



Article

Prob-POS: A Framework for Improving Visual Explanations from Convolutional Neural Networks for Remote Sensing Image Classification

Xianpeng Guo , Biao Hou *, Zitong Wu , Bo Ren, Shuang Wang and Licheng Jiao

The Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China; guoxp@stu.xidian.edu.cn (X.G.); wuzitong@stu.xidian.edu.cn (Z.W.); boren@xidian.edu.cn (B.R.); shwang@mail.xidian.edu.cn (S.W.); lchjiao@mail.xidian.edu.cn (L.J.)

* Correspondence: houbiao@mail.xidian.edu.cn

Abstract: During the past decades, convolutional neural network (CNN)-based models have achieved notable success in remote sensing image classification due to their powerful feature representation ability. However, the lack of explainability during the decision-making process is a common criticism of these high-capacity networks. Local explanation methods that provide visual saliency maps have attracted increasing attention as a means to surmount the barrier of explainability. However, the vast majority of research is conducted on the last convolutional layer, where the salient regions are unintelligible for partial remote sensing images, especially scenes that contain plentiful small targets or are similar to the texture image. To address these issues, we propose a novel framework called Prob-POS, which consists of the class-activation map based on the probe network (Prob-CAM) and the weighted probability of occlusion (wPO) selection strategy. The proposed probe network is a simple but effective architecture to generate elaborate explanation maps and can be applied to any layer of CNNs. The wPO is a quantified metric to evaluate the explanation effectiveness of each layer for different categories to automatically pick out the optimal explanation layer. Variational weights are taken into account to highlight the high-scoring regions in the explanation map. Experimental results on two publicly available datasets and three prevalent networks demonstrate that Prob-POS improves the faithfulness and explainability of CNNs on remote sensing images.

Keywords: convolutional neural networks (CNNs); visual explanation; remote sensing image



Citation: Guo, X.; Hou, B.; Wu, Z.; Ren, B.; Wang, S.; Jiao, L. Prob-POS: A Framework for Improving Visual Explanations from Convolutional Neural Networks for Remote Sensing Image Classification. *Remote Sens.* **2022**, *14*, 3042. <https://doi.org/10.3390/rs14133042>

Academic Editors: Sidike Paheding, Matthew Maimaitiyiming, Zahangir Alom and Maitiniyazi Maimaitijiang

Received: 10 May 2022

Accepted: 21 June 2022

Published: 24 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Remote sensing image scene classification has various applications in a wide range of scenarios, including environment monitoring, urban planning, spatial-temporal data analysis and smart cities. Owing to the development of image equipment for earth observation [1–3], remote sensing images with different resolutions are increasing daily, which boosts the demand for intelligent interpretation of land use and land cover scenes.

Remote sensing image scene classification is a popular research field in satellite image analysis. During the past decades, numerous effective methods have emerged to achieve continuous improvement. Earlier methods, such as bag-of-visual words (BoVW) [4], HOG [5] and SIFT [6], generally require rigorous design from experts to extract valuable features, which imposes certain limitations on feature extraction. More recently, algorithms based on deep learning have achieved satisfactory results in many application fields owing to their strong feature representation power and powerful graphical processing units. Among the deep learning methods, convolutional neural networks (CNNs) have been successfully applied in remote sensing image scene classification [7–14]. Compared with traditional unsupervised feature learning methods, feature representations with more complex patterns can be learned via deep architecture neural networks. Some extensively

used pre-trained networks, such as AlexNet [15] and VGG-VD [16], have been proved to achieve excellent performance on remote sensing image scene classification [17].

However, CNNs are considered as a black-box operation mechanism, which means humans cannot easily understand the decision-making process. Therefore, the explainability of the model, which is considered as a barrier to developing artificial intelligence, has attracted increasing attention. When exploiting a machine learning model, considering the interpretability of the system can lead to the correction of model deficiencies and also improve implementability. For example, the improvement of explainability can help ensure overall impartiality in many applications involving ethics. To address this issue, many attempts have been made from various aspects [18–27]. A deep generator network (DGN), which generates the most representative image for a given output neuron, was proposed in [18]. The authors in [20] used a different approach called network dissection to quantify the interpretability of the latent representations of CNNs. In the field of remote sensing, long short-term memory (LSTM) recurrent neural networks were used by the authors in [26] to discover the interpretability of crop yield estimation. A synthetic aperture radar (SAR) specific deep learning framework called Deep SAR-Net that considers complex-valued SAR images to learn both spatial texture information and backscattering patterns of objects, was proposed in [27].

Compared with the aforementioned explainability techniques, local explanation methods which provide a visual saliency map for each specific decision are more widely studied. The saliency map, where the relevance score of each spatial position indicates its contribution to the prediction, is generally presented as a heat map. Consequently, it is more intuitive and comprehensible. One of the seminal works in this category was [19]. The authors of [19] utilized a global average pooling layer to modify and substitute the fully-connected layers. A class activation map (CAM) that highlights the image regions that are relatively important for a specific object class was achieved by projecting back the weights of the output layer on the convolutional feature maps. However, the CAM alters models' architectures, making it difficult to apply to a wide variety of CNN model families. Later, authors [28] proposed a gradient-weighted class activation mapping (Grad-CAM), which uses the gradients of any target concept, flowing into the final convolutional layer, to produce a coarse localization map. This approach avoids modifying the models' architectures.

Existing visual explanation methods, such as those mentioned above, have the identical feature that they almost solely pay attention to the saliency map of the final convolutional layer of CNNs. Some of these algorithms cannot visualize middle or shallow layers owing to the change of the original model, such as CAM. The feature maps of the final layer draw more attention, mainly because the receptive field of the high layer is larger than that of the shallow layer, which means the saliency map obtained by the high layer may contain more whole objects or object parts and may be more visually comprehensive. Besides, most visual explanation research is carried out based on natural images, in which the size of the salient object is relatively large. In contrast, the scales of discriminative objects in remote sensing image vary greatly from class to class, and some of them are very small, for instance, the cars in a parking lot and the mobile homes in a mobile home park. Beyond that, there are certain scenes that do not contain distinct objects and are more like texture images in visual research, such as the agricultural area, chaparral and forest. For illustration, in the three examples are shown in Figure 1, it can be seen that the salient regions generated by CAM and Grad-CAM irregularly emerge in the above scenes. The jumbled highlighted areas are still hard to understand, so the client may wonder why these parts are crucial for decision-making but not other similar parts.

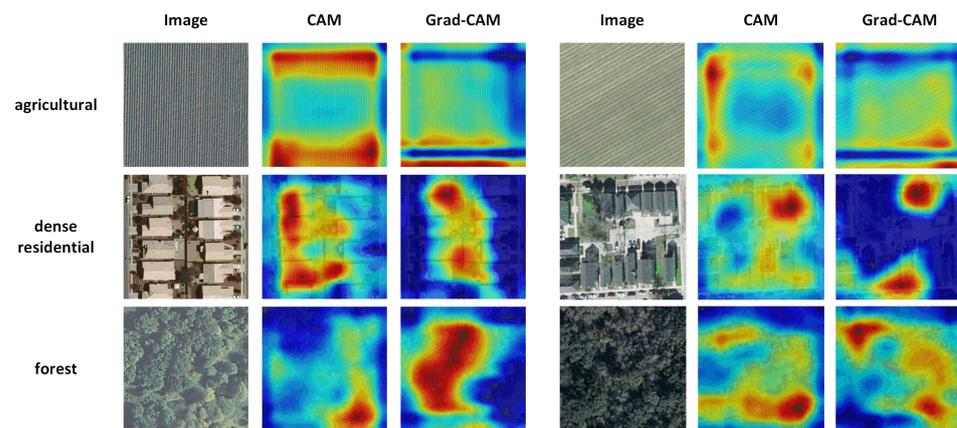


Figure 1. CAM and Grad-CAM on several scenes of remote sensing image.

