RSIs often encounter complex scenarios involving the distribution of various objects, where multicategory scenes account for the vast majority [23]. In addition, RSI frequently necessitates the

Manuscript received 19 January 2023; revised 19 March 2023 and 29 April 2023; accepted 16 June 2023. Date of publication 28 June 2023; date of current version 12 July 2023. This work was supported by the National Natural Science Foundation of China under Grant 42071297 and Grant 41871235. *(Corresponding author: Xueliang Zhang.)*

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

Digital Object Identifier 10.1109/TGRS.2023.3290242

extraction of multiclass geographic information. For example, land cover mapping calls for multicategory semantic segmentation method.

It is easy to think that multiclass WSSS tasks can be solved by adopting a binary WSSS method several times, but it could be faced with three problems. First, compared with training one multiclass WSSS model, training multiple binary WSSS models apparently requires a multiplicate increase in time and space costs. Second, when combining different binary results into a unified multiclass result, it is difficult to impartially determine the final category for the confusing pixels with different predictions. Third, the binary WSSS model can only identify a single category and is unable to comprehensively consider the differences and relationships between different categories.

Superb research progress has been made in terms of multiclass WSSS for natural images, in which the methods mainly follow the process of generating CAMs, generating and optimizing pixel-level pseudo masks from CAMs, and training segmentation networks with pseudo masks. Compared with natural images, multiclass WSSS for RSIs faces more challenges. On the one hand, RSI patches are faced with more serious co-occurrence problems,[1] which is a frequently discussed issue in the computer vision field, seriously affecting the discriminative ability of classification networks. On the other hand, RSIs contain a variety of geographical categories, resulting in complex distribution of targets within a single image patch, increasing the difficulty for the classification network to identify objects accurately and thus generating pseudo masks with more labeling errors. A pioneering study on multiclass WSSS of RSIs was proposed based on explicit pixel-level constraints [23] and achieved good accuracies on three RSI datasets, which has, however, much room to improve due to the lack of consideration of the aforementioned issues. In addition, for multiclass WSSS, it seems that adopting a binary WSSS method multiple times can also achieve goals. Then, what is the difference between using the multiclass WSSS method and the binary WSSS methods for image-level multiclass WSSS?

To solve the abovementioned problems, we propose a comprehensive framework for image-level multiclass WSSS of RSIs, including image-level label generation, pixel-level pseudo mask generation, and segmentation network iterative training. We design a co-occurrence matrix (CM) to determine the appropriateness of image-level labels, as well as a label filtering strategy, to train classification networks with more discrimination. Furthermore, we suggest a pseudo mask optimization method, specifically driven by multiclass CAMs, to evaluate pixel-level uncertainty and weights, which can provide more accurate supervision to the segmentation network and improve the effectiveness of the subsequent iterative training. Experiments on three remote sensing datasets demonstrate that the proposed WSSS framework can achieve comparable multiclass accuracies with FSSS. Moreover, we discover that training a multiclass classification network can generate better CAMs than training multiple binary classification networks, which can provide new inspiration for WSSS of RSIs.

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

1) We propose a comprehensive and effective framework for multiclass WSSS of RSIs based on image-level labels. Evaluations on three datasets demonstrate that the WSSS framework can achieve superb results, i.e., more than 90% accuracies of FSSS.
2) Considering the particularity of RSIs, we propose a data filtering method to filter out improper training samples and an image-level label evaluation method called CM, helping train a classification network with better discriminative ability.
3) We propose a pixel-level uncertainty evaluation method driven by multiclass CAM, which is used for reweighting pseudo masks to mitigate the interference of labeling noise and further strengthen the training of segmentation networks in iterative training.
4) We discover that, compared with using multiple binary CAMs for multiclass WSSS, a unified multiclass CAM can achieve results with higher quality, which can not only indicate the meaning of multiclass WSSS but also provide more inspiration for binary WSSS tasks with image-level labels.

II. Related Work

A. WSSS for Remote Sensing Images

In the field of remote sensing, using weak label-based semantic segmentation for geographic information extraction has attracted much attention [24], in which image-level labels are used the most. On the one hand, some studies utilized CAMs directly for RSI semantic segmentation. Cropland segmentation was performed by simply thresholding binary CAMs [21]. An attention mechanism was utilized in the classification network combined with conditional random field (CRF) postprocessing to detect destruction areas [25]. The global convolutional pooling operation was proposed to enhance the representation of feature maps [9], and the pooling layers were removed when predicting CAM to improve the resolution. Multiple feature maps were extracted to capture a sequence of class-specific hierarchical saliency maps, which were then fused based on superpixel and low-rank matrix recovery to extract residential areas [26].

On the other hand, a large number of studies used CAMs to generate pseudo masks as pixel-level supervision to train segmentation networks for WSSS. Superpixels were used to improve the pseudo masks for water/cloud extraction [8] and building detection [11], [15]. Different usages of CRF were introduced to different stages and the main factors of WSSS for RSIs were also illustrated [13]. Adversarial climbing [27] and gated convolution were combined to improve CAMs to generate pseudo masks [12], which were then refined with AffinityNet and random walk strategy [28]. A deep generative model was used to improve the unbalanced distribution of foreground and background samples and an uncertainty-aware joint optimization training strategy was used to mitigate the negative impacts of noisy pseudo masks [29]. However, the

[1] An image assigned with a label of 'car' often include roads (because cars are always parking or running on the roads), which will make the classification network recognize the roads wrongly as cars.

abovementioned studies are limited to binary WSSS tasks, i.e., based on generating a single CAM of a certain category, which are difficult to apply to multiclass cases. Zhou et al. [23] explored the multicategory CAMs of RSIs for the first time. They proposed explicit pixel-level constraints based on the self-supervised equivariant attention mechanism (SEAM) framework, which preliminarily demonstrated the feasibility of image-level multiclass WSSS of RSIs.

In addition to image-level labels, other forms of weak labels are explored for WSSS of RSIs, in which the vast majority of studies still concerned with binary cases. For instance, point label is used for cropland segmentation [21], road extraction [16], and water delineation [10]. Scribble is used for road detection [17], [18], with centerline-like open source data such as the Open Street Map, and building detection [30]. Bounding box is used for building extraction [31], [32], reaching comparable performance with the fully supervised case. In addition to the abovementioned binary WSSS tasks, multiclass WSSS with these kinds of labels has also been explored for land cover mapping based on point labels [33], [34] or scribbles [35].

B. Multiclass WSSS With Image-Level Labels

Image-level WSSS has attracted much attention in the field of computer vision, in which the algorithms were developed in terms of multicategory datasets, such as VOC2012 [36] and COCO [37]. Currently, WSSS based on image-level labels mainly focuses on three aspects: optimization of CAM generation, optimization of pseudo mask generation, and improvement of segmentation network training.

Generation of CAMs is the first step to achieve WSSS with image-level labels, which is significant as the foundation of subsequent procedures. Benefiting from self-supervision methodology, imposing consistency regularization is an effective technique to improve the fineness of CAM. By applying consistency regularization on CAMs from transformed images [38], separate patches [39], complementary patches [40], local patches [41], recalibration features [42], or feature prototypes [43], more construction features and supervisory signals can be utilized to improve the CAMs. In addition, CAMs usually focus on discriminative object parts, leading many efforts to concentrate on improving the response integrity of objects. The approaches include using dilated convolution [44], random inactivation [45], erasing mechanisms [46], [47], [48], imposing restraints or disturbances [27], [49], and other techniques [50], [51], [52], [53].

Optimization of pseudo mask generation aims at refining the seed masks extracted from CAMs to generate pseudo masks with better quality. A famous refinement approach is the affinity-based method achieved by the random walk algorithm [28], [54], [55], [56], [57], which propagates the initial object regions to semantically similar pixels in the neighborhood. In addition, methods, including seed region growing [58], [59], region erasing [60], [61], pixel-level network training [49], [62], and multiple inference [48], [63], have also been proposed for generating high-quality pseudo masks.

