面向SAR目标识别深度网络可理解的类激活映射方法

崔宗勇 杨致远 蒋 阳 曹宗杰* 杨建宇
(电子科技大学信息与通信工程学院 成都 611731)

摘要：随着深度学习方法在合成孔径雷达(SAR)图像解译领域的广泛应用，SAR目标识别深度网络可理解性问题逐渐受到学者的关注。类激活映射(CAM)作为常用的可理解性算法，能够通过热力图的方式，直观展示对识别任务起作用的显著性区域。然而作为一种事后解释的方法，其只能静态展示当次识别过程中的显著性区域，无法动态展示当输入发生变化时显著性区域的变化规律。该文将扰动的思想引入类激活映射，提出了一种基于SAR背景杂波特性类激活映射方法(SCC-CAM)，通过对输入图像引入同分布的全局扰动，逐步向SAR识别深度网络施加干扰，使得网络判决发生翻转，并在此刻计算网络神经元输出激活值的变化程度。该方法既能解决添加扰动可能带来的扰动传染问题，又能够动态观察和度量目标识别网络在识别过程中显著性区域的变化规律，从而增强深度网络的可理解性。在MSTAR数据集和OpenSARShip-1.0数据集上的试验表明，该文提出的算法具有更加精确的定位显著性区域的能力，相比于传统方法，在平均置信度下降率、置信度上升比例、信息量等评估指标上，所提算法具有更强的可理解性，能够作为通用的增强网络可理解性的方法。

关键词：SAR目标识别；网络可理解性；SAR杂波特性；类激活映射；面积约束置信度下降率
中图分类号：TN959.72 **文献标识码：**A **文章编号：**2095-283X(2024)02-0428-15
DOI: 10.12000/JR23188

引用格式：崔宗勇, 杨致远, 蒋阳, 等. 面向SAR目标识别深度网络可理解的类激活映射方法[J]. 雷达学报(中英文), 2024, 13(2): 428–442. doi: 10.12000/JR23188.

Reference format: CUI Zongyong, YANG Zhiyuan, JIANG Yang, *et al*. Explainability of deep networks for SAR target recognition via class activation mapping[J]. *Journal of Radars*, 2024, 13(2): 428–442. doi: 10.12000/JR23188.

Explainability of Deep Networks for SAR Target Recognition via Class Activation Mapping

CUI Zongyong YANG Zhiyuan JIANG Yang CAO Zongjie* YANG Jianyu
(*School of Information and Communication Engineering*, *University of Electronic Science and Technology of China*, *Chengdu* 611731, *China*)

Abstract: With the widespread application of deep learning methods in Synthetic Aperture Radar (SAR) image interpretation, the explainability of SAR target recognition deep networks has gradually attracted the attention of scholars. Class Activation Mapping (CAM), a commonly used explainability algorithm, can visually display the salient regions influencing the recognition task through heatmaps. However, as a post hoc explanation method, CAM can only statically display the salient regions during the current recognition process and cannot dynamically show the variation patterns of the salient regions upon changing the input. This study introduces the concept of perturbation into CAM, proposing an algorithm called SAR Clutter Characteristics CAM (SCC-CAM). By introducing globally distributed perturbations to the input image, interference is gradually applied

收稿日期：2023-10-04；改回日期：2024-01-13；网络出版：2024-02-05
*通信作者： 曹宗杰 zjcao@uestc.edu.cn *Corresponding Author: CAO Zongjie, zjcao@uestc.edu.cn
基金项目：国家自然科学基金(62271116, 61971101)
Foundation Items: The National Natural Science Foundation of China (62271116, 61971101)
责任主编： 张增辉 Corresponding Editor: ZHANG Zenghui

to deep SAR recognition networks, causing decision flips. The degree of change in the activation values of network neurons is also calculated. This method addresses the issue of perturbation propagation and allows for dynamic observation and measurement of variation patterns of salient regions during the recognition process. Thus, SCC-CAM enhances the explainability of deep networks. Experiments on the MSTAR and OpenSARShip-1.0 datasets demonstrate that the proposed algorithm can more accurately locate salient regions. Compared with traditional methods, the algorithm in this study shows stronger explainability in terms of average confidence degradation rates, confidence ascent ratios, information content, and other evaluation metrics. This algorithm can serve as a universal method for enhancing the explainability of networks.

Key words: SAR target recognition; Network explainability; SAR clutter characteristics; Class Activation Mapping (CAM); Area constrained confidence decline rate

1　引言

合成孔径雷达(Synthetic Aperture Radar, SAR)是一种可以实现高分辨率的微波主动观测系统，具备全天时、全天候以及大范围的观测能力，在海陆探测、国防军事等领域有着重要应用。SAR图像自动目标识别(Automatic Target Recognition, ATR)技术能够从SAR图像中获取目标的位置、类别等关键信息，自SAR诞生以来就受到大量关注与研究[1–5]。近年来，随着深度学习相关技术的持续发展与广泛应用，深度网络也被用于SAR图像目标检测与识别任务中，并在检测精度与识别准确率上超越了基于特征工程的传统方法[6–8]。尽管深度学习技术大幅提升了SAR目标检测识别效率，但其属于基于数据驱动的监督学习方法，主要依赖大量的标注数据来拟合网络中的参数，形成一个有曲折边界的复杂模型，其内部神经元的激活、工作或决策逻辑难以被人类理解，系统的决策边界也难以被掌握[9]，常被诟病为黑盒子。

对深度学习技术可理解性[10]的研究旨在洞悉深度网络内部工作机制、理解模型的决策，扮演人类与深度网络模型间的接口角色，帮助人们理解深度网络从数据中学到了什么特征、如何依据学到的特征进行检测识别、如何构建一个可理解的网络模型以及模型的输出是否合理与可靠等[11–13]。

现有多数可理解性方法的研究主要集中于光学图像处理领域，在面对SAR图像输入时，虽然理解结果具有一定的有效性，但是忽略了SAR图像特有的属性。因此SAR目标识别网络的可理解性研究，应有独特的切入点，许多学者也开始了有意义的探索。Datcu等人[14]利用SAR目标的物理特性和空间纹理，构建了针对SAR目标的深度学习框架，该方法保留了复值SAR数据的全部信息，使网络框架更具可理解性；Li等人[15]提出DeepSAR-Net，根据SAR目标特点微调网络结构，取得了更好的识别效果；Zhao等人[16]提出基于对比度正则化的深度网络，从复值数据中学习SAR目标的散射特征；Huang等人[17–19]将卷积神经网络(Convolutional Neural Networks, CNN)提取的图像空间特征与从SAR复值信号中提取的物理特征融合用于SAR目标分类；并提出物理信息引导网络模型与物理注入网络模型，将从散射信号中提取的物理信息精炼成先验知识，引导网络学习更具可理解性的特征或注入网络模块中提升网络性能；Li等人[20]引入属性散射中心模型和成分分析融合CNN特征，用于SAR目标自动识别。

本文将借助类激活映射方法(Class Activation Mapping, CAM)，首次将扰动的思想引入其中，并结合SAR图像特性，提出一种新的类激活映射方法，能够动态观察和度量目标识别网络在识别过程中显著性区域的变化规律，从而增强深度网络的可理解性。

2　相关工作

在研究深度神经网络的可理解性问题时，最希望能够总结得出神经网络的各层或每一个神经元学习到的具体物理信息。然而输入图像经过卷积和池化等操作，从神经网络各通道输出的特征已经变成高度抽象的特征向量，难以直观地从中观测出人类可理解的物理信息。Zeiler等人[21]经过对神经网络中神经元输出进行反卷积、反池化等一系列操作，将神经网络各层各通道的输出映射至输入空间，得出该神经元学习到的图像的像素区域；同时也证明了随着卷积和池化等操作，虽然特征图变得越来越抽象，但是其中包含的位置信息并没有丢失。如果能确定对网络决策起重要作用的高层次特征，就能定位到输入图像中的区域，在像素空间观测对网络决策起重要作用的内容，这种方法称为类激活映射，由Zhou等人[22]提出。为了能够直接比较线性层的权重，在特征提取层与分类器之间加入一个全局池化层(Global Average Pooling, GAP)，所以该类激活映射算法也被称为GAP-CAM。

梯度加权类激活映射(Gradient-weighted CAM,

Grad-CAM)[23]是对GAP-CAM的改进。在GAP-CAM中，因为必须加入全局平均池化层且不能有任何的线性层，所以模型的性能和可理解性之间总是需要平衡。与基本的CAM使用分类器的组合系数作为深层特征对最终决策结果的贡献度不同的是，Grad-CAM采用从某个决策结果流入最后一层卷积层的梯度，来定位图像中对决策起着关键作用的重要区域。Grad-CAM是一种较为通用的方法，因为只要获得梯度信息，它可以用于理解任意层的激活输出。

Grad-CAM方法虽然泛化了基本的CAM算法，能够在无需重新训练或改动网络结果的情况下适用于任何结构的CNN网络。但是当输入图像中有多个同类目标时，Grad-CAM并不能准确定位出每一个目标。另外，Grad-CAM对梯度的平均方式为简单的几何平均而非加权平均，这可能导致算法定位出的仅仅是目标的局部而非整个目标。针对上述问题，Chattopadhay等人[24]提出Grad-CAM++，主要将梯度的平均方式改为了加权平均。