With the aforementioned considerations, this paper aims to search for an acceptable visual explanation algorithm for remote sensing image classification. For this to occur, the CAM based on the probe network and weighted probability of occlusion selection (Prob-POS) framework is proposed. Essentially, visual explanation algorithms based on CAM aim to find optimal combinations of feature maps. The proposed probe network can elaborately discriminate between feature maps with different classes in any layer of a pending CNNs. Consequently, the weight of a particular class in the probe layer precisely locates the distinct target. The generated CAM based on probe network (Prob-CAM) possesses a more understandable explanation in a specific layer, especially a shallow layer. Subsequently, a metric called weighted probability of occlusion (wPO) is presented to automatically select the most suitable layer to be explained for different categories. As mentioned above, visualizing the explanation map on the last convolutional layer is not satisfactory for every remote sensing scene category, especially scenes that contain plentiful small targets or are similar to texture images. The explanation maps of middle or shallow layers are more likely to locate small targets or realistically describe texture information in visual research. In light of the above, we put forward wPO to quantitatively evaluate the explanation effectiveness of each convolutional layer for different categories to find out the optimal target layer.

The main contributions of this paper are as follows.

1. We propose a fundamental framework called Prob-POS to visualize the decision regions from CNNs for remote sensing image classification. Prob-POS provides accessible visual explanation patterns for remote sensing scene images, especially scenes that contain plentiful small targets or scene similar to the texture image.
2. A probe network is proposed to generate Prob-CAM. The weight of a particular class in the probe layer precisely locates the distinct target; therefore, the achieved explanation map extracts discriminative objects for classifying a specific scene category in any layer.
3. We develop a metric called wPO to quantitatively evaluate the explanation effectiveness of each layer for different categories. For a specific class, the most suitable layer to be explained can be automatically picked.

2. Related Work

Within remote sensing images, the class activation map (CAM) [19] has been utilized to accomplish various tasks. The authors in [29] proposed a unified feature fusion framework by applying Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm which pushes networks to focus on the most salient parts of the image. The CAM was used to extract segmentation predictions from an intermediate CNNs layer in the segmentation of remote sensing images [30]. A weakly supervised deep learning method based on CAM for multiclass geo-spatial object detection was proposed in [31]. The CAM was modified

to extract the discriminative parts of aircrafts of different categories in [32]. While the CAM has achieved some applications in various tasks of remote sensing images, most of these methods only used or slightly modified CAM to complete the extraction of the target area, and they also focused only on the explanation effect of the last convolutional layer. For the characteristics of remote sensing images rich in low-level semantic information, the mechanism of CAM needs to be further improved if better visual explanation is to be achieved. In the following subsection, we briefly review the related methods of CAM and Grad-CAM.

2.1. Class Activation Map

The authors in [19] found that the global average pooling (GAP) [33] layer can not only be used as regularizing training but also locate deep representation that exposes the implicit attention of CNNs on an image. The CAM for a particular category was proposed by combining convolutional feature maps to indicate the discriminative image regions. Given a network architecture, the layers after last conv-layer are removed. A GAP layer is added after the convolutional feature maps and the spatial average of feature map of each unit is output to compute regression or other losses. Similarly, the weighted sum of the last convolutional feature maps is calculated to obtain class activation maps.

Specifically, for a given image of class c , let $f_k(x, y)$ indicate the activation of the k -th unit in the last convolutional layer at location (x, y) . The feature map after global average pooling F_k is $\sum_{x,y} f_k(x, y)$. Thus, the input to the softmax S_c is calculated as

$$S_c = \sum_k w_k^c F_k = \sum_{x,y} \sum_k w_k^c f_k(x, y) \quad (1)$$

where w_k^c denotes the weight corresponding to class c for k -th unit. To a certain degree, w_k^c is considered as the importance of F_k for class c , so the class activation map M_c of each pixel for class c is given by

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \quad (2)$$

It should be noted that M_c actually describes the importance of the activation at location (x, y) leading to the classification of the specific image to class c . Finally, the saliency map that reveals the most relevant regions to a particular category is obtained by upsampling the class activation map to the size of the input image.

2.2. Gradient-Weighted Class Activation Mapping

The CAM approach has an obvious drawback that it could only visualize the last convolutional layer, thus CAM is restricted to particular kind of CNN architectures performing global average pooling over convolutional maps immediately prior to prediction. Moreover, the change in architectures may cause damage to the network performance. An advanced approach called gradient-weighted class activation map (Grad-CAM), which is applicable to any CNN-based model, was proposed in [28]. For those kinds of fully convolutional architectures, Grad-CAM can be reduced to CAM.

For a given image of class c , let A^k and Y^c represent the feature map of the k -th unit and the final score before softmax respectively. First, the neuron importance weights w_k^c in Grad-CAM is redefined by computing the gradient of score for class c as:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \quad (3)$$

where Z is the total number of pixels in the activation map. This weight w_k^c represents a partial linearization of the deep network downstream from A , and also indicates the

importance of A^k for class c . Next, each spatial location in the specific class saliency map is calculated as:

$$L_{Grad-CAM}^c = \text{ReLU} \left(\sum_k w_k^c A_{ij}^k \right) \tag{4}$$

Specifically, a Relu operation is applied to the linear combination of maps because features with a positive influence on a specific class are of interest. To obtain fine-grained pixel-scale saliency maps, $L_{Grad-CAM}^c$ is similarly upsampled to the size of the input image.

We illustrate the flow chart of applying CAM and Grad-CAM in visualizing the explanation map of a storage tank image in Figure 2.

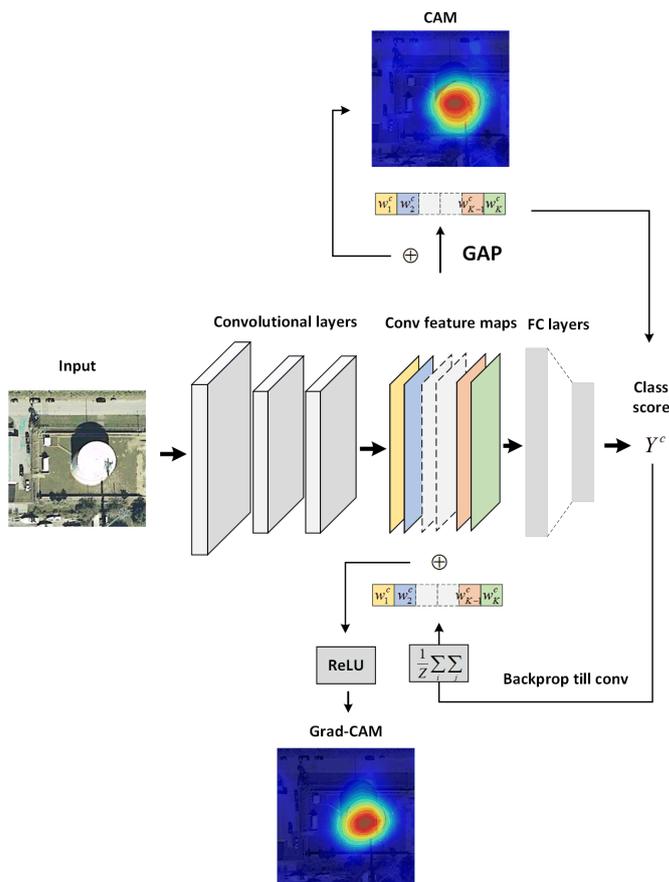


Figure 2. The flow chart of the class activation map (CAM) and gradient-weighted class activation map (Grad-CAM).

3. Proposed Method

3.1. Probe Network

In accordance with the characters mentioned above, let A represent the feature map of an input image with class c . Specifically, $A^k \in \mathbb{R}^{M \times N}$ denotes the feature map after the k -th unit in l -th convolutional layer, where $M \times N$ indicates the size of the feature map.

When studying the visual explanations of l -th convolutional layer, as shown in Figure 3, a probe network composed of a probe layer, a ReLU layer and a softmax layer following behind are built and trained. The feature maps of l -th convolutional layer are employed as input data and the output of the probe network is also the category number. The proposed probe layer is a conventional 2d-convolution. For this to occur, the size of kernel matrix \mathcal{F} of the probe layer should be set as $M \times N \times K \times C$, where K and C denote the depth of l -th convolutional layer and the total number of categories, respectively. The score input to the

softmax $S_c = \sum_k A^{l_k} \otimes \mathcal{F}_k^c$, where \otimes denotes the convolution operation. Thus, the output of the softmax for class c is $P_c = \frac{\exp(S_c)}{\sum_c \exp(S_c)}$.