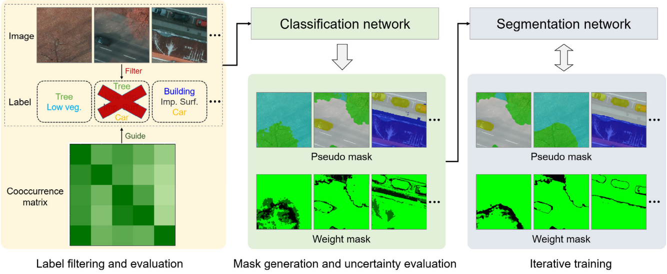


Fig. 1. Illustration of the proposed multiclass WSSS framework.

Improvement of segmentation network training aims at making full use of the supervision information of imperfect pseudo masks to train a robust segmentation model. Iterative training is a commonly used training mode [64] that can gradually improve the boundary accuracy and segmentation quality. In addition, many studies designed regularized loss functions to mine more information or mitigate the impact of labeling errors of pseudo masks, such as constraint-to-boundary loss [65], CRF loss [66], [67], and metric learning loss [68]. Considering the nature of deep networks tending to first fit the clean annotations before eventually memorizing error labels, an adaptive early learning correction method [69] was proposed to pretend overfitting and refine pseudo masks.

III. Methodology

The proposed multiclass WSSS framework is shown in Fig. 1, which consists of three steps: image-level label filtering and evaluation, pixel-level pseudo mask generation and uncertainty evaluation, and segmentation network iterative training driven by uncertainty. Given the image-level WSSS dataset, we first filter out the improper training samples, followed by a dataset evaluation method, CM, to measure whether the dataset is appropriate to train a discriminative classification network. Then, a classification network is trained with the image-level dataset to generate CAMs, by which pixel-level pseudo masks and uncertainty masks are derived. Finally, the segmentation network is trained iteratively with the pseudo masks and reweighted by the weight masks generated from the uncertainty masks to mitigate the effects of labeling noise. It is assumed that the trained segmentation network can output better pseudo masks, which can gradually improve the segmentation network through iterative training.

A. Image-Level Label Filtering and Evaluation

Given the WSSS dataset $\boldsymbol{D} = \{\boldsymbol{I}, \boldsymbol{L}\}$, consisting of N images $\boldsymbol{I} \in \boldsymbol{R}^{3\times H\times W}$and the corresponding image-level labels $\boldsymbol{L} \in \boldsymbol{R}^{1\times C}$, where C denotes the number of classes, it is inappropriate to directly train a classification network with all N samples. The RSI patches are obtained by clipping, other than the object-centric imaging modality of natural images. Hence, it is difficult to guarantee that the context information of every category is complete. For the samples with only a couple of pixels being a certain category (consider an extreme scenario where a solitary pixel of a building occupies the corner of the image), it is unreasonable to regard this kind of samples as owning the category; otherwise, it will obscure the learning of the classification network. Hence, we exclude samples with a small number of pixels for certain categories to train the classification network. The filtered dataset $\boldsymbol{D}' = \{\boldsymbol{I}', \boldsymbol{L}'\}$ can support the classification network to learn class characteristics sufficiently.

The filtered dataset $\boldsymbol{D}'$ is still not necessarily suitable for training a classification network because a severe co-occurrence problem may exist. Ideally, a dataset should possess adequate volume and abundant category distribution. In practical applications, however, the data volume is always limited, and the object distribution in RSIs often presents certain rules, such as cars always being situated on roads. Therefore, the challenge of co-occurrence in RSI datasets may hamper the classification network to distinguish different categories, such as cars and roads. We propose an evaluation criterion, called the CM, to measure the category co-occurrence degree of the dataset and determine its suitability for training a classification network.

The process of generating the CM is shown in Fig. 2. In particular, for each image $I \in \boldsymbol{I}'$, a matrix $A_{ij}(I)$ is calculated according to its image-level label L as (1), indicating the existing relationship of class i and j, i.e., $A_{ij} = 1$ if classes i and $j$both exist in the image; otherwise, 0

$$A_{ij} = \begin{cases} 1, & L_i = 1 \text{ and } L_j = 1 \\ 0, & \text{else} \end{cases} \tag{1}$$

Then, the CM of $\boldsymbol{D}'$ is calculated as (2), in which N' represents the total number of samples in $\boldsymbol{D}'$

$$\mathrm{CM}_{ij}(\boldsymbol{D}') = \frac{\sum_{N'} A_{ij}}{\sum_{N'} A_{ii}}, \quad i, j \in C. \tag{2}$$

Taking the final CM in Fig. 2 as an example, it can be seen that CM_{51} possesses a high value of 0.99, representing that for all the images with cars in the dataset, 99% of them also have the class of impervious surface, which may cause the car to tend to be judged as impervious surface. If the overall CM values of a dataset present a high level, it is difficult to train a discriminative classification network on the dataset. In this case, additional operations should be taken to relieve this, such as extending the data volume. Based on the RSI dataset with limited data volume, we generate more appropriate training samples under the guidance of the CM and achieve better results, which will be illustrated in Section IV.

B. Generating Pixel-Level Pseudo Mask and Evaluating Label Uncertainty

Leveraging the filtered dataset $\boldsymbol{D}'$, the classification network can be trained, and the CAMs are generated by applying the fully connected layer weights to the final feature map [20]. It is noted that numerous methods for improving the CAM and pseudo mask have been proposed, as summarized in Section II. In this study, we improve pseudo masks by adopting fully connected CRF processing [70] to the CAMs, as shown in (3), in which $\boldsymbol{P}$ indicates the final pseudo masks and $\bar{C}$ means the appeared categories in the CAMs $\boldsymbol{M}$

$$\boldsymbol{P} = \underset{\bar{C}}{\mathrm{Argmax}}(\mathrm{Crf}(\boldsymbol{M})). \tag{3}$$

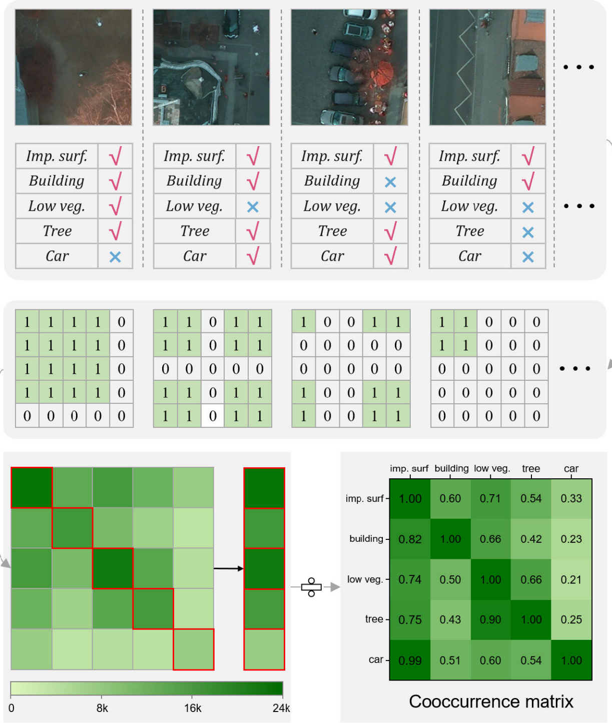


Fig. 2. Process of generating CM. “Imp. surf.” refers to impervious surface and “low veg.” refers to low vegetation.

