remote sensing

MDPI

Article

C-RISE: A Post-Hoc Interpretation Method of Black-Box Models for SAR ATR

Mingzhe Zhu [1], Jie Cheng [1,*], Tao Lei [2], Zhenpeng Feng [1], Xianda Zhou [3], Yuanjing Liu [1] and Zhihan Chen [1]

[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China; zhumz@mail.xidian.edu.cn (M.Z.); zpfeng_1@stu.xidian.edu.cn (Z.F.); liuyuanjing@stu.xidian.edu.cn (Y.L.); chenzhihan@stu.xidian.edu.cn (Z.C.)
[2] Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China; leitao@sust.edu.cn
[3] National Key Laboratory of Science and Technology on Aerospace Intelligence Control, Beijing Aerospace Automatic Control Institute, Beijing 100854, China; zhouxianda999@gmail.com
* Correspondence: agentcj@stu.xidian.edu.cn

Abstract: The integration of deep learning methods, especially Convolutional Neural Networks (CNN), and Synthetic Aperture Radar Automatic Target Recognition (SAR ATR) has been widely deployed in the field of radar signal processing. Nevertheless, these methods are frequently regarded as black-box models due to the limited visual interpretation of their internal feature representation and parameter organization. In this paper, we propose an innovative approach named C-RISE, which builds upon the RISE algorithm to provide a post-hoc interpretation technique for black-box models used in SAR Images Target Recognition. C-RISE generates saliency maps that effectively visualize the significance of each pixel. Our algorithm outperforms RISE by clustering masks that capture similar fusion features into distinct groups, enabling more appropriate weight distribution and increased focus on the target area. Furthermore, we employ Gaussian blur to process the masked area, preserving the original image structure with optimal consistency and integrity. C-RISE has been extensively evaluated through experiments, and the results demonstrate superior performance over other interpretation methods based on perturbation when applied to neural networks for SAR image target recognition. Furthermore, our approach is highly robust and transferable compared to other interpretable algorithms, including white-box methods.

Keywords: Convolutional Neural Networks (CNN); Synthetic Aperture Radar Automatic Target Recognition (SAR ATR); C-RISE; cluster; Gaussian blur

Citation: Zhu, M.; Cheng, J.; Lei, T.; Feng, Z.; Zhou, X.; Liu, Y.; Chen, Z. C-RISE: A Post-Hoc Interpretation Method of Black-Box Models for SAR ATR. *Remote Sens.* **2023**, *15*, 3103. https://doi.org/10.3390/rs15123103

Academic Editor: Lionel Bombrun

Received: 12 April 2023
Revised: 31 May 2023
Accepted: 8 June 2023
Published: 14 June 2023

1. Introduction

Synthetic Aperture Radar (SAR) is a kind of active earth-observation system which can produce high-resolution image all day, has been widely used in ground observation and military reconnaissance. One of its primary applications is the detection and identification of various military targets [1,2]. With the enhancement of SAR data acquisition capability, Synthetic Aperture Radar Automatic Target Recognition (SAR ATR) [3] has become a key technology and research hotspot of radar signal processing. Traditional SAR image recognition methods, such as template matching [4], feature-based approaches [5,6], and CAD model-based methods [7], predominantly rely on the statistical and physical characteristics inherent in the image data. This methodology offers robust interpretability, as the identified features and models possess well-defined statistical or physical interpretations. Nevertheless, manual modeling approaches [8] merely rely on artificial experience for feature extraction and selection, which lead to a certain degree of subjectivity and bias when confronted with the intricacies and variabilities present in SAR imagery, thereby limiting their practical applicability and performance capabilities. Additionally, it is challenging to guarantee the effectiveness of recognition results. In recent years, deep learning

methods [9], especially Convolutional Neural Networks (CNN), have been extensively used in computer vision [10,11] and demonstrating remarkable achievements. Meanwhile, based on deep learning, the image processing method has also been successfully extended to the field of remote sensing images [12,13], presenting a new direction and breakthrough for SAR target recognition [10,14,15].

At present, CNN has become one of the most effective network architecture for image recognition tasks. As the earliest CNN network, LeNet-5, proposed by LeCun et al. [16] in 1998 for handwritten digit recognition, was regarded as the first CNN structure. Over time, researchers have continuously refined and optimized the classic CNN architecture and its features, leading to the design of more complex and high-performing CNNs, such as Alexnet [17], GoogLeNet [18], VGGNet [19], Resnet [20], etc. Despite the outstanding performance achieved by classic CNN structures, the neural network has a low level of transparency and is also known as the black boxes [13] due to the lack of a clear visual explanation for the representation of internal features and parameter organization. These limitations significantly constrain people's ability to understand and interpret the internal workings of neural networks, consequently restricting their potential applications in specialized fields, such as medicine, finance, transportation, military, and other domains [21,22]. There are currently two primary research directions for interpretability, which are Intrinsic Explanation and Post-hoc Explanation [23]. Intrinsic Explanation aims to enhance the interpretability of the model itself, enabling users to understand the calculating process and rationale without requiring additional information or algorithms. In contrast, Post-hoc Explanation mainly focuses on explaining the behavior and decision-making process of black-box models [24]. Retraining the model can be too costly in terms of time and resources since the model has already been trained and deployed. As such, the Post-hoc Explanation approach is often more appropriate in such cases. Representation visualization, as an intuitive method in post-hoc interpretation, mainly involves combining the input, middle layer parameters, and output information of the pre-trained model to achieve an interpretation of the decision results. Gradient-based methods, Perturbation, and Class Activation Map (CAM) are three widely adopted methods for achieving representation visualization [23,25].

The gradient-based method [25–31] backpropagates the gradients of a specific class into the input image to highlight image regions that contribute positively or negatively to the result. The methods are fast computation and high resolution of the generated images but usually suffer from excessive noise. CAM is one class of the most important methods specifically designed for CNNs [32–38]. The method utilizes the form of a heatmap to visually highlight the regions most relevant to the particular category. The CAM-based method was first proposed by Zhou et al. [32] in 2016. They believed that with the deepening of CNN layers, the feature map of the intermediate layer contains less and less irrelevant information, and the last convolutional layer of the CNN achieves the highest-level semantic information. After that, numerous CAM methods have been proposed, including Grad-CAM [33], Grad-CAM++ [34], Grad-CAM [35], Group-CAM [36], Score-CAM [37], Ablation-CAM [38], etc. Although these methods have demonstrated good performance in image interpretation, they may suffer from low resolution and spatial precision in some cases. Interpretability methods based on perturbation [39–42] typically utilize the element-wise product of generated masks and the original image to obtain the perturbed input images, which are then fed into the model to observe the changes in the prediction result. The information generated is used to optimize the weighted mask to obtain the final interpretation result image. Among them, RISE [42] randomly generates a large number of masks through Monte Carlo sampling method to occlude different parts of the input image. And the final saliency map is generated by the weighted sum of the masks and the scores predicted by the base model on the masked images.

In this paper, we propose a post-hoc interpretation method of black-box models for SAR ATR called Randomized Input Sampling for explanation based on Clustering (C-RISE). We demonstrate the effectiveness of C-RISE through extensive experimental validation

and comparative analysis. Specifically, our method exhibits superior performance when dealing with SAR images that suffer from severe noise interference, as well as cases where adjacent pixels exhibit mutual influence and dependence. C-RISE offers several advantages over other neural network interpretable algorithms, including white-box methods:

1. The method is a black-box interpretation method, and the calculation process does not need to use the weight, gradient, feature map and other information of the model so that it has better robustness and transferability. Furthermore, the approach avoids errors caused by unreasonable weight selection and information loss during feature map upsampling in Class Activation Mapping (CAM) methods;
2. Compared with RISE, our algorithm can group mask images that capture similar fusion features into different groups by clustering strategy. This allows for the concentration of more energy in the heatmap on the target area, thereby increasing the interpretability of the model.
3. C-RISE employs Gaussian blur to process masked regions, as opposed to simply setting occluded pixels to 0. This technique ensures the consistency and integrity of the original image structure while covering certain areas. As a result, it reduces the deviation of network confidence caused by the destruction of spatial structure, leading to more credible results when compared to other perturbation-based interpretation methods.

The contents of this article are organized as follows: In Section 2, we introduce the principle of the RISE algorithm and CAM methods. Section 3 elaborates on the details of the C-RISE algorithm. Section 4, we verify the effectiveness and robustness of the proposed method through both qualitative judgment and quantitative description. Finally, in Section 5, we discuss the experimental results, clarify any confusion, and explore potential future work.

2. Related Work

In this section, we first review the existing classical methods of CAM [32–38] and the RISE [42] algorithm. Since both CAM methods and RISE interpretation methods display in the form of heatmaps, we focus our subsequent experiments [42] on comparing the effects of different CAM methods, RISE, and C-RISE. This chapter provides theoretical support for the design and experimentation of C-RISE.

2.1. CAM Methods

Zhou et al. [32] proposed the Class Activation Map (CAM) method which utilizes the final convolutional layer of CNN to extract the most abstract target-level semantic information. Its corresponding feature map contained the most abstract target-level semantic information and each channel detected different activated parts of the target. Thus, the class activation map relevant to the recognition result of class c can be generated by the channel-wise weighted summation of the final feature maps. The formal representation of this process can be expressed as follows:

$$L_{CAM}^{c} = \mathrm{ReLU}\left(\sum_{k=1}^{n} w_k^c \boldsymbol{A}_k^L\right) \tag{1}$$

where w_k^c represents the connection weight of the kth neuron pair classified as class c in the Softmax layer, and $\boldsymbol{A}_k^L$ represents the feature map of the kth channel in the lth convolutional layer. The disadvantage of this method is that it can only be applied to the last layer feature map and the full connection is GAP operation. Otherwise, it requires the user to modify the network and retrain, and such costs are sometimes substantial. To overcome the disadvantages, Selvaraju et al. [33] proposed a method named Grad-CAM and updated the weight generation method in Equation (1) as follows:

$$w_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c(\boldsymbol{x})}{\partial \boldsymbol{A}_{k,i,j}^L} \tag{2}$$

where the sum element is the gradient of the calculated class score ($y^c(x)$) with respect to the pixel values at each position of A_k^L, and Z represents the normalization factor. Compared to the CAM method, Grad-CAM is more generalized and can be used for different model structures. Both Grad-CAM++ [34] and XGrad-CAM [35] are improved algorithms based on Grad-CAM method. The basic form of Grad-CAM++ is the same as Grad-CAM, but the difference is that the combination of higher-order gradients is used as the channel weight in Grad-CAM, which improves the visualization effect of multi-object images and the positioning is more accurate. XGrad-CAM achieves better visualization of CNN decisions through a clear mathematical interpretation.