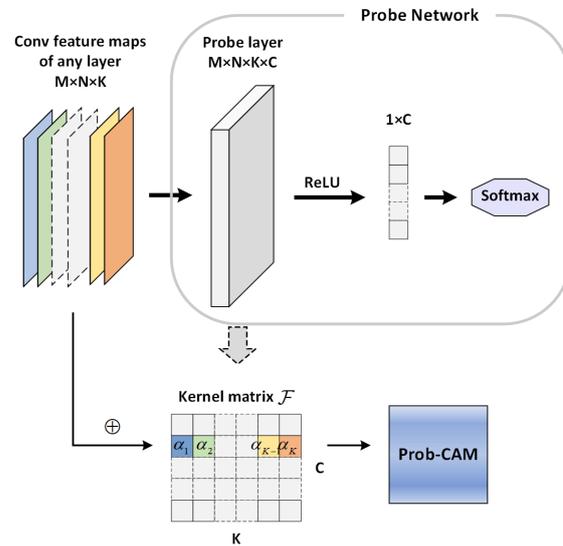


Figure 3. The flow chart of Prob-CAM.

To some extent, the concept of probe layer is similar to the global average pooling layer. It not only plays a role as regularizing training but also keeps the spatial location of the potential salient target. Furthermore, the probe network is more lightweight owing to the lack of a need for another fully-connected layer for final score. The weights of class c in probe layer imply the importance of the input feature map A^{l_k} for the current class. The weight is computed as the average value of the k -th unit for class c .

$$\alpha_k^c = \frac{1}{MN} \sum_{m=1}^M \sum_{n=1}^N \mathcal{F}_k^c \quad (5)$$

Subsequently, the salient region map for each spatial location of class c is performed as a weighted combination of feature maps after l -th convolutional layer

$$L_{Prob-CAM}^c = \sum_{k=1}^K \alpha_k^c A_{i,j}^{l_k} \quad (6)$$

The size of the achieved saliency map is the same as that of a feature map with $M \times N$, but each unit is supposed to be activated within its receptive field. The $L_{Prob-CAM}^c$ at different spatial locations forms a heat map of the current convolutional layer, so enlarging the heat map to the original image size can help obtain more details of the target, making it more intuitive and comprehensive for humans. Finally, the visual explanation image is obtained by simply upsampling the heat map.

3.2. Weighted Probability of Occlusion

The visual saliency maps of particular layers can be acquired by the proposed probe network, afterwards the next task is picking out the most appropriate layer in which the saliency map with a reasonably favourable explainable results. The concept of faithfulness is used for reference to address this problem. For visual explanation, faithfulness is generally quantified by computing the difference of predictions between the saliency map and original image, where the saliency map is usually occluded or blurred to mask out high-energy regions. In the remote sensing image classification task, there are kinds of scene categories similar to texture images. Blurring leads will lead to the prediction of the blurred

image; in particular, largely blurred images tend to fall into several specific categories. Consequently, occlusion is implemented on saliency maps in this paper.

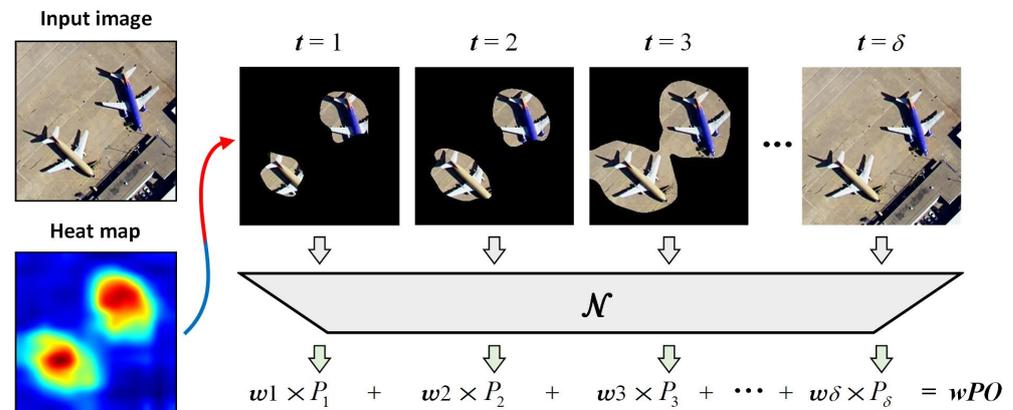


Figure 4. Illustration of the weighted probability of occlusion (wPO).

The weighted probability of occlusion (wPO) is proposed to evaluate the explanation effectiveness of each convolutional layer for different categories. The procedure of wPO is illustrated in Figure 4; here, we take one specific layer for example. For each image I , the pixels sorted by the score in the saliency heat map are equally divided into δ portions. The original image is first all occluded, and then we compute the probabilities by recovering the high-score clean pixels portion by portion. It is noted that the probabilities here are computed by the original whole network. Thus, the probability of the t times operation is

$$P_t = \mathcal{N}(\Phi_t(I, t/\delta)) \quad t = 1, 2, \dots, \delta \quad (7)$$

where \mathcal{N} denotes the network inference, and Φ is the occluded image recovered by adding t/δ clean pixels.

There is another factor that cannot be neglected: the high-scoring regions of the saliency map that embody the effect of visual explanation. Namely, these highlighted image regions imply the explainability of the algorithm. Therefore, the probabilities of occluded images with fewer high-score clean pixels should be given larger weights to enhance their impact. The weight for probability of t time is defined as $2(\delta + 1 - t)/(\delta + 1)\delta$. Finally, the wPO is computed as a linear combination of weighted probabilities

$$\text{wPO} = \frac{1}{\|I\|_0} \sum_I \sum_t \frac{2(\delta + 1 - t)}{(\delta + 1)\delta} P_t \quad (8)$$

where I denotes the images of a particular class. Besides, $\sum_t 2(\delta + 1 - t)/(\delta + 1)\delta = 1$.

At this point, the weighted probability of occlusion that corresponds to every class at each convolutional layer is available, and let wPO_l^c represent it. l denotes the layer number of networks. A larger value means a better explanation result. For class c , the most favourable explanation layer $\hat{l} = \arg \max \text{wPO}_l^c, l = 1, 2, \dots, L$.

4. Experimental Results

In this section, we evaluate the proposed method on two datasets designed for remote sensing image classification and three widely adopted CNNs. We first demonstrate the experimental setup in Section 4.1, including introductions to datasets, implementation details, evaluation criteria, etc. In Section 4.2, we first report the quantitative results of our algorithm on two datasets in detail. Then, we show the evaluation on faithfulness in Section 4.3. Next, in Section 4.4, the explainability is evaluated from two aspects, the intuitive visual presentation of results and the verification of localization capability. Finally, we separately show the explainability of shallow layers on the scene categories that are

similar to the texture image in Section 4.5. The influential and state-of-art visual explanation algorithms CAM, Grad-CAM, and Grad-CAM++ are implemented for comparison.

4.1. Experimental Setup

4.1.1. Datasets Description

Two publicly available datasets designed for remote sensing image classification with a different quantity of images were adopted in our experiments. The first one is UC Merced land-use dataset which consists of images of 21 land-use scene categories selected from aerial orthoimagery with a pixel resolution of 1 foot [34]. Each class contains 100 images with the size of 256×256 pixels. Figure 5 shows some example images randomly selected from UCM. We implemented data augmentation, including flipping and rotating, owing to the limited amount of images. The other one is NWPU-RESISC45 dataset which are extracted from Google Earth (Google Inc., Mountain View, Santa Clara County, CA, USA) [1], covering more than 100 countries and regions all over the world. The NWPU-RESISC45 dataset contains 31,500 images, containing 45 scene classes with 700 images in each class. Every image is 256×256 pixels in the red–green–blue (RGB) color space. The spatial resolution varies from about 30 to 0.2 m per pixel for most scene classes. Compared with UCM, NWPU-RESISC45 not merely covers the most categories of UCM, such as airplane, baseball diamond, forest, freeway, intersection, storage tank, etc., but also includes more generalized scene classification data such as commercial area, mountain, church, palace, cloud, sea ice, etc., making the classification task more challenging. Figure 6 shows some example images from NWPU-RESISC45 with one sample per class.