Score-CAM由Wang等人[25]提出，是一种不依靠梯度信息，而是通过计算前向传播过程中输入对输出的贡献分数，来获取目标层各个神经元的组合权重。

给定一个模型$\boldsymbol{Y}=f(\boldsymbol{X})$, $\boldsymbol{X}=[\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n]^{\mathrm{T}}$, $\boldsymbol{Y}$是模型输出的预测向量，$\boldsymbol{X}$是模型输入的图片集。想要计算$\boldsymbol{x}_i$对模型输出$\boldsymbol{Y}$的贡献程度，可以选择一个已知输出的基准输入$\boldsymbol{X}_b$，将其第i个向量替换为$\boldsymbol{x}_i$，再比较更改前后神经网络输出的变化即可得到。由此，神经网络模型第l层的第k个通道输出的特征图为$\boldsymbol{A}_l^k$, $\boldsymbol{A}_l^k$对模型预测输出$\boldsymbol{Y}$的贡献可以定义为

$$C(\boldsymbol{A}_l^k)=f(\boldsymbol{X}^{\circ}\boldsymbol{H}_l^k)-f(\boldsymbol{X}_b) \tag{1}$$

$$\boldsymbol{H}_l^k=s(\mathrm{Up}(\boldsymbol{A}_l^k)) \tag{2}$$

其中，$\mathrm{Up}(\cdot)$表示将特征图上采样至输入样本尺寸，$s(\cdot)$为归一化函数，将输入向量中的元素映射至区间$[0,1]$，$(\cdot)^{\circ}$表示矩阵的哈达玛积，即矩阵对应元素分别做乘法运算，$\boldsymbol{H}_l^k$表示神经网络模型第l层的第k个通道的重要性分数。

在得到了$\boldsymbol{A}_l^k$对网络决策的贡献程度后，替换基本CAM框架中特征图的组合系数，于是Score-CAM可以定义为

$$\boldsymbol{L}_{\mathrm{Score-CAM}}^c=\mathrm{ReLU}\left(\sum_k \alpha_k^c\boldsymbol{A}_l^k\right) \tag{3}$$

其中，$\alpha_l^c=C\left(\boldsymbol{A}_l^k\right)$，由式(1)计算得出。

Score-CAM获取特征图权重的方式类似于控制变量法，将一个网络高层卷积核输出的特征图上采样至输入图像尺寸，处理输入图像，保留该卷积核所关注的信息，观测更改前后网络输出分数的变化，并以此来表征该卷积核及其感受野在网络决策中的贡献程度。

Feng等人[26]提出Self-Matching CAM方法可以精确地突出显示SAR图像中与目标最相关的区域，较为适用于分辨率低的SAR图像。Self-Matching CAM最初受到Score-CAM的启发，但旨在生成一组与输入图像匹配的新特征图，而不是对权重进行复杂操作。但是该方法将Score-CAM传统的上采样替换为下采样，可能存在对小目标信息丢失的问题。

对于CAM和其变体(如Grad-CAM)等这类基于梯度的方法，特别容易受到输入扰动的干扰。因为上述方法依赖于特征图的梯度信息，而扰动可能会改变梯度，从而影响了对通道重要性的计算。而基于模型的预测分数(Score)和类激活映射(CAM)的组合来生成显著图的(Score-CAM)方法也存在“扰动传染”[27,28]的问题。之前一系列CAM方法都试图在避免噪声对生成显著图的扰动，而本文主动将扰动加入深度网络中，进而得出特征显著图的动态变化规律。

针对SAR目标识别深度网络，扰动就需要考虑SAR图像的特性，否则在引入扰动时会导致扰动传染到其他通道，从而影响对通道重要性的判断。因此，本文根据SAR图像的背景杂波特性，在CAM系列算法[22–25]的基础上，提出一种新的类激活映射方法，增强以SAR图像为输入的识别网络模型的可理解性，有助于剖析SAR图像中目标识别的作用机理。

3 基于SAR背景杂波特性的类激活映射

本文提出的面向SAR目标识别深度网络可理解的、基于SAR背景杂波特性的类激活映射算法，其整体流程图如图1所示。首先，通过对输入图像施加与背景杂波分布相似的扰动，该扰动会随着迭代逐步增强，迫使训练好的网络发生决策错误；然后，依据该判决翻转过程中网络神经元输出激活值的变化程度，衡量该通道的重要性，并以此作为类激活映射方法中的组合系数；最后，再将特征图反向映射至输入空间，可以更准确地计算出输入图像中对网络正确决策起重要作用的显著性区域，从而增强目标识别深度网络的可理解性。

因此，本节主要从SAR背景杂波模型、类激活映射、判决翻转、通道重要性计算与更新等环节，对本文提出的方法做出详细介绍。

3.1 基于瑞利分布的SAR背景杂波模型

两个正交高斯噪声信号之和的包络服从瑞利分布[29–32]，最常见的是用于描述平坦衰落信号接收包

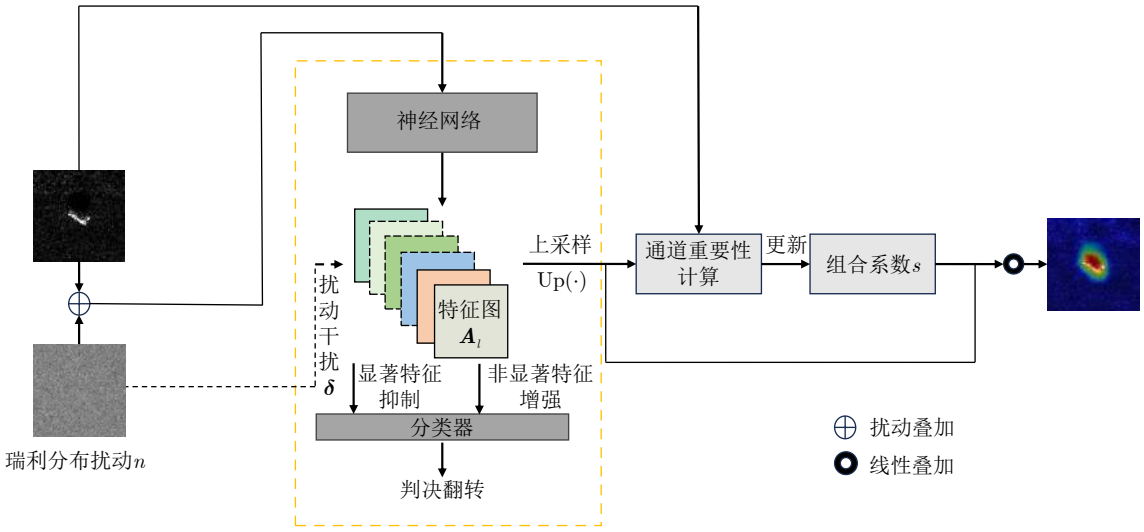


图 1 基于SAR背景杂波特性的类激活映射算法整体流程图

Fig. 1 The flowchart of class activation mapping algorithm based on SAR background clutter characteristics

络或独立多径分量接收包络的统计时变特性。瑞利分布是一种比较常用的经验分布模型，较为适合对SAR图像背景噪声的建模，瑞利分布的概率密度函数如下：

$$p(v)=\frac{2v}{\sigma}\exp\left(-\frac{v^2}{\sigma}\right) \tag{4}$$

其中，v是像素灰度值，σ是瑞利分布式中的参数。瑞利分布的σ参数的设置也需要进行考虑，根据数据的性质进行初步估计，然后使用拟合的方法，近似得出参数的范围，最后用最大似然估计来进一步优化参数的选择，在MSTAR-SOC (the Moving and Stationary Target Acquisition and Recognition-Standard Operating Conditions)数据集实验中σ参数选用的1.3，在OpenSARShip数据集实验中σ参数选用的1.5。

事实上，描述SAR背景杂波统计模型有多种，如K分布、Weibull分布、Gamma分布等。本文的研究对象设定为常见的单视SAR数据，因此选取瑞利分布作为SAR背景杂波模型。

本文方法的思路是在不改变原图像分布的基础上加入扰动，迫使网络决策发生翻转，那么加入的扰动应该与图像的分布相同。如若加入扰动的方式为乘法，表示如下：

$$\boldsymbol{x}_i^*=\boldsymbol{x}_i*\boldsymbol{\delta} \tag{5}$$

其中，$\boldsymbol{x}_i$代表将要加入扰动的图像，$\boldsymbol{x}_i^*$代表加入扰动的图像，$\boldsymbol{\delta}$为对输入图像施加的扰动，是一个尺寸与输入图像一致，数据服从瑞利分布的二维矩阵。这种干扰方式就会改变这个原有的分布，使之不再服从瑞利分布，那么就有可能会发生扰动传染的问题。

添加干扰方式：

$$\boldsymbol{x}_i^*=\boldsymbol{x}_i+\lambda\boldsymbol{\delta} \tag{6}$$

其中，λ表示扰动的强度。由于加法不会改变分布(同分布相加不改变分布，称为分布的可折叠性[33,34])，所以加入扰动的图像$\boldsymbol{x}_i^*$仍服从瑞利分布。从而避免了后续网络中扰动传染的发生。

3.2 基于SAR背景杂波特性的类激活映射

基于SAR背景杂波特性的类激活映射(SAR Clutter Characteristics-CAM, SCC-CAM)继承了Score-CAM算法的框架，将深度网络高层卷积核输出的特征图叠加映射至输入空间，以得出对网络决策起了关键作用的显著性区域。