Different from the improvement idea based on gradient, Score-CAM [37] is a gradient-free algorithm for visualizing CNN decisions. It defines the concept of Increase of Confidence (CIC), which measures the increment of confidence relative to a baseline image. The CIC score for a particular feature map A_k^L is computed as:

$$C\left(A_k^L\right) = f\left(\boldsymbol{X} \circ A_k^L\right) - f(\boldsymbol{X}_b) \tag{3}$$

where $\boldsymbol{X}$ is the input image, $\circ$ represents the Hadamard product, and $\boldsymbol{X}_b$ is the baseline image, which can be set to an all-0 matrix with the same size as the original image. $f(\cdot)$ denotes the neural network's output score for the target class. The algorithm then computes CIC scores for all feature maps in a particular layer and updates the scores using the Softmax operation. These updated scores are used as the weights for the corresponding feature maps. Finally, the different feature maps are weighted and summed to generate a visual image.

The CAM approach has been demonstrated to be effective in visualizing the important regions of objects in various optical image datasets. However, when applied to Synthetic Aperture Radar (SAR) images, several challenges arise such as gradient dispersion, energy unconcentration, and inaccurate positioning. These challenges are primarily due to the unique characteristics of SAR images which include:

1. SAR images are often characterized by low resolution and low Signal-to-Noise Ratio (SNR), which makes it challenging to visualize important features and information accurately. Additionally, the imaging principle of SAR images is based on active imaging, which introduces a significant amount of interference spots in the image, thereby making SAR images significantly different from optical images. These interference spots can significantly impact the visualization process, leading to inaccurate feature localization and reduced effectiveness of CAM-based visualization methods;
2. The relatively small difference between different categories in SAR image datasets poses a challenge to visualization techniques such as CAM, which heavily rely on distinguishing features between different categories. Furthermore, the target area of SAR images is often highly localized, which makes accurate positioning critical for the interpretation of visualizations. However, different CAM methods typically use feature maps to upsample to the size of the original image, which can introduce positioning deviations. Despite ongoing efforts to generate high-resolution feature maps, the visualization effect of SAR images using CAM methods remains suboptimal.

2.2. RISE

Randomized Input Sampling for Explanation (RISE) [42] is a perturbation-based method that aims to generate a heatmap highlighting important regions in an input image concerning the prediction of a black-box model. The architectural details of RISE are illustrated in Figure 1. Initially, the Monte Carlo sampling method is employed to generate a considerable number of masks, which have the same size as the original image. Subsequently, the element-wise product of these masks and the original image is computed to derive the corresponding perturbed images. These masked images are then inputted into the black-box model to acquire prediction probabilities for the inferred category. Finally, the prediction probabilities serve as weights for aggregating the masks, facilitating the

superimposition of regions within the original image that significantly contribute to the specified category. RISE has demonstrated effectiveness in providing local interpretability for various image classification models. Furthermore, Score-CAM [37] is a gradient-free method that draws inspiration from RISE.

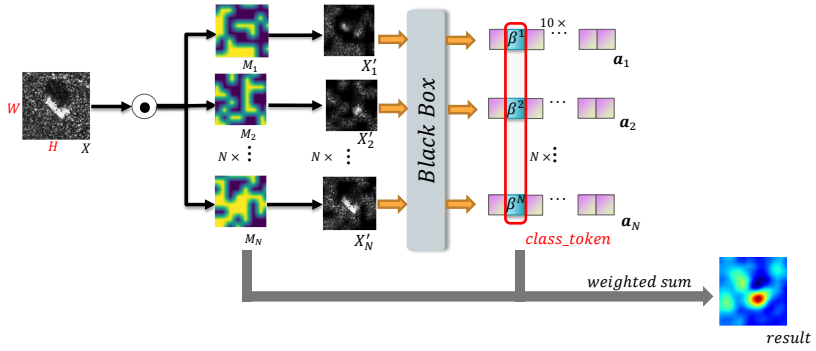

Figure 1. The flowchart of RISE method.

RISE is a black-box interpretation technique that circumvents the need for utilizing weight, gradient, and feature map information during the calculation process. Utilizing the Monte Carlo sampling method, which is a stochastic approximation inference technique, RISE achieves approximate calculations of complex integrals or expected values through sampling. Let x denote a specific input condition, and z represent a random variable. By employing random sampling from the probability distribution $p(z \mid x)$, a collection of independent and identically distributed samples $\{z_1, z_2, ..., z_N\}$ can be generated, with N representing the total number of random samples.The expression for the expected value of the function $f(z)$ under the complex probability distribution $p(z \mid x)$ is presented in Equation (4).

$$E_{z|x}[f(z)] = \int p(z \mid x) f(z) dz \cong \frac{1}{N} \sum_{i=1}^{N} f(z_i) \tag{4}$$

where $z_i (i = 1, 2, \ldots, N)$ represents the i-th sample obtained after random sampling.

In the RISE algorithm, the predicted probability of the black-box model for the category to which the perturbed image belongs can be viewed as the importance of the region retained by the mask. Then the importance of the prominent region of the final generated image can be viewed as the expectation obtained from all masks, as shown in Equation (5).

$$S_{I,f}(\lambda) = E_M[f(I \circ M) \mid M(\lambda) = 1] \tag{5}$$

where λ represents the pixel coordinate in the mask M. The expression $M(\lambda) = 1$ indicates that the pixel at coordinate λ in the mask has a value of 1, implying it is one of the important or retained regions in the mask. And the notation $E_M[\cdot]$ denotes the expectation operator with respect to the random variable M. $S_{I,f}(\lambda)$ in Equation (5) represents the expected score obtained by averaging the predictions of the model $f(\cdot)$ over all masks where the pixel at coordinate λ is retained.

Then, the expression can be expanded according to the definition of expectation as follows:

$$\begin{aligned} S_{I,f}(\lambda) &= \sum_m f(I \circ m) P[M = m \mid M(\lambda) = 1] \\ &= \sum_m f(I \circ m) \frac{P[M = m, M(\lambda) = 1]}{P[M(\lambda) = 1]} \\ &= \frac{1}{P[M(\lambda) = 1]} \sum_m f(I \circ m) P[M = m, M(\lambda) = 1] \end{aligned} \tag{6}$$

where

$$P[M = m, M(\lambda) = 1] = \begin{cases} 0, & \text{if } m(\lambda) = 0 \\ P[M = m], & \text{if } m(\lambda) = 1 \end{cases} = m(\lambda)P[M = m] \tag{7}$$

In Equations (6) and (7), m represents an individual binary mask, determining the retained and non-retained pixels in the mask, and capturing different configurations of important regions in the image.

By substituting Equations (6) and (7), we can get:

$$S_{I,f}(\lambda) = \frac{1}{P[M(\lambda) = 1]} \sum_m f(I \circ m) \cdot m(\lambda) \cdot P[M = m] \tag{8}$$

Since the mask m follows a 0-1 distribution, we can further simplify the equation, we can obtain Equation (9):

$$P[M(\lambda) = 1] = E[M(\lambda)] \tag{9}$$

$$\therefore S_{I,f}(\lambda) = \frac{1}{E[M(\lambda)]} \sum_m f(I \circ m) \cdot m(\lambda) \cdot P[M = m] \tag{10}$$

It is noted that the heatmap can be obtained by summing the masks obtained from random sampling with weighting based on the predicted probabilities of the perturbed images. When masks are sampled using uniform sampling, $P[M = m]$ can be expressed as:

$$P[M = m] = \frac{1}{N} \tag{11}$$

where N represents the total number of masks. When employing the Monte Carlo sampling method to obtain a set of masks denoted as $\{M_i\}, i = 1, 2, \ldots, N$, the Equation (10) can be refined as follows:

$$S_{I,f}(\lambda) \approx \frac{1}{E[M] \cdot N} \sum_{i=1}^{N} f(I \circ M_i) \cdot M_i(\lambda) \tag{12}$$

The equation suggests that the importance score $S_{I,f}(\lambda)$ is obtained by summing the weighted predicted probabilities over all sampled masks. The weight for each term is given by $M_i(\lambda)$, which indicates the importance of the pixel at coordinate λ in each mask. The sum is then normalized by the product of the expected value $E[M]$ and the total number of masks N.

Furthermore, taking into account that pixel-wise masks can result in significant variations in the model's prediction and the exponential computational cost associated with sampling pixel-level masks, a strategy is employed during mask generation. Initially, smaller masks are created and subsequently upsampled to match the image size, ensuring a smoother transition. This approach aims to balance the importance of capturing fine-grained details while managing computational complexity.

3. Our Method

As a post-hoc interpretation algorithm based on perturbation, RISE algorithm has a more intuitive and understandable presentation than the visual interpretation method based on back propagation. At the same time, RISE also overcomes the limitations of general CAM methods by avoiding the generation of unreasonable weights and the problem of small feature maps during the up-sampling process. However, the effectiveness of RISE and other optical image-based interpretive methods in SAR ATR scenarios is limited. This is because the active imaging mechanism of SAR images results in multiplicative noise, which causes problems such as noise, energy dispersion, and inaccurate positioning when applying optical image-based interpretive methods to SAR image recognition [3,8]. To address this issue, we propose an algorithm based on RISE, called Randomized Input Sampling for Explanation based on Clustering (C-RISE), which is a post-hoc interpretation

method for black-box models in SAR ATR. Our algorithm considers the structural consistency and integrity of SAR images and highlights the regions that contribute to category discrimination in SAR images. Figure 2 illustrates the workflow of our proposed approach.