Figure 5. Example images from the UCM dataset: (1) agricultural; (2) airplane; (3) baseball diamond; (4) beach; (5) building; (6) chaparral; (7) dense residential; (8) forest; (9) freeway; (10) golfcourse; (11) harbor; (12) intersection; (13) medium residential; (14) mobile home park; (15) overpass; (16) parking lot; (17) river; (18) runway; (19) sparse residential; (20) storage tank; (21) tennis court.

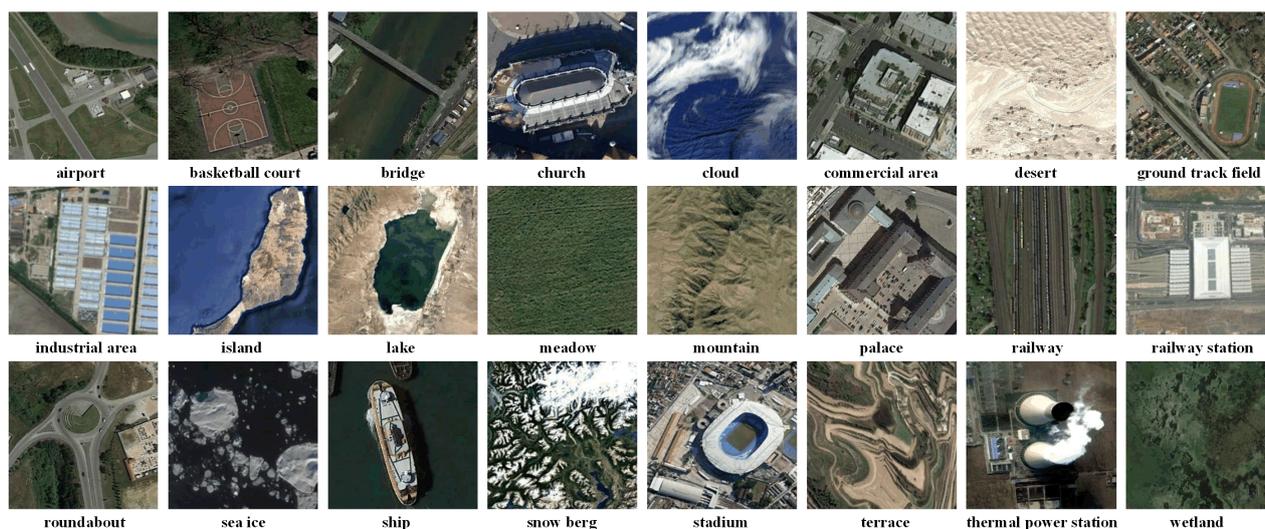


Figure 6. Example images from the NWPU-RESISC45 dataset.

4.1.2. Networks and Parameters Setup

We implemented three practical CNNs with different types of structures on the proposed framework: VGG-16 [16], AlexNet [15] and ResNet-50 [35], and the corresponding numbers of convolutional layers are 5, 16, and 50, respectively. For VGG-16, we took ‘pool1’, ‘pool2’, ‘pool3’, ‘conv4-3’ and ‘conv5-3’ as examples to verify the ability of the proposed method, and ‘pool1’, ‘conv2’, ‘conv3’, ‘conv4’, ‘conv5’ are selected for AlexNet. For ResNet-50, we took ‘pool1’ and the last convolutional layer of each residual block, namely ‘conv2-3’, ‘conv3-4’, ‘conv4-6’ and ‘conv5-3’. When train the probe network, the feature maps directly after the pooling layer and the ReLU layer of the convolutional layer were employed as input.

The three networks are supposed to be re-trained on two remote sensing image datasets at first. The images were randomly divided into subsets for training, validation and testing. The ratios of UCM and NWPU-RESISC45 were set as 60%, 20%, 20% and 20%, 10%, 70%, respectively. These ratios and the identical images in respective subsets were also adopted in the period of training the probe network and computing the weighted probability of occlusion. All the networks including probe networks were trained using SGD [36], with a batch size of 64. During re-training the networks, the initial learning rate was set to 0.05 and divided by 5 every 5 epochs, while the average value of the objective function over the validation set stopped decreasing. The training epochs for three networks are 40, 20 and 50, respectively. As for training of the probe networks, the initial learning rate was set to 0.001 and is divided by 5 every 5 epochs, while the training epochs was set to 20.

The weighted probability of occlusion (wPO) for each category was obtained on the training subsets. The only hyper-parameter δ in wPO was set to 32 in the following experiments.

4.1.3. Evaluation Criteria

The effectiveness of visual explanation is usually measured from two aspects: faithfulness and explainability. The faithfulness objectively evaluates the fidelity of the explanation to the judgment of the original model, while the explainability subjectively describes how much humans can understand the explanation results.

To quantify the faithfulness, two metrics proposed in [37] were adopted in this paper: (i) Average Drop and (ii) Increase in Confidence. The Average Drop compares the average % drop in the model’s confidence of a particular category in an image. It is obtained by calculating the percentage of original probability occupied by the probability drop between original image and occluded image. Thus, a lower Average Drop means higher faithfulness. The Increase in Confidence considers another scenario, where the probability

of the highlighted explanation map might exceed the original one. This index computes the proportion of abovementioned cases over the total number of images, so a higher Increase in Confidence is better. Note that these two evaluation indexes are gained on the testing subset.

Quantifying the explainability of algorithms on remote sensing image classification is tough because there is not a dataset that marks every distinct object in each scene category, and it is hard to mark a specific target in the scenes similar to the texture image. As a consequence, we evaluate the explainability from two aspects. First, considering that the explainability is a subjective measurement, we directly display the explanation maps generated by the proposed method and comparison methods as well. The gray level maps of salient heat maps corresponding to shallow layers are also shown to demonstrate the proposed algorithm's explainability on scene categories similar to the texture image. Second, we refer to the measurement method in [19,28], a weakly supervised localization is implemented to indirectly verify the explainability.

4.1.4. Hardware Equipment

All the experiments were run on an HP Z8 workstation with two Intel Xeon Gold 5120 CPUs with 14 cores, two NVIDIA GeForce RTX 2080 Ti GPUs and a 128-GB memory.

4.2. Optimal Explanation Layer

In this subsection, we report the proposed weighted probability of occlusion (wPO) and generated optimal explanation layer achieved on two datasets and three networks. The wPO of each layer on UCM from VGG-16, AlexNet and ResNet-50 is shown in Tables 1–3. The maximum wPO of each category in all layers is highlighted in bold and marked with a check mark for the ease of comparing. In the meantime, The optimal explanation layer from three networks for each class of NWPU-RESIS45 dataset is presented in Tables 4–6 respectively. The wPO of NWPU-RESISC45 is not presented in detail because of paper length limits.