Labeling errors are inevitably presented in pseudo masks. Using the inaccurate pseudo masks directly will make the segmentation network memorize the errors and thus affect the subsequent iterative training. A labeling noise mitigation method was proposed [71] by applying weights to the cross-entropy loss according to the calculated pixel-level uncertainty. However, they evaluated pixel uncertainty with the segmentation network, which is still trained with noisy pseudo masks. In this study, we propose a pseudo mask uncertainty evaluation method according to multiclass CAM, which can mitigate the impact of labeling noise since the initial training stage of the segmentation network and thus improve the performance in successive iterative trainings. Inspired by [71] that noise is always related to the response range of foregrounds, we discover that by scaling operation with CRF on CAMs, the variance of different scales has larger responses to the uncertain areas, which can reflect the labeling noise in the noisy pseudo masks. Driven by this observation, we evaluate pixel-level uncertainty of pseudo masks to guide noise mitigation training.

The framework for evaluating pixel-level uncertainty is shown in Fig. 3. Given image I and its CAM M, different scaling processes via exponential functions with varied power factors are applied to each class $c{\in}\bar{C}$ of M. Then, we have the processed $K{\times}\bar{C}$ CAMs M_c^k, in which $k \in K$, $c{\in}\bar{C}$, and K is the number of scales. After that, different foreground masks under different scaling processes can be extracted with the CRF and argmax operations as follows:

$$T_c^k = \left[\mathrm{Argmax}\left(\mathrm{Crf}\left(M_c^k\right)\right) == c\right], \quad \forall c \in \bar{C}, k \in K. \tag{4}$$

After obtaining scaled foreground masks, the uncertainty can be estimated by calculating variances, as shown

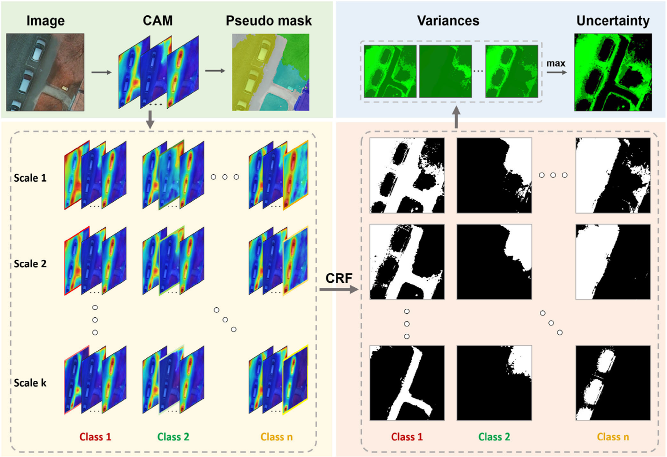


Fig. 3. Framework for evaluating pixel-level uncertainty from CAM.

in the following:

$$\mathrm{Var}(T) = \frac{1}{K-1}\sum_{k=1}^{K}\left(T_c^k - \boldsymbol{E}(T_c)\right)^2, \quad \forall c \in \bar{C}. \tag{5}$$

Finally, the max operation is applied to Var(T) to obtain the final pixel-level uncertainty, which is then normalized to [0, 1], as shown in the following:

$$U = \mathrm{Norm}\left(\underset{\bar{C}}{\mathrm{Max}}(\mathrm{Var}(T))\right). \tag{6}$$

C. Training Segmentation Network Iteratively Guided by Label Uncertainty

Iterative training is utilized when training the segmentation network [64], which means that the trained segmentation network is used to generate new pseudo masks for training a new network. For the iterative training, uncertainty is also estimated in terms of the pseudo mask generated from the segmentation networks [71]. The higher uncertainty of the pixel represents the greater probability of being wrongly labeled in the pseudo mask. Hence, the weight mask is calculated as (7), in which assigning low weights to pixels with high uncertainty can mitigate the negative effects of label noise

$$W = 1 - U. \tag{7}$$

Instead of taking the weight mask directly as the final weights, we assign the pixel-level weights Y, as shown in (8), to the cross-entropy loss of the segmentation network, in which t represents the iteration round and$t = 0$ means training with the pseudo masks generated with CAMs. The reweighting operation in (8) is based on two observations: 1) pixels with low values in W tend to be incorrectly labeled, which deserve to be ignored; and 2) pixels with higher W values tend to be correctly labeled, which are critical for training the segmentation network. It should be noted that, when training the segmentation network, we used the images in dataset D and the corresponding pseudo masks P, other than the filtered dataset $\boldsymbol{D}'$, because although the images in $\boldsymbol{D} - \boldsymbol{D}'$ are not appropriate to train the classification network, they can provide abundant information to supervise the training of the segmentation network

$$Y_{ij} = \begin{cases} 1, & 0.5 < W_{ij} < 1 \\ w_t, & 0.2 < W_{ij} < 0.5 \quad t = 0, 1, 2, \ldots \\ 0, & 0 < W_{ij} < 0.2 \end{cases} \tag{8}$$

IV. Results

A. Experimental Setup

1) Dataset Description: Three multicategory RSI datasets are used to evaluate the proposed WSSS framework, which are described as follows.

1) *ISPRS Potsdam Dataset:* It is composed of 38 orthorectified images with a size of 6000 × 6000 pixels and a spatial resolution of 5 cm. Three bands of IRRG (infrared, red, and green) are extracted from the original four-band IRRGB (infrared, red, green, and blue). The class of clutter is ignored for both training and testing, and thus, the dataset includes five land cover types: impervious surfaces (imp. surf.), building, low vegetation (low veg.), tree, and car. Following the official data division, 23 images except for the image named 7_10 (removed due to its error annotations) are used for training, and 14 images are used for testing. The images are cropped into 256 × 256 pixels with 128 pixels overlapped for augmentation. The patches for training the classification network are filtered by removing the samples with pixels of any category smaller than 10% (2.5% for the small object of car) of the total pixels. In terms of the segmentation network, we use nonoverlapping patches for both training and testing to ensure time savings and evaluation fairness.
2) *ISPRS Vaihingen Dataset:* It contains 33 IRRG images with an average size of 2494 × 2064 pixels and a spatial resolution of 9 cm. The categories are the same as the Potsdam dataset. According to the official data division method, 16 images are used for training and 17 images are used for testing. The images are cropped into 128 × 128 pixels with 64 overlapping pixels, with other settings the same as the Potsdam dataset.
3) *DeepGlobe Land Cover Classification Challenge Dataset:* It is the first public dataset with high-resolution submeter satellite imagery focusing on rural areas. There are 803 RGB images with a size of 2448 × 2448 pixels published with pixel-level labels. The dataset is annotated with seven categories: urban, agriculture, rangeland, forest, water, barren, and unknown. We split the dataset into training/validation/test sets with 563/120/120 images based on the naming order. The images are cropped into 306 × 306 pixels without overlay. The filtering ratio is set as 10% for all the categories.

2) Implementation Details: For the classification network used for generating CAMs and pseudo masks, several networks are used to demonstrate the reliability of the proposed WSSS framework. The methods include the original CAM (Ori-CAM) [20] with ResNet101 [72] as

TABLE I

COMPARISON OF THE PSEUDO MASK mIoU (%) FROM MULTICLASS AND BINARY CAMs

Dataset	SEAM		Ori-CAM		Puzzle-CAM		DRS		VWL	
	Multiclass	Binary	Multiclass	Binary	Multiclass	Binary	Multiclass	Binary	Multiclass	Binary
Potsdam	65.27	61.54	63.35	60.55	66.16	47.69	62.15	41.57	67.94	66.42
Vaihingen	60.66	55.23	58.49	58.68	61.60	55.37	48.78	34.25	60.72	59.52
DeepGlobe	70.99	67.53	73.76	71.09	70.74	56.66	62.77	60.28	70.39	67.41

the backbone network, the SEAM framework [38] with ResNet38 [73] as the backbone network, Puzzle-CAM [39] with ResNet50, discriminative region suppression (DRS) method [49] with VGG16, and visual words learning (VWL) method [51] with ResNet50. In the training stage, the batched stochastic gradient descent (SGD) optimizer is utilized for 20 epochs, with momentum as 0.9 and weight decay as 0.0005. The initial learning rate is set as 0.01, with the learning rate decay strategy of "poly." For the segmentation network, DeepLabV3+ [74] is used, with output stride 8 and ResNet50 as the backbone. The Adam optimizer is used for 30 epochs, with an initial learning rate of 0.0002. All the backbones are initialized by the pretrained model on ImageNet to enhance feature extraction abilities. All the experiments are conducted on a computer with an Intel Core i7-11700K CPU, one NVIDIA GeForce RTX 3080 GPU, and 64-GB memory.