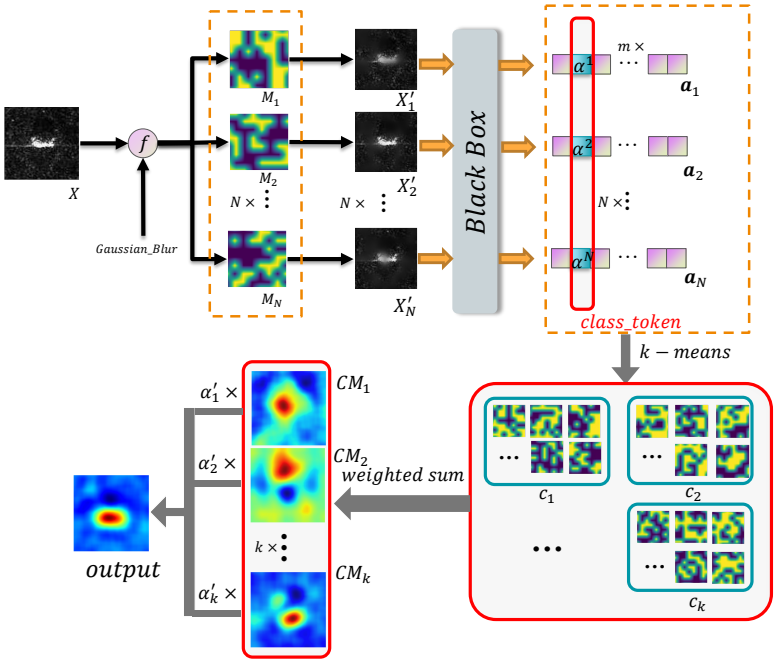

Figure 2. The flowchart of C-RISE.

3.1. Mask Generation

As shown in Section 2.2, pixel-level occlusion may have a huge impact on the model, and the computational complexity of sampling is high. Therefore, in order to ensure the smoothness and the consistency of the target space structure when generating masks, small masks are generated first and then upsampled back to the image size. The upsampling process inherently fills in the gaps and creates smoother transitions between mask elements. The upsampling process distributes the mask values more smoothly across the image, reducing the visibility of sharp edges and promoting a gradual transition between occluded and non-occluded regions. The basic process is shown in Figure 3. Formally, the process of generating masks can be described as follows:

1. N binary masks $\{grid_1, grid_2, \ldots, grid_N\}$ are randomly generated based on Monte Carlo sampling, where $grid_i \in \mathbb{R}^{s \times s}, i = 1, 2, \ldots, N$. s is smaller than image size H and W. In $grid_i$, each element independently to 1 with probability p and to 0 with the remaining probability;
2. Upsample $grid_i$ to $grid'_i \in \mathbb{R}^{(s+1)H \times (s+1)W}$;
3. A rectangular area was randomly selected from $grid'_i$ as M_i, where $M_i \in \mathbb{R}^{H \times W}$, $i = 1, 2, \ldots, N$.

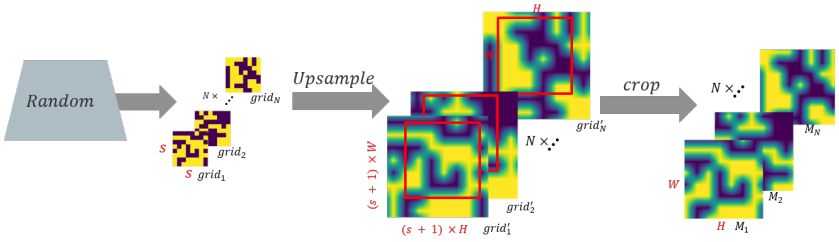

Figure 3. The flowchart of generating masks.

The parameter p governs the density of generated masks and plays a crucial role in balancing the preservation and obscuration of information within the saliency map. Lower values of p entail a higher degree of masking, potentially accentuating more localized and specific salient regions while sacrificing some contextual information.

From Figure 4, the impact of different values of s on the quality of the generated mask can be observed. It provides visual evidence of how varying s affects the characteristics of the binary mask *grid* and the resulting mask M. A larger s value means that each mask covers a smaller portion of the image, leading to fragmented and disjointed masked regions. Therefore, selecting an appropriate value for s requires careful consideration of the trade-off between granularity and the preservation of spatial structure. Empirically, we generally choose $10 \leq \lfloor \min(H, \mathrm{W})/\mathrm{s} \rfloor \leq 15$ to ensure that the resulting saliency map provides meaningful and coherent interpretations while minimizing potential errors introduced by excessive fragmentation.

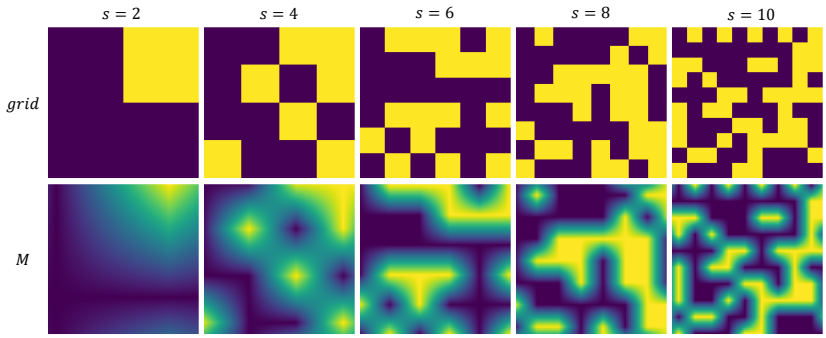

Figure 4. Influence of different s on mask generation.

After obtaining N masks, we introduce Gaussian blur to the occluded part of the original image, which is in order to make the image after the mask processing can retain the maximum consistency of the original image, and smoothly occlusion of the region. Gaussian blur is an image blurring filter that computes the transformation of each pixel in an image with a normal distribution. The normal distribution equation in 2-dimensional space can be written as:

$$G(X) = \frac{1}{2\pi\sigma^2} e^{-\left(u^2+v^2\right)/\left(2\sigma^2\right)} \tag{13}$$

where (u, v) denotes the pixel position and σ means the standard deviation of the normal distribution. It is worth noting that in 2-dimensional space, the contours of the surface generated by Equation (13) are normally distributed concentric circles from the center. The value of each pixel is a weighted average of the neighboring pixel values. The value of the original pixel has the largest Gaussian distribution value, so it has the largest weight, and the neighboring pixels get smaller and smaller as they get farther from the original pixel. The Gaussian blur preserves the edge effect more than other equalization blur filters, which is equivalent to a low-pass filter.

Based on Gaussian blur, We can use Equation (14) to obtain the image after mask processing:

$$X_i' = X \circ M_i + G(X) \circ (\mathbf{1}^{H \times \boldsymbol{W}} - M_i), \quad i = 1, 2, .., N \tag{14}$$

where $X \in \mathbb{R}^{H \times W}$ denotes the original image, $\mathbf{1}^{H \times \boldsymbol{W}} \in \mathbb{R}^{H \times W}$ means an all-1 matrix and its shape is $H \times W$.

SAR images often contain important structural details and information that need to be preserved during the mask generation process. The application of Gaussian blur to the occluded part of the image ensures a smooth transition between the occluded and non-occluded regions. This smooth occlusion helps to maintain the coherence and consistency of the image by preventing abrupt changes or sharp boundaries between the occluded

and non-occluded areas. It creates a visually pleasing interpretation by ensuring that the occluded regions blend smoothly with the rest of the image.

3.2. Clustering

The masked image $\{X'_1, X'_2, \ldots, X'_N\}$ are input to the black-box model $f(\cdot)$ to obtain the output vector $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$. Moreover, we use $\boldsymbol{a}_i \in \mathbb{R}^{1\times m}, i = 1, 2, \ldots, N$ as the feature vectors to cluster M_i by *k-means*. m is the number of categories. The process is shown in Equations (15)–(17).

$$\boldsymbol{a}_i = f\left(X'_i\right), \quad i = 1, 2, \ldots, \mathrm{N} \tag{15}$$

$$(\boldsymbol{c}_1; \boldsymbol{c}_2; \ldots; \boldsymbol{c}_k) = k - means([(\mathrm{M}_1, \boldsymbol{a}_1), (\mathrm{M}_2, \boldsymbol{a}_2), \ldots, (\mathrm{M}_N, \boldsymbol{a}_N)]) \tag{16}$$

$$\boldsymbol{c}_i = \left\{M_j^i\right\}, \quad i = 1, 2, .., k; j = 1, 2, ..., N_i \tag{17}$$

where c_i denotes the ith cluster, M_j^i denotes the jth mask in ith cluster, k and N_i represent the number of clusters and the number of elements in the ith cluster, respectively.

If the original image is identified as class l after the black-box model, we can obtain the contribution of the jth mask in the ith cluster to the model:

$$\alpha_j^i = \boldsymbol{a}_j^i[l], \quad i = 1, 2, .., k; j = 1, 2, ..., N_i; l \leq m \tag{18}$$

where l represents the prediction category of the model for the input image and $\boldsymbol{a}_j^i$ denotes the feature vector obtained by inputting the image masked by M_j^i into the black box model and $a_i^j[l]$ represents the value of the l-th dimension of the feature vector, i.e., ,the confidence that the model identifies the masked image as class l. After that, we use α_j^i to estimate the weight of a specific mask and calculate the weighted sum in each cluster CM_i as follows:

$$CM_i = \sum_{j=1}^{N_i} \alpha_j^i \cdot M_j^i, \quad i = 1, 2, .., k \tag{19}$$

After that, we calculated the CIC value of CM_i through Equation (3) and used it as the classificatory information that CM_i was concerned about. Finally, the final result H^{C-RISE} is generated by weighted summation of the feature maps of different clusters. The process is formulated as Equations (20) and (21). The pseudo-code is presented in Algorithm 1.

$$\alpha'_i = [f(X \circ CM_i) - f(X_b)]_l, \quad i = 1, 2, ..., k \tag{20}$$

$$H^{C-RISE} = \sum_{i=1}^{k} \alpha'_i \cdot CM_i \tag{21}$$

The clustering strategy employed in C-RISE plays a crucial role in concentrating more energy in the heatmap on the target area. Through the clustering process, masks that exhibit similar patterns or characteristics are grouped together in the same cluster. The clustering helps to identify a subset of masks that collectively represent important features associated with the target category. By focusing on these specific masks, C-RISE effectively filters out irrelevant or less informative regions, allowing it to concentrate more energy on the target area. In the original RISE algorithm, weights are calculated individually for each mask, leading to a dispersed distribution of energy in the saliency map. In contrast, the clustering strategy in C-RISE introduces a grouping scoring strategy. Once the masks are clustered, C-RISE assigns weights to the individual masks based on their contribution to the model's output. Masks that are more influential in determining the target category receive higher weights. By aggregating the masks based on their weights and cluster

assignments, C-RISE produces a heatmap that highlights the regions of the image that are most relevant to the target category. This concentration of energy in the heatmap on the target area is achieved by prioritizing and emphasizing the masks within the clusters that capture essential discriminative features.