Table 1. wPO(%) of each layer from VGG-16 on UCM.

	pool1	pool2	pool3	conv4-3	conv5-3
agricultural	23.88(✓)	12.52	2.13	2.66	12.64
airplane	28.86	16.69	16.85	20.83	42.34(✓)
baseballdiamond	47.41	9.92	4.58	8.62	85.49(✓)
beach	9.81	11.28	3.11	6.63	48.89(✓)
buildings	57.41(✓)	33.09	32.81	28.33	43.73
chaparral	87.41	91.99(✓)	6.55	1.67	19.17
denseresidential	41.67	46.82(✓)	26.66	17.51	14.88
forest	68.90(✓)	63.25	20.59	2.83	4.80
freeway	43.32	23.04	23.01	21.17	45.82(✓)
golfcourse	46.70	23.21	90.92	96.60(✓)	82.09
harbor	63.37	76.63	70.94	79.54	82.31(✓)
intersection	35.19(✓)	22.54	11.60	12.72	20.82
mediumresidential	6.02	13.17	16.14	22.75(✓)	17.41
mobilehomepark	16.18	60.21	64.73(✓)	48.77	37.14
overpass	25.45	12.55	24.97	22.68	44.73(✓)
parkinglot	17.11	18.71	19.96	58.03(✓)	51.75
river	49.63	31.24	16.19	24.83	54.13(✓)
runway	4.27	2.15	0.61	1.15	5.17(✓)
sparseresidential	34.75	9.66	17.49	16.57	51.88(✓)
storagetanks	64.82	48.55	47.19	67.20	81.76(✓)
tenniscourt	26.13	12.50	28.16	35.77	65.34(✓)

Table 2. wPO(%) of each layer from AlexNet on UCM.

	pool1	conv2	conv3	conv4	conv5
agricultural	6.51(✓)	5.43	0.95	2.82	3.32
airplane	13.41	22.10	16.82	15.85	36.47(✓)
baseballdiamond	19.85	26.05	48.39	55.01(✓)	52.71
beach	6.96	11.87	4.56	8.38	13.73(✓)
buildings	36.34	38.72(✓)	34.71	35.56	38.42
chaparral	2.29(✓)	1.67	0.29	0.48	0.44
denseresidential	17.35(✓)	16.26	7.63	5.23	6.84
fores	61.61(✓)	43.30	17.63	15.13	20.59
freeway	5.56	12.37	8.95	9.38	13.30(✓)
golfcourse	9.19	11.95	33.53	40.68	43.72(✓)
harbor	27.28	36.35	50.74	52.08	78.82(✓)
intersection	23.22(✓)	20.70	11.96	9.96	16.54
mediumresidential	5.97	4.79	4.95	4.62	6.65(✓)
mobilehomepark	24.95	30.48(✓)	12.49	9.82	12.64
overpass	13.43	18.43	15.55	15.83	38.89(✓)
parkinglot	3.01	11.07	15.34(✓)	11.33	12.30
river	19.75	19.73	15.51	28.26	36.07(✓)
runway	28.77	56.86	42.05	65.33(✓)	56.59
sparseresidential	17.10	15.04	13.56	19.02	19.54(✓)
storagetanks	61.80	65.05	72.36	78.00	79.11(✓)
tenniscourt	20.07	25.73	38.48	40.73	41.13(✓)

Table 3. wPO(%) of each layer from ResNet-50 on UCM.

	pool1	conv2-3	conv3-4	conv4-6	conv5-3
agricultural	18.92(✓)	7.47	0.60	0.54	1.32
airplane	10.60	5.50	14.82	19.70	38.94(✓)
baseballdiamond	77.08	65.56	53.87	67.20	80.04(✓)
beach	36.57	38.12	37.17	39.86	46.18(✓)
buildings	13.87(✓)	10.47	3.21	2.38	5.01
chaparral	19.32(✓)	4.74	3.97	3.50	3.71
denseresidential	11.88(✓)	2.23	5.51	7.69	10.14
forest	45.26(✓)	35.68	38.84	34.66	28.01
freeway	43.36	19.12	19.67	26.07	58.73(✓)
golfcourse	11.83	3.63	43.35	64.51	51.31(✓)
harbor	43.36	19.12	19.67	26.07	58.73(✓)
intersection	17.97(✓)	17.92	4.57	4.74	3.63
mediumresidential	11.83	3.63	43.35	64.51(✓)	51.31
mobilehomepark	58.12	62.78(✓)	43.09	42.26	33.21
overpass	5.17	9.59	8.04	8.69	14.18(✓)
parkinglot	9.44	18.48	19.44	24.24(✓)	21.40
river	12.76	18.43	19.85	17.39	22.73(✓)
runway	2.11	1.25	1.91	2.72	35.22(✓)
sparseresidential	26.68	25.86	24.68	16.59	27.93(✓)
storagetanks	19.61	29.97	31.78	33.43	35.61(✓)
tenniscourt	6.31	11.48	47.81	51.63	61.83(✓)

Table 4. The optimal explanation layer from VGG-16 for each class of NWPU-RESIS45 dataset.

pool1	chaparral, commercial area, dense residential, desert, forest, harbor, meadow, railway, railway station, snowberg
pool2	
pool3	parking lot, roundabout, wetland
conv4-3	golf course, intersection, island, medium residential, mobile home park
conv5-3	airplane, airport, baseball diamond, basketball court, beach, bridge, church, circular farmland, cloud, freeway, ground track field, industrial area, lake, mountain, overpass, palace, rectangular farmland, river, runway, sea ice, ship, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station

Table 5. The optimal explanation layer from AlexNet for each class of NWPU-RESIS45 dataset.

pool1	dense residential, desert, forest, harbor, meadow, snowberg
conv2	bridge, chaparral, commercial area, freeway, medium residential, mobile home park, mountain, railway, railway station, wetland
conv3	parking lot
conv4	airport, baseball diamond, beach, church, circular farmland, golf course, ground track field, industrial area, rectangular farmland, roundabout, stadium, terrace
conv5	airplane, basketball court, cloud, intersection, island, lake, overpass, palace, river, runway, sea ice, ship, sparse residential, storage tank, tennis court, thermal power station

Table 6. The optimal explanation layer from ResNet-50 for each class of NWPU-RESIS45 dataset.

pool1	chaparral, desert, forest, harbor, meadow, mountain, railway, railway station, snowberg, wetland
conv2-3	commercial area, dense residential, industrial area, parking lot, terrace
conv3-4	medium residential, roundabout
conv4-6	beach, circular farmland, golf course, intersection, mobile home park, rectangular farmland
conv5-3	airplane, airport, baseball diamond, basketball court, bridge, church, cloud, freeway, ground track field, island, lake, overpass, palace, river, runway, sea ice, ship, sparse residential, stadium, storage tank, tennis court, thermal power station

From these results, we made the following observations. First, not all scene classes identified the last convolutional layer as the optimal explanation layer, which proves the argument of this article. The scene classes containing clear objects of a large scale generally chose the last layer, such as airplane, storage tanks, tennis court and so on. In contrast, the scene classes similar to the texture image, such as agricultural, chaparral, forest and meadow, had the obvious trend that the explainability of the lower layer was stronger than the higher layer. Moreover, the middle layer was most likely to be chosen by the scene classes containing targets with a small scale, such as parking lot and mobile home park. This could be reasonably explained: the shallow layers in CNNs capture more detailed features, whereas the deep layers learn more holistic features of objects owing to the larger receptive fields. The kernels in the first convolutional layer usually represent texture information. Therefore, the low and middle layers provide more discriminative features for the abovementioned categories. In other words, the shallow layers are more suitable to be explained for these classes. The visual explanations are shown in the following subsections.

Second, although there were small differences in some classes owing to the diversity of networks and datasets, the optimal explanation layer of identical class basically remained the same between different networks and datasets. This indicates that our argument possesses a certain universality. There are small differences in some classes on account of the diversity of networks and datasets.

4.3. Faithfulness

The quantitative evaluation on the faithfulness of the proposed Prob-POS and comparison methods is reported in this subsection. Two different metrics proposed in [37] are adopted: (i) Average Drop and (ii) Increase in Confidence. For comparison, we implement CAM, Grad-CAM (GCAM) and Grad-CAM++ (GCAM++) on the two datasets and evaluate their faithfulness. Additionally, we also apply our proposed weighted probability of occlusion (wPO) selection strategy on Grad-CAM (GCAM&POS) and Grad-CAM++ (GCAM++&POS) to verify the effectiveness. It is noted that the Grad-CAM and Grad-CAM++ are implemented on the same retrained networks that are used by Prob-POS. Nevertheless, CAM changes the network architecture, so we sufficiently train the CAM networks according to the original reference on both datasets.

Table 7. Quantitative evaluation of faithfulness on UCM.