3.2.1 基于扰动的判决翻转策略

在式(1)中，当输入由$\boldsymbol{X}$转变为$\boldsymbol{X}\circ\boldsymbol{H}_l^k$时，网络中传递的数据不仅仅只有$\boldsymbol{A}_l^k$发生了变化，而是所有层所有通道的输入与输出都会相应地发生改变，也就是说根据一个通道对图像做出的扰动，影响会传染到网络全局。这是因为训练完毕的神经网络模型虽然每个神经元各有关注重点，但相互之间并非独立，而是一个“牵一发而动全身”的整体。

给定图像集$\boldsymbol{X}$与其对应的标签类别集$\boldsymbol{L}=\{\text{lable}_1,\text{lable}_2,\cdots,\text{lable}_n\}$，对于图像$\boldsymbol{x}\in\boldsymbol{X}$与其对应的标签$\text{lable}_1$，当类别$\text{lable}_1$的特征相对于其他类别在图像$\boldsymbol{x}$中更显著时，$\boldsymbol{x}$会被分类为$\text{lable}_1$，反之亦然。于是：

(1) 图像$\boldsymbol{x}$不被分类为lable_1，则$\boldsymbol{x}$中类别lable_1的特征不显著。

(2) 图像$\boldsymbol{x}$中类别lable_1的特征更不显著时，$\boldsymbol{x}$不被分类为lable_1。

基于以上推理，可以通过干扰特征来间接寻找影响正确分类的重要特征。

假设一张能被训练好的模型正确分类的输入图像，逐步向其施加扰动，迫使网络分类错误。根据网络高层特征对分类决策起正向作用或误导作用，可以将它们划分为显著特征[35,36]和非显著特征。那么在逐步加入扰动使得网络决策错误的过程中，显著特征会被抑制，而非显著特征会被增强。

3.2.2 通道重要性计算

在确定了上述的逻辑关系之后，需要进一步对输入图像施加的扰动进行定义。在输入图像加入扰动来误导网络模型在深度学习的研究中是一种常见的方法，如生成对抗网络中的对抗样本。然而扰动的强度需要精心考虑，过于轻微的扰动可能并不能破坏模型的鲁棒性，过强的干扰会完全破坏输入以至于无法确定决策变化的边界。因此需要确定一个足以翻转模型的预测结果的最小扰动。现有的相关文献中，加入的扰动常为随机值[37]或是直接使用灰度值为0[38]。

针对SAR图像，为了有效地影响图像特征而最小程度地扰动输入图像的数据分布，本文加入的扰动服从瑞利分布，这与SAR图像的背景相干噪声的统计分布一致。于是对于图像$\boldsymbol{x}$，模型f和类别y，最小扰动可以定义为

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} D(\boldsymbol{\delta}) \tag{7}$$

s.t.

$$f(\boldsymbol{x}_i) = y \tag{8}$$

$$f(\boldsymbol{x}_i - \boldsymbol{\delta}) \neq y \tag{9}$$

$$f(t(\boldsymbol{x}_i)) = f(\boldsymbol{x}_i) = y \tag{10}$$

$$f(t(\boldsymbol{x}_i - \boldsymbol{\delta})) \neq y \tag{11}$$

其中，$D(\cdot)$是L_1范数，$t(\cdot)$代表旋转、放缩等图像变换。上述目标函数与约束条件保证了输入在正确的范围内，且加入的扰动是能使原始网络判决错误，且对原始图像的破坏是最小的扰动。

对于网络模型$y = f(\boldsymbol{x})$，将网络第l层第j个卷积核输出激活值的变化程度c_j^i定义如式(12)所示：

$$c_j^i = \frac{f_l(\boldsymbol{x}_i)[j] - f_l(\boldsymbol{x}_i - \boldsymbol{\delta}^*)[j]}{f_l(\boldsymbol{x}_i)[j]} \tag{12}$$

其中，$\boldsymbol{x}_i$为输入图像，$f_l(\boldsymbol{x}_i)[j]$为网络第l层第j个卷积核输出激活值，$f_l(\boldsymbol{x}_i - \boldsymbol{\delta}^*)[j]$则为对图像加入扰动$\boldsymbol{\delta}^*$后，对应输出激活值。

神经元输出激活值变化的剧烈程度代表了在网络决策从正确到错误的过程中，该神经元所表征的特征被破坏的程度，也从侧面说明了该特征对网络能够正确决策的贡献程度。从前向传播角度而言，是该神经元特征被严重破坏，引起了网络决策错误。也就是说，c_j^i也表征了模型对原始输入图像做出正确决策时，网络第l层第j个通道的通道重要性。为了在后续实验中能生成更加平滑的显著性图，将c_j^i归一化至$[0,1]$区间，得出通道重要性分数：

$$s_j^i = \frac{c_j^i - \min c_j^i}{\max c_j^i - \min c_j^i} \tag{13}$$

3.2.3 更新组合系数

在求得通道的重要性分数后，便可以在Score-CAM结构的基础上，用通道重要性分数更新各特征图映射至输入图像时的组合系数。给定一个目标类标签y，网络模型的l层的所有通道生成的SCC-CAM可以定义为

$$\boldsymbol{L}^y = \sum_k \omega_k^y \boldsymbol{A}_l^k \tag{14}$$

$$\omega_k^y = s_k^i = s\left(\frac{f_l(\boldsymbol{x}_i)[k] - f_l(\boldsymbol{x}_i - \boldsymbol{\delta})[k]}{f_l(\boldsymbol{x}_i)[k]}\right) \tag{15}$$

其中，$s(\cdot)$为式(11)代表的归一化处理。至此，便得到了新的可以替代基本类激活映射方法中组合权重的方案，能适合任何形式的网络结构，且对图像的扰动不用传染影响其他神经元节点。

3.3 算法求解流程

所提方法中的关键是如何求解式(7)中的最小扰动$\boldsymbol{\delta}^*$，由于约束条件中包含了神经网络，目标函数难以求得其解析解。为了得到最佳的扰动$\boldsymbol{\delta}^*$，需要使用迭代的方法求解。具体来说，对于训练好的网络模型和测试图像，从一个极其小的扰动开始，该扰动不会破坏网络的鲁棒性，逐渐迭代增大扰动的强度，同时观测受扰图像通过网络后判决的类别，直至网络决策发生错误，以此逼近判决翻转最小扰动的近似解。

求解流程如算法1所示，其中需要的输入信息是输入图像$\boldsymbol{I}_{\text{src}}$、训练完毕的网络模型$f$、感兴趣的类别标签$y$、尺度因子$s$以及服从瑞利分布的扰动矩阵$\boldsymbol{n}$。尺度因子用于控制扰动变化的细腻程度，尺度因子越小，寻找到的决策边界越准确，但同时迭代的次数也可能相应地增加。通过q来控制扰动的强度，用施加扰动后的图像来作为模型的输入，若网络的输出标签l仍然与原始标签y一致，则增加扰动强度重复此步骤。

算法 1　SCC-CAM求解算法流程

Alg. 1　SCC-CAM algorithm flow

Data: SAR图像$\boldsymbol{I}_{\mathrm{src}}$，模型$f(\cdot)$，目标类别$y$，尺度因子$s$，扰动矩阵$\boldsymbol{n}$

Result: SCC-CAM显著性图

初始化;

$q \leftarrow 0$;

$\mathrm{lable} \leftarrow f(\boldsymbol{I}_{\mathrm{src}})$;

$\boldsymbol{\delta}^* \leftarrow 0$;

while $\mathrm{lable} = y$ and $q < 60$ do

$\boldsymbol{\delta}^* = q * \boldsymbol{n} * s$;

$\boldsymbol{I}_{\mathrm{src}} = \boldsymbol{I}_{\mathrm{src}} + \boldsymbol{\delta}^*$;

$l = f(\boldsymbol{I}_{\mathrm{src}})$;

$q = q + 1$;

end

$s_j^i = \dfrac{f_l(\boldsymbol{x}_i)[j] - f_l(\boldsymbol{x}_i - \boldsymbol{\delta}^*)[j]}{f_l(\boldsymbol{x}_i)[j]}$;

$\boldsymbol{A}_l^j \leftarrow f_l(\boldsymbol{x}_i)[j]$;

$\mathrm{SCC\_CAM} \leftarrow \sum_j s_l^j \mathrm{Up}\left(\boldsymbol{A}_l^j\right)$;

循环退出的条件是网络决策发生变化，或者强度q超过预设的上限，此时就近似求解出了使网络决策错误的最小扰动。设置q上限是为了防止在特征被完全破坏时，神经网络无从判决而倾向于一直推理为某个类别从而导致算法进入死循环。算法的输出为以通道重要性作为组合系数的SCC-CAM。

图2以VGG16网络为例，展示算法的迭代过程。通过对输入图像施加扰动，能够动态展示显著性区域随着扰动的强度增加而逐渐变化的情况。图2从左到右，依次经过欠扰动、判决翻转最小扰动、过度扰动3个阶段。初始时刻扰动的强度q为0，随着扰动逐渐加强，当扰动强度q为45时，达到判决翻转，在这之前的过程为欠扰动阶段，具体表现为显著区域逐渐向目标处集中，杂波背景区域的显著性逐渐降低；判决翻转之后，扰动强度q继续增加到上限60，对应着图2中的后半部分，即过度扰动阶段，具体表现为特征图逐渐恶化，最终显著性区域完全消失，代表深度网络对于过度干扰的输入完全失去判断能力。