Algorithm 1: C-RISE.

Input: SAR image X, black-box model $f(\cdot)$,randomly mask $grid_i$
Output: H^{C-RISE}
# masked image and feature vector generation;
for $i = 1 : N$ **do**
  # mask generation;
  $M_i \leftarrow crop(Upsampling(grid_i))$;
  # $G(\cdot)$ means Gaussian blur;
  $X_i' \leftarrow X \circ M_i + G(X) \circ (\mathbf{1}^{H \times W} - M_i)$;
  $\boldsymbol{a}_i \leftarrow f(X_i')$;
end
# clustering;
for $i = 1 : N$ **do**
  $(\boldsymbol{c}_1; \boldsymbol{c}_2; \ldots; \boldsymbol{c}_k) = k - means([(\mathrm{M}_1, \boldsymbol{a}_1), (\mathrm{M}_2, \boldsymbol{a}_2), \ldots, (\mathrm{M}_N, \boldsymbol{a}_N)])$;
end
# calculate the subheatmap and CIC score in each group;
for $i = 1 : k$ **do**
  $CM_i = \sum_{j=1}^{N_i} \alpha_j^i \cdot M_j^i$;
  $\alpha_i' = C(CM_i) = [f(X \circ CM_i) - f(X_b)]_l$;
end
generate final heatmap;
$H^{C-RISE} = \sum_{i=1}^{k} \alpha_i' \cdot CM_i$;

In conclusion, the C-RISE algorithm offers the following advantages:

1. Perturbation-based Interpretation: The algorithm utilizes the concept of perturbation, where multiple randomly sampled masks are combined to generate a saliency map highlighting important regions. This approach is considered a black-box interpretation method, as it does not rely on model-specific information such as weights, gradients, or feature maps. By avoiding the need for feature maps, it also circumvents errors stemming from unsuitable upsampling and weight selection methods employed by CAM series methods;
2. Smoother Mask Generation: The upsampling process distributes the mask values more smoothly across the image, reducing the visibility of sharp edges and promoting a gradual transition between occluded and non-occluded regions.This approach balances global and local considerations and ensures smooth and coherent mask generation. Additionally, the random generation of each mask ensures representativeness, diversity, and prevents bias resulting from a specific sampling mode;
3. Feature-based Clustering: The algorithm leverages the confidence vectors obtained under each mask as feature vectors. These vectors effectively evaluate the importance of unmasked areas for a particular image, the feature vectors enable clustering of masks based on their fusion features. This clustering approach serves two purposes: first, masks with similar fusion features are grouped, reducing redundancy in calculations; second, the original RISE algorithm's weight calculation for each mask is enhanced through a grouping scoring strategy. This improvement leads to a more concentrated saliency map generation.

4. Experiment

4.1. Experimental Settings

This study employs SAR images of ten vehicle target types under standard operating conditions (SOC) from the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [43] as the experimental data. The dataset comprises 5172 SAR images with dimensions of $1 \times 100 \times 100$, with 2536 images used for training and 2636 for testing. The ten target categories include 2*S*1, *BRDM*2, *BTR*60, *D*7, *SN*_132, *SN*_9563, *SN*_*C*71, *T*62, *ZIL*131, and *ZSU*_23_4. Figure 5 displays ten representative SAR images for each category.

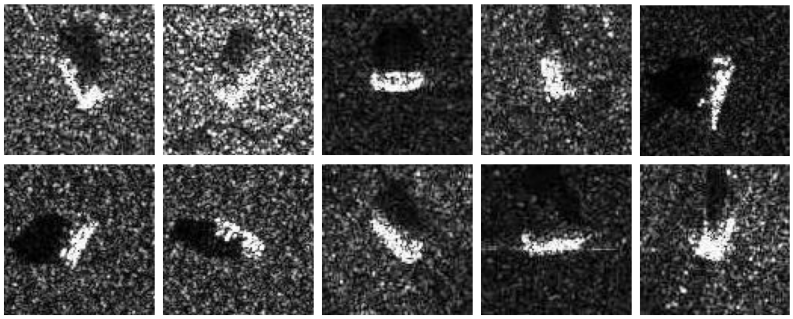

Figure 5. 10 typical SAR images for each category in MSTAR. The first row depicting random images from 2*S*1, *BRDM*2, *BTR*60, *D*7, and *SN*_132, and the second row showing randomly selected images from *SN*_9563, *SN*_*C*71, *T*62, *ZIL*131 and *ZSU*_23_4.

During the experiment, the Alexnet model [9] was utilized as a classifier, and its structure is depicted in Figure 6. It is worth mentioning that, as the C-RISE algorithm is primarily tailored for black-box models, alternative efficient models may be employed in place of Alexnet. After conducting multiple iterations of training, the neural network achieved a recognition rate of 97.6%, which indicates the effectiveness of using various methods to generate saliency maps. However, since this paper primarily focuses on interpreting and analyzing the network structure using different visualization methods, the training techniques and processes are not extensively discussed. During the implementation of the C-RISE algorithm, several parameters were set, including $k = 4$, $N = 2000$, $s = 8$, $p = 0.5$. $p = 0.5$ ensures an equitable probability for each pixel to be masked or preserved. The selection embodies a balanced masking strategy that seeks to strike a reasonable equilibrium between information preservation and obscuration within the resulting saliency map. Empirically, we generally choose $10 \leq \lfloor \min(H, \mathrm{W})/\mathrm{s} \rfloor \leq 15$. The impact of different values for N and k on the experimental results will be thoroughly examined and discussed in Section 4.5.

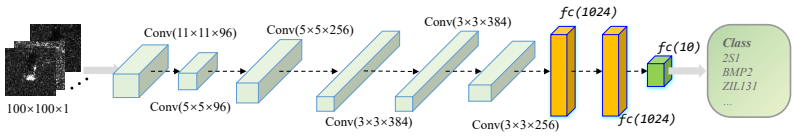

Figure 6. The structure of Alexnet.

4.2. Class Discriminative Visualization

Since the class activation map generated by CAM methods and the saliency map generated by C-RISE algorithm are presented in the form of heatmap, we focus on comparing the experimental effects of different CAM methods, RISE algorithm and C-RISE algorithm in the following experimental part, referring to the comparison method in [42]. In this section, we randomly selected ten graphs that were correctly classified in different networks from the testset, and used Grad-CAM [33], Grad-CAM++ [34], XGrad-CAM [35], Score-CAM [37], RISE [42] and C-RISE to visually analyze the model recognition process, and the comparison is shown in Figure 7.

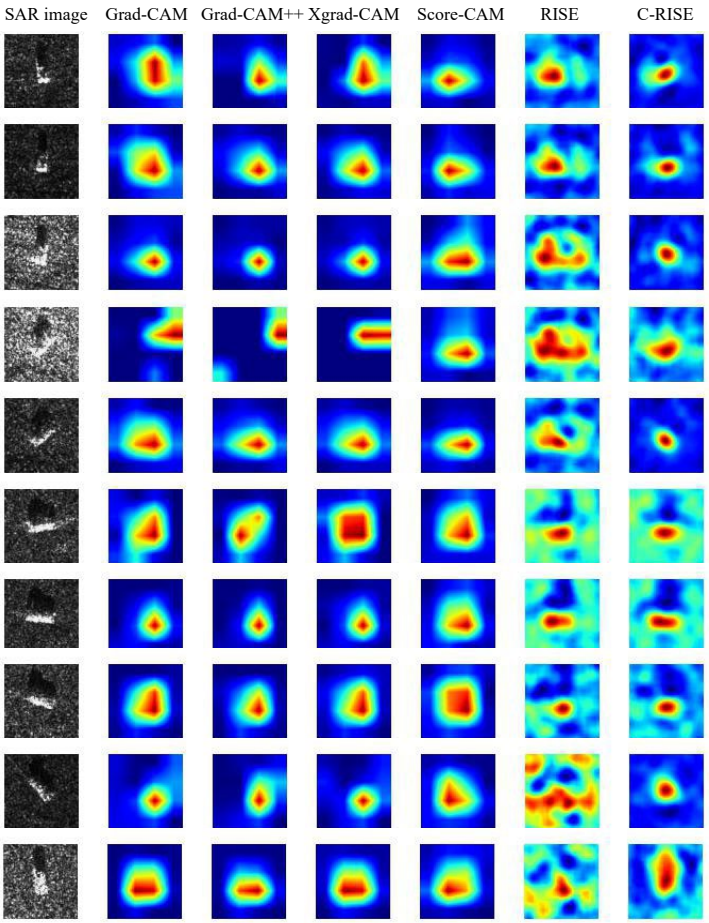

Figure 7. Comparison of Grad-CAM, Grad-CAM++, XGrad-CAM, Score-CAM, RISE, C-RISE. The first column is the SAR images of ten classes. The rest of columns are corresponding heatmaps generated by each method respectively.

We can verify the fairness and localization ability of the C-RISE algorithm from a qualitative and quantitative perspective. The employed interpretation methods produce saliency heatmaps, which serve the purpose of highlighting the crucial regions or features within an image. These heatmaps undergo visual comparison to evaluate the efficacy of the interpretations. Qualitative analysis encompasses the crucial task of evaluating the consistency between interpretation results and human perception. This evaluation process serves to determine the extent to which the methods align with human intuition and offer meaningful interpretations. It can be intuitively seen from Figure 7 that CAM methods, RISE and C-RISE are all heatmap-based network interpretation methods that visually analyze the importance of different image regions. Heatmaps represent an image with varying color intensities, where the intensity of each pixel corresponds to its level of importance. Compared to gradient-based CAM algorithms like Grad-CAM and Grad-CAM++, the heatmaps generated by Score-CAM, RISE and C-RISE exhibit higher concentration of relative energy within the target area of the original image.

CAM methods tends to produce divergent activation areas that are roughly located around the target and its surroundings. However, it often suffers from positioning deviations and lacks granularity. These errors can be attributed to the improper weight selection method as well as the positioning deviation introduced when upsamping the feature map to match the original image size.In the original RISE algorithm, weights are calculated individually for each mask, leading to a dispersed distribution of energy in the saliency map.In terms of visual effect, the C-RISE algorithm outperforms the RISE algorithm and the CAM method in localizing the heatmap. It exhibits more concentrated energy, better category discrimination, and produces more stable saliency map results. These advantages

contribute to the superior performance of C-RISE in generating high-quality heatmaps for interpreting the target regions of interest.