	Average Drop %			Incr. in Confidence %		
	VGG-16	AlexNet	ResNet-50	VGG-16	AlexNet	ResNet-50
Prob-POS	32.37	48.24	37.28	9.29	4.95	6.86
Prob-CAM	40.41	55.70	42.45	6.29	3.14	6.23
CAM	42.38	51.14	49.62	3.96	4.09	5.73
GCAM	52.55	60.36	55.78	3.19	2.48	3.28
GCAM++	53.89	59.52	54.74	4.06	2.37	4.77
GCAM&POS	48.97	56.96	50.49	4.10	2.96	4.31
GCAM++&POS	49.33	54.14	52.16	4.83	2.8	5.25

Table 8. Quantitative evaluation of faithfulness on NWPU-RESISC45.

	Average Drop %			Incr. in Confidence %		
	VGG-16	AlexNet	ResNet-50	VGG-16	AlexNet	ResNet-50
Prob-POS	58.45	65.42	59.33	8.79	4.77	6.58
Prob-CAM	67.97	74.49	66.41	4.10	2.40	3.63
CAM	70.42	72.23	69.21	3.92	3.25	4.47
GCAM	73.15	79.73	71.27	2.93	2.18	3.55
GCAM++	69.22	79.52	70.06	3.48	3.67	4.26
GCAM&POS	68.17	76.71	66.93	3.99	3.29	5.13
GCAM++&POS	62.53	74.43	65.44	4.75	4.16	4.92

The results of faithfulness on the UCM and NWPU-RESISC45 dataset are shown in Tables 7 and 8, respectively. The lower Average Drop and higher Increase in Confidence indicate the better faithfulness. It can be seen that the proposed Prob-POS achieved the best performance in both metrics and both datasets. This is because the Prob-POS chooses the layers with the highest explainability for each category and the fidelity of global explanation consequently outperforms the explanation methods that only concentrate on the last layer. When we only employed the proposed method on the last layer (Prob-CAM in the Table), the performance was also better than that of the comparison methods on two

datasets with VGG-16. However, CAM achieved better faithfulness on AlexNet, a possible conjecture is that CAM adds another convolutional layer with 1024 kernels after original networks, enhancing the feature representation power of CNNs with a small volume, such as AlexNet, and so the explainability was boosted. In addition to the above two conclusions, the Average Drop and Increase in Confidence were both promoted by applying the wPO selection strategy on Grad-CAM and Grad-CAM++; hence, the wPO selection strategy is conducive to improving the faithfulness of visual explanation algorithms.

4.4. Explainability

Explainability subjectively describes how much humans could understand the explanation results. This is a direct metric evaluating the performance of a visual explanation algorithm. In this subsection, we respectively show the generated explanation maps from the last convolutional layer and middle-shallow layers to demonstrate the effectiveness of Prob-POS and Prob-CAM. If the salient regions of the generated explanation maps involve more specific semantic meanings, the explainability is higher.

4.4.1. Explainability of Last Layer

We sampled images whose optimal explanation layer was the last convolutional layer from UCM and NWPU-RESISC45 datasets, and display their corresponding explanation results on the last layer generated by the proposed Prob-CAM, CAM, Grad-CAM and Grad-CAM++ in Figure 7. It can be observed that the explanation maps from the last layer mainly locate discriminative targets such as the airplane, the ground track field and the runway, or locate the whole area that contains many specific targets, for example, the pier berthing with rows of boats, shown in the harbor scene. These salient regions of the last layer are generally associated with targets of a relatively large scale.

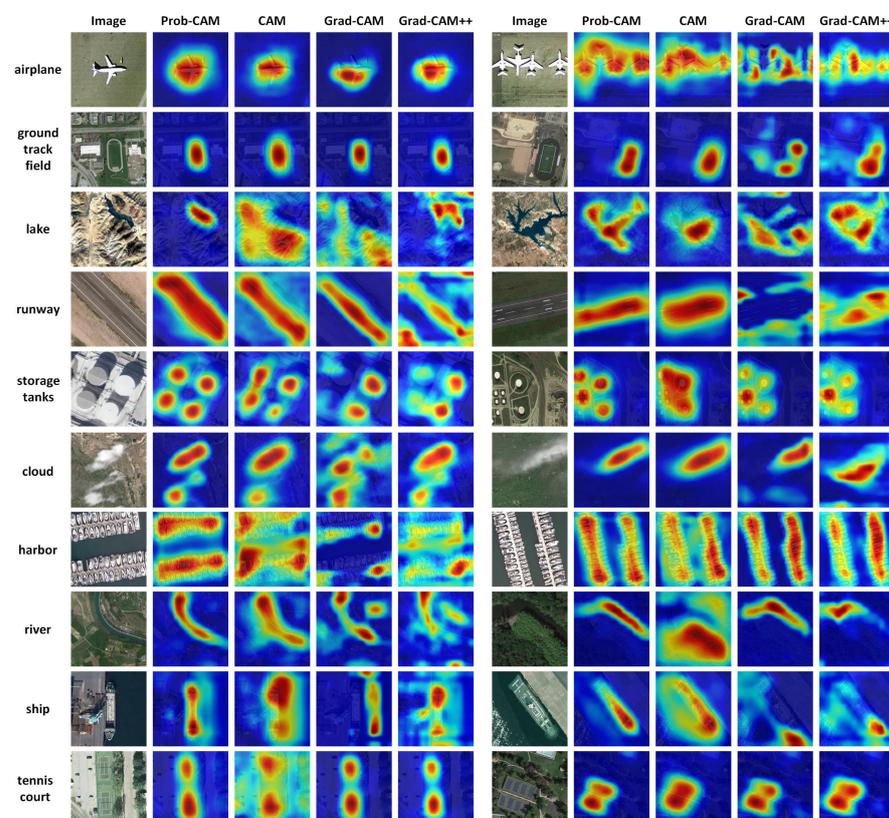


Figure 7. The explanation maps for sampled images from different scene class generated by Prob-CAM, CAM and Grad-CAM. They were all obtained on the last layer, which was the optimal explanation layer for the shown categories.

Furthermore, the Prob-CAM achieved similar explainability as CAM, Grad-CAM and Grad-CAM++ on some scene classes, such as the airplane and the ground track field. However, in some categories, the explainability of Prob-CAM was better than that of the other three methods. For example, the salient region of runway generated by Prob-CAM includes more complete runway area than that of the others, and the explanation map on storage tanks generated by our method precisely locates four storage tanks, yet the location of CAM has errors, Grad-CAM and Grad-CAM++ both omit two storage tanks. On the whole, the explainability of Prob-CAM on last layer is equal to or better than the other three methods on remote sensing images.

4.4.2. Explainability of Middle-Shallow Layer

Similarly, we sampled images whose optimal explanation layer was the middle or shallow layer from two datasets. The results of corresponding explanation maps obtained by the proposed method and comparison methods are shown in Figure 8, where the 'Prob-POS' indicates the explanation maps generated by the Prob-POS on the optimal explanation layer for each scene class, the 'Prob-CAM' refers to the explanation maps obtained by the Prob-CAM on the last layer, the 'CAM', 'Grad-CAM' and 'Grad-CAM++' also represent the explanation maps on last layer. The following conclusions could be drawn by observing these results.

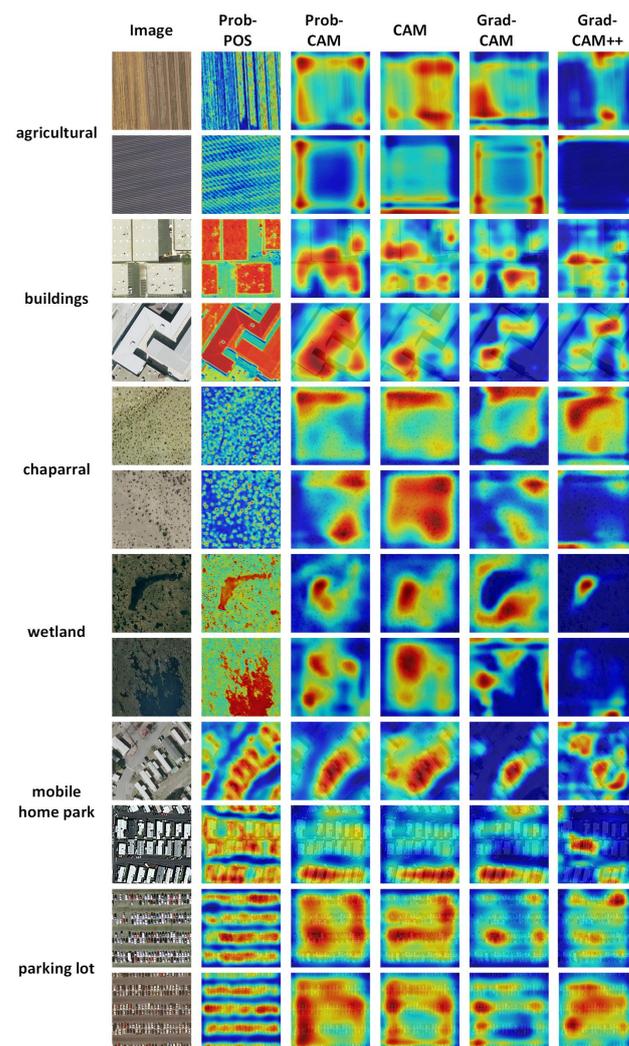


Figure 8. The explanation maps for sampled images from different scene classes generated by Prob-POS, CAM and Grad-CAM. They were obtained on middle or shallow layer, which was the optimal explanation layer for the category shown.