4　试验验证与分析

4.1　试验数据集

本文的研究对象是面向SAR目标识别任务的深度网络，为了验证本文提出的方法的有效性，试验将使用MSTAR数据集[39]。MSTAR数据集的分辨率是0.3 m×0.3 m，采用HH极化方式，用于显著性区域试验和所有的定量试验。SOC条件下的MSTAR数据集，包含10种目标在不同方位角和俯仰角下的SAR单视复数据，通常俯仰角17°作为训练样本，俯仰角15°作为测试样本。表1展示了MSTAR数据集中用于训练与测试的各类别目标的数量情况。

同时本文也使用了OpenSARShip-1.0数据集[40]分辨率在2.7 m×22 m至3.5 m×22 m的3类目标，主要用于显著性区域试验和定量试验，数据使用情况如表2所示。

4.2　SAR目标识别深度网络设计

为验证SCC-CAM的泛化性能，试验从两个方面进行网络选择和设计：一是选取经典的网络结构如VGG16[41]和ResNet18[42]，分别在MSTAR数据集和OpenSARShip-1.0数据集上重新训练其特征提取器和分类器；二是针对当前数据集自行搭建网络结构，使其达到较高识别率。

4.2.1　VGG16网络结构

如图3(a)所示，VGG16网络结构由5个卷积模块、2个全连接层和1个输出的softmax层组成。卷积模块中前2个卷积模块都是连续2个卷积层后接1个池化层，后3个卷积模块是连续3个卷积层后进

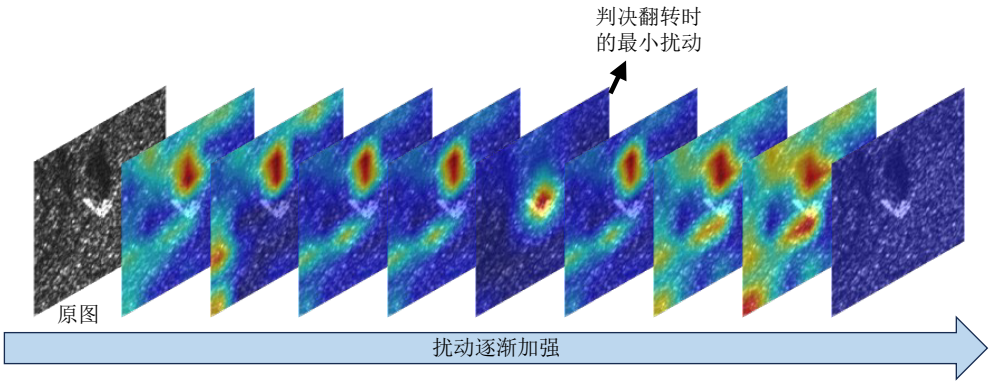


图 2 随着扰动强度增加，在VGG16网络的最后一个MaxPooling层上使用SCC-CAM展示显著性区域的变化

Fig. 2 As the perturbation intensity increases, variations in the saliency regions displayed using SCC-CAM on the last MaxPooling layer of the VGG16 network

表 1 MSTAR-SOC数据集样本选取情况
Tab. 1 The sample selection situation of the MSTAR-SOC dataset

类别	训练样本	测试样本
2S1	299	274
BMP2	233	195
BRDM2	298	274
BTR60	256	195
BTR70	233	196
D7	299	274
T62	298	273
T72	232	196
ZIL131	299	274
ZSU23-4	299	274

表 2 OpenSARShip-1.0数据集样本选取情况
Tab. 2 The sample selection situation of the OpenSARShip-1.0 dataset

类别	训练样本	测试样本
BulkCarrier	160	40
Cargo	160	40
Container	160	40

行池化操作。针对MSTAR数据集，最终分类正确率达到95.60%；针对OpenSARShip-1.0数据集，最终分类正确率达到98.77%。

4.2.2 ResNet18网络结构

与VGG16网络类似，由于ResNet18网络结构的全连接层参数量巨大，本文试验中保留了其特征提取层，对分类器做了调整，使其最终的输出维度为10，以适应试验数据集。将ResNet18网络分别在MSTAR数据和OpenSARShip-1.0数据集上进行测试，最后分类正确率达到96.31%和98.73%。

4.2.3 自建网络

VGG16和ResNet18网络结构复杂，网络层数多，为了验证本文所提方法的适用性，本节还自建了一个小型的网络结构，网络结构如图3(b)所示，共有10层(不计算池化层)，其中前7层为特征提取层，后3层是全连接层，特征提取层的卷积核尺寸均为3×3。该网络在MSTAR数据集上的识别正确率可以达到98.25%，针对OpenSARShip-1.0数据集中3类目标的识别正确率可以达到99.12%。

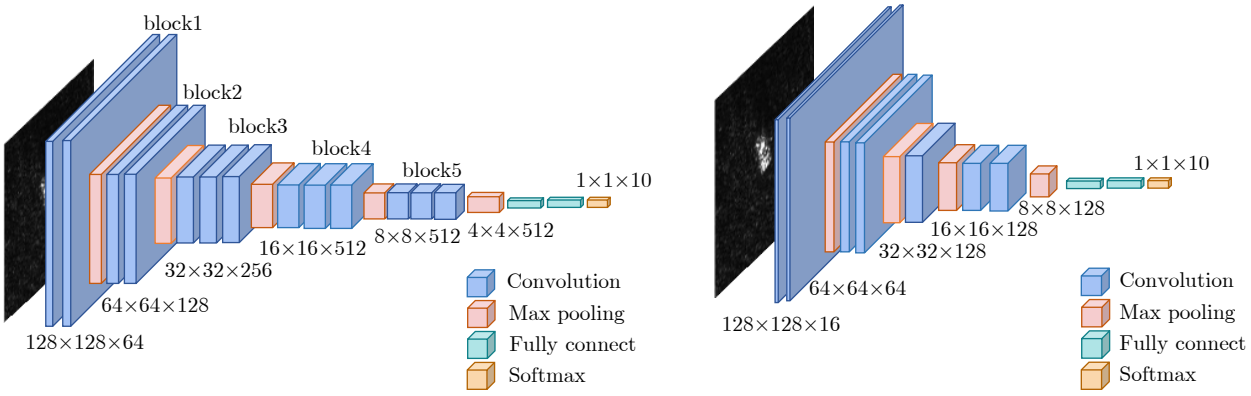
4.3 试验结果与分析

将训练好的VGG16, ResNet18和自建网络，分别使用Grad-CAM++, Score-CAM和SCC-CAM进行显著性区域提取，然后在目标背景分离[43,44]的基础上，分别采用平均置信度下降率、置信度上升比例、显著性区域分类性能等评价指标对不同算法进行评估。

4.3.1 显著性区域对比

图4展示了在VGG16网络结构下，3种算法得出的输入图像中对图像正确分类起关键作用的显著性区域。其中，图4(a)、图4(e)为输入图像，图4(b)、图4(f)与图4(c)、图4(g)分别是Grad-CAM++和Score-CAM算法生成的显著性图，图4(d)、图4(h)则是SCC-CAM得出的决定网络正确分类的关键区域。3种方法的结果都显示出，图像中的目标区域是决定了网络决策的显著性区域，但是在细节处仍有不同，Grad-CAM++生成的显著图在图像的4个角落仍有较高的能量。

