Which CAM is Better for Extracting Geographic Objects? A Perspective From Principles and Experiments

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

***Abstract*—As a method of deep learning interpretability, class activation mapping (CAM) is efficient and convenient for extracting geographic objects supervised by image-level labels. However, in addition to the inherent problem of inaccuracy and incompleteness of CAM, we have to deal with the spectral and spatial variance of geographic objects when applying CAM methods to remote sensing images. To explore the capabilities of CAM methods on extracting various geographic objects, we make a comprehensive comparison of five commonly-used CAM methods, including original CAM, GradCAM, GradCAM++, SmoothGradCAM++, and Score-CAM, in four aspects: efficiency; accuracy; effectiveness on dealing with the spectral and spatial variance; and performance of delineating different geographic object categories. The results demonstrate that the original CAM, GradCAM, and GradCAM++ achieves the highest efficiency, accuracy, and integrity for extracting geographic objects, respectively, which can help us choose the appropriate CAM methods according to the specific requirements of different extraction tasks. Benefiting from the capability in extracting various geographic objects and adaptability in complex scenes, GradCAM achieves the best performance in dealing with the spectral and spatial variance problem and shows the advantage of capturing object details and keeping object completeness at the same time. In addition to the comparison experiments and suggestions, we also provide the principle explanations of the performance differences. The findings of this article could contribute to a deep understanding of different CAM methods and benefit to selecting suitable CAM methods for extracting geographic objects from the perspectives of both principles and experiments.**

***Index Terms*—Class activation mapping (CAM), convolutional neural network (CNN), deep learning, geographic objects, weakly-supervised semantic segmentation.**

I. Introduction

IN RECENT years, fully convolutional network (FCN) [1] with the supervision of rich pixel-level labels has been widely used for extracting numerous geographic objects, e.g., buildings [2]–[4], roads [5]–[7], vegetation [8]–[10], clouds [11]–[13], vehicles [11], [12], as well as land cover mapping [13]–[15]. However, collecting the huge amount of pixel-level labels for training an FCN is time-consuming and even demands expertise and fieldwork, especially when extracting geographic objects with large spatial ranges. To solve the high-cost problem, Zhou *et al.* [16] visualized deep features in convolutional neural network (CNN) by mining the high-level semantic information and proposed the class activation mapping (CAM) method with CNN interpretability. Instead of training FCN with pixel-level labels, CAM uses only image-level weak labels and a classification network to localize geographic objects in pixel-level, which provides an effective way for extracting geographic objects with low cost and high efficiency.

1) *CNN Interpretability:* It has been demonstrated that weakly supervised object extraction without spatial priors can be realized by mining the interpretability of CNNs [16]–[20]. In other words, CNN interpretability can obtain the object spatial information by visualizing the semantic information from CNN feature maps without pixel-level labels. Bazzani *et al.* [17] proposed a self-taught object localization technique that involves shadowing regions of the image to identify the regions causing the maximal activations to localize objects. Cinbis *et al.* [18] proposed to localize objects by combining multi-instance learning with CNN features. Oquab *et al.* [20] introduced a method for transferring mid-level image representations by global maximal pooling (GMP) and demonstrated that object localization can be achieved by superimposing a part of feature maps in the same convolutional layer.

Instead of utilizing GMP, Zhou *et al.* [16] adopted global average pooling (GAP) and proposed the original CAM method that can localize objects in a single forward pass by an end-to-end training mode. The end-to-end training mode has the advantages of both integrity and simplicity, which means that the CAM method can be transplanted to various object extraction tasks. After that, CAM quickly became popular for object extraction because, on the one hand, the weighted heatmaps generated for each feature map can be applied for accurate discriminative localization, and on the other hand, only image-level labels are needed to train a CNN model for generating CAM, which allows CAM to be adopted for performing weakly-supervised object extraction tasks directly [21], [22] or indirectly [23], [24].

Manuscript received 28 April 2022; revised 14 June 2022; accepted 29 June 2022. Date of publication 5 July 2022; date of current version 21 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297 and Grant 41871235, in part by the Fundamental Research Funds for the Central Universities under Grant 020914380095, and in part by the High-Level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China. *(Corresponding author: Xueliang Zhang.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: su17826062068@163.com; zxl@nju.edu.cn; xiaopf@nju.edu.cn; lzhenshi@outlook.com; wwy13594137516@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3188493

Although CAM has the characteristics of weakly labeling and high efficiency, weakly labeling will discourage the network from learning rich features, and the classification network will lead to coarse-grained semantic features, both of which will do harm to the accuracy and integrity of object extraction. Accordingly, the subsequently modified CAM methods were devoted to improving these two shortcomings.

2) *Modified CAM Methods:* As an initial method, CAM adds a GAP layer and calculates the weight between the fully connected layer and the corresponding class. To solve the problem that CAM cannot be generated in some networks with the lack of GAP layer, e.g., VGG16, GradCAM was proposed by Selvaraju *et al.* [25]. GradCAM records the gradient information during backward propagation, and then takes the average of the gradient as the weight. Compared to CAM, GradCAM is more versatile because of the "GAP free." Based on GradCAM and CAM, Chattopadhyay *et al.* [26] proposed a method called GradCAM++ that can extract multiple objects in an image. Thus, GradCAM++ has better coverage, but it also faces the potential problem of over-extraction. Omeiza *et al.* [27] further proposed the SmoothGradCAM++ method, which introduces gradient smoothening into GradCAM++. Smoothening requires adding noise to the sample images of interest and averaging all the gradient matrices generated by each noise image. Hence, attention maps generated by SmoothGradCAM++ are of low noise but time-consuming. Different from the gradient-based CAM (GradCAM, GradCAM++, and SmoothGradCAM++), ScoreCAM proposed by Wang *et al.* [28] belongs to the gradient-free CAM. As the name implies, ScoreCAM eliminates the dependence on gradients by obtaining the weight of each heatmap through its forward passing score on object class, which can also eliminate the noise caused by gradients. Modified CAM methods were proposed for objects in natural images in the computer vision domain, which have been demonstrated to be effective. However, considering the difference between the objects in natural images and the geographic objects in remote sensing images, whether the improvement of CAM methods to objects in natural images is applicable to geographic objects in remote sensing images is still unknown.

3) *CAM Methods for Extracting Geographic Objects:* Recently, CAM methods have been widely used for extracting geographic objects by both direct and indirect ways due to the characteristics of concise principles and high efficiency. As for the direct way of utilizing CAM by thresholding the heatmap, it was successfully applied to extract the densely distributed artificial objects, such as oil tanks, aircraft, and boats [29]–[31]. The natural objects were also reported to be extracted by CAM with high accuracy, such as vegetation [32] and clouds [33]. Comprehensively, Abitbol and Karsai [34] utilized GradCAM to extract various objects in the whole of Paris. Apparently, the core of directly utilizing CAM to extract geographic objects is to select an appropriate CAM method. However, the current related works ignored to compare different CAM methods and selected methods at will, which may have a considerable impact on extracting geographic objects. As for the indirect way of utilizing CAM, it is used as a pseudo-mask to further train an FCN model for extracting geographic objects. For example, Fu *et al.* [35] proposed WSF-NET with CAM to extract water and clouds. Chen *et al.* [36] proposed SPMF-Net with CAM to extract buildings. The CAM was also used as pseudo-mask for extracting buildings [37] and cars [38]. No matter in which way the CAM methods are used, the accuracy and completeness of CAM methods serve as the key to the effectiveness of extracting geographic objects. In addition, Fu *et al.* [35] and Wang *et al.* [39] found that CAM has different effects on multiple scales and multiple objects that commonly appear in remote sensing images.