3) Evaluation Metrics: To evaluate the multiclass WSSS performance, the mean intersection over union (mIoU) and the mean F1-score (mF1) are adopted, with the following formulas:

$$\text{mIoU} = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{\text{TP}}{\text{TP}+\text{FN}+\text{FP}} \tag{9}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$\text{m}F1 = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{12}$$

where TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively, and N_c is the number of categories.

B. Performance of Multiclass WSSS

1) Comparison Between Multiclass and Single-Class WSSS: Multiclass WSSS tasks may seem solvable by training multiple binary WSSS models, which have seen considerable development. However, compared to using binary models for multiclass WSSS, it is obvious that one multiclass method tends to require much lower costs in time and space. More importantly, we demonstrate in this section that a multiclass method can achieve better performance than binary methods. Specifically, the accuracies of pseudo masks from a multiclass classification network and binary networks are compared, which are decisive for WSSS performance [13]. The multiclass and binary networks are consistent except for the number of output layers: N_c for multiclass and 1 for binary. The training and testing datasets are also kept consistent for fairness, i.e., the datasets used for the binary tasks are derived from the multiclass dataset with the same image-level labels (0 for absence and 1 for presence of certain class). In the multiclass case, a pseudo mask is generated by applying argmax to the multiclass CAM, which is directly generated by a single multiclass classification network. In the binary case, CAMs of different categories are obtained by multiple binary classification networks, which are concatenated followed by the argmax operation to generate pseudo masks.

TABLE II

COMPARISON OF THE FOREGROUND IoU (%) OF DIFFERENT CLASSES FROM THRESHOLDING SINGLE-CLASS CAMs AND APPLYING ARGMAX TO MULTICLASS CAMs OF THE ISPRS DATASET. THE CAM METHOD IS SEAM

	Potsdam		Vaihingen	
	Thresholding	Argmax	Thresholding	Argmax
Imp. surf.	56.64	69.65	63.09	68.60
Building	62.68	82.79	71.61	77.30
Low veg.	54.43	67.26	49.53	54.70
Tree	58.64	63.33	60.13	63.71
Car	49.41	43.30	46.81	38.97
Average	56.36	65.27	58.23	60.66

Table I presents the pseudo mask accuracies from different CAM methods for three datasets. The accuracies of pseudo masks generated by the multiclass network are apparently higher than those generated by the binary networks, whether for different datasets or classification networks, demonstrating the advantage of training a unified multiclass classification network for identifying different categories, which can be attributed to two reasons. On the one hand, the classification network trained by multicategory labels can identify different classes, whereas the binary network can only identify a single category and is unable to comprehensively consider the differences and relationships between different categories. On the other hand, the CAMs of different classes from the multiclass network are obtained through the same network backbone, but the binary networks may generate CAMs at different levels, which are improper to be integrated together.

For binary image-level WSSS, pseudo masks are always generated by introducing a threshold to the CAM to extract foregrounds and backgrounds [12], [13]. For the multicategory CAM, however, pseudo masks can be generated by considering different classes, rather than just thresholding a single-class CAM. Tables II and III compare the foreground accuracies

TABLE III

COMPARISON OF THE FOREGROUND IoU (%) OF DIFFERENT CLASSES FROM THRESHOLDING SINGLE-CLASS CAMs AND APPLYING ARGMAX TO MULTICLASS CAMs OF THE DeepGlobe DATASET. THE CAM METHOD IS ORI-CAM

	Thresholding	Argmax
Urban	63.72	72.56
Agriculture	82.79	89.32
Rangeland	45.52	56.32
Forest	79.80	84.81
Water	57.58	71.46
Barren	63.60	76.97
Unknown	58.22	64.87
Average	64.46	73.76

TABLE IV

COMPARISON OF THE INFERENCE TIME (MS) FOR AN IMAGE WITH 256 × 256 PIXELS USING MULTICLASS AND BINARY CAM METHODS

	Ori-CAM	SEAM	Puzzle-CAM	DRS	VWL
Multiclass	8.7	20.4	4.4	5.6	7.9
Binary	9.0	20.6	4.3	5.3	7.2

of different classes obtained through argmax and thresholding on the ISPRS and DeepGlobe datasets, respectively, in which the thresholds are set as the optimal values leading to the best IoU of the foregrounds. The results show that, across all three datasets, the argmax operation is more effective in extracting foregrounds with higher accuracies, in which the largest mIoU gap of 8.91% is achieved for the ISPRS Potsdam dataset, demonstrating that leveraging the interaction of multiple CAMs can result in pseudo masks with higher quality. This not only proves the benefits of the multiclass WSSS model over binary models for identifying different categories but also provides new inspiration to improve binary WSSS.

In addition, we have calculated the inference time for various classification networks in both multiclass and binary scenarios, as shown in Table IV. The only distinction between these two types of networks is the number of output layers, i.e., N_c for multiclass and 1 for binary. It is worth noting that there is virtually no discernible difference in prediction time between the two cases. Hence, the multiclass method can apparently save more time than the binary method because the latter needs inferencing multiple times.

2) Comparison Between WSSS and FSSS: The performance difference between WSSS and FSSS can clearly reflect the effectiveness of the WSSS method. Table V shows the final accuracies of the proposed WSSS framework compared with that of FSSS on the ISPRS Potsdam and Vaihingen datasets. The mIoU gaps are 5.06% and 7.53% for Potsdam and Vaihingen, respectively, and the WSSS accuracies can achieve 94.23% and 90.77% of FSSS, as indicated by mIoU. Such a small accuracy gap reflects the applicability of the proposed multiclass WSSS method. The class of building presents the best extraction accuracy for both WSSS and FSSS, and the class of car presents the lowest accuracy due to the co-occurrence problem, i.e., cars always coexist with impervious surfaces, making them more inclined to be misclassified as impervious surfaces. In general, the WSSS method can achieve comparable performance with FSSS in terms of all the classes. Figs. 4 and 5 show the visual WSSS results of the ISPRS Potsdam and Vaihingen datasets, respectively, which demonstrates that the proposed WSSS framework can not only achieve high segmentation accuracies but also draw clear boundaries.

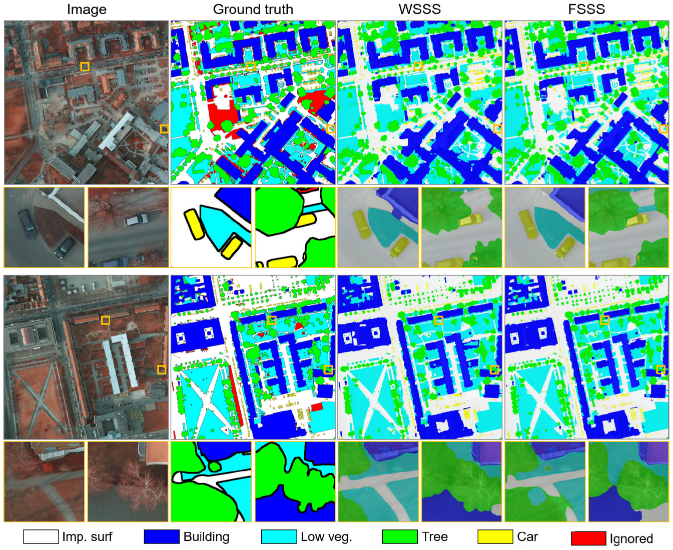


Fig. 4. Examples of segmentation results from WSSS and FSSS methods on the ISPRS Potsdam dataset.