图5展示在ResNet网络结构下，Grad-CAM++, Score-CAM和SCC-CAM得到的反映网络决策的显著性区域。可以看出，图像中目标区域对网络将输

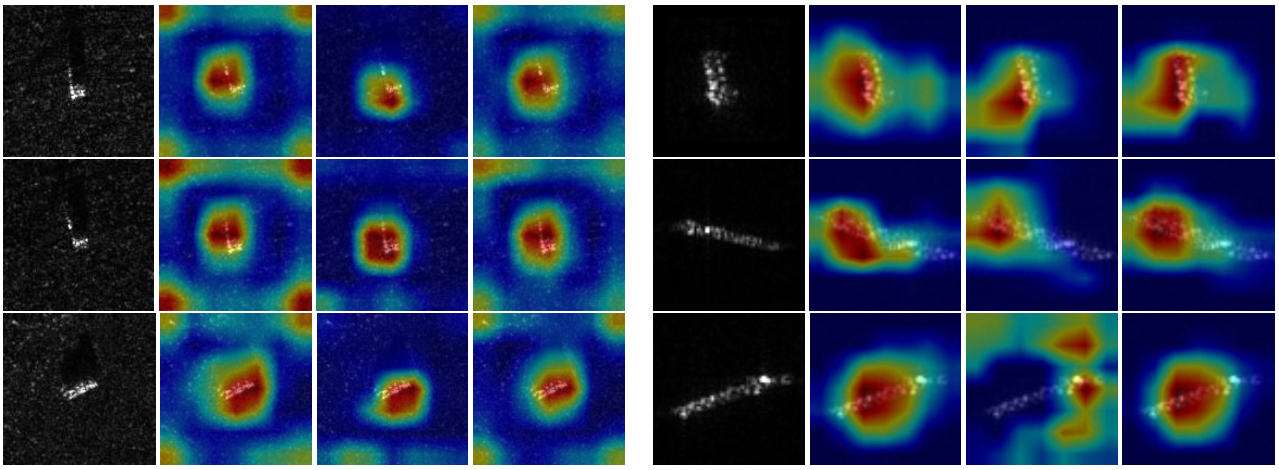


(a) VGG16网络结构
(a) Network structure of VGG16

(b) 自建网络结构
(b) Self-built network structure

图 3 试验选取的网络结构

Fig. 3 The network structure selected in the experiment

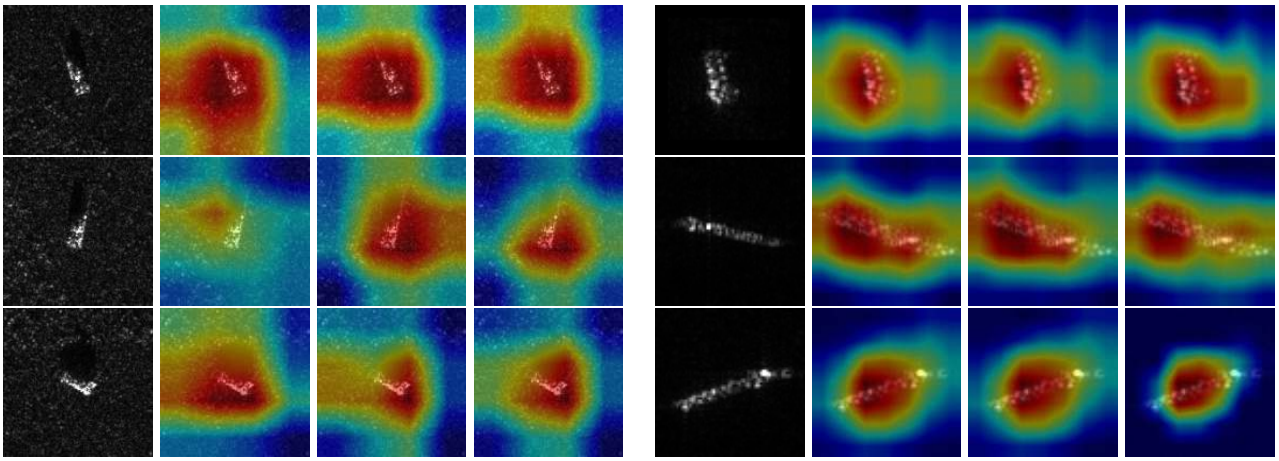


(a) 输入SAR图像　(b) Grad-CAM++　(c) Score-CAM　(d) SCC-CAM　(e) 输入SAR图像　(f) Grad-CAM++　(g) Score-CAM　(h) SCC-CAM

(a) Input SAR image　(b) Grad-CAM++　(c) Score-CAM　(d) SCC-CAM　(e) Input SAR image　(f) Grad-CAM++　(g) Score-CAM　(h) SCC-CAM

图 4 VGG16网络显著性区域对比(左侧为MSTAR，右侧为OpenSARShip-1.0)

Fig. 4 Comparison of saliency area of VGG16 (the left is MSTAR, the right is OpenSARShip-1.0)



(a) 输入SAR图像　(b) Grad-CAM++　(c) Score-CAM　(d) SCC-CAM　(e) 输入SAR图像　(f) Grad-CAM++　(g) Score-CAM　(h) SCC-CAM

(a) Input SAR image　(b) Grad-CAM++　(c) Score-CAM　(d) SCC-CAM　(e) Input SAR image　(f) Grad-CAM++　(g) Score-CAM　(h) SCC-CAM

图 5 ResNet网络显著性区域对比(左侧为MSTAR，右侧为OpenSARShip-1.0)

Fig. 5 Comparison of saliency area of ResNet (the left is MSTAR, the right is OpenSARShip-1.0)

入图像判定为所属类别标签起到了主要作用。但是Grad-CAM++得到的显著性区域面积更大，并且在第2张图像作为输入时，Grad-CAM++算法出现了梯度消失的问题，无法定位图中的显著性区域，这是基于梯度的方法可能存在的潜在问题。针对MSTAR数据集，Score-CAM与SCC-CAM生成的结果相似，但是Score-CAM生成的显著性区域在边界处存在拖尾现象，SCC-CAM对于区域的范围控制更好。针对OpenSARShip-1.0数据集，Score-CAM生成的显著性区域难以捕捉完整的目标区域，而SCC-CAM生成的显著性区域能更好地定位目标。

图6展示了在自建网络中，Grad-CAM++, Score-CAM以及SCC-CAM生成的显著性图对比。可以看出，无论是MSTAR还是OpenSARShip-1.0数据集，自建网络都能更好地关注到目标区域。这可能是由于自建网络的层数较浅，每一层的神经元个数也较少，更适合于纹理信息不丰富的SAR图像。

但是每种方法生成的显著性图在细节处仍有不同，显著性区域的面积以及像素分布均有差异。如图6第1张测试图像(第1行)，Grad-CAM++算法认为对网络决策起重要作用的区域为图像中目标的中下部分，Score-CAM算法框选出目标的中下部分和目标左上方的阴影部分，而SCC-CAM则认为目标整体和阴影部分共同决定了网络的正确决策。

为了探讨深度网络不同层的推理能力，针对VGG16, ResNet18和自建网络3种网络，采用本文提出的SCC-CAM算法，分别提取3种网络在发生判决翻转时，其5个不同层的显著性区域，结果如图7所示。可以看到，随着网络的层数变深，显著性区域从零散的区域逐渐变为集中的区域，但是在不同网络结构下，最后显著性区域略有差别，对于VGG16和自建网络，显著性区域集中于目标区域；而对于ResNet18网络，显著性区域集中于目标和阴影以及周围的背景区域。

图8对比了3种CAM方法SCC-CAM, Grad-CAM++和Score-CAM，在VGG16网络下提取的

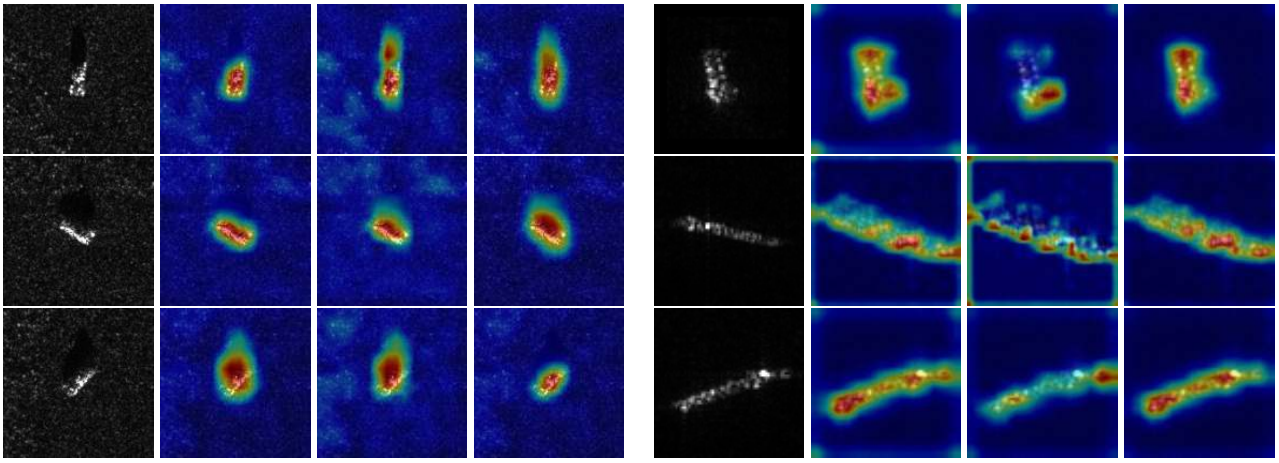


图 6 自建网络显著性区域对比(左侧为MSTAR，右侧为OpenSARShip-1.0)

Fig. 6 Comparison of saliency area of self-built network (the left is MSTAR, the right is OpenSARShip-1.0)