As described above, even though CAM methods cannot achieve extraction accuracy as high as that of fully supervised FCN models, they received a lot of attention for extracting geographic objects because of the low cost and high efficiency. We have several choices from the different CAM methods, and we also need to select a proper CAM method for extracting variant geographic objects. However, which method is more suitable for a specific extraction task of geographic objects? To address the question, the following problems should be illustrated. First, since the large-size remote sensing image takes much longer processing time than natural images, the efficiency of CAM methods is very important. Second, different from the natural images with relatively single scales and higher category saliency, the large spectral and spatial variance of geographic objects is very common in remote sensing images, such as multiple scales, multiple objects, low inter-class separability, and high intra-class heterogeneity. Therefore, the effect of CAM methods on dealing with the large variety needs to be comprehensively evaluated in terms of the accuracy and completeness. Third, previous studies reported that the performance of CAM methods on extracting different geographic objects varies greatly, and a systematic comparison is thus lacking to illustrate the potential of CAM methods for extracting different categories of geographic objects.

Hence, because of the lack of a clear and comprehensive perception of the practicability of different CAM methods for different geographic objects, we are not sure which CAM method is suitable for specific tasks. In this article, we attempt to clarify the applicability of CAM methods for extracting geographic objects by comparing the efficiency, accuracy, and completeness of different CAM methods and exploring their spatial response capabilities to different geographic objects. The main contributions of this article can be expected as follows.

1) By comparing the efficiency and accuracy of different CAM methods, we demonstrate that the original CAM, GradCAM, and SmoothGradCAM++ achieve the best performance in terms of efficiency, accuracy, and completeness for extracting geographic objects, respectively. Furthermore, we illustrate the reasons for the performance difference from the perspective of principles, and provide guidance for the selection of CAM methods according to the demands of different geographic extracting tasks.
2) By comparing the capability of different CAM methods in dealing with the spectral and spatial variance problem of geographic objects, we demonstrate that GradCAM

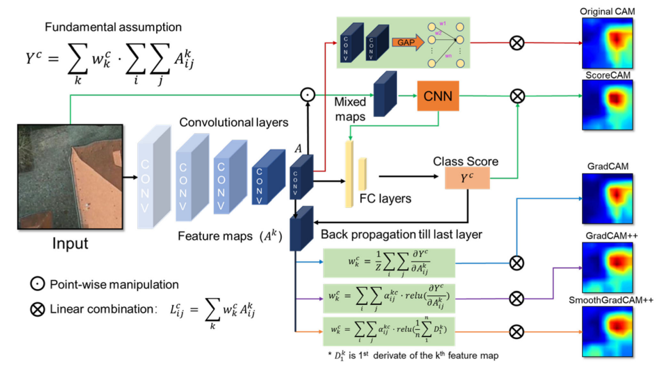

Fundamental assumption
$Y^c = \sum_k w_k^c \cdot \sum_i \sum_j A_{ij}^k$
Input
CONV
Convolutional layers
Feature maps (A^k)
A
GAP
w1
w2
wn
Mixed maps
CNN
FC layers
Class Score
Y^c
Back propagation till last layer
$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$
$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \cdot relu(\frac{\partial Y^c}{\partial A_{ij}^k})$
$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \cdot relu(\frac{1}{n}\sum_1^n D_1^k)$
* D_1^k is 1st derivate of the kth feature map
⊙ Point-wise manipulation
⊗ Linear combination: $L_{ij}^c = \sum_k w_k^c A_{ij}^k$
Original CAM
ScoreCAM
GradCAM
GradCAM++
SmoothGradCAM++

Fig. 1. Flow chart of the five CAM methods.

achieves better performance in terms of the capability in extracting various geographic objects and the adaptability in complex scenes, which indicates the advantage of GradCAM on better capturing the fine-grained object features and keeping the object completeness at the same time.

3) We further evaluate the performance of GradCAM for extracting different categories of geographic objects and it is illustrated that GradCAM performs better for geographic objects with low intra-class heterogeneity, clear boundaries, and not too small size. This shows that using a weakly supervised method based on GradCAM is an effective solution for practical applications when extracting geographic objects with the above features.

The rest of this article is organized as follows. In Section II, we introduce five different CAM methods and their applications in geographic object extraction. In Section III, we introduce the datasets, implementation details and experimental design. Section IV experimentally compares the performance of different CAM methods. In Section V, we discuss the reasons why GradCAM performs best and how to further improve CAM methods. Finally, Section VI concludes article.

II. Methods

Here, we describe the principles of different CAM methods and analyze the applicability of different CAM methods from the perspective of principle. A heatmap would be generated by applying a CAM method to a remote sensing image, in which brighter pixels indicate a larger probability of belonging to the target geographic object. As a result, the heatmap can then be effectively used to extract the spatial information of geographic objects.

A. Original CAM

The original CAM method was proposed by [16], which uses GAP on the feature maps (A^k) in the last convolutional layer and calculates the object class weight (w_k^c) from the fully connected layer, as shown in Fig. 1. Specifically, the weight w_k^c is derived from the feature maps and contains a linear relationship with the extracted feature maps, which can be regarded as the contributions of the feature maps to the object class. In a word, the final class heatmap (M^c) is obtained through the fusion of the feature maps and the weight (w_k^c), which is formulized as follows:

$$M_{ij}^c = \sum_k w_k^c \, A_{ij}^k \, . \tag{1}$$

GAP plays a key role in generating original CAM, which not only achieves data reduction but also prevents overfitting. However, if the network does not contain a GAP layer, GAP cannot be implemented unless the network is changed, which goes against the original intention for easy operation and high efficiency.

B. GradCAM

Selvaraju *et al.* [25] proposed the GradCAM method without using the GAP layer. Let Y^c represent the final classification score and A^k represent the feature maps in the last convolutional layer, GradCAM records the gradient information $\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right)$ obtained during the CNN backward propagation. Then, the gradient is averaged as the weight w_k^c and thus bypasses the GAP layer for weight acquisition. The weight w_k^c is formulized as (2), where Z is the number of pixels in A^k

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}. \tag{2}$$

Similar to original CAM, GradCAM also uses the principle of the linear combination of feature maps and weights, as shown in Fig. 1. In short, GradCAM realizes deep feature extraction without the GAP layer, which is more universal and extensive.

C. GradCAM++

Different from obtaining the weight by averaging, Chattopadhyay *et al.* [26] suppose that the weight w_k^c in GradCAM makes the same contributions in each pixel, which leads to the lack of object completeness. In this regard, GradCAM++ adds a coefficient $\alpha_{i,j}^{kc}$ (weighting coefficients for the pixel-wise gradients for class c and convolutional feature maps A^k) for calculating weight w_k^c. The weight of GradCAM++ is formulized as follows:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot \text{relu}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{3}$$

where relu(·) is the rectified linear unit activation function.

By integrating the additional coefficient, GradCAM++ achieves high coverage and a low omission rate of multiple units when the object class contains multiple units in an image. However, the extra coefficient magnifies the noise in the image and reduces the extracting accuracy.

D. SmoothGradCAM++

Compared with GradCAM++, SmoothGradCAM++ [27] is devoted to eliminating noise, as shown in Fig. 1. It creatively adds noise into the original image to achieve the purpose of "removing noise by adding noise" [40]. Smoothing operation makes the heatmap contain both the noise attached to the gradient and that artificially added. After adding noise, the noise attached to the gradient is weakened due to the randomness of its location. Therefore, we can obtain the denoised heatmap by manually eliminating the artificial noise. Specifically, let D_1^k represent the first derivate of the *k*th feature map and *n*the number of times to add noise. The weight of SmoothGradCAM++ is formulized as follows:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot \text{relu}\left(\frac{1}{n}\sum_1^n D_1^k\right). \quad (4)$$

This method eliminates the noise interference generated by gradient to a certain extent, but it cannot completely solve the problem of the low accuracy of GradCAM++.