First, the explanation maps generated by the proposed method on shallow layer could provide higher explainability for specific scene classes such as agricultural, chaparral and wetland. As shown in Figure 8, the explanation maps on the shallow layer clearly depict the texture information for above mentioned scene images. Compared with the jumbled and irregular salient regions of explanation maps on the last layer, the results of shallow layer are obviously more understandable for humans in visual appearance. Moreover, there was a surprising phenomenon where the buildings scene that contained apparent targets of a large scale identified the shallow layer as the optimal explanation layer. We give a possible conjecture that the buildings are extremely regular shapes so the marginal information learned by shallow layers could represent these shapes more properly.

Second, the scene classes that contain plenty of targets of a small scale were better visually explained on middle layers. As shown in Figure 8, the mobile homes in the mobile home park and the cars in parking lot are precisely located in the explanation maps on the middle layer, yet the salient regions in explanation maps on the last layer omit more or fewer targets, depending on methods applied. Considering the two points mentioned above, we may draw another conclusion that the explanation maps on last layer trend to describe centralized salient regions, whereas the tiny discrete objects and detailed information are more likely described by explanation maps on middle and shallow layers respectively.

Finally, the ideal visual explanation results rely on understandable explanation maps generated by the Prob-CAM as well as the optimal explanation layer chosen by the wPO selection strategy. These two aspects jointly verify the effectiveness of the proposed Prob-POS framework.

4.4.3. Evaluating Localization

In order to provide quantitative comparison results on explainability, we referred to the operations in [19,28] and performed a weakly supervised localization in the context of scene classification. The localization capability helps to indirectly reflect the explainability. Here, we took VGG-16 as the backbone to implement the experiment. The NWPU-RESISC45 dataset was used as the training dataset. The NWPU VHR-10 dataset [38], which is a publicly available 10-class geospatial object detection dataset, was used for testing. Except from vehicle class, the other nine classes of NWPU VHR-10, which are airplane, ship, oil tank, baseball diamond, tennis court, basketball court, ground track field, harbor and bridge, are all included in the NWPU-RESISC45 dataset.

When inputting a testing scene image, to generate bounding boxes from explanation results, we also used a simple thresholding technique to segment the saliency heatmap. First, the regions of which the value is above 20% of the max value of the heatmap were segmented. Then, the bounding boxes were drawn on the segments that cover the connected component. We performed this procedure on the proposed Prob-POS and comparison methods, respectively. Note that the explanation heatmap of the Prob-POS was generated from the optimal convolutional layer. The comprehensive metric average precision (AP), in which the area overlap ratio was set to 0.5, was adopted to measure the location performance. The experimental results are shown in Table 9.

Table 9. Localization capability of the Prob-POS and comparison methods in terms of the evaluation metric of AP(%)

	Airplane	Ship	Oil Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge
Prob-POS	6.49	1.36	6.67	12.61	0.40	0.29	2.23	2.17	0.22
CAM	5.36	0.16	2.84	12.27	0.13	0.11	1.79	0.07	0.04
GCAM	4.47	0.27	4.62	10.52	0.34	0.21	1.48	1.08	0.07
GCAM++	5.85	0.31	4.33	10.07	0.41	0.26	1.61	1.15	0.07

It can be observed that the Prob-POS shows the best localization capability in most scenes. In the scenes of ship, oil tank and harbor, the Prob-POS achieves clear advantages. These quantitative comparison results further verify the explainability of Prob-POS. It should be noted that the average precision in this experiment is generally lower than that with the method designed for object detection. This is because there are many factors we had not taken into account, such as the diversity of scale and the segmentation of dense objects. But this implementation is capable and fair to verify the explainability.

4.5. Explainability on Texture Information

To further demonstrate explainability on texture information of the explanation maps obtained by the proposed method, we simply display the gray level maps of Prob-CAM on the shallow layer of networks. Just like the above operation, here we also sampled images whose optimal explanation layer was first layer from corresponding networks. The results are shown in Figure 9. We took the gray level maps generated by Grad-CAM, Grad-CAM++ and the texture image obtained by the entropy filter for comparison.

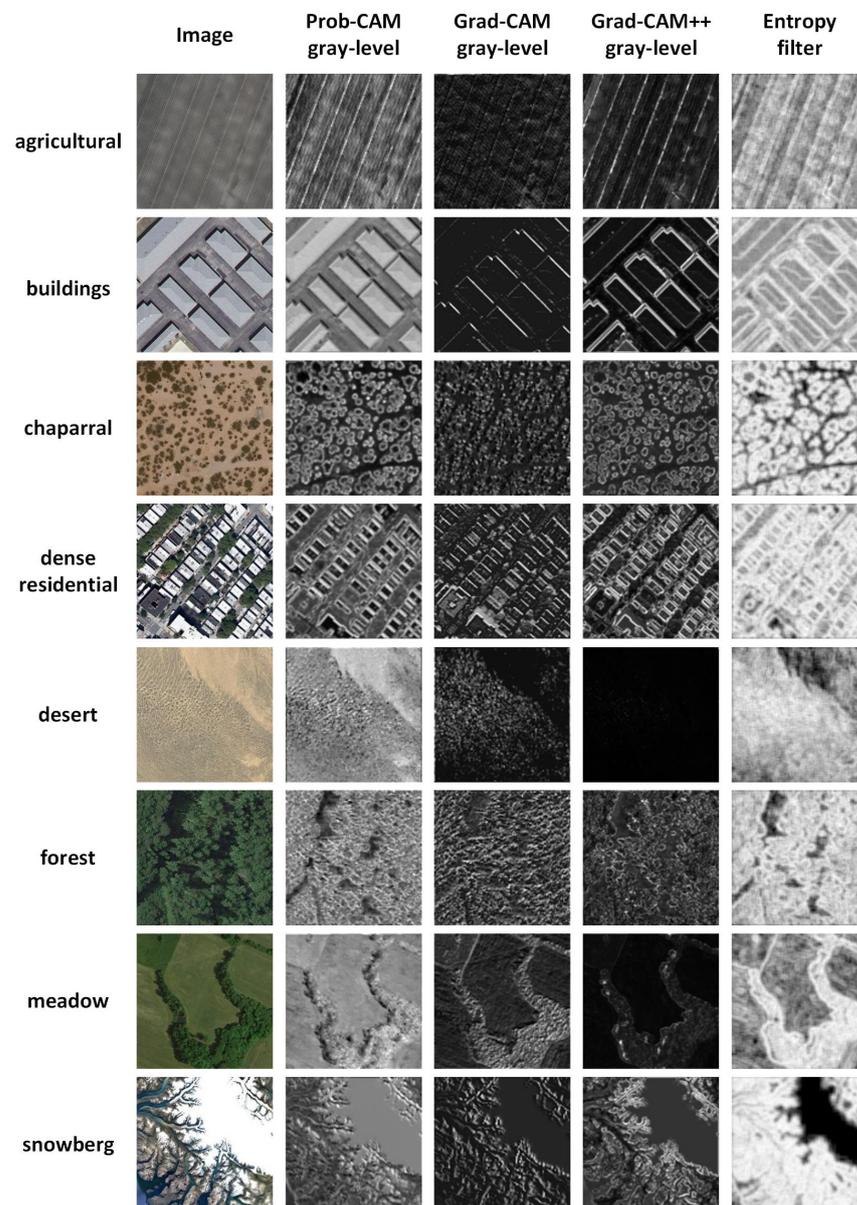


Figure 9. The gray level maps of Prob-CAM and Grad-CAM on the shallow layer, and the texture image obtained by the entropy filter.