Subsequently, we perform a quantitative analysis of the heatmap by examining its energy characteristics. Our focus lies in evaluating the amount of energy present within the bounding box of the target object in the saliency map. Therefore, we adopted a similar measure to [37], the specific process is shown in Figure 8. The evaluation process involved the following steps:

Firstly, we manually annotated the bounding boxes of the objects in all the images within the test set. This step involved marking the regions that corresponded to the target objects of interest. After obtaining the bounding box annotations, we binarized the images based on a specific rule. The inner region of the bounding box was set to 1, representing the target area, while the outer region was set to 0, representing the background. We then computed the sum of the energy values within the target bounding box in these heatmaps.This step quantified the concentration of energy within the specified region of interest. In parallel, we also calculated the sum of the energy values in the entire heatmap, including both the target bounding box and the background regions. *proportion* was calculated by dividing the sum of energy within the bounding box by the sum of energy within the bounding box plus the sum of energy in the background, which served as a quantitative measure to evaluate the localization and recognition capabilities of different interpretation methods. Higher Proportion values indicated that more energy was concentrated within the target area. By employing this quantitative evaluation approach, we were able to compare and analyze the effectiveness of different interpretation methods in terms of their ability to concentrate energy within the target bounding box and provide accurate saliency maps. The mathematical expression of *proportion* is shown in Equation (22).

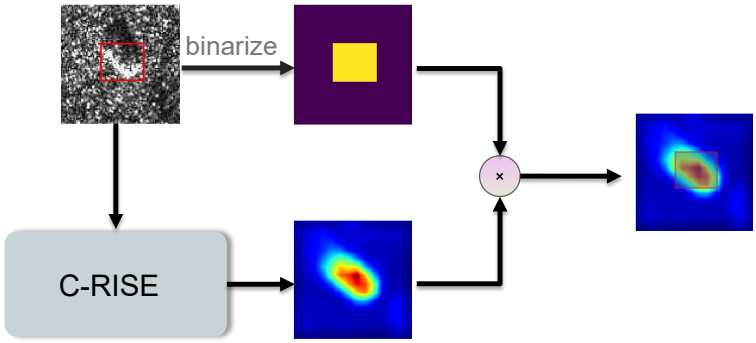

Figure 8. The flowchat of calculating *proportion*.

$$Proportion = \frac{\sum E_{(i,j)\in bbox}}{\sum E_{(i,j)\in bbox} + \sum E_{(i,j)\notin bbox}} \tag{22}$$

where $E_{(i,j)}$ denotes the energy value of the pixel at position (i, j) in the heatmap.

It is worth mentioning that the information contained in each image in the MSTAR dataset is a single target. And in different pictures, the position occupied by the target is usually a large area of the image, which facilitates us to label each subset. Figure 9 shows the binarization results of ten groups of data randomly selected. We calculate *proportion* of images in each category of the testset separately, and the results are shown in Table 1. Based on the experimental results, it has been observed that the C-RISE algorithm demonstrates higher accuracy in weak target supervision and positioning compared to other methods in various types of SAR images. Additionally, the degree of matching between the energy distribution in the heatmap and the actual target is notably higher. These findings provide empirical evidence supporting the effectiveness of the C-RISE algorithm in SAR image interpretation.

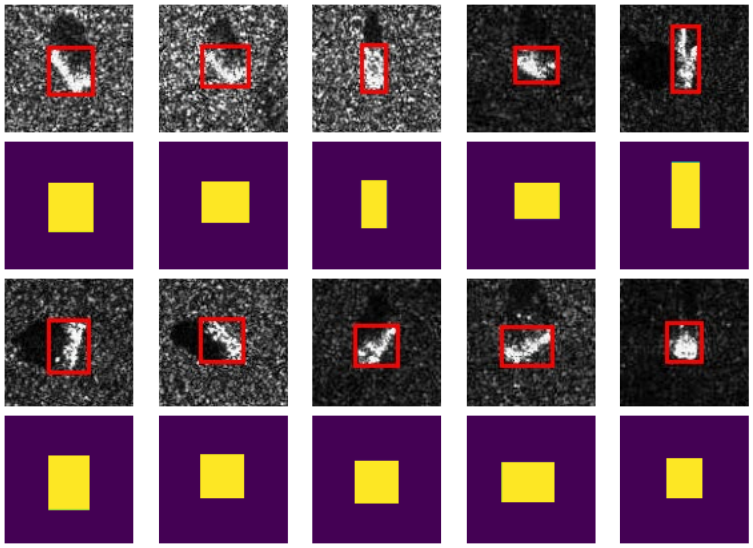

Figure 9. The first and third rows represent randomly selected images with bounding boxes from 10 categories in the test set and the results of binarization of each images are shown as the second and fourth rows.

Table 1. The *proportion* of images in each category. The best records are marked in bold.

	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	RISE	C-RISE
2S1	0.5764	0.4252	0.5785	0.5524	0.3483	**0.5876**
BRDM_2	0.5881	0.5138	0.5970	**0.6230**	0.3621	0.5930
BTR_60	0.4355	0.3744	0.4553	0.3892	0.1024	**0.4731**
D7	0.3782	0.6225	0.3920	0.5425	**0.6406**	0.4394
SN_132	0.3820	**0.5579**	0.4168	0.4915	0.4797	0.4723
SN_9563	**0.4895**	0.4024	0.4851	0.4421	0.2964	0.4817
SN_C71	0.4121	0.2868	0.4409	0.3823	0.0856	**0.4494**
T62	0.4975	0.3894	0.5158	0.4886	0.3374	**0.5233**
ZIL131	0.5420	0.3984	**0.5559**	0.5265	0.4254	0.5498
ZSU_23_4	0.4018	**0.5315**	0.4298	0.4616	0.5209	0.4474
average	0.4758	0.4555	0.4918	0.4976	0.3726	**0.5060**

4.3. Conservation and Occlusion Test

In Section 4.3, we conducted quantitative analyses of different methods' localization capability using occlusion and conservation tests [35,43]. The occlusion test involved selectively discarding specific areas of the input images, while the conservation test aimed to maintain certain regions intact. These experiments evaluated the effectiveness of energy-concentrated regions in heatmaps by inputting mask-processed or reverse mask-processed images into the black-box model and observing the resulting score changes. The resulting maps were then binarized at various thresholds to obtain masks or reverse masks. These tests provided insights in identifying and highlighting relevant regions in SAR images, providing insights into its effectiveness in understanding the decision-making process of the black box model. The way masks generated is shown as Equations (23) and (24).

$$M_{threshold}(i,j) = \begin{cases} 1, \text{ if } H^{C-RISE}(i,j) \geq threshold \\ 0, \qquad\qquad otherwise \end{cases} \tag{23}$$

$$\bar{M}_{threshold} = \mathbf{1}^{H \times W} - M_{threshold} \tag{24}$$

where $threshold \in [0,1]$, H^{C-RISE} denotes the pixel value of the heatmap from C-RISE. $M_{threshold}$ and $\bar{M}_{threshold}$ mean the masks/reverse masks, respectively.

Based on Equations (23) and (24), we could use the element-wise product to get the processed images $I/\bar{I}$ after masked/reverse masked and the results after masked/reverse masked are shown in Figure 10.

$$I = M_{threshold} \circ X \tag{25}$$

$$\bar{I} = \bar{M}_{threshold} \circ X \tag{26}$$

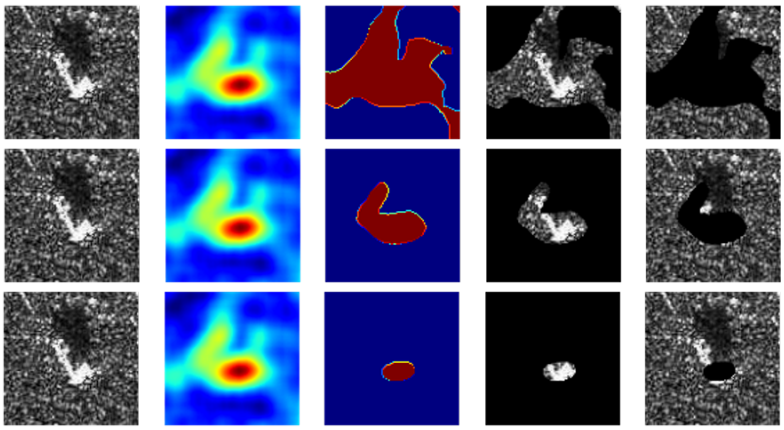

Figure 10. The first column represents a randomly selected image from 2S1, the second column represents H^{C-RISE}, the third column represents $M_{threshold}$, and the fourth and fifth columns represent images after masked/reverse masked, respectively. The *threshold* selected in the three lines were 0.25, 0.50 and 0.75, respectively.

The conservation test examines the ability of methods to identify and preserve the important visual features or regions that contribute to the classification decision made by the black box model. It verifies whether the algorithm highlights the significant regions consistently which the model depends on. And the occlusion test evaluates the robustness of the C-RISE algorithm by assessing its resistance to occlusion or masking of specific regions in the input image. In this test, different parts of the input image are occluded, and the algorithm's response is analyzed to determine if it correctly identifies the occluded regions as significant contributors to the classification decision. The occlusion test helps validate the interpretability of methods by examining its sensitivity to important image regions.

However, directly replacing some pixels with black may produce high-frequency sharp edges [8], and these artificial traces may also lead to changes in the prediction probability, which cannot guarantee the fairness and objectivity of the model recognition process. In order to solve the above problems, we improved the original experiment and proposed two new measures, namely, introducing multiplicative noise and Gaussian blur to the occluded region. The follow two experiments show the effectiveness and rationality of our algorithm.