E. ScoreCAM

The above methods only modified the object class weight acquisition in the previous CAM methods, but did not involve the improvement of feature maps. Different from the methods based on GradCAM, ScoreCAM [28] no longer uses the gradient to obtain weights but returns to the original gradient-free CAM method. Moreover, ScoreCAM no longer directly obtains the heatmap by multiplying the feature maps and weights. Before the linear calculation, ScoreCAM performs up-sampling and standardization on the feature maps. Then, ScoreCAM multiplies the feature maps with the original image pixel by pixel. The new mixed image not only presents the original image information but also presents the information of feature maps.

In the second stage, ScoreCAM masks the information of the original feature maps in the mixed image and puts the masked image into the CNN again to obtain values representing the importance of different feature maps. Due to the introduction of confidence, the heatmap focuses more on the target object than other CAM methods. Nevertheless, the focus characteristics of ScoreCAM may lead to insufficient integrity, and the complexity of the weight acquisition process is very high. Hence, the effectiveness of ScoreCAM needs to be further evaluated in terms of both efficiency and accuracy.

F. Geographic Objects Extraction by CAM

After training a CNN by image-level labels, we can obtain the heatmap from the trained network by the CAM methods, as shown in Fig. 2. A thresholding strategy is then applied to the heatmap to directly extract the corresponding geographic objects. Specifically, we set the pixels larger than the threshold as the foreground, and set the remaining pixels as the background. Since the optimal thresholds of different object classes are different, we use the prior knowledge of the size and shape of the geographic objects to preselect several different thresholds for accuracy comparison and pick out the best threshold. After thresholding, the result is further processed by a sequence labeling algorithm named fully connected conditional random field (CRF) [41]. CRF comprehensively considers the contextual information to calculate the class probability of the target pixel and can thus improve the boundaries and completeness of objects.

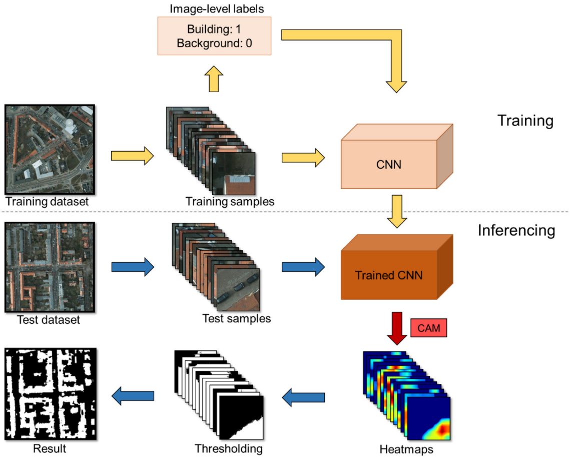


Fig. 2. Pipeline of extracting geographic objects by using CAM in the direct way.

It is noted that the extracted results by thresholding the heatmaps could be used as the pseudo-mask to train an FCN model for semantic segmentation, which refers to the indirect way of utilizing CAM. The higher quality of the pseudo-mask would result in the higher segmentation accuracy [37]. Hence, we focus on evaluating the results after thresholding, which could also be beneficial to selecting a proper CAM method for training FCN.

III. Experiment Setup

A. Datasets

To comprehensively evaluate and compare the effectiveness of different CAM methods, three datasets are adopted for the experiments, which cover remote sensing images with different spatial and spectral resolutions, and geographic objects with different types, sizes, shapes, as well as large variances. The details of each dataset are presented as follows.

1) ISPRS-Potsdam Dataset: There are 38 aerial images of 6000 × 6000 pixels with 5 cm spatial resolution in ISPRS-Potsdam dataset [42]. We use 20 images for training and the other 18 images for testing. The dataset covers abundant geographic objects with pixel-level labels of buildings, roads, vegetation, cars, and water. Compared with other public multiclass datasets, such as NWPU VHR-10 [42]–[44] and DIOR [45], ISPRS-Potsdam dataset has higher spatial resolution and the annotated geographic objects are more common. Therefore, ISPRS-Potsdam dataset is used to evaluate the effectiveness of CAM methods on extracting buildings, vegetation, and cars, as shown in Fig. 3.

2) WDCD Dataset: The WDCD dataset proposed by [33] is adopted to evaluate the performance of CAM methods on cloud detection, as shown in Fig. 4. The dataset consists of Gaofen-1 PMS images with a spatial resolution of 8 m. It provides over 200 000 image-level training patches with 128 × 128 pixels for

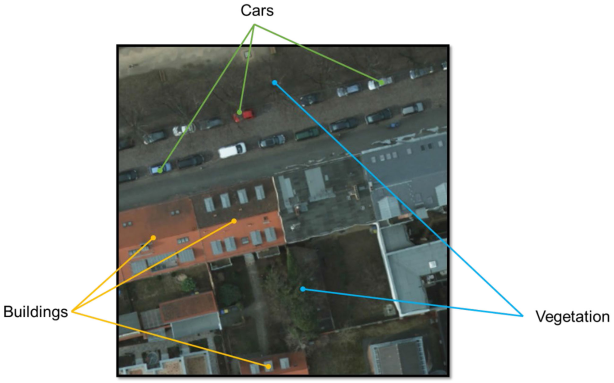


Fig. 3. Sample image in ISPRS-Potsdam dataset with different geographic objects.

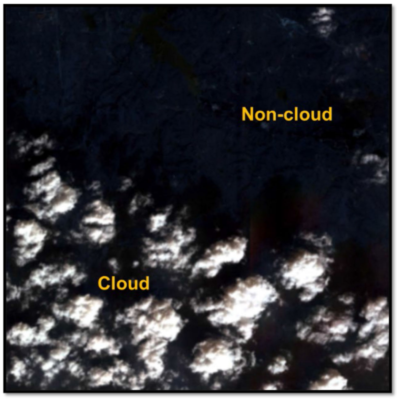


Fig. 4. Sample image in WDCD dataset for cloud detection.

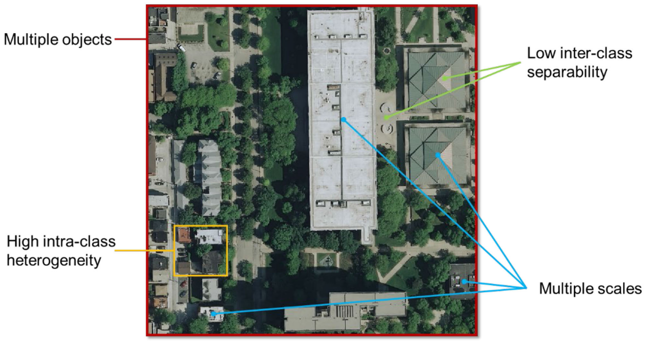


Fig. 5. Sample image in Inria Building dataset with large spectral and spatial variance of buildings.

training and 25 large-size images (approximately 6000 × 6000 pixels) with pixel-level labels for testing. Even though cloud is indeed not a geographic object, as for optical remote sensing images, accurate cloud extraction is very helpful for extracting geographic objects, because the correct detection of clouds can exclude the interference of clouds to extract geographic objects. Hence, the cloud extraction task is involved in this article.

3) Inria Building Dataset: The Inria Building dataset consists of 360 aerial images with a spatial resolution of 0.3 m, as shown in Fig. 5. Each image contains 5000 × 5000 pixels with pixel-level labels of each building [46]. We select 100 images for training and 25 images for testing. The images cover dissimilar urban settlements, ranging from densely populated areas (e.g., San Francisco's financial district) to alpine towns (e.g., Lienz in Austrian Tyrol). Hence, Inria Building dataset is very suitable for evaluating the potential of dealing with the large variance problem of geographic objects.