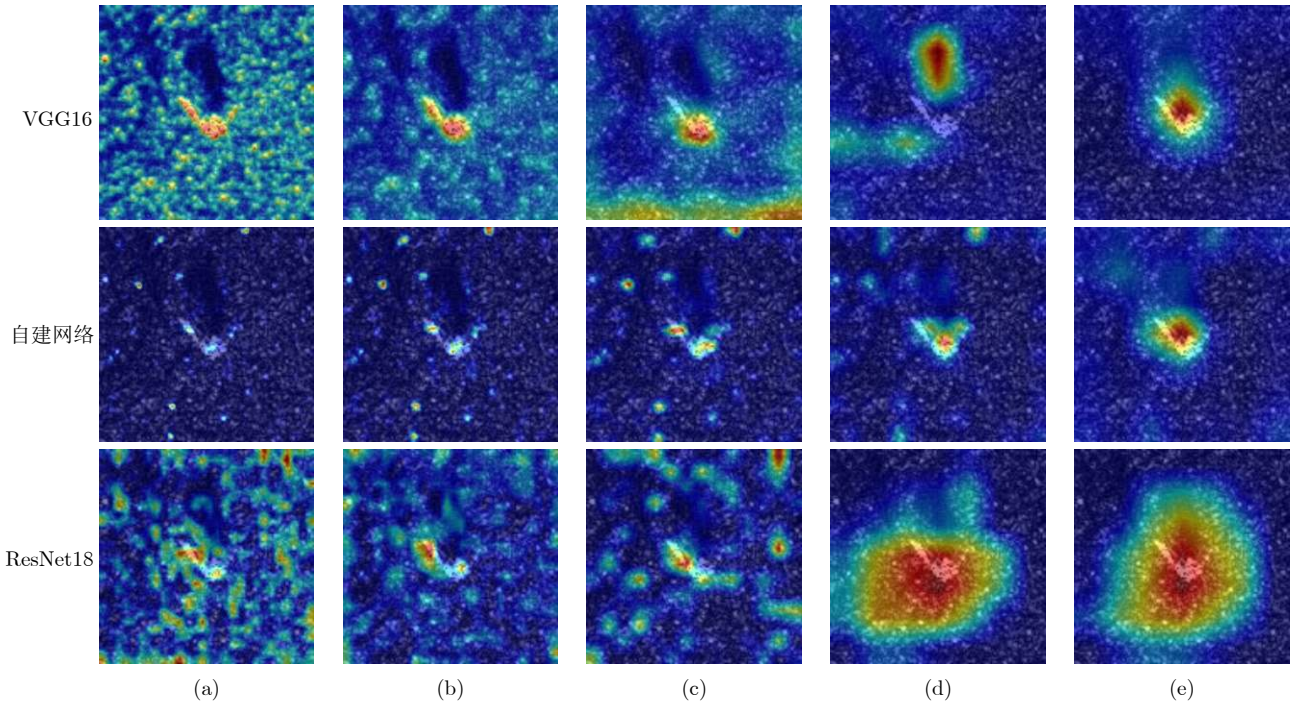


图 7 VGG16, ResNet18和自建网络发生判决翻转时采用SCC-CAM提取的不同层显著性区域(第1行和第2行的(a)～(e)分别对应VGG16和自建网络的第1到第5个最大池化层；第3行的(a)～(e)对应ResNet18的layer1到layer3的第4个卷积层以及layer4的第2和第4个卷积层)

Fig. 7 When decision flipping occurs for VGG16, ResNet18, and the self-built network, different salient regions are extracted using SCC-CAM from various layers (for the first and second rows, (a)～(e) correspond to the first through fifth max-pooling layers of VGG16 and the self-built network. In the third row, (a)～(e) correspond to the fourth convolutional layer of ResNet18's layer1 to layer3, and the second and fourth convolutional layers of layer4)

不同层显著性区域。可以看到，随着网络的层数变深，3种算法均出现显著性区域从零散的区域逐渐变为集中的区域的现象。Grad-CAM++和Score-CAM算法在第4个和第5个最大池化层处的显著性区域关注到图片四周和背景区域，而本文提出的方法在不同层显著性区域更加集中于目标附近。

综上所述，3种算法在不同的网络结构中，选定的显著性区域有所差异。但在相同网络结构下，都能生成在视觉效果上相似的显著性图，要评价方法的质量，需要进一步开展量化评估试验。

4.3.2 量化评估试验

为了检验SCC-CAM算法选取显著性区域的准确性，本节开展量化评估试验，首先将目标区域与背景区域分离，观测在目标或背景遮挡的情况下，网络输出的变化。此处的目标与背景是指CAM方法认定的显著性区域和图像的其他区域，而非图像中的目标的实际区域和其所处的场景。

具体的做法是，当CAM方法生成显著性图像时，为目标层的每一张特征图赋予一个权重，最后组成了输入图像中像素的重要分数，设定一个阈值，记录高于此的像素位置并将其分割为显著性区域，其余部分为非显著性区域。图9展示了2S1类别中一张图像的分离效果，其中图9(c)和图9(d)分别为遮挡显著性区域和遮挡背景的互补图像。

4.3.2.1 平均置信度下降率

在遮挡了图像的非显著性区域后，再送入使用完整图像训练的网络中，相较于使用原始图像作为

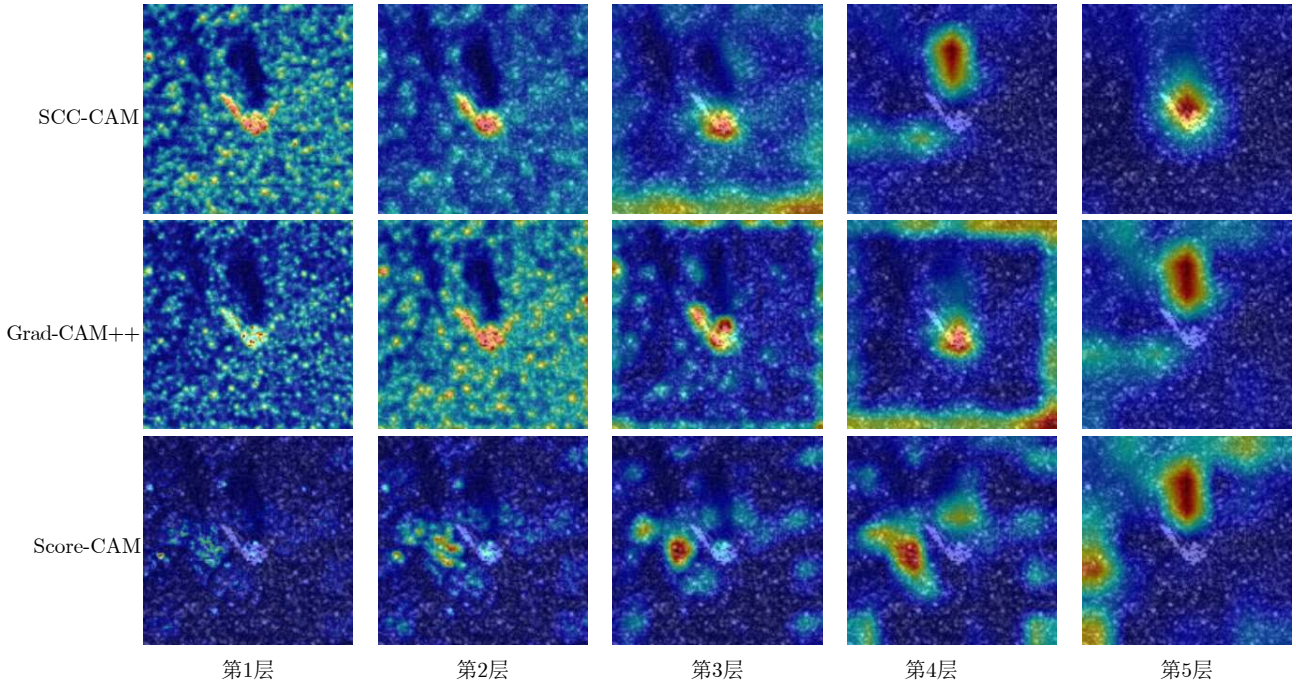
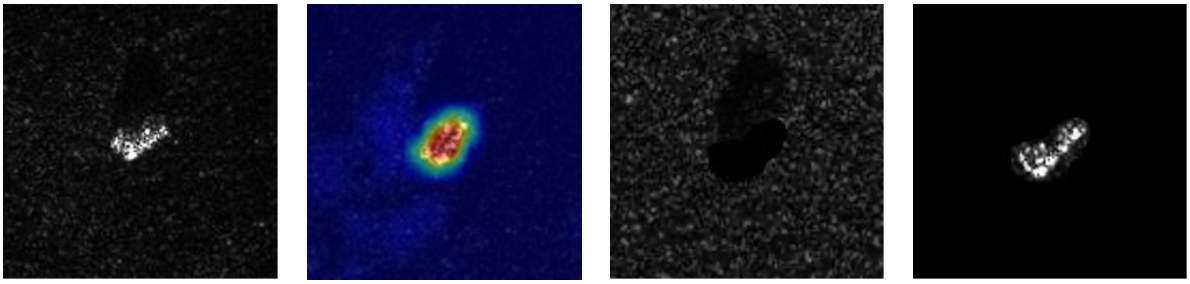


图 8 SCC-CAM, Grad-CAM++和Score-CAM在VGG16网络下提取的不同层显著性区域

Fig. 8 Displays the salient regions extracted by SCC-CAM, Grad-CAM++, and Score-CAM



(a) 输入SAR图像
(a) Input SAR image

(b) 显著性区域定位
(b) Location of saliency area

(c) 遮挡显著性区域
(c) Mask saliency area

(d) 遮挡背景
(d) Mask background

图 9 显著性与非显著性区域分离

Fig. 9 Split of saliency area and non-saliency area

输入，网络的判断会错误，或者即使还能预测为正确类别，其置信度也会下降[45]，如从95%下降至51%。所以本节使用平均置信度下降率来衡量SCC-CAM的准确性。使用灰度值为0的像素把非显著性区域遮挡住，仅含有显著性区域的图像通过网络时，置信度下降率的幅度可以度量该区域对于网络决策的重要性。

$$d = \frac{1}{N}\sum_{i}\frac{\max(0, y_i^c - m_i^c)}{y_i^c} \tag{16}$$

其中，y_i^c为数据集中第i张原始输入图像通过网络时，网络输出的标签c的得分，m_i^c为将仅保留显著区域，遮挡其他区域的图像输入网络后，计算得到的标签c的得分。使用$\max(0, y_i^c - m_i^c)$来处理仅保留显著性区域网络预测置信度反而升高的情况，是因为试验主要关注遮挡前后置信度的下降，而对于升高的情况，可以认为其置信度下降率为0，忽略其上升部分。显然，仅保留显著性区域作为网络输入，置信度相较于完整图像输入时下降越少，说明该区域在网络正确决策时的贡献度越高，显著性区域越准确，算法的性能更好。

表3展示了当仅保留Grad-CAM++, Score-CAM和SCC-CAM选定的显著性区域，分别通过VGG16, ResNet18以及自建网络时，相比于原始完整图像，判定为正确类别的平均置信度下降率。两个数据集均表明，SCC-CAM在3种网络结构下，输入保留的显著性区域引起的置信度下降率，都小于Grad-CAM++和Score-CAM算法。