4.3.1. Based on Multiplicative Noise

In the experiments, we firstly add multiplicative noise to the occluded region and updated Equations (23) and (24) to Equations (27) and (28). The reason for adding multiplicative noise is based on the physical scattering mechanism of SAR coherent imaging. We believe that the intensity of each resolved element of SAR image is modulated by the Radar Cross Section (RCS) [3] of the ground object in the element and a multiplicative noise whose intensity follows the exponential distribution of unit mean (mean = 1). So we can consider the SAR image as the product of the RCS of the ground object in the scene and the noise of the unit mean exponential intensity distribution. Therefore, in the process of signal

processing, we generally consider the noise of SAR image as multiplicative noise [3,9]. Figure 11 shows the above processing of the same image.

$$I = M_{threshold} \circ X + \bar{M}_{threshold} \circ Noise(X) \tag{27}$$

$$\bar{I} = \bar{M}_{threshold} \circ X + M_{threshold} \circ Noise(X) \tag{28}$$

where $Noise(X)$ denotes add high-variance Gaussian multiplicative noise to the input image X.

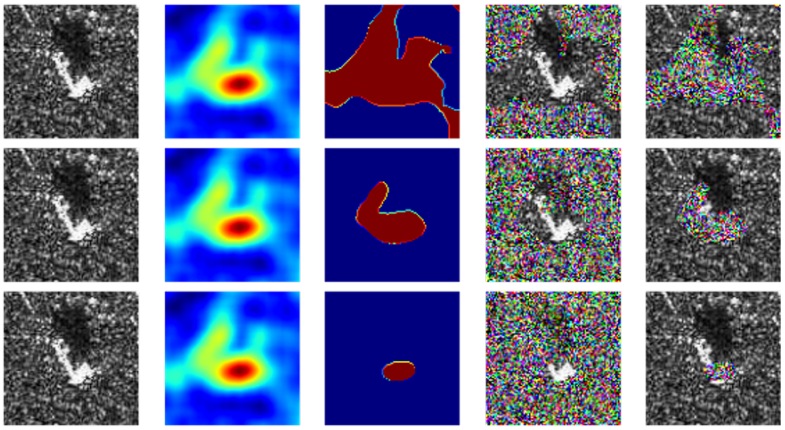

Figure 11. The first column represents a randomly selected image from 2*S*1, the second column represents H^{C-RISE}, the third column represents $M_{threshold}$, and the fourth and fifth columns represent images after masked/reverse masked based on multiplicative noise, respectively. The *threshold* selected in the three lines were 0.25, 0.50 and 0.75, respectively.

Then we define $confidence\_drop(a, b)$ to represent the divergence in the confidence that the processed image b and the original image a are classified into the same category. The mathematical expression of $confidence\_drop(a, b)$ is shown in Equation (29).

$$confidence\_drop(a, b) = \frac{S^c(a) - S^c(b)}{S^c(a)} \tag{29}$$

where $S^c(x)$ is used to represent the score of the input image x being classified as class c. Based on this, we use $confidence\_drop^{con}(X, I)$ and $confidence\_drop^{occ}(X, \bar{I})$ to represent the scores in the conservation and occlusion test, respectively. The process is shown as Equations (30) and (31).

$$confidence\_drop^{con}(X, I) = \frac{S^c(X) - S^c(I)}{S^c(X)} \tag{30}$$

$$confidence\_drop^{occ}(X, \bar{I}) = \frac{S^c(X) - S^c(\bar{I})}{S^c(X)} \tag{31}$$

It is worth noting that the smaller $confidence\_drop^{con}(X, I)$, the greater the difference between the values of $S^c(X)$ and $S^c(I)$, and the generated heatmap can be considered to be located in the salient feature part of the target. Similarly, the larger the $confidence\_drop^{occ}$, the lager the difference between the values of $S^c(X)$ and $S^c(\bar{I})$, and the main features after image processing can be considered to be preserved.

The $confidence\_drop^{con}(X, I)$ and $confidence\_drop^{occ}$ of various methods under different thresholds including Grad-CAM, Grad-CAM++, XGrad-CAM, Score-CAM, RISE and C-RISE, are shown in Tables 2 and 3.

Table 2. $confidence\_drop^{con}(X, I)$ of Different Methods in Conservation and Occlusion Test Based on Multiplicative Noise. The best records are marked in bold.

Threshold	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	RISE	C-RISE
0.25	0.6975	0.6731	0.6949	0.7017	0.7364	**0.6672**
0.50	0.6750	0.7063	0.6760	0.6776	0.8257	**0.6658**
0.75	0.7620	0.7691	0.7644	0.7615	0.7646	**0.6626**

Table 3. $confidence\_drop^{occ}$ of Different Methods in Conservation and Occlusion Test Based on Multiplicative Noise. The best records are marked in bold.

Threshold	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	RISE	C-RISE
0.25	**0.7008**	0.6434	0.6973	0.6427	0.4372	0.4934
0.50	0.3524	0.3287	0.4791	0.4804	0.1867	**0.5361**
0.75	0.1306	0.0475	0.1026	0.1359	0.1537	**0.2637**

4.3.2. Based on Gaussian Blur

From Tables 2 and 3, we can see that compared with other methods, C-RISE achieved relatively optimal performance under different thresholds. Similarly, we can also use high-variance Gaussian blur to process the masked area, and the processed results are shown in Figure 12. Experimental indicators are shown in Tables 4 and 5 respectively. The mathematical expressions are updated from Equations (23) and (24) to Equations (32) and (33).

$$I = M_{threshold} \circ X + \bar{M}_{threshold} \circ G(X) \tag{32}$$

$$\bar{I} = \bar{M}_{threshold} \circ X + M_{threshold} \circ G(X) \tag{33}$$

where $G(X)$ denotes introduce high-variance Gaussian blur to the input image X.

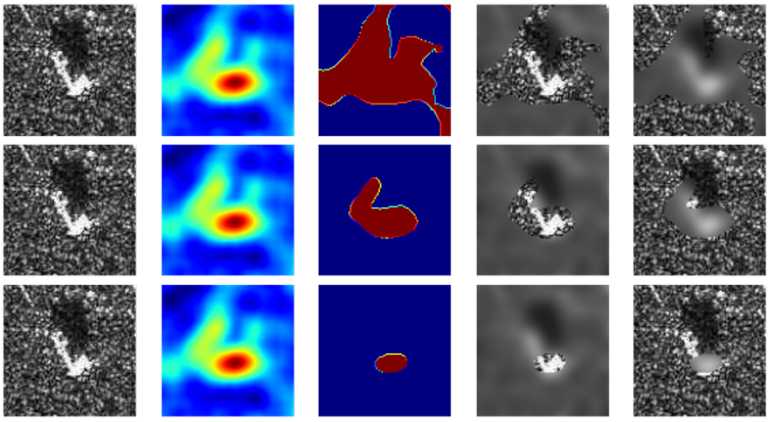

Figure 12. The first column represents a randomly selected image from 2*S*1, the second column represents H^{C-RISE}, the third column represents $M_{threshold}$, and the fourth and fifth columns represent images after masked/reverse masked based on Gaussian blur, respectively. The *threshold* selected in the three lines were 0.25, 0.50 and 0.75, respectively.

Table 4. $confidence\_drop^{con}(X, I)$ of Different Methods in Conservation and Occlusion Test Based on Gaussian blur. The best records are marked in bold.

Threshold	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	RISE	C-RISE
0.25	0.0665	0.1038	0.0768	0.0205	0.0137	**0.0064**
0.50	**0.0285**	0.2391	0.1764	0.0944	0.0924	0.1692
0.75	0.3147	0.3721	0.3249	0.2893	0.2466	**0.1631**

Table 5. $confidence\_drop^{occ}$ of Different Methods in Conservation and Occlusion Test Based on Multiplicative Noise. The best records are marked in bold.

Threshold	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	RISE	C-RISE
0.25	0.2805	0.2250	0.2682	0.3283	0.3898	**0.3985**
0.50	0.1634	0.0968	0.1519	0.2217	0.2513	**0.2870**
0.75	0.0350	0.0119	0.0305	0.0556	0.0906	**0.1663**

4.4. Insertion and Deletion Test

In this experiment, we compared different methods by insertion-deletion test [42]. The experiment is a metric used to evaluate visual interpretation methods and measures the ability of visual interpretation to capture important pixels. During the deletion experiment, the k most important pixels in the heatmap are successively removed, and then we calculate the degree of change in the prediction probability. The insertion curve is the opposite. The curves are shown in Figure 13, with smaller AUC of deletion curves and higher AUC of insertion curves indicative of a better explanation. We randomly select an image from the test set for demonstration and plot its deletion and insertion curves of different algorithms. The results are shown in Figure 14. We calculate AUC of both curves and the $Over\_All$ score [36] ($AUC(insertion) - AUC(deletion)$) of all images from the test set as a quantitative indicator. The average results over 2636 images is reported in Table 6. We found that C-RISE achieves splendid results, indicating that the pixel importance revealed by the visualization method is in high agreement with the model and has great robustness.

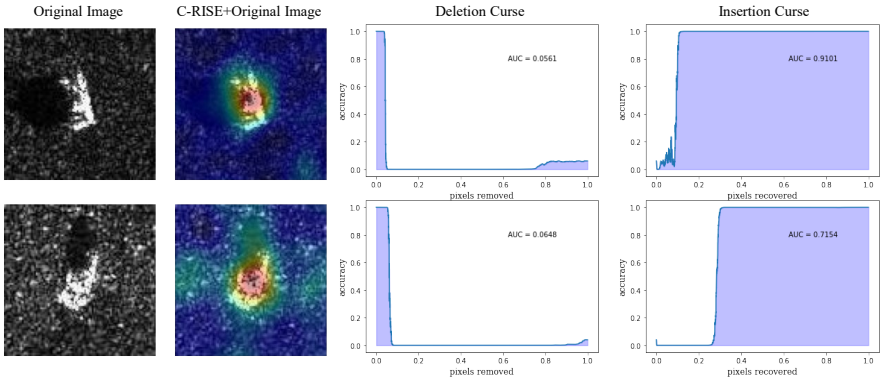

Figure 13. The heatmap generated by C-RISE(second column) for two representative images (first column) with deletion (third column) and insertion (fourth column) curves.

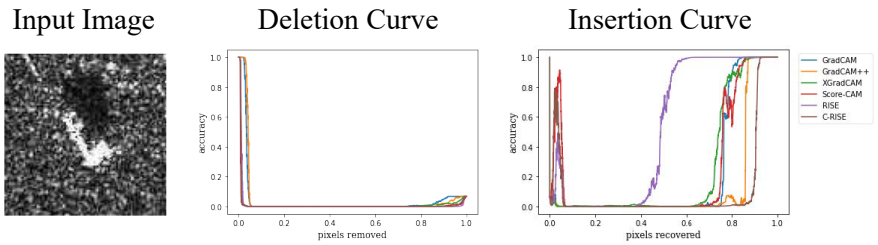

Figure 14. Grad-CAM, Grad-CAM++, XGrad-CAM, Score-CAM, RISE and C-RISE generated saliency maps for a seleted image randomly (firstly column) in terms of deletion (second column) and insertion curves (third column).