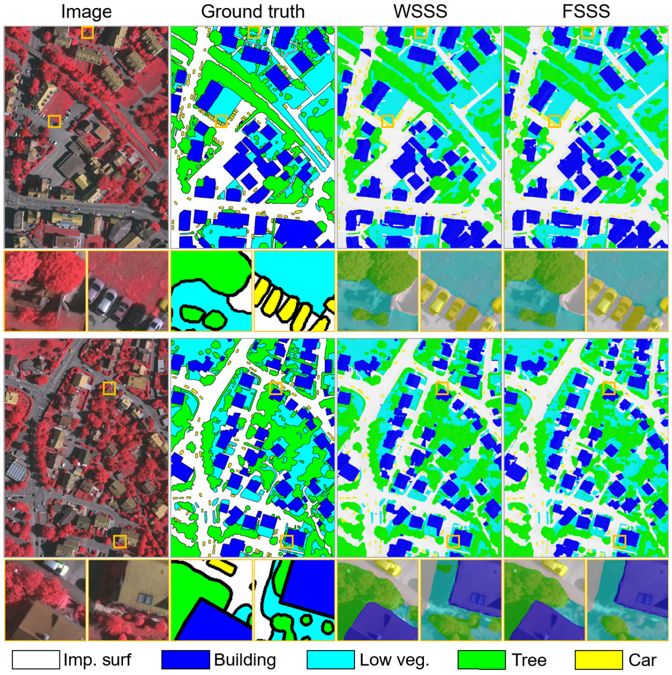


Fig. 5. Examples of segmentation results from WSSS and FSSS methods on the ISPRS Vaihingen dataset.

The accuracies of WSSS and FSSS on the DeepGlobe validation (val) and test sets are shown in Table VI. The results reveal a very small accuracy gap between the two methods,

TABLE V

ACCURACIES AND PER CLASS IoU (%) OF THE PROPOSED WSSS FRAMEWORK ON THE ISPRS POTSDAM AND VAIHINGEN DATASETS COMPARED WITH THOSE OF FSSS. THE CAM METHOD IS SEAM

Dataset	Supervision	Imp. surf.	Building	Low veg.	Tree	Car	mIoU	mF1
Potsdam	WSSS	86.31	91.97	74.95	74.42	85.86	82.70	90.38
	FSSS	90.13	95.06	79.53	79.67	94.43	87.76	93.34
Vaihingen	WSSS	82.46	87.60	64.98	74.48	60.55	74.01	84.67
	FSSS	86.14	90.99	72.23	80.80	77.56	81.54	89.69

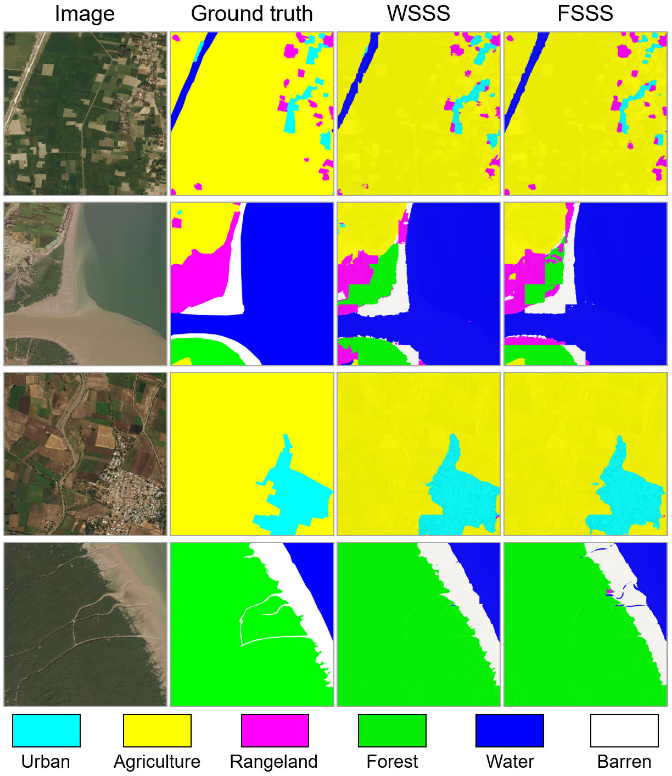


Fig. 6. Examples of segmentation results from WSSS and FSSS methods on the DeepGlobe dataset.

with only 0.72%/1.31% mIoU and 0.52%/0.78% mF1 for the val/test set. We assume that this tiny accuracy gap is due to the coarse pixel-level ground truth labels, which can be shown in the second row of Fig. 6. Notably, the pseudo masks generated by the WSSS framework demonstrate similar or even better performance than the ground truth, resulting in comparable segmentation performance, as shown in Fig. 6. Such a negligible gap not only demonstrates the effectiveness of the proposed WSSS framework but also highlights the huge potential of using WSSS for coarse-grained land cover mapping, that is, it may be more efficient to simply label image-level weak annotations than to spend significant time producing coarse pixel-level labels.

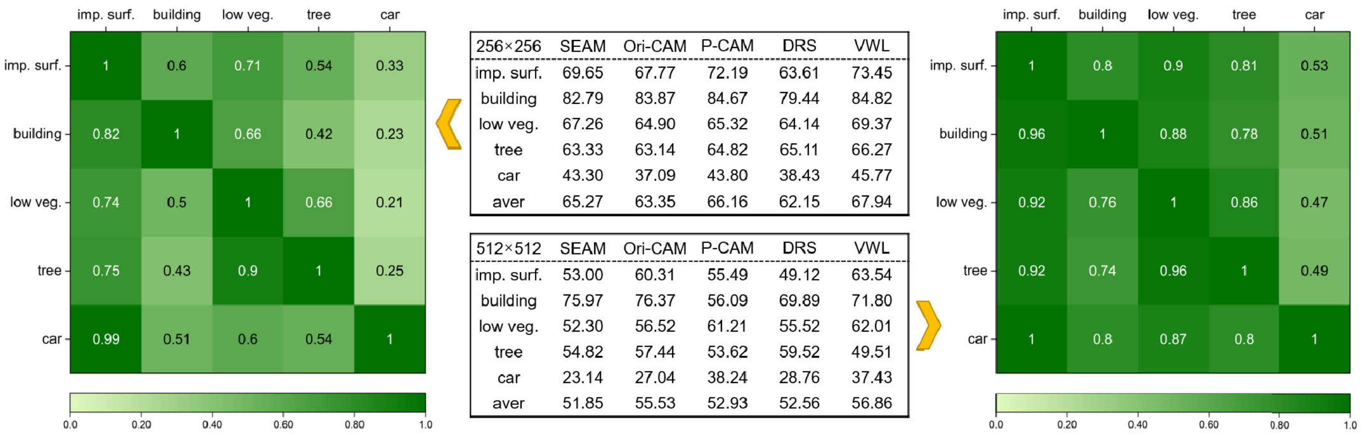
C. Component Analysis

1) Ablation Study: The proposed WSSS framework comprises multiple processes. To verify the effectiveness of different modules, extensive experiments are conducted with varying settings. Table VII shows the WSSS accuracies on different datasets from the segmentation network trained by different pseudo masks, in which "base" indicates the pseudo mask from argmax, "CRF" indicates the improvement for the base mask, "mask-w" utilizes weight mask for the pseudo mask, "iteration" indicates segmentation network training iteratively, and "iteration-w" utilizes weight mask in iterative training.