The images are cropped into 256 × 256 patches with a sliding stride of 32 for both training and test, except for the training samples of the WDCD dataset. The training patches are augmented by image rotation, random horizontal flipping, and adding random noise. The image-level labels are generated by counting the proportion of foreground pixels in the patch, where the training patch with over 25% foreground pixels is viewed as positive and that with fewer than 5% foreground pixels is viewed as negative. The proportion of positive and negative training samples is further balanced by filtering. The specific numbers of samples of each dataset are given in Table I.

B. Implementation Details

1) Computing Environment: The experiment is conducted on PyTorch 1.8.0 and Python 3.8. The network is trained on a computer with an Intel Core i7-10700F CPU, one NVIDIA GeForce RTX 3080 GPU, and 64 GB memory.

2) Evaluation Metrics: Several commonly used accuracy measures are derived: including precision, recall, overall accuracy (OA), F1-score, and intersection of union (IoU), which are formulized as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{5}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{6}$$

$$\text{OA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{7}$$

$$\text{F1}-\text{score} = 2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{8}$$

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

where TP and TN refer to truly predicted pixels in the positive and negative samples, while FP and FN refer to falsely predicted pixels in the positive and negative samples. Precision represents the proportion of the number of correctly predicted pixels in all predicted pixels, and Recall represents the proportion of correctly predicted pixels in the reference image. OA, F1-score, and IoU indicate the OA.

3) CNN Architecture: We choose ResNet101 as the CNN architecture for three reasons. First, compared with early CNN architectures, such as AlexNet [47] and VGG [48], ResNet [49] contains the GAP layer, which generates original CAM without changing the network structure. In other words, CNN networks like VGG need to add the GAP layer to generate original CAM. Moreover, compared with novel CNN architectures such as EfficientNet [50] and HRNet [51], ResNet has fewer parameters and a lighter network architecture. Second, ResNet has "shortcut connections." Identity shortcut connections increase

TABLE I
LIST OF THE NUMBERS OF SAMPLES FOR EACH DATASET

Number of samples	Dataset		
	ISPRS-Potsdam	WDCD	Inria Building
Train-Positive	5475	83792	17911
Train-Negative	5836	98528	17470
Test-Positive	1199	13280	4177
Test-Negative	1145	10874	3938

TABLE II
ACCURACIES OF DIFFERENT CNN ARCHITECTURES ON ISPRS-POTSDAM DATASETS

	Original CAM	GradCAM	GradCAM++	SmoothGradCAM++	ScoreCAM
VGG16	0.493	0.629	0.541	0.552	0.603
ResNet18	0.584	0.661	0.556	0.609	0.619
ResNet101	0.634	0.678	0.612	0.620	0.635

TABLE III
LIST OF HYPERPARAMETERS FOR TRAINING RESNET101

Hyperparameter	Value
Learning rate	0.0001
FC-layer learning rate	0.0010
Batch size	10
Shuffle	True
Weight decay	0.0005
Momentum	0.9000
Epoch	50

TABLE IV
EFFICIENCY COMPARISON OF DIFFERENT CAM METHODS

Method	Original CAM	GradCAM	GradCAM++	SmoothGradCAM++	ScoreCAM
Time (seconds)	25.920	70.824	70.745	418.067	1245.246
FLOPS (G)	10.229	24.700	24.700	372.004	327.328
Params (M)	42.504	42.500	42.500	1062.504	1360.128

neither extra parameters nor computational complexity, which allows the information obtained in the previous residual block to flow into the next residual block without any loss, and avoids the problem of gradient degradation. Third, ResNet contains different versions, including ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. ResNet101 weighs the training efficiency, precision, and hardware configuration, which is suitable for this article. To verify the superiority of ResNet101 in extracting geographic objects, we design a CNN architecture performance comparison including VGG16 [48], ResNet18, and ResNet101. As a typical geographic object, we choose buildings in ISPRS-Potsdam dataset and calculate the IoU of five CAM methods for comparison, as given in Table II. Benefiting from the advanced and deeper network structure, ResNet101 performs best among these CNN architectures. Limited from the depth of network, ResNet18 ranks second. Because of the naïve architecture and transplanted GAP layer, the performance of VGG16 is the worst. Consequently, we choose ResNet101 to do follow-up study.

4) CNN Training: The effectiveness of CNN training guarantees the quality of the generated heatmap by CAM methods. The selection of hyperparameters influences the training effects, which requires a series of comparative experiments, and the final settings are given in Table III. In order to achieve faster and better iterations of the network, the classification network is initialized by the weights of ResNet101 pretrained on ImageNet [52]. The stochastic gradient descent optimizer and cross-entropy loss function are utilized for training.

C. Experimental Design

In order to achieve a comprehensive comparison of CAM methods for extracting geographic objects, we design the following four experiments: efficiency comparison; accuracy comparison; comparison of dealing with the large variance problem; and comparison of the localization ability for different geographic objects.

1) Efficiency Comparison: The amount of time it takes to generate the heatmap is an important indicator of CAM methods, especially for processing the large-size remote sensing images. In order to compare the efficiency of different CAM methods in generating heatmaps, the time spent in generating heatmap is recorded and averaged.

2) Accuracy Comparison: We use the five measures (precision, recall, OA, F1-score, and IoU) to quantitatively evaluate the accuracy of the extracted geographic objects from the heatmaps generated by different CAM methods.

3) Comparison of Dealing With the Large Variance Problem: This comparison aims at illustrating the effectiveness of different CAM methods on dealing with the large variance problem that is common for geographic objects, including multiple scales, multiple objects, high intraclass heterogeneity and low interclass separability.

4) Comparison of the Localization Ability for Different Categories of Geographic Objects: The most suitable CAM method for extracting geographic objects could be selected by the above three experiments. However, due to the various characteristics among different object classes, the effectiveness of the CAM method varies for different object classes. Taking the selected CAM method as an example, this comparison aims at illustrating the performance difference of the CAM methods in terms of different object classes, especially the effects of the spectral heterogeneity, boundary, and size of geographic objects.

IV. Results

A. Efficiency Comparison

Totally 53 full-size images (20 images from ISPRS-Potsdam dataset, 13 images from WDCD dataset, and 20 images from Inria Building dataset) are used to obtain the average time for each CAM method to generate heatmaps, as given in Table IV. The original CAM is the most efficient, which takes only 25.9 s to generate a heatmap. GradCAM and GradCAM++ are less efficient than original CAM, and take 70.8 and 70.7 s to generate a heatmap, respectively. The efficiency of SmoothGradCAM++ is significantly reduced, which takes 418.0 s. ScoreCAM has the lowest efficiency, taking up to 1245.2 s, which is approximately 3 times as long as SmoothGradCAM++, 18 times that of GradCAM, and 48 times that of original CAM. To further compare the efficiency of different CAM methods, we compared the time complexity (FLOPS) and space complexity (Params) of different CAM methods, and the results have a similar pattern to the running time.

In general, the main reason for the efficiency difference of different CAM methods is their different mapping principles. Original CAM directly uses the existing information in CNN architecture as the weight, while GradCAM and GradCAM++ need to calculate the gradient to obtain the weight. For SmoothGradCAM++, to make the heatmaps "smoother," the image needs to be passed through CNN multiple times, which leads to a greatly increased time consumption to generate a heatmap. ScoreCAM needs to fuse multiple feature maps with the original images and re-feed them into CNN to calculate the weights, which makes it take the longest time among all the CAM methods. Through the efficiency comparison, we can infer that the original CAM extracts geographic objects with the least computation. However, the comparison results also prove that the original CAM uses less feature information in the extracting process, which may affect the fineness of the extracting results.