4.3.2.2 基于面积约束的平均置信度下降率

平均置信度下降率对于评估不同算法提取同一张图像、或者同一算法提取不同图像的显著性区域是否准确，是比较有效的。但是针对不同的算法，平均置信度下降率并不能精确评估与比较算法的性能，因为遮挡的面积也是影响网络决策的关键因素。被遮挡的面积越大，可提供给网络模型推理的信息就越少，网络预测为正确类的置信度就可能明显下降；反之，当遮挡面积较少时，模型正确判决的置信度较高，模型的分类错误率更低。

表 3 不同网络模型的平均置信度下降率(%)

Tab. 3 Average confidence degradation rates across different network models (%)

数据集	网络模型	Grad-CAM++	Score-CAM	SCC-CAM
MSTAR-SOC	VGG16	59.60	59.01	57.20
	ResNet18	60.54	55.91	52.77
	自建网络	46.00	43.29	42.14
OpenSARShip-1.0	VGG16	44.27	39.13	37.40
	ResNet18	46.94	42.17	41.84
	自建网络	41.49	37.89	33.66

考虑一种极端情况，若一种算法认为整张图像均为对决策起重要作用的显著性区域，那么在遮挡非显著性区域时，几乎没有任何像素会被遮挡，处理后的图像与原始图像几乎无异，网络决策时，对其所属标签类的决策置信度的下降程度影响甚微，然而这并不能说明当前方法有更好的理解效果。如图10所示，其中图10(a)在仅保留了显著性区域后通过网络，网络的置信度下降率相比于图10(b)的更低，但其却圈定了远超后者面积的区域。

因此在平均置信度下降率的基础上，本文提出了一种新的评价指标：基于面积约束的平均置信度下降，对显著性区域的面积加以反向约束，以更加精确地度量算法的准确性。对于输入图像$\boldsymbol{x}_i$，面积为S_i，显著性区域面积为A_i，对式(16)进行修正为

$$r_i = \frac{A_i}{S_i} \tag{17}$$

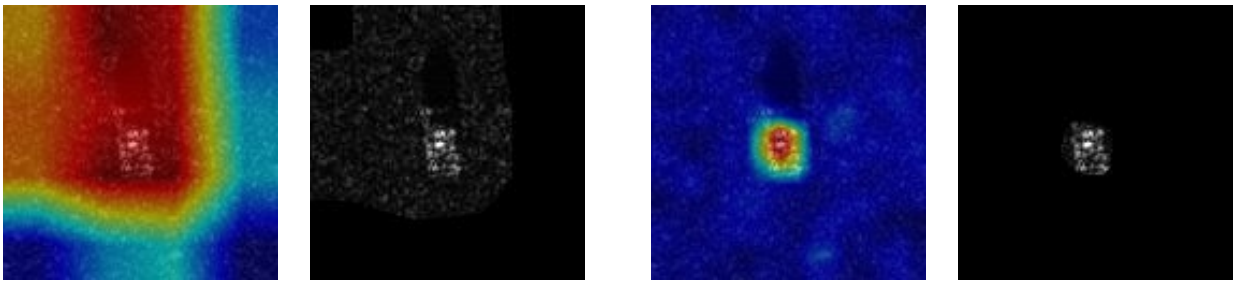
$$d_{\text{Area}} = \frac{1}{N}\sum_{i}\frac{\max(0, y_i^c - m_i^c)}{y_i^c}\cdot r_i \tag{18}$$

其中，r_i表示选中的显著性区域的面积与输入图像面积的比值，在式(18)中起到惩罚项的作用，当算法选中的显著性区域的面积占比较大时，虽然平均置信度下降率会较小，但此时r_i会拉高d_{Area}的数值。同时，当仅用少量面积导致置信度下降稍高时，也不直接认为对应方法质量不好，因为r_i会使整体d_{Area}更小。

如图10的例子，图10(a)显著性面积占比大，其r_i约为0.63，图10(b)的r_i约为0.08。将面积约束作为对平均置信度下降率的惩罚因子，得到基于面积约束的平均置信度下降率d_{Area}，分别为6.93%和0.24%。可以看到，在未考虑面积约束时，图10(a)的置信度下降率相比图10(b)的较小，在考虑了面积约束后，修正后置信度下降率相比图10(b)的变大。

经修正后的d_{Area}表征了算法定位对决策起重要作用的区域的准确性，越小的d_{Area}值代表在保留了较小的区域，通过原先的网络模型，仍有较高的置信度将其判断为正确的目标类别，从而避免了片面地评价算法的质量，保证了评估显著性区域定位的精准性。

从表3和表4可以看到，经过面积约束修正后，平均置信度下降率更能衡量算法的准确性。针对自建网络，在未考虑面积约束时Score-CAM比Grad-CAM++算法的平均置信度下降率更小，在考虑了面积约束后，由于Score-CAM选中的显著性面积更大，导致Score-CAM比Grad-CAM++算法的面



(a) 置信度下降率11%
(a) 11% drop in confidence

(b) 置信度下降率30%
(b) 30% drop in confidence

图 10 不同面积的显著性区域下置信度对比

Fig. 10 Comparison of confidence scores under different area sizes of salient regions

积修正后的平均置信度下降率更大。而本文提出的SCC-CAM，其确定的显著性区域所引发的置信度下降率，远小于Score-CAM及Grad-CAM++算法。

4.3.2.3 置信度上升比例

置信度上升比例[21]用作对平均置信度下降率的补充评价指标。多数图像在被遮挡一部分后，网络预测为原来类别的置信度会降低，但是仍然会出现只保留显著性区域最终预测为正确类的置信度上升的情形，尤其是目标之外的背景可能会造成干扰时，遮挡干扰部分反而更有助于模型的决策。置信度上升比例定义为测试集中，仅保留显著性区域通过网络，预测为正确类置信度上升的样本占总体的比例：

$$I=\sum_{i=1}^{N}\frac{1_{Y_i^c<O_i^c}}{N} \tag{19}$$

其中，N是测试集中样本数量，1_x是一个二值函数，当条件为真时，返回1，否则为0，用于置信度上升样本计数。

表5展示了Grad-CAM++, Score-CAM和SCC-CAM保留的显著性区域通过网络时的置信度上升样本占全部样本的比例。本文使用的MSTAR数据集中测试集样本共2425张图像。在VGG16网络中，Grad-CAM++算法置信度上升样本共347张，占比14.31%，Score-CAM算法置信度上升样本394张，占比16.25%，SCC-CAM置信度上升样本共414张，占比17.07%。

OpenSARShip-1.0数据集中测试集样本共120张，在VGG16网络中，Grad-CAM++算法置信度上升样本共16张，占比13.33%，Score-CAM算法置信度上升样本17张，占比14.17%，SCC-CAM置信度上升样本共19张，占比15.83%。

针对ResNet18网络和自建网络的测试结果类似，相比于Grad-CAM++和Score-CAM算法，SCC-CAM能够取得最高的置信度上升比例。

4.3.2.4 显著性区域分类性能

为了对比不同算法确定的显著性区域是否包含了决定网络分类的足够信息，本节使用分离出来的图像显著性区域和非显著性区域分别作为训练集，测试数据保持不变。基于显著性区域的分类性能越好，说明算法提取的显著性区域包含了更多对分类有用的信息，也就说明了算法所选定的网络决策关键区域越准确。

表6展示了使用3种算法提取的显著性区域用作训练集，对原网络结构进行重新训练后，模型的分类正确率。针对MSTAR-SOC数据集，使用2425张样本用作测试，使用SCC-CAM选取的显著性区域，重新训练VGG16, ResNet18网络和自建网络后，分类正确率分别为76.00%, 77.20%和81.24%。在

表 4 不同网络模型的基于面积约束的平均置信度下降率(%)

Tab. 4 Average confidence degradation rates based on area constraints across different network models (%)

数据集	网络模型	Grad-CAM++	Score-CAM	SCC-CAM
MSTAR-SOC	VGG16	7.19	5.74	4.82
	ResNet18	17.87	14.61	12.97
	自建网络	1.82	2.06	1.54
OpenSARShip-1.0	VGG16	6.20	6.02	4.10
	ResNet18	17.56	15.55	13.09
	自建网络	2.14	3.45	1.53

表 5 不同网络模型的置信度上升比例(%)

Tab. 5 Confidence ascent ratios across different network models (%)

数据集	网络模型	Grad-CAM++	Score-CAM	SCC-CAM
MSTAR-SOC	VGG16	14.31	16.25	17.07
	ResNet18	15.55	16.74	17.69
	自建网络	19.22	21.40	21.94
OpenSARShip-1.0	VGG16	13.33	14.17	15.83
	ResNet18	16.71	17.08	19.86
	自建网络	17.50	19.17	20.83

表 6 显著性区域用作训练集的分类性能(%)

Tab. 6 The performance of saliency area is used as the training set (%)

数据集	网络模型	Grad-CAM++	Score-CAM	SCC-CAM
MSTAR-SOC	VGG16	71.01	74.97	76.00
	ResNet18	70.31	75.34	77.20
	自建网络	78.89	80.08	81.24
OpenSARShip-1.0	VGG16	77.50	81.67	83.33
	ResNet18	78.33	80.00	80.00
	自建网络	80.00	82.25	85.00