Based on the pseudo masks improved by CRF, introducing the pixel-level weights from the uncertainty masks can reduce the negative effects of labeling errors and thus enhance the WSSS accuracies. As shown in Table VII, the weight mask is effective for all the datasets, in which 0.74% and 1.03% improvements are yielded for the ISPRS Potsdam and Vaihingen datasets, respectively. The iterative training works for the ISPRS Potsdam and Vaihingen datasets, and iterative training with the help of weight masks can achieve the best accuracies for the two datasets, leading to the best mIoU of 82.70% and 74.01% for Potsdam and Vaihingen, respectively. However, iterative training does not work for the DeepGlobe dataset. This is because the segmentation network can only generate coarse-grained results on this dataset, and the pseudo masks have already attained the upper limit, which can hardly be further improved.

2) Effectiveness of Generating Appropriate Image-Level Labels: For image-level WSSS, the image-level annotation provides the fundamental supervision information, which determines the final WSSS performance. For multiclass WSSS of RSIs, it is critical to yield appropriate image-level annotations for training a classification network with discriminative ability. In terms of generating appropriate image-level labels, we propose a label filtering method and a label co-occurrence estimation method named the CM, the effects of which are illustrated in this section.

The accuracy of the pseudo masks determines the segmentation network training and, thus, the performance of WSSS [13]. Table VIII shows the accuracies of the pseudo masks generated from the classification network trained by the image-level labels before and after filtering. In general, label filtering can apparently improve the accuracies of pseudo masks for different CAM methods on all three datasets (except for the DRS method on the Vaihingen dataset which presents extremely low accuracies for both cases), demonstrating the effectiveness of label filtering for improving the identification ability of the classification network. The biggest improvement comes from the DeepGlobe dataset with an Ori-CAM of



	imp. surf.	building	low veg.	tree	car
imp. surf.	1	0.6	0.71	0.54	0.33
building	0.82	1	0.66	0.42	0.23
low veg.	0.74	0.5	1	0.66	0.21
tree	0.75	0.43	0.9	1	0.25
car	0.99	0.51	0.6	0.54	1

256×256	SEAM	Ori-CAM	P-CAM	DRS	VWL
imp. surf.	69.65	67.77	72.19	63.61	73.45
building	82.79	83.87	84.67	79.44	84.82
low veg.	67.26	64.90	65.32	64.14	69.37
tree	63.33	63.14	64.82	65.11	66.27
car	43.30	37.09	43.80	38.43	45.77
aver	65.27	63.35	66.16	62.15	67.94

512×512	SEAM	Ori-CAM	P-CAM	DRS	VWL
imp. surf.	53.00	60.31	55.49	49.12	63.54
building	75.97	76.37	56.09	69.89	71.80
low veg.	52.30	56.52	61.21	55.52	62.01
tree	54.82	57.44	53.62	59.52	49.51
car	23.14	27.04	38.24	28.76	37.43
aver	51.85	55.53	52.93	52.56	56.86

	imp. surf.	building	low veg.	tree	car
imp. surf.	1	0.8	0.9	0.81	0.53
building	0.96	1	0.88	0.78	0.51
low veg.	0.92	0.76	1	0.86	0.47
tree	0.92	0.74	0.96	1	0.49
car	1	0.8	0.87	0.8	1

Fig. 7. CM of different sets of image-level labels and the corresponding pseudo mask accuracies (IoU, %) on the Potsdam dataset. "P-CAM" indicates "Puzzle-CAM".

TABLE VI

Accuracies and Per Class IoU (%) of the Proposed WSSS Framework on the DeepGlobe Dataset Compared With That of FSSS, in Which "Val" and "Test" Represent the Validation Data and the Test Set, Respectively. The CAM Method Is Ori-CAM

Dataset	Supervision	Urban	Agriculture	Rangeland	Forest	Water	Barren	Unknown	mIoU	mF1
DeepGlobe-val	WSSS	75.10	86.90	35.74	78.06	72.02	66.37	46.62	65.83	78.03
	FSSS	74.49	87.59	37.94	78.44	76.41	66.35	44.64	66.55	78.55
DeepGlobe-test	WSSS	75.19	86.22	40.88	79.62	76.82	56.84	53.24	66.97	79.14
	FSSS	77.30	87.51	43.18	82.60	81.34	57.34	48.68	68.28	79.92

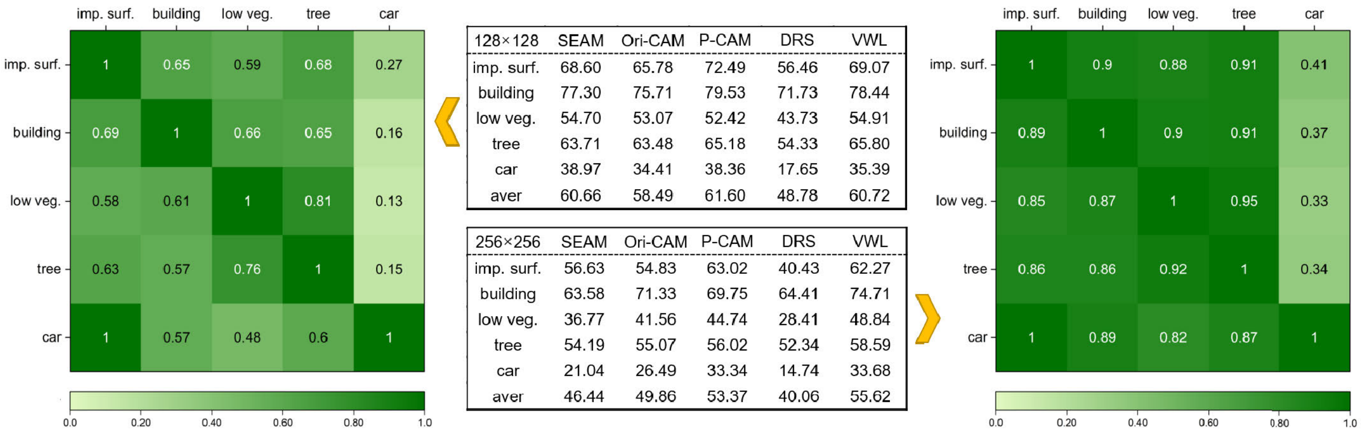
6.01% mIoU. Table VIII shows that the Vaihingen dataset obtains the lowest improvement among the three datasets, which can be attributed to the small data volume of this dataset.

Although the image-level label after filtering includes clearer category characteristics, the co-occurrence problem may still exist. We propose the method of CM to evaluate the co-occurrence of the image-level labels, which can be used to guide the generation of better image-level labels. To illustrate its effectiveness, we crop the images of the two ISPRS datasets into patches of different sizes to build different datasets. Specifically, we crop the images of the Potsdam dataset into sizes of 256 × 256 and 512 × 512 pixels and those of the Vaihingen dataset into sizes of 128 × 128 and 256 × 256 pixels. The pseudo mask accuracies and the corresponding CM of different cropped datasets are shown in Figs. 7 and 8 in terms of the Potsdam and Vaihingen datasets, respectively.

As shown in Figs. 7 and 8, the dataset with a larger patch size faces a more severe co-occurrence phenomenon, indicated by the larger values in the CM. For all the CAM methods, image-level labels with more severe co-occurrence problems lead to pseudo masks with lower accuracies in terms of both datasets. For example, for the Potsdam/Vaihingen dataset, the image-level labels with less co-occurrence yield pseudo masks with 13.42%/14.22% higher mIoU for SEAM, demonstrating the impact of co-occurrence on training the fine classification network. In a word, the proposed CM can indicate the co-occurrence problem of the image-level labels and thus provide guidance to generate appropriate weak labels for land cover mapping of RSIs.