B. Accuracy Comparison

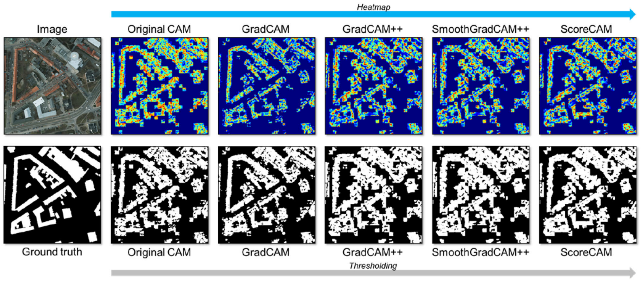


Fig. 6. Sample heatmap and binarization results of buildings from ISPRS-Potsdam dataset by different CAM methods.

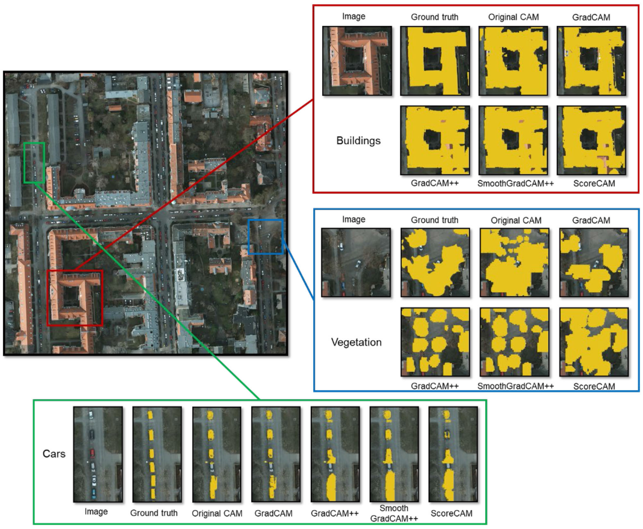


Fig. 7. Sample extraction results of buildings, vegetation, and cars from ISPRS-Potsdam dataset by different CAM methods.

Five evaluation measures are used to comprehensively compare the accuracies of different CAM methods. The 13 test images from ISPRS-Potsdam dataset and 13 images from WDCD dataset are used together for this comparison, involving the geographic objects of buildings, vegetation, and cars, as well as clouds. Taking building extraction as an example, we first generate the heatmap results of different CAM methods, as shown in Fig. 6. Compared to other CAM methods, GradCAM shows more accurate foreground extraction and smoother edges. According to the shape and size of geographic objects, we threshold the heatmap and obtain the binarization results. On the basis of binarization results, we calculate the five evaluation measures for each object class and then their averages for comparison, and each result is averaged by 12 independent experiments in terms of random initialization of training and test samples, as shown in Table V and Table VI. GradCAM performs the best in terms of precision, OA, F1-score, and IoU. In addition to achieving higher accuracy than other CAM methods, GradCAM also has an advantage in the capability of localizing object boundaries through visual evaluation, as shown in Figs. 7 and 8. Considering the object completeness, SmoothGradCAM++ performs best in terms of recall and GradCAM++ also performs well.

The accuracy comparison results show that the improvement of CAM methods to objects in natural images is not completely applicable to geographic objects in remote sensing images. The improvement from CAM to GradCAM is applicable because the accuracy and completeness of GradCAM to extract the geographic objects are both improved. It can be seen that the gradient used by GradCAM has achieved a good performance.

TABLE V
ACCURACIES OF DIFFERENT CAM METHODS ON ISPRS-POTSDAM DATASETS

Method	*precision*	*recall*	*OA*	*F1-score*	*IoU*
Original CAM	0.5287±0.0026	0.7030±0.0039	0.8848±0.0006	0.6021±0.0030	0.4357±0.0030
GradCAM	**0.5621**±0.0022	0.7309±0.0038	**0.9006**±0.0006	**0.6343**±0.0026	**0.4716**±0.0026
GradCAM++	0.4099±0.0016	0.7569±0.0038	0.8319±0.0004	0.5297±0.0022	0.3670±0.0019
SmoothGradCAM++	0.4095±0.0016	**0.7633**±0.0033	0.8307±0.0005	0.5309±0.0021	0.3679±0.0018
ScoreCAM	0.4626±0.0021	0.7192±0.0045	0.8648±0.0005	0.5623±0.0028	0.3992±0.0025

TABLE VI
ACCURACIES OF DIFFERENT CAM METHODS ON WDCD DATASETS

Method	*precision*	*recall*	*OA*	*F1-score*	*IoU*
Original CAM	0.6353±0.0008	0.8227±0.0007	0.9554±0.0001	0.7170±0.0006	0.5588±0.0008
GradCAM	**0.6731**±0.0009	0.3009±0.0017	**0.9631**±0.0002	**0.7703**±0.0011	**0.6264**±0.0014
GradCAM++	0.5458±0.0004	0.9169±0.0012	0.9419±0.0001	0.6843±0.0006	0.5201±0.0007
SmoothGradCAM++	0.5425±0.0004	**0.9255**±0.0008	0.9413±0.0001	0.6841±0.0005	0.5199±0.0005
ScoreCAM	0.5225±0.0004	0.9254±0.0015	0.9368±0.0001	0.6679±0.0006	0.5013±0.0007

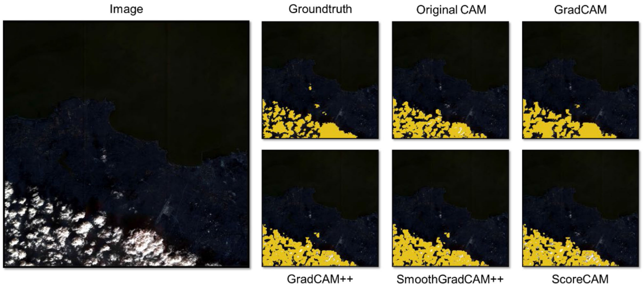


Fig. 8. Sample extraction results of clouds from WDCD dataset by different CAM methods.

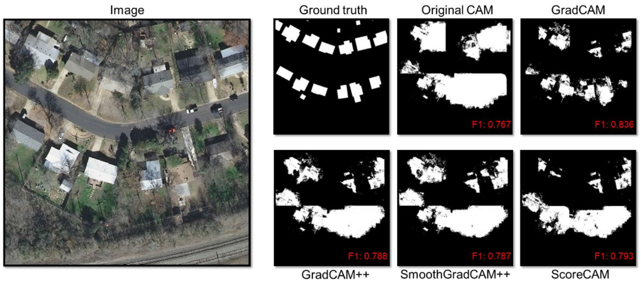


Fig. 9. Sample extraction results by different CAM methods for illustrating the ability of dealing with the problem of multiple objects in Inria Building dataset.

GradCAM++ and SmoothGradCAM++ can improve the extraction completeness, but at the cost of losing much more precision, which finally leads to the decrease of OA indicated by OA, F1-score, and IoU.

It shows that GradCAM stands out for its excellent mapping principles when dealing with geographic objects. Compared with CAM, which only uses the GAP to obtain the weight for heatmap, GradCAM makes use of the more abundant gradient information to obtain the weight and thus achieves better performance. Compared with GradCAM++ and SmoothGradCAM++, GradCAM uses the weights of all the pixels in feature maps for extracting geographic objects, which enables GradCAM to pay balanced attention to each pixel. In contrast, GradCAM++ and SmoothGradCAM++ pay more attention to the foreground regions. Although it works for improving completeness, the background pixels are more inclined to be recognized as the foreground, leading to the decreased precision and OA. Meanwhile, the boundary pixels are not given a lower weight due to the lack of information in feature maps. Hence, the boundary accuracy and object localization capabilities are simultaneously reduced in the results of GradCAM++ and SmoothGradCAM++.