It can be seen from Figure 9 that the gray level maps of Pro-CAM show more clear and detailed texture information than the other methods on most scene classes. Although the gray level maps of Grad-CAM and Grad-CAM++ also present visual texture information, some detailed elements are missing, for example part of the buildings are hard to identify in the buildings scene and dense residential scene. In addition, the contrast of Grad-CAM and Grad-CAM++ is generally low, so the present images are fuzzy. This is because the hypothesis proposed in Grad-CAM and Grad-CAM++ is based on the final convolutional layer. However, the proposed Prob-CAM does not suffer from this limitation, and thus it works better on shallow layers.

5. Conclusions

In this paper, we proposed a framework for improving visual explanations of remote sensing images called Prob-POS. The framework consists of two parts, the class activation map based on probe network (Prob-CAM) and the weighted probability of occlusion (wPO) selection strategy.

First, we used the probe network, which is composed of a probe layer, a ReLU layer as well as a softmax, which is a simple but effective algorithm to elaborately discriminate between the feature maps with different classes. We utilized the weights of a particular class in a well-trained probe layer to linearly combine the feature maps, and then the Prob-CAM is obtained. The Prob-CAM can be applied to convolutional feature maps after any layer.

Second, we developed a metric called wPO to measure which layer was the most suitable to be explained for different class. The variational weights were added to the probability of occlusion considering that the high-scoring regions in the explanation map embody the explainability of the visual image. The wPO provided a quantitative evaluation of the explanation effectiveness of each layer for different categories, and further, the optimal explanation layer can be automatically picked.

Finally, extensive experiments were carried out over publicly available datasets, UCM and NWPU-RESISC45, as well as the prevalent networks VGG-16 and AlexNet. The experimental results showed that the faithfulness of Prob-POS was better than that of other algorithms that only concentrated on the last convolutional layer. Moreover, the Prob-POS also achieved better explainability on remote sensing images, as the proper layer was selected to generate explanation maps. Thus, compared with visual results on the last layer, the Prob-CAM on middle or shallow layers provided accessible visual explanation patterns for remote sensing scene images, especially scenes that contained plentiful small targets or were similar to the texture image.

Some aspects of this study need further research. First, the gradients in CNNs were shown to be helpful for locating distinct objects; thus, it is worth utilizing the gradients to assist in the generation of explanation maps. Second, the location ability of objects is generally used to compute the explainability in existing methods, yet this is infeasible in some remote sensing categories. Consequently, a quantified and fair metric should be proposed to evaluate the explainability of explanation algorithms for remote sensing images.

Author Contributions: Conceptualization, X.G. and B.H.; methodology, X.G.; software, X.G. and Z.W.; validation, X.G., Z.W. and B.R.; formal analysis, X.G. and Z.W.; investigation, B.H.; resources, B.R.; data curation, B.R.; writing—original draft preparation, X.G.; writing—review and editing, B.H. and Z.W.; supervision, S.W. and L.J.; project administration, L.J.; funding acquisition, B.H., S.W. and L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62171347, 61877066, 61771379, 62001355, 62101405; the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61621005; the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY03-05 and 2021ZDLGY02-08; the Science and Technology Program in Xi'an of China under Grant XA2020-RGZNTJ-0021; 111 Project.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [[CrossRef](#)]
2. Wang, S.; Guan, Y.; Shao, L. Multi-granularity canonical appearance pooling for remote sensing scene classification. *IEEE Trans. Image Process.* **2020**, *29*, 5396–5407. [[CrossRef](#)]
3. Wan, Y.; Ma, A.; Zhong, Y.; Hu, X.; Zhang, L. Multiobjective hyperspectral feature selection based on discrete sine cosine algorithm. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3601–3618. [[CrossRef](#)]
4. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.; Zhang, L. Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. 2513443. [[CrossRef](#)]
5. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [[CrossRef](#)]
6. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
7. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [[CrossRef](#)]
8. Pan, X.; Zhao, J. High-Resolution Remote Sensing Image Classification Method Based on Convolutional Neural Network and Restricted Conditional Random Field. *Remote Sens.* **2018**, *10*, 920. [[CrossRef](#)]
9. Shen, J.; Yu, T.; Yang, H.; Wang, R.; Wang, Q. An Attention Cascade Global&Local Network for Remote Sensing Scene Classification. *Remote Sens.* **2022**, *14*, 2042. [[CrossRef](#)]
10. Zhao, X.; Zhang, J.; Tian, J.; Zhuo, L.; Zhang, J. Residual Dense Network Based on Channel-Spatial Attention for the Scene Classification of a High-Resolution Remote Sensing Image. *Remote Sens.* **2020**, *12*, 1887. [[CrossRef](#)]
11. Gu, S.; Zhang, R.; Luo, H.; Li, M.; Feng, H.; Tang, X. Improved SinGAN Integrated with an Attentional Mechanism for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 1713. [[CrossRef](#)]
12. Shi, C.; Zhao, X.; Wang, L. A Multi-Branch Feature Fusion Strategy Based on an Attention Mechanism for Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*. [[CrossRef](#)]
13. Shi, C.; Zhang, X.; Sun, J.; Wang, L. Remote Sensing Scene Image Classification Based on Dense Fusion of Multi-level Features. *Remote Sens.* **2021**, *13*, 4379. [[CrossRef](#)]
14. Li, Q.; Yan, D.; Wu, W. Remote Sensing Image Scene Classification Based on Global Self-Attention Module. *Remote Sens.* **2021**, *13*, 4542. [[CrossRef](#)]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
17. Nogueira, K.; Penatti, O.A.; Santos, J.A.D. Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification. *Pattern Recognit.* **2016**, *61*, 539–556. [[CrossRef](#)]
18. Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3387–3395.
19. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
20. Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6541–6549.
21. Fong, R.C.; Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3429–3437.
22. Dong, Y.; Su, H.; Zhu, J.; Zhang, B. Improving interpretability of deep neural networks with semantic information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4306–4314.
23. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
24. Zhang, Q.; Nian Wu, Y.; Zhu, S.C. Interpretable convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8827–8836.
25. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. *arXiv* **2018**, arXiv:1810.03292.

26. Pérez-Suay, A.; Adsuara, J.E.; Piles, M.; Martínez-Ferrer, L.; Díaz, E.; Moreno-Martínez, A.; Camps-Valls, G. Interpretability of Recurrent Neural Networks in Remote Sensing. In Proceedings of the IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 3991–3994.
27. Huang, Z.; Datcu, M.; Pan, Z.; Lei, B. Deep SAR-Net: Learning objects from signals. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 179–193. [[CrossRef](#)]
28. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
29. Li, J.; Lin, D.; Wang, Y.; Xu, G.; Zhang, Y.; Ding, C.; Zhou, Y. Deep Discriminative Representation Learning with Attention Map for Scene Classification. *Remote Sens.* **2020**, *12*, 1366. [[CrossRef](#)]
30. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 207. [[CrossRef](#)]
31. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *Isprs J. Photogramm. Remote Sens.* **2018**, *146*, 182–196. [[CrossRef](#)]
32. Fu, K.; Dai, W.; Zhang, Y.; Wang, Z.; Yan, M.; Sun, X. MultiCAM: Multiple Class Activation Mapping for Aircraft Recognition in Remote Sensing Images. *Remote Sens.* **2019**, *11*, 544. [[CrossRef](#)]
33. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400
34. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Bottou, L. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7700, pp. 421–436.
37. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
38. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *Isprs J. Photogramm. Remote. Sens.* **2014**, *98*, 119–132. [[CrossRef](#)]