3) Analysis of Pseudo Mask Uncertainty: The weight masks are generated from the uncertainty masks, which can mitigate the effects of noisy labels in the pseudo masks. The weight mask is used in two specific areas: one is the pseudo mask generated from the CAM and the other is that from the segmentation network in iterative training. Fig. 9 shows the examples of the visualization for the weight masks in the two usages. It can be seen that the low weight pixels tend to be incorrectly labeled in the pseudo masks, illustrating the effectiveness of weight masks for mitigating interference from labeling errors. The pseudo masks from the CAMs (iter 0 in Fig. 9) are the initial masks and thus contain more error labels. After the iterative training by the segmentation network, the error pixels decrease as the iteration grows, and the pseudo masks become increasingly accurate. The object boundaries always possess low weights, which are more likely to be incorrectly labeled.

4) Effects of Iterative Training: Iterative training is a commonly used strategy that can gradually improve the pseudo mask. Li et al. [71] introduced uncertainty-driven mask weights into segmentation network iterative training to mitigate the impacts of noisy labels, but their initial pseudo masks for training the segmentation network are still noisy, which hampers the segmentation network from learning accurate information and thus interferes with the subsequent iterative training. We refer to the iterative training of [71] but introduce



	imp. surf.	building	low veg.	tree	car
imp. surf.	1	0.65	0.59	0.68	0.27
building	0.69	1	0.66	0.65	0.16
low veg.	0.58	0.61	1	0.81	0.13
tree	0.63	0.57	0.76	1	0.15
car	1	0.57	0.48	0.6	1

128×128	SEAM	Ori-CAM	P-CAM	DRS	VWL
imp. surf.	68.60	65.78	72.49	56.46	69.07
building	77.30	75.71	79.53	71.73	78.44
low veg.	54.70	53.07	52.42	43.73	54.91
tree	63.71	63.48	65.18	54.33	65.80
car	38.97	34.41	38.36	17.65	35.39
aver	60.66	58.49	61.60	48.78	60.72

256×256	SEAM	Ori-CAM	P-CAM	DRS	VWL
imp. surf.	56.63	54.83	63.02	40.43	62.27
building	63.58	71.33	69.75	64.41	74.71
low veg.	36.77	41.56	44.74	28.41	48.84
tree	54.19	55.07	56.02	52.34	58.59
car	21.04	26.49	33.34	14.74	33.68
aver	46.44	49.86	53.37	40.06	55.62

	imp. surf.	building	low veg.	tree	car
imp. surf.	1	0.9	0.88	0.91	0.41
building	0.89	1	0.9	0.91	0.37
low veg.	0.85	0.87	1	0.95	0.33
tree	0.86	0.86	0.92	1	0.34
car	1	0.89	0.82	0.87	1

Fig. 8. CM of different sets of image-level labels and the corresponding pseudo mask accuracies (IoU, %) on the Vaihingen dataset. “P-CAM” indicates “Puzzle-CAM”.

TABLE VII

Ablation Study for the Proposed WSSS Framework, Showing the mIoU (%) on Different Datasets From the Segmentation Network Trained by Different Pseudo Masks

Base	CRF	Mask-w	Iteration	Iteration-w	Potsdam	Vaihingen	Deepglobe-val	Deepglobe-test
✓					75.67	70.37	64.41	63.92
✓	✓				78.53	71.27	65.33	66.85
✓	✓	✓			79.27	72.30	**65.83**	**66.97**
✓	✓	✓	✓		81.77	73.99	65.02	62.88
✓	✓	✓		✓	**82.70**	**74.01**	64.15	63.57

TABLE VIII

Accuracies (mIoU, %) of the Pseudo Masks Generated From the Classification Network Trained by the Image-Level Labels Before and After Filtering

Dataset	SEAM		Ori-CAM		Puzzle-CAM		DRS		VWL	
Filtering	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Potsdam	62.28	65.27	61.46	63.35	61.93	66.16	61.78	62.15	66.15	67.94
Vaihingen	60.53	60.66	56.38	58.49	58.88	61.60	49.56	48.78	60.06	60.72
DeepGlobe	65.66	70.99	67.75	73.76	68.26	70.74	62.38	62.77	66.98	70.39

the weight masks, which are obtained from multiclass CAMs, into the initial pseudo masks to mitigate the noise effect from the fundamental process. In this section, the effectiveness of iterative training in the proposed framework is illustrated.

The accuracies of the segmentation network iterative training on the Potsdam and Vaihingen datasets are shown in Fig. 10. It is noted that the iterative training does not work for the DeepGlobe dataset. It can be found in Fig. 10 that, by comparing the red line with the blue line, introducing uncertainty weight masks into the initial pseudo masks can improve the segmentation performance to a great extent for both datasets, which can demonstrate the benefits of the proposed noise mitigation strategy. Overlooking the noise interference of the initial pseudo masks will reduce the upper limit of iterative training because of the error accumulation effect. Furthermore, comparing the red line with the yellow line shows that the introduction of uncertainty weight masks into the iterative training process will further improve the performance of WSSS. Therefore, combining the uncertainty mask weights in terms of the initial pseudo masks and iterative training can lead to better segmentation network.

V. Discussion

Multiclass-oriented tasks can be decomposed into several binary tasks, which are often easier to handle. In contrast to binary-class scenarios, WSSS faces multiple challenges when dealing with multiclass image-level labels of RSIs, including complex object distribution, indistinct object characteristics, and prevalent co-occurrence issues. As a result, some people may stereotype that multiclass WSSS can be resolved by repeatedly utilizing binary WSSS models. In this study, however, we reveal a surprising finding: directly training a

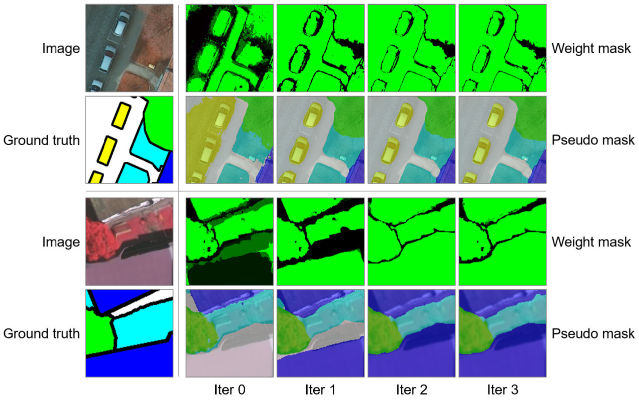


Fig. 9. Visualization of weight masks and pseudo masks. Iter 0 means the pseudo mask from CAM, and iter 1–3 mean that from the segmentation network of different iteration rounds. Darker pixel in the weight mask indicates lower weights.

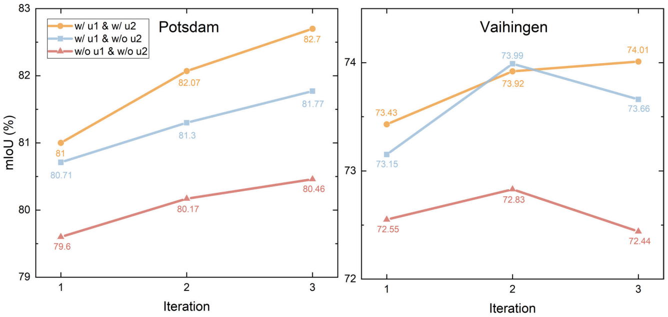


Fig. 10. Accuracies of segmentation network iterative training on Potsdam and Vaihingen datasets. “u1” and “u2” indicate using uncertainty weights in the initial pseudo masks from CAMs and that in the segmentation network iterative training.

multiclass classification network can generate pseudo masks with higher accuracies by comparison with that from multiple binary classification networks. Besides, we discovered that higher accuracies of foregrounds for different categories can be always obtained by considering multiple classes, rather than considering only one class such as adopting thresholding.