C. Comparison of Dealing With the Large Variance Problem

The Inria Building dataset is used for comparing different CAM methods on the ability of dealing with the large variance problem of geographic objects, as already shown in Fig. 5. The achieved extraction accuracies by different CAM methods are given in Table VII. We find that the accuracies of different CAM methods are lower than those in Table V because of the high complexity of buildings in the Inria building dataset. However, the performance gaps among different CAM methods in Table VII remain similar to those in Table V, where GradCAM still comes out on top in terms of precision and overall accuracies, again demonstrating the excellent object localization capability of GradCAM. In addition, GradCAM also has superior performance than other CAM methods in dealing with each of large variance problems, including multiple scales, multiple objects, high intra-class heterogeneity, and low inter-class separability, which is illustrated as below.

In dealing with the problem of multiple objects, we found it difficult to separate the gathered objects by CAM methods, especially in the case of small-scale geographic objects with the disturbance of surrounding objects. However, as shown in Fig. 9, although GradCAM also reduces its localization capability due to the disturbance of surrounding vegetation, it can effectively isolate the small buildings, rather than identifying them as a whole by other CAM methods, which shows the priority of GradCAM on capturing the fine-grained object features.

GradCAM also demonstrates a better localization capability than other CAM methods in extracting multiscale objects, as shown in Fig. 10. Overall, all CAM methods perform well in extracting large-size buildings completely, but GradCAM produces results with shape closer to the ground truth than any

TABLE VII
ACCURACIES OF DIFFERENT CAM METHODS ON INRIA BUILDING DATASET

Method	*precision*	*recall*	*OA*	*F1-score*	*IoU*
Original CAM	0.496	0.737	0.798	0.591	0.421
GradCAM	**0.506**	0.752	**0.806**	**0.604**	**0.435**
GradCAM++	0.482	**0.757**	0.787	0.586	0.416
SmoothGradCAM++	0.497	0.745	0.798	0.594	0.424
ScoreCAM	0.487	0.754	0.793	0.590	0.420

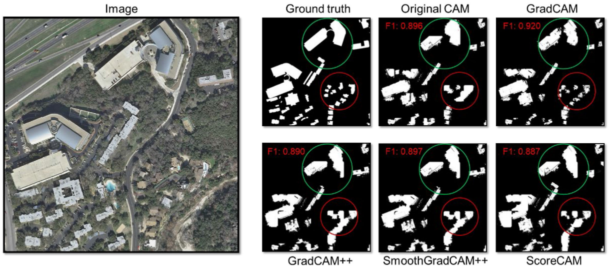


Fig. 10. Sample extraction results by different CAM methods for illustrating the ability of dealing with the problem of multiple scales in Inria Building dataset.

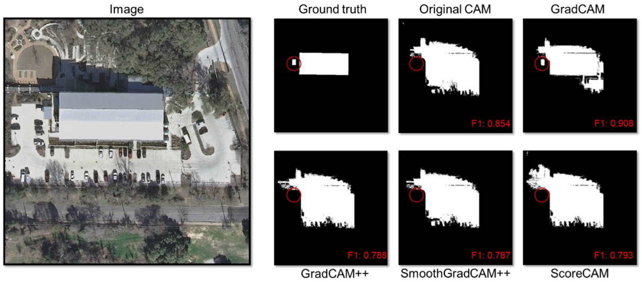


Fig. 11. Sample extraction results by different CAM methods for illustrating the ability of dealing with the problem of low inter-class separability in Inria Building dataset.

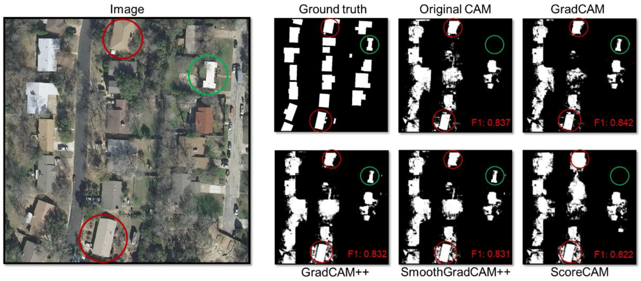


Fig. 12. Sample extraction results by different CAM methods for illustrating the ability of dealing with the problem of high intra-class heterogeneity in Inria Building dataset.

other CAM methods, as marked by the green circles in Fig. 10. In addition, GradCAM has the unique localization capability for small-scale geographic objects, as marked by the red circles in Fig. 10, where GradCAM successfully identifies all the small buildings while the other CAM methods tend to identify them as a whole and thus aggregate into a large-scale fake object. This further shows the advantage of GradCAM on capturing details.

In the case of low inter-class separability, a common problem for building extraction is the confusion with other impervious surfaces. As shown in Fig. 11, the building roof is spectrally very similar to its surrounding parking lot, which results in the extraction errors for all the CAM methods. However, it is visually obvious that the error of GradCAM is much lower than other CAM methods. Additionally, as marked by the red circle, GradCAM can even overcome the confusion and successfully extract the small-scale building, which is missed by all other CAM methods. This could be caused by the better ability of GradCAM on capturing object details. It is noted that all the CAM methods identify the shadow as building. This is caused by the co-occurrence of building and shadow, which is difficult to separate under the supervision of image-level labels.

As for the high intra-class heterogeneity problem, it includes two types of heterogeneities: spectral and shape. GradCAM also achieves the best performance among the CAM methods, as shown in Fig. 12. In terms of the spectral heterogeneity, the typical error would be the omitted buildings, such as those marked as green circles in Fig. 12 by original CAM and ScoreCAM. The shape heterogeneity would lead to errors of the extracted buildings, such as those marked as red circles in Fig. 12 by GradCAM++, SmoothGradCAM++, and ScoreCAM.

According to the above comparisons, we can know that GradCAM achieves better performance in dealing with the high variance problem in terms of both the accuracies and each specific high variance case. The results demonstrate that GradCAM has the advantage of capturing more fine-grained object features than other CAM methods, as well as has a good performance on capturing the complete object extent.

The following reasons can explain why GradCAM performs best among the five CAM methods in large variance problem: On the one hand, the weights obtained from gradient information contain more geographic feature information than the GAP operation, which is more conducive to the extracting results. On the other hand, the extracting results of each feature map become more fine-grained because the gradients information would be fed back to each pixel in feature maps.

D. Comparison of the Localization Ability for Different Categories of Geographic Objects

Through the above comparisons, we can safely conclude that GradCAM performs best among the CAM methods and could be preferred for extracting geographic objects, taking into account both efficiency and accuracy. Until now, the performance of GradCAM on different geographic object classes has remained unclear. To explore the localization capabilities of CAM methods represented by GradCAM to different geographic objects, the factors of the spectral heterogeneity, boundary, and size are mainly considered.