Table 6. Comparative evaluation in terms of deletion (lower *AUC* is better) and insertion (higher AUC is better) *AUC*. The *Over_All* score (higher *AUC* is better) shows that C-RISE outperform other related methods significantly. The best records are marked in bold.

AUC	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	RISE	C-RISE
Insertion	0.2768	0.3011	0.4145	0.5512	0.4659	**0.6875**
Deletion	0.1317	0.1676	0.1255	**0.0246**	0.0420	0.1317
Over_All	0.1451	0.1335	0.2890	0.5266	0.4239	**0.5558**

4.5. Ablation Study

A comprehensive ablation study was conducted to assess the impact of the number of clusters k and the number of generated masks N on the performance of the C-RISE algorithm. The study was conducted on a subset of 100 randomly sampled images from the testset. Figure 15 and Table 7 highlight the *AUC* mentioned in Section 4.4 of Insertion, Deletion curve and the Over_All scores obtained for different combinations of k and N. A single image was randomly selected from the dataset, which was correctly recognized by the network and the results are presented in Figure 16, which illustrates the saliency maps obtained using different combinations of N and k. Each saliency map highlights the important regions and features in the image, providing insights into the network's decision-making process.By comparing the saliency maps generated with different N and K values, we can observe the variations in the interpretation results.

When the parameter k is frozen, a higher value of N $(N = 1000, 2000)$ enables the generation of a larger number of masks, resulting in more detailed interpretations. This allows for better coverage of the image space and the potential identification of smaller or more subtle features. Conversely, a lower value of N leads to more generalized interpretations that emphasize larger regions or prominent features. However, generating a larger number of masks ($N = 4000$) may introduce redundancy and impose a heavy computational burden without significantly improving the quality of the interpretations. With a small value of N $(N = 200, 500)$, the diversity and representativeness of the generated masks may be limited and result in inconsistent localization of important areas in the saliency maps. The interpretation results may not fully capture the range of important regions or features in the image. This can lead to incomplete or biased interpretations, potentially missing out on crucial information.The interpretation results may exhibit high variability or instability, making it challenging to identify reliable and consistent regions of interest, which can hinder the interpretability and reliability of the algorithm.

Selecting an appropriate value for k is essential to achieve effective clustering and meaningful interpretation results. A large value of k $(k = 8, 16, 32)$ may cause over-segmentation of the generated masks. This means that the masks within each cluster may capture fine-grained or localized details, resulting in fragmented interpretations. The saliency maps may exhibit numerous small regions of interest, making it challenging to extract meaningful insights or identify coherent patterns. On the other hand, a small value of k $(k = 2)$ may result in under-segmentation of the generated masks. This means that the masks within each cluster may capture broader regions or features, potentially overlooking finer details or localized variations. The saliency maps may exhibit larger, less specific regions of interest, making it challenging to distinguish important elements within them. Furthermore, it is essential to take into account the computational efficiency when selecting the value of k. As k increases, the number of clusters and associated computations also increase. It is necessary to strike a balance between interpretability and computational efficiency by choosing a value that provides meaningful results without excessive computational overhead.

Through this ablation study, a comprehensive understanding of the relationship between k, N, and the interpretability of the algorithm was achieved. The results serve as a valuable reference for researchers and practitioners working with the C-RISE algorithm,

providing guidance on selecting appropriate values for k and N to achieve the desired interpretation outcomes.

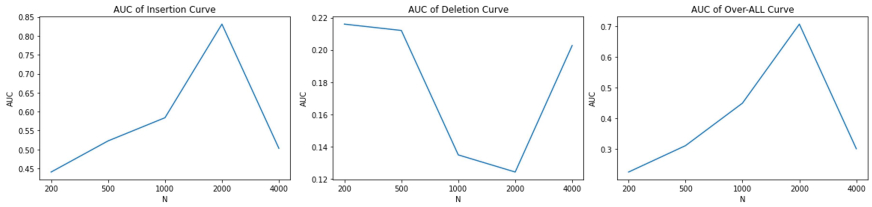

Figure 15. Ablation studies of N with $k = 4$ in terms of insertion (**higher *AUC is better***), deletion (**lower *AUC is better***) curve and the over-all scores (**higher *AUC is better***) on a subset of 100 randomly sampled images from the testset.

Table 7. Ablation studies of k with $N = 2000$ in terms of insertion (**higher *AUC is better***), deletion (**lower *AUC is better***) curve and the over-all scores (**higher *AUC is better***) on a subset of 100 andomly sampled images from the testset.

AUC	$k = 2$	$k = 4$	$k = 8$	$k = 16$	$k = 32$
Insertion	0.5687	**0.7140**	0.5076	0.5329	0.4619
Deletion	0.0942	**0.0910**	0.1697	0.1753	0.1783
Over_All	0.4745	**0.6230**	0.3379	0.3576	0.2836

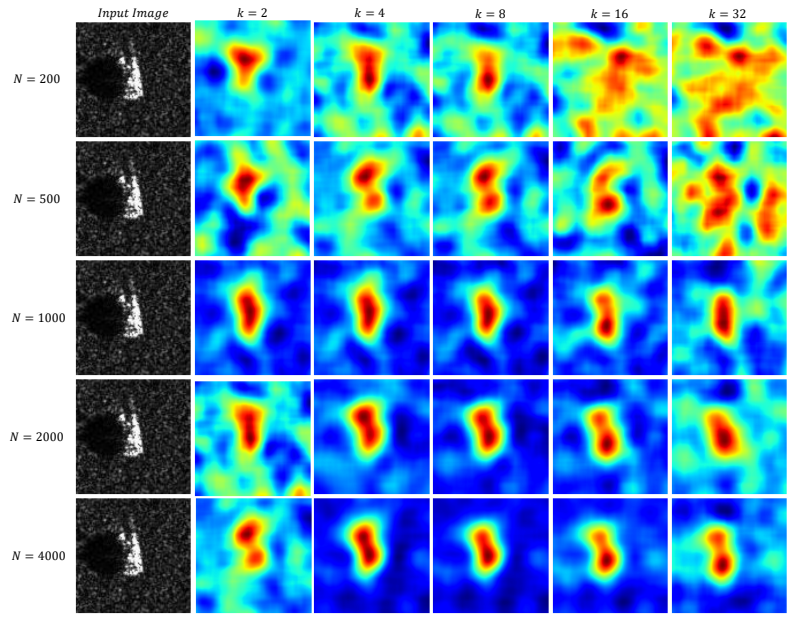

Figure 16. The saliency maps which generated by a randomly selected image obtained using different combinations of N and k.

4.6. Generalization Analysis

Although the experimental results presented so far have been obtained using the AlexNet architecture, the C-RISE algorithm exhibits remarkable generalization capabilities across a range of commonly used SAR target recognition networks. In this section, we extend our evaluation to two well-known network models: VGG16 and ResNet50. These models achieve classification accuracies of 95.05% and 92.24%, respectively, after training. Figure 17 illustrates the results obtained by applying the C-RISE algorithm to ten randomly selected images from ten different categories, using each of the aforementioned networks. To facilitate a comprehensive analysis of target localization performance, we overlay the original images with their corresponding saliency heatmaps.

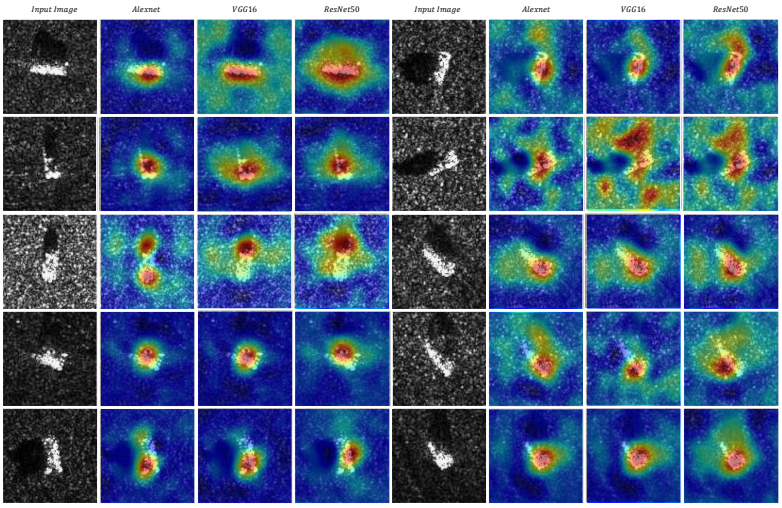

Figure 17. The results obtained by applying the C-RISE algorithm to ten randomly selected images from ten different categories, using each of the aforementioned networks.

The observations reveal that the C-RISE algorithm demonstrates consistent transferability across different networks, with the generated heatmaps consistently highlighting the target regions. However, the distribution patterns of energy vary among the different networks. Notably, ResNet50 exhibits greater divergence in energy organization, which correlates with its relatively lower recognition rate. In contrast, VGG16 achieves higher recognition rate, with increased concentration of energy across different types of heatmaps. This superior performance can be attributed to VGG16's ability to capture and effectively utilize the advanced semantic information inherent in SAR images.

SAR images possess distinct characteristics compared to traditional optical images, arising from the imaging modality and physical properties of the radar system. These images exhibit rich texture and structural details due to the interaction of radar waves with the target and surrounding environment. The deeper architecture and larger receptive field of VGG16 enable it to effectively capture and leverage these intricate details, enabling accurate recognition of subtle variations in texture and structure. On the other hand, ResNet, with its deeper architecture and skip connections, may encounter challenges in capturing and modeling the diverse and intricate energy organization present in SAR images. This can result in less accurate localization and recognition of targets, leading to lower overall performance.

It is important to note that SAR images have unique characteristics compared to optical images, including different scattering mechanisms, imaging geometry, and signal properties. These differences can impact the recognition mechanisms of different methods, especially CNNs. While deeper network architectures have demonstrated advantages in various computer vision tasks, the specific attributes of SAR images may require different network architectures or adaptations to fully leverage the SAR-specific information. The design of ResNet may not be optimized for these specific SAR image characteristics, which could explain its relatively lower performance compared to other networks. This highlights the need for further investigation and exploration to better understand and adapt SAR ATR architectures to SAR image recognition tasks.