The proposed WSSS framework can achieve high-quality results comparable to the fully supervised case, mainly due to three important factors. First, inappropriate training samples are filtered out to build a better dataset with distinct category characteristics, leading to the trained classification network being able to accurately recognize different objects, which lays a good foundation for generating accurate pseudo masks. Second, the interference of the inevitably existing noise in pseudo masks is mitigated by using uncertainty-driven weighted masks to help the trained segmentation network better identify objects. Third, iterative training with weighted masks can gradually improve the pseudo masks with a high upper limit because the segmentation network is trained with the weighted masks from the beginning and can obtain further improvements in iterative training.

All the strategies and methods in the proposed framework can provide references to use image-level WSSS for land cover mapping. In particular, it is essential to filter the improper training samples to ensure that the fine characteristics of each category are preserved, rather than using all samples at once. The CM could be used to evaluate the co-occurrence situation of the training data. For datasets with severe co-occurrence, measures, such as data expansion or cropping into smaller image patches, could be useful. The uncertainty mask used for reweighting the pseudo mask is directly obtained from multiclass CAM, which is tailored for the multiclass WSSS task and easy to generate without introducing extra modules. Iterative training with weighted masks can gradually improve the pseudo masks and thus the segmentation network, which may not be suitable for certain datasets.

VI. Conclusion

This study proposed an elaborate framework for image-level WSSS of RSIs in multiclass scenarios. By introducing sample filtering to eliminate unsuitable samples and guided by the proposed CM, image-level training samples with more suitability for multiclass WSSS of RSIs can be collected. To mitigate the noise interference of noisy pseudo masks, an uncertainty-driven pixel-level weight mask generation method based on multiclass CAMs is proposed, which can significantly improve the upper limit of segmentation network iterative training and thus the final segmentation accuracy. In the context of the prevalence of binary image-level WSSS of RSIs, we discover that directly training a multiclass classification network can generate better pseudo masks by comparison with training multiple binary classification networks, demonstrating the benefits of studying multiclass WSSS methods. Elaborate experiments demonstrate that the proposed framework can achieve high-quality semantic segmentation with image-level labels comparable to that of pixel-level labels, which can attain, specifically, exceed 90% for ISPRS Potsdam and Vaihingen datasets and near 99% for the DeepGlobe dataset in terms of mIoU. We expect that our method and discovery can provide a reference for technology and ideas to achieve large-scale high-quality WSSS of RSIs for land cover mapping. In future work, we are devoted to developing specific methods to solve the co-occurrence problem and end-to-end WSSS methods for RSIs.

References

[1] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, “EarthNets: Empowering AI in Earth observation,” 2022, *arXiv:2210.04936*.

[2] X. Yuan, J. Shi, and L. Gu, “A review of deep learning methods for semantic segmentation of remote sensing imagery,” *Exp. Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.

[3] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[4] J. Yue et al., “Optical remote sensing image understanding with weak supervision: Concepts, methods, and perspectives,” *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 250–269, Jun. 2022.

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

[6] 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, Sep. 2020.

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

[8] 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.

[9] 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.

[10] M. Lu, L. Fang, M. Li, B. Zhang, Y. Zhang, and P. Ghamisi, "NFANet: A novel method for weakly supervised water extraction from high-resolution remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617114.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] R. Lian and L. Huang, "Weakly supervised road segmentation in high-resolution remote sensing images using point annotations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4501013.

[17] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602312.

[18] M. Zhou, H. Sui, S. Chen, J. Liu, W. Shi, and X. Chen, "Large-scale road extraction from high-resolution remote sensing images based on a weakly-supervised structural and orientational consistency constraint network," *ISPRS J. Photogramm. Remote Sens.*, vol. 193, pp. 234–251, Nov. 2022.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[20] 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.

[21] 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.

[22] 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.

[23] R. Zhou et al., "Weakly supervised semantic segmentation in aerial imagery via explicit pixel-level constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634517.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Y. Liu and L. Zhang, "Weakly supervised region of interest extraction based on uncertainty-aware self-refinement learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628416.

[30] 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.

[31] R. Guo et al., "JMLNet: Joint multi-label learning network for weakly supervised semantic segmentation in aerial images," *Remote Sens.*, vol. 12, no. 19, p. 3169, Sep. 2020.

[32] M. U. Rafique and N. Jacobs, "Weakly supervised building segmentation from aerial images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3955–3958.

[33] L. Wu, L. Fang, J. Yue, B. Zhang, P. Ghamisi, and M. He, "Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 7419–7434, 2022.

[34] W. Zhang, P. Tang, T. Corpetti, and L. Zhao, "WTS: A weakly towards strongly supervised learning framework for remote sensing land cover classification using segmentation models," *Remote Sens.*, vol. 13, no. 3, p. 394, Jan. 2021.

[35] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

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

[38] 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.

[39] S. Jo and I. 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] F. Zhang, C. Gu, C. Zhang, and Y. Dai, "Complementary patch for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7222–7231.

[41] P. Jiang, Y. Yang, Q. Hou, and Y. Wei, "L2G: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16865–16875.

[42] 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, vol. 36, no. 2, pp. 2117–2125.

[43] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Weakly supervised semantic segmentation by pixel-to-prototype contrast," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4310–4319.

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

[45] 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.

[46] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[47] H. Kweon, S. Yoon, H. Kim, D. Park, and K. Yoon, "Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6974–6983.

[48] W. Sun, J. Zhang, and N. Barnes, "Inferring the class conditional response map for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2653–2662.

[49] B. Kim, S. Han, and J. Kim, "Discriminative region suppression for weakly-supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1754–1761.

[50] Z. Chen, T. Wang, X. Wu, X. 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.

[51] L. Ru, B. Du, Y. Zhan, and C. Wu, "Weakly-supervised semantic segmentation with visual words learning and hybrid pooling," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 1127–1144, Apr. 2022.

[52] 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.

[53] T. Zhou, M. Zhang, F. Zhao, and J. Li, "Regional semantic contrast and aggregation for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4289–4299.

[54] 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.

[55] 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.

[56] 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)*, vol. 12371, 2020, pp. 347–362.

[57] 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.

[58] Y. Chong, X. Chen, Y. Tao, and S. Pan, "Erase then grow: Generating correct class activation maps for weakly-supervised semantic segmentation," *Neurocomputing*, vol. 453, pp. 97–108, Sep. 2021.

[59] 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.

[60] A. Chaudhry, P. K. Dokania, and P. H. S. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," 2017, *arXiv:1707.05821*.

[61] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6488–6496.

[62] P. Jiang, Q. Hou, Y. Cao, M. Cheng, Y. Wei, and H. Xiong, "Integral object mining via online attention accumulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2070–2079.

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

[64] Y. Wei et al., "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[65] 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.

[66] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11220, 2018, pp. 524–540.

[67] 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.*, vol. 34, Apr. 2020, pp. 12765–12772.

[68] T.-W. Ke, J.-J. Hwang, and S. X. Yu, "Universal weakly supervised segmentation by pixel-to-segment contrastive learning," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[69] S. Liu, K. Liu, W. Zhu, Y. Shen, and C. Fernandez-Granda, "Adaptive early-learning correction for segmentation from noisy annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2596–2606.

[70] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[71] 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, vol. 36, no. 2, pp. 1447–1455.

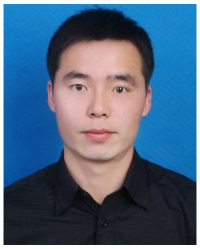
[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[73] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.

[74] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211, 2018, pp. 833–851.

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.



Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information system 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.

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.