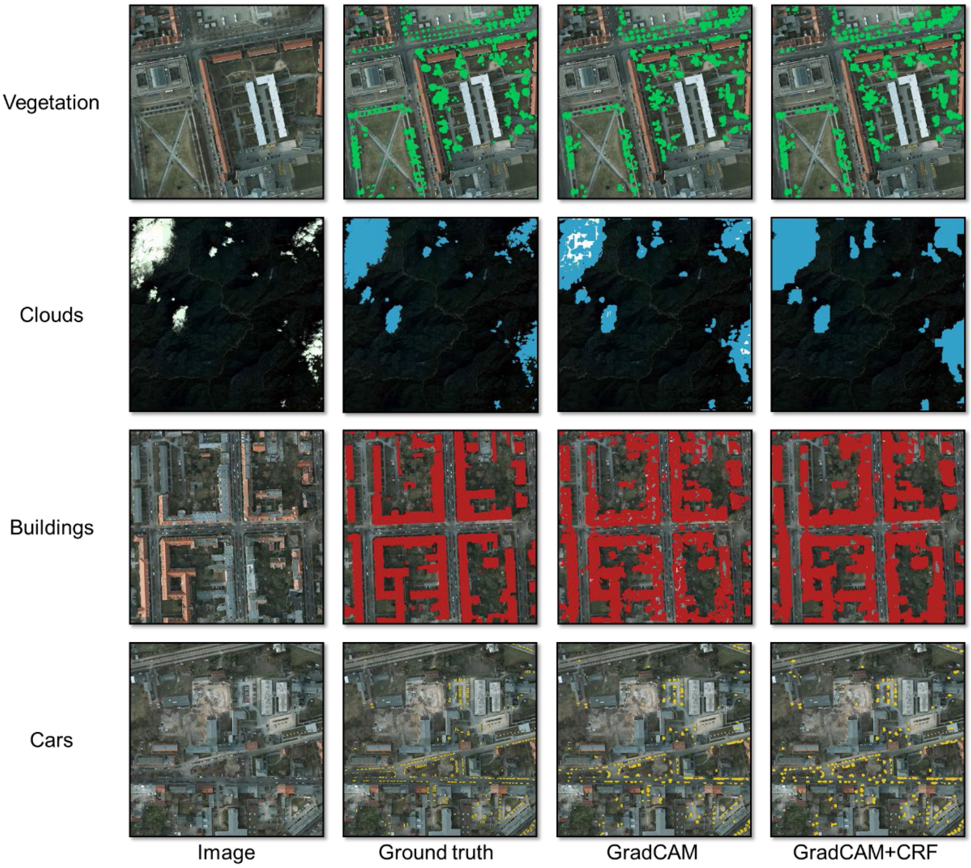


Fig. 13. Extraction results of various geographic objects by GradCAM with and without CRF post-processing.

TABLE VIII
EXTRACTION ACCURACIES OF VARIOUS GEOGRAPHIC OBJECTS BY GRADCAM WITH/WITHOUT CRF POSTPROCESSING

Geographic objects	Methods	*precision*	*recall*	*OA*	*F1-score*	*IoU*
Vegetation	GradCAM	0.6058±0.0014	0.7190±0.0024	0.8911±0.0005	0.6558±0.0015	0.4898±0.0017
	+ CRF	0.6976±0.0010	0.7102±0.0012	0.9042±0.0003	0.7039±0.0009	0.5431±0.0010
Clouds	GradCAM	0.6731±0.0009	0.3009±0.0017	0.9631±0.0002	0.7703±0.0011	0.6264±0.0014
	+ CRF	0.7297±0.0007	0.9101±0.0011	0.9707±0.0001	0.8100±0.0007	0.6806±0.0100
Buildings	GradCAM	0.6564±0.0020	0.8441±0.0018	0.8403±0.0012	0.7385±0.0018	0.5854±0.0022
	+ CRF	0.7405±0.0009	0.8902±0.0014	0.8656±0.0007	0.8085±0.0010	0.6785±0.0014
Cars	GradCAM	0.4240±0.0032	0.6297±0.0073	0.9707±0.0002	0.5068±0.0045	0.3394±0.0040
	+ CRF	0.5526±0.0109	0.5745±0.0131	0.9754±0.0006	0.5633±0.0112	0.3922±0.0109

The geographic objects of vegetation, buildings, and cars in ISPRS-Potsdam dataset and the objects of clouds in WDCD dataset are specifically used for comparison, to illustrate the effects of the spectral heterogeneity, boundary, and size on extracting geographic objects by GradCAM. The accuracies for each object class by GradCAM are given in Table VIII.

The effect of object size on GradCAM performance is illustrated by the much lower accuracies of the small-size cars compared with those of other object classes in Table VIII. Even though GradCAM has been demonstrated to be able to capture fine-grained object features, the results here show that it is still not fine enough to capture the small car objects, which could be due to the loss of details by the convolutional network.

The accuracy difference among the vegetation, clouds, and buildings in Table VIII illustrates the effect of spectral heterogeneity of geographic objects. Clouds have the smallest intra-class spectral heterogeneity and thus the highest accuracies. Usually, vegetation would have smaller intra-class spectral heterogeneity than buildings because of the roofs made up of the abundant man-made materials. However, most of the leaves of the vegetation were fallen at the imaging date. The fallen leaves allow the complex background being imaged and thus increase the spectral heterogeneity of the vegetation, as shown in Fig. 13. Hence, the accuracies of vegetation are lower than those of buildings.

The effect of the object boundaries on GradCAM is reflected by the accuracy gaps between the results with and without CRF post-processing in Table VIII. Compared with the smooth boundary of vegetation, the boundaries of clouds, buildings, and cars are clearer, as shown in Fig. 13. Hence, the accuracy improvements by CRF post-processing of the objects with clear boundaries are apparently higher than those of vegetation. Comparing the accuracy improvements of cars with those of clouds and buildings, since the small size of cars leads to the boundaries less clear than those of clouds and buildings, we can see that the accuracy improvements of cars by CRF post-processing are smaller.

According to the above results, we can safely conclude that GradCAM tends to produce better extraction results for objects with low intra-class heterogeneity, clear boundaries, and not too small size. The extraction results by GradCAM with and without CRF post-processing in Fig. 13 further illustrate the different effects of the above three factors as well as the effectiveness of GradCAM for extracting various geographic objects.

V. Discussion

Through a variety of comparative experiments, it is found that GradCAM performs better than original CAM because of the utilization of the abundant gradient information of convolutional layers. In addition, the OA of GradCAM is even higher than the modified GradCAM++ and SmoothGradCAM++ for extracting geographic objects from remote sensing images, which is opposite to previous studies on natural images [26], [27]. This phenomenon can be explained from the following perspectives. In natural images with relatively small-scale variance, the α factor in GradCAM++ and SmoothGradCAM++ weight can improve the completeness of objects. However, the geographic object class covers a large range of scales, with large-size and small-size objects co-occurring in remote sensing images. Therefore, the α factor tends to merge the isolated small-size objects into fake large-size ones, resulting in low accuracy. GradCAM, on the other hand, achieves the foreground and background balance and therefore stands out on extracting geographic objects in remote sensing images.

CAM methods only utilize image-level labels and a classification network to achieve object localization. It has the advantage in training efficiency, but inferior accuracy than the FCN model for semantic segmentation. Therefore, the extracted results of CAM methods are always used as pseudo-masks for training a segmentation network to further improve accuracies [37]. In this case, we also need to carefully consider the selection of a suitable CAM method, which could be benefited from the findings of this article. For example, if the pseudo-mask is required to have labeling errors as few as possible, we can choose CAM methods with high precision, such as GradCAM, and then the completeness of objects can be made up by training the segmentation network. If the objects are continuous and preferred to be extracted completely, then we can choose CAM methods with high integrity, such as GradCAM++ or SmoothGradCAM++, in which the influence of labeling errors could be removed by improving the training of segmentation network on noisy labels.