5. Conclusions

This paper introduces C-RISE, a novel post-hoc interpretation method for black-box models in SAR ATR, which builds on the RISE algorithm. We compare the interpretation effects of different methods and C-RISE algorithm using both qualitative analysis and quantitative calculation. C-RISE offers several advantages, including its ability to group mask images that capture similar fusion features using a clustering strategy, which allows for concentration of more energy in the heatmap on the target area. Additionally, Gaussian

blur is used to process the masked area, ensuring the consistency and integrity of the original image structure and taking into account both global and local characteristics. Compared with other neural network interpretable algorithms and even white box methods, C-RISE's black-box model-oriented characteristics make it more robust and transferable. Furthermore, C-RISE avoids the error that can be caused by the unreasonable weight generation method in general CAM methods and the small feature map in the CNN model during the up-sampling process to the original image size. In addition to SAR ATR, the C-RISE algorithm has several other potential applications:

1. Weak Supervised Target Location (WSOL): Currently, both the CAM algorithm and RISE algorithm have gained widespread usage in WSOL scenarios. Similarly, C-RISE can be applied to object detection tasks, assisting in identifying important areas in an image that contribute to the presence of specific objects. By generating saliency maps and highlighting distinctive regions, C-RISE aids in object localization and provides insights into the decision-making process of black-box object detection models;
2. Weakly supervised semantic segmentation (WSSS) : WSSS methods primarily rely on CAM for target localization using image-level labels. However, CAM tends to focus on the most salient part of the object while exhibiting false activations in the background region, resulting in inadequate target activation and excessive background activation. C-RISE algorithm presents a similar form to CAM but offers improved target localization capabilities. The paper demonstrates that C-RISE achieves more accurate target positioning, effectively activating the target area, and providing guidance for precise semantic segmentation by leveraging the semantic information of category attributes;
3. Adversarial sample detection: Most existing neural network interpretability methods focus on real samples and lack explanations for classification results on adversarial samples. Adversarial sample attacks expose the vulnerability and limitations of neural network models, emphasizing the need for reliable and robust models. By generating saliency maps, C-RISE helps identify key features or patterns that contribute to detecting fraudulent activities or anomalies in large datasets. The C-RISE algorithm can qualitatively and quantitatively evaluate the reasons for misclassifications of adversarial samples by neural networks, offering effective visual explanations for the network's incorrect judgments and novel insights for adversarial sample detection.

In addition to conducting experiments on the MSTAR (SOC) database, we will apply the C-RISE algorithm to experiments involving extended operating conditions (EOC) and other SAR image datasets. By conducting experiments on diverse datasets, we can assess the applicability and performance of the C-RISE algorithm across different scenarios and datasets, thereby confirming its portability. In our future research, our objective is to concentrate on the specific application of C-RISE in the aforementioned scenarios and explore its potential in detecting improper behaviors manifested by black-box models. We aim to leverage the insights provided by C-RISE to guide parameter adjustments, thus enabling us to systematically investigate the capabilities of our proposed approach in identifying and diagnosing the sources of model inaccuracies. By devising strategies to enhance the performance of black-box models, our research endeavors will contribute to bolstering the interpretability and robustness of these models across diverse practical applications.

Author Contributions: Conceptualization, M.Z., J.C. and T.L.; methodology, J.C.; software, J.C. and T.L.; validation, M.Z. and T.L.; formal analysis, J.C. and Z.F.; investigation, X.Z.; resources, M.Z. and Y.L.; data curation, Y.L. and Z.C.; writing—original draft preparation, J.C. and Z.C.; writing—review and editing, M.Z., J.C. and T.L.; visualization, J.C. and Z.F.; supervision, X.Z.; project administration, M.Z. and X.Z.; funding acquisition, M.Z. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and technology project of Xianyang city, grant number: 2021ZDZX-GY-0001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this paper, the SAR images of ten types of vehicle targets under standard operating conditions (SOC) in the MSTAR dataset are selected as experimental data. The dataset contains 5172 SAR images with the size of $1 \times 100 \times 100$ and the training set contains 2536 images, and 2636 are used for testing. These vehicle targets are: 2*S*1, *BRDM*2, *BTR*60, *D*7, *SN*_132, *SN*_9563, *SN_C*71, *T*62, *ZIL*131, *ZSU*23_4. Readers can get the dataset from the author by email (agentcj@stu.xidian.edu.cn).

Acknowledgments: The authors would like to thank all the reviewers and editors for their great help and useful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lin, M.; Chen, S.; Lu, F.; Xing, M.; Wei, J. Realizing Target Detection in SAR Images Based on Multiscale Superpixel Fusion. *Sensors* **2021**, *21*, 1643.
2. Wang, Z.; Wang, S.; Xu, C.; Li, C.; Yue, B.; Liang, X. SAR Images Super-resolution via Cartoon-texture Image Decomposition and Jointly Optimized Regressors. In Proceedings of the 2017 International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 1668–1671.
3. Kong, L.; Xu, X. A MIMO-SAR Tomography Algorithm Based on Fully-Polarimetric Data. *Sensors* **2019**, *19*, 4839. [CrossRef] [PubMed]
4. Greenspan, M.; Pham, L.; Tardella, N. Development and evaluation of a real time SAR ATR system. In Proceedings of the 1998 IEEE Radar Conference, RADARCON'98, Challenges in Radar Systems and Solutions, Dallas, TX, USA, 14 May 1998; Cat. No. 98CH36197; pp. 38–43.
5. Clausi, D.A. Comparison and fusion of co-occurrence, Gabor and MRF texture features for classification of SAR sea-ice imagery. *Atmos.-Ocean* **2001**, *39*, 183–194. [CrossRef]
6. Belkin, M.; Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 585–591.
7. Potter, L.C.; Moses, R.L. Attributed scattering centers for SAR ATR. *IEEE Trans. Image Process.* **1997**, *6*, 79–91. [CrossRef]
8. Novak, L.M.; Benitz, G.R.; Owirka, G.J.; Bessette, L.A. ATR performance using enhanced resolution SAR. *Algorithms Synth. Aperture Radar Imag. III* **1996**, *2757*, 332–337.
9. Ding, B.; Wen, G.; Huang, X.; Ma, C.; Yang, X. Data augmentation by multilevel reconstruction using attributed scattering center for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 979–983. [CrossRef]
10. Wang, Y.; Zhang, Y.; Qu, H.; Tian, Q. Target Detection and Recognition Based on Convolutional Neural Network for SAR Image. In Proceedings of the 2018 11th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics, Beijing, China, 13–15 October 2018; pp. 1–5.
11. Mohsenzadegan, K.; Tavakkoli, V.; Kyamakya, K. A Deep-Learning Based Visual Sensing Concept for a Robust Classification of Document Images under Real-World Hard Conditions. *Sensors* **2021**, *21*, 6763. [CrossRef] [PubMed]
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
13. Dong, Y.P.; Su, H.; Wu, B.Y. Efficient Decision-based Black-box Adversarial Attacks on Face Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
14. Cai, J.L.; Jia, H.G.; Liu, G.X.; Zhang, B.; Liu, Q.; Fu, Y.; Wang, X.W.; Zhang, R. An Accurate Geocoding Method for GB-SAR Images Based on Solution Space Search and Its Application in Landslide Monitoring. *Remote Sens.* **2021**, *13*, 832. [CrossRef]
15. Cho, J.H.; Park, C.G. Multiple Feature Aggregation Using Convolutional Neural Networks for SAR Image-Based Automatic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2018**, *56*, 1882–1886. [CrossRef]
16. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

20. 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.
21. Giacalone, J.; Bourgeois, L.; Ancora, A. Challenges in aggregation of heterogeneous sensors for Autonomous Driving Systems. In Proceedings of the 2019 IEEE Sensors Applications Symposium, Sophia Antipolis, France, 11–13 March 2019; pp. 1–5.
22. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
23. Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.R. Layer-Wise Relevance Propagation: An Overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, K.R., Eds.; Springer: Cham, Switzerland, 2019; pp. 14–15.
24. Zhu, C.; Chen, Z.; Zhao, R.; Wang, J.; Yan, R. Decoupled Feature-Temporal CNN: Explaining Deep Learning-Based Machine Health Monitoring. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
25. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Proceedings of the 2nd International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
26. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
27. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 3319–3328.
28. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smooth-grad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
29. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
30. Srinivas, S.; Fleuret, F. Full-gradient representation for neural network visualization. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019 ; pp. 4126–4135.
31. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
32. Zhou, B.; Khosla, K.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
33. Ramprasaath, R.S.; Michael, C.; Abhishek, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv* **2015**, arXiv:1610.02391v4.
34. Aditya, C.; Anirban, S.; Abhishek, D.; Prantik, H. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *arXiv* **2018**, arXiv:1710.11063v34.
35. Fu, H.G.; Hu, Q.Y.; Dong, X.H.; Guo, Y.I.; Gao, Y.H.; Li, B. Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. In Proceedings of the 2020 31th British Machine Vision Conference (BMVC), Manchester, UK, 7–10 September 2020.
36. Zhang, Q.; Rao, L.; Yang, Y. Group-cam: Group score-weighted visual explanations for deep convolutional networks. *arXiv* **2021**, arXiv:2103.13859.
37. Saurabh, D.; Harish, G.R. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020.
38. Wang, H.F.; Wang, Z.F.; Du, M.N. Methods for Interpreting and Understanding Deep Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020.
39. Fong, R.; Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017 ; pp. 3449–3457.
40. Fong, R.; Patrick, M.; Vedaldi, A. Understanding deep networks via extremal perturbations and smooth masks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2950–2958.
41. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
42. Petsiuk, V.; Das, A.; Saenko, K. RISE: Randomized input sampling for explanation of black-box models. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; pp. 151–168.
43. Wissinger, J.; Ristroph, R.; Diemunsch, J.R.; Severson, W.E.; Fruedenthal, E. MSTAR's extensible search engine and model-based inferencing toolkit. In Proceedings of the Algorithms for Synthetic Aperture Radar Imagery VI, Orlando, FL, USA, 5–9 April 1999; Volume 3721, pp. 554–570.