In addition to directly applying CAM methods for extracting geographic objects, a lot of attention has been paid to modify the training of CAM methods in weakly supervised semantic segmentation to improve segmentation accuracy. Among the different CAM methods, it is more inclined to modify based on original CAM, rather than GradCAM or other CAM methods. There could be the following reasons for the preference of original CAM. First of all, compared to the GradCAM series, original CAM requires the GAP operation, in which GAP was modified into GMP [20] to enhance the structure of foreground, or global convolutional pooling to enhance the ability of the feature map to represent spatial variance of objects [33]. Second, CAM does not need to obtain weight through gradient, so it will not affect the efficiency of obtaining weights due to the increase in image resolution. For example, we are able to remove part of the pooling layer in CNN to pursue high spatial resolution by using original CAM [33]. If GradCAM series are used in the same situation, the computational complexity of gradient would increase exponentially, which leads to a significant decrease in the efficiency and is not conducive to extract geographic objects in large-size remote sensing images. Third, since training segmentation network requires a large number of labeled samples, the efficiency of generating pseudo-mask is important for sample collection, so the original CAM with the highest efficiency is also suitable. Hence, the original CAM is widely used to be modified as a pseudo-mask for further training segmentation network because of its efficiency and flexibility. However, since GradCAM achieves better accuracy and acceptable efficiency among the different CAM methods, it is preferable to be used for both directly extracting geographic objects and providing pseudo-masks for training segmentation network.

VI. Conclusion

In this article, we made a comprehensive comparison on the capability of different CAM methods including the original CAM, GradCAM, GradCAM++, SmoothGradCAM++, and ScoreCAM, which contributes to a deep understanding of different CAM methods and benefits to selecting a suitable CAM method for extracting geographic objects from the perspectives of both principles and experiments.

From the perspective of experiments, the results demonstrate that the original CAM has the highest efficiency and flexibility among these methods, which is suitable. GradCAM++ achieves the best performance on completeness and thus can minimize the leakage of geographic objects. GradCAM can satisfy the extraction of geographic objects in most scenes because of the best performance in accuracy and the acceptable efficiency. In addition, benefiting from the outstanding capability in extracting various geographic objects and excellent adaptability in complex scenes, GradCAM holds the advantage of capturing object details and keeping high-level object completeness, which makes it perform better in dealing with the large variance problems. Accordingly, GradCAM is preferred to be used for both directly extracting geographic objects and providing pseudo-masks for training segmentation networks.

From the perspective of principles, we illustrated the principles of different CAM methods in detail, especially the key difference between them. We found that the rational use of the gradient information with a large amount of parameters is the key for GradCAM to achieve fine-grained extraction while reducing the classification errors of foreground and background.

Furthermore, we explored the localization ability of GradCAM on different geographic object classes from three aspects: spectral heterogeneity; boundary; and size. We found that GradCAM tends to produce better extraction results for objects with low intra-class heterogeneity, clear boundaries, and not too small size.

In the future, we should pay more attention to the following two aspects. First, since CAM methods still have a big accuracy gap with fully-supervised methods, we would pay more attention to modifying the CAM methods from the perspective of network training and CAM itself to improve the accuracy. Second, we should expand the applications of different CAM methods for extracting geographic objects and further verify the effectiveness.

References

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

[2] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, “Building extraction at scale using convolutional neural network: Mapping of the United States,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[3] K. Bittner, S. Cui, and P. Reinartz, “Building extraction from remote sensing data using fully convolutional networks,” *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-1/W1, pp. 481–486, May 2017.

[4] S. Shrestha and L. Vanneschi, “Improved fully convolutional network with conditional random fields for building extraction,” *Remote Sens.*, vol. 10, no. 7, Jul. 2018, Art. no. 1135.

[5] Y. Wei, Z. Wang, and M. Xu, “Road structure refined CNN for road extraction in aerial image,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 709–713, May 2017.

[6] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks,” *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.

[7] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, “RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[8] T. Kattenborn, J. Eichel, and F. E. Fassnacht, “Convolutional neural networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery,” *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, Art. no. 17656.

[9] A. Safonova, S. Tabik, D. Alcaraz-Segura, A. Rubtsov, Y. Maglinets, and F. Herrera, “Detection of fir trees (Abies sibirica) damaged by the bark beetle in unmanned aerial vehicle images with deep learning,” *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 643.

[10] S. Hartling, V. Sagan, P. Sidike, M. Maimaitijiang, and J. Carron, “Urban tree species classification using a worldview-2/3 and LiDAR data fusion approach and deep learning,” *Sensors*, vol. 19, no. 6, Mar. 2019, Art. no. 1284.

[11] X. Feng, J. Han, X. Yao, and G. Cheng, “TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6946–6955, Aug. 2021.

[12] 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.

[13] 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.

[14] P. Wei, D. Chai, T. Lin, C. Tang, M. Du, and J. Huang, “Large-Scale rice mapping under different years based on time-series sentinel-1 images using deep semantic segmentation model,” *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 198–214, Apr. 2021.

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

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[17] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, “Self-taught object localization with deep networks,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[18] R. G. Cinbis, J. Verbeek, and C. Schmid, “Weakly supervised object localization with multi-fold multiple instance learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1717–1724.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free? - Weakly-supervised learning with convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 685–694.

[21] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *Proc. Workshop at Int. Conf. Learn. Representations*, 2014.

[22] M. Menikdiwela, C. Nguyen, H. Li, and M. Shaw, “CNN-based small object detection and visualization with feature activation mapping,” in *Proc. Int. Conf. Image Vis. Comput. New Zealand*, 2017, pp. 1–5.

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

[24] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, “Weakly-supervised semantic segmentation network with deep seeded region growing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7014–7023.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[26] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Improved visual explanations for deep convolutional networks,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[27] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, “Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models,” *CoRR*, abs/1908.01224, 2019.

[28] H. Wang *et al.*, “Score-CAM: Score-weighted visual explanations for convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 24–25.

[29] B. Vasu, F. U. Rahman, and A. Savakis, “Aerial-CAM: Salient structures and textures in network class activation maps of aerial imagery,” in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop*, 2018, pp. 1–5.

[30] R. Yang, X. Xu, Z. Xu, C. Ding, and F. Pu, “A class activation mapping guided adversarial training method for land-use classification and object detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019. pp. 9474–9477.

[31] J. Li, D. Lin, Y. Wang, G. Xu, and C. Ding, “Deep discriminative representation learning with attention map for scene classification,” *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1366.

[32] W. Castro *et al.*, “Deep learning applied to phenotyping of biomass in forages with UAV-based RGB imagery,” *Sensors*, vol. 20, no. 17, Aug. 2020, Art. no. 4802.

[33] 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, pp. 112045, Dec. 2020.

[34] J. L. Abitbol and M. Karsai, “Interpretable socioeconomic status inference from aerial imagery through urban patterns,” *Nature Mach. Intell.*, vol. 2, no. 11, pp. 684–692, Nov. 2020.

[35] K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun, “MultiCAM: Multiple class activation mapping for aircraft recognition in remote sensing images,” *Remote Sens.*, vol. 11, no. 5, Mar. 2019, Art. no. 544.

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

[37] 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, Mar. 2021, doi: 10.1109/JSTARS.2021.3063788.

[38] E. Kilic and S. Ozturk, "An accurate car counting in aerial images based on convolutional neural networks," in *Proc. J. Ambient Intell. Humanized Comput.*, 2021, Art. no. 237746064.

[39] 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, pp. 207, Jan. 2020.

[40] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[41] 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.

[42] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[43] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[44] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-Class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[45] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J.Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[46] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[47] 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.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–14.

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

[50] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[51] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

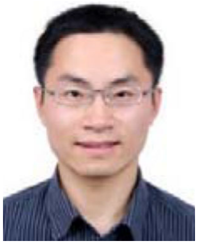
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

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

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

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. He is currently an Associate Professor with the Department of Geographic Information Science, Nanjing University. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.



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

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor, from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. He was a Visiting Scholar with the Department of Geography, University of Giessen, Frankfurt, 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. He has authored 4 books and more than 60 articles. His current research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.

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 in 2022 from Nanjing University, Nanjing, China, where he is currently working toward the Ph.D. degree in remote sensing of resources and environment from Nanjing University.

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

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

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