OpenSARShip-1.0数据集上进行相似的试验配置，挑选出120张样本用作测试集，使用SCC-CAM选取的显著性区域，训练VGG16, ResNet18网络和自建网络后，分类正确率分别为83.33%, 80.00%和85.00%。

可以看出，在相同网络结构下，利用SCC-CAM的分类正确率均高于Grad-CAM++算法和Score-CAM算法，说明了SCC-CAM定位的显著性区域包含了更多的驱使网络正确分类的信息，进而验证了本文所提方法的有效性。

5 结语

本文提出了一种面向SAR目标识别深度网络可理解任务的类激活映射方法SCC-CAM，通过扰动图像迫使网络决策发生错误，且对图像的扰动尽量不传染其他神经元节点，通过观测该过程中网络特征输出变化的程度来测量网络通道的重要性，并以通道重要性替换基本CAM算法框架的系数。为了最小程度地减小对图像数据分布的破坏，对图像施加的扰动为描述SAR图像背景杂波特性的瑞利分布扰动，并迭代求解足以使网络决策发生改变的最小扰动。

为验证SCC-CAM的有效性，在VGG16, ResNet18网络结构和自建网络结构上进行了试验。在提取对网络决策起到关键作用的显著性区域的基础上，使用平均置信度下降率、基于面积约束的置信度下降率、置信度上升比例以及显著性区域分类性能等来评估算法的质量。试验结果表明，在所有的评估指标下，本文提出的算法具有更精确的定位显著区域的能力，相比于传统方法具有更强的理解性。由于本文方法在多个网络上均验证了有效性，因此可以作为一种通用的增强网络可理解性的技术方案。

利益冲突 所有作者均声明不存在利益冲突

Conflict of Interests The authors declare that there is no conflict of interests

参 考 文 献

[1] PANATI C, WAGNER S, and BRÜGGENWIRTH S. Feature relevance evaluation using grad-CAM, LIME and SHAP for deep learning SAR data classification[C]. 2022 23rd International Radar Symposium (IRS), Gdansk, Poland, 2022: 457–462. doi: 10.23919/irs54158.2022.9904989.

[2] SU Shenghan, CUI Ziteng, GUO Weiwei, *et al*. Explainable analysis of deep learning methods for SAR image classification[C]. IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022: 2570–2573. doi: 10.1109/igarss46834.2022.9883815.

[3] 李玮杰, 杨威, 刘永祥, 等. 雷达图像深度学习模型的可解释性研究与探索[J]. 中国科学: 信息科学, 2022, 52(6): 1114–1134. doi: 10.1360/SSI-2021-0102.

LI Weijie, YANG Wei, LIU Yongxiang, *et al*. Research and exploration on the interpretability of deep learning model in radar image[J]. *Scientia Sinica Informationis*, 2022, 52(6): 1114–1134. doi: 10.1360/SSI-2021-0102.

[4] 金亚秋. 多模式遥感智能信息与目标识别: 微波视觉的物理智能[J]. 雷达学报, 2019, 8(6): 710–716. doi: 10.12000/JR19083.

JIN Yaqiu. Multimode remote sensing intelligent information and target recognition: Physical intelligence of microwave vision[J]. *Journal of Radars*, 2019, 8(6): 710–716. doi: 10.12000/JR19083.

[5] KEYDEL E R, LEE S W, and MOORE J T. MSTAR extended operating conditions: A tutorial[C]. SPIE 2757, Algorithms for Synthetic Aperture Radar Imagery III, Orlando, USA, 1996: 228–242. doi: 10.1117/12.242059.

[6] ZHAO Juanping, GUO Weiwei, ZHANG Zenghui, *et al*. A coupled convolutional neural network for small and densely clustered ship detection in SAR images[J]. *Science China Information Sciences*, 2019, 62(4): 42301. doi: 10.1007/s11432-017-9405-6.

[7] 杜兰, 王兆成, 王燕, 等. 复杂场景下单通道SAR目标检测及鉴别研究进展综述[J]. 雷达学报, 2020, 9(1): 34–54. doi: 10.12000/JR19104.

DU Lan, WANG Zhaocheng, WANG Yan, *et al*. Survey of research progress on target detection and discrimination of single-channel SAR images for complex scenes[J]. *Journal of Radars*, 2020, 9(1): 34–54. doi: 10.12000/JR19104.

[8] 徐丰, 王海鹏, 金亚秋. 深度学习在SAR目标识别与地物分类中的应用[J]. 雷达学报, 2017, 6(2): 136–148. doi: 10.12000/JR16130.

XU Feng, WANG Haipeng, and JIN Yaqiu. Deep learning as applied in SAR target recognition and terrain classification[J]. *Journal of Radars*, 2017, 6(2): 136–148. doi: 10.12000/JR16130.

[9] 郭炜炜, 张增辉, 郁文贤, 等. SAR图像目标识别的可解释性问题探讨[J]. 雷达学报, 2020, 9(3): 462–476. doi: 10.12000/JR20059.
GUO Weiwei, ZHANG Zenghui, YU Wenxian, *et al*. Perspective on explainable SAR target recognition[J]. *Journal of Radars*, 2020, 9(3): 462–476. doi: 10.12000/JR20059.

[10] FENG Sijia, JI Kefeng, WANG Fulai, *et al*. Electromagnetic scattering feature (ESF) module embedded network based on ASC model for robust and interpretable SAR ATR[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5235415. doi: 10.1109/tgrs.2022.3208333.

[11] 吴飞, 廖彬兵, 韩亚洪. 深度学习的可解释性[J]. 航空兵器, 2019, 26(1): 39–46. doi: 10.12132/issn.1673-5048.2018.0065.
WU Fei, LIAO Binbing, and HAN Yahong. Interpretability for deep learning[J]. *Aero Weaponry*, 2019, 26(1): 39–46. doi: 10.12132/issn.1673-5048.2018.0065.

[12] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071–2096. doi: 10.7544/issn1000-1239.2019.20190540.
JI Shouling, LI Jinfeng, DU Tianyu, *et al*. Survey on techniques, applications and security of machine learning interpretability[J]. *Journal of Computer Research and Development*, 2019, 56(10): 2071–2096. doi: 10.7544/issn1000-1239.2019.20190540.

[13] DHURANDHAR A, CHEN Pinyu, LUSS R, *et al*. Explanations based on the missing: Towards contrastive explanations with pertinent negatives[C]. 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 590–601.

[14] DATCU M, ANDREI V, DUMITRU C O, *et al*. Explainable deep learning for SAR data[C]. Φ-week, Frascati, Italy, 2019.

[15] LI Yang, WANG Jiabao, XU Yulong, *et al*. DeepSAR-Net: Deep convolutional neural networks for SAR target recognition[C]. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 2017: 740–743. doi: 10.1109/icbda.2017.8078734.

[16] ZHAO Juanping, DATCU M, ZHANG Zenghui, *et al*. Contrastive-regulated CNN in the complex domain: A method to learn physical scattering signatures from flexible PolSAR images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(12): 10116–10135. doi: 10.1109/tgrs.2019.2931620.

[17] HUANG Zhongling, DATCU M, PAN Zongxu, *et al*. Deep SAR-Net: Learning objects from signals[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 161: 179–193. doi: 10.1016/j.isprsjprs.2020.01.016.

[18] HUANG Zhongling, DUMITRU C O, and REN Jun. Physics-aware feature learning of SAR images with deep neural networks: A case study[C]. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 2021: 1264–1267. doi: 10.1109/igarss47720.2021.9554842.

[19] HUANG Zhongling, YAO Xiwen, LIU Ying, *et al*. Physically explainable CNN for SAR image classification[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 190: 25–37. doi: 10.1016/j.isprsjprs.2022.05.008.

[20] LI Yi, DU Lan, and WEI Di. Multiscale CNN based on component analysis for SAR ATR[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5211212. doi: 10.1109/tgrs.2021.3100137.

[21] ZEILER M D and FERGUS R. Visualizing and understanding convolutional networks[C]. 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 818–833. doi: 10.1007/978-3-319-10590-1_53.

[22] ZHOU Bolei, KHOSLA A, LAPEDRIZA A, *et al*. Learning deep features for discriminative localization[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 2921–2929. doi: 10.1109/cvpr.2016.319.

[23] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. IEEE International Conference on Computer Vision, Venice, Italy, 2017: 618–626. doi: 10.1109/iccv.2017.74.

[24] CHATTOPADHAY A, SARKAR A, HOWLADER P, *et al*. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks[C]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, USA, 2018: 839–847. doi: 10.1109/wacv.2018.00097.

[25] WANG Haofan, WANG Zifan, DU Mengnan, *et al*. Score-CAM: Score-weighted visual explanations for convolutional neural networks[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, USA, 2020: 111–119. doi: 10.1109/cvprw50498.2020.00020.

[26] FENG Zhenpeng, ZHU Mingzhe, STANKOVIĆ L, *et al*. Self-matching CAM: A novel accurate visual explanation of CNNs for SAR image interpretation[J]. *Remote Sensing*, 2021, 13(9): 1772. doi: 10.3390/rs13091772.

[27] SUNDARARAJAN M, TALY A, and YAN Qiqi. Axiomatic attribution for deep networks[C]. 34th International Conference on Machine Learning, Sydney, Australia, 2017: 3319–3328.

[28] MONTAVON G, SAMEK W, and MÜLLER K R. Methods

for interpreting and understanding deep neural networks[J]. *Digital Signal Processing*, 2018, 73: 1–15. doi: 10.1016/j.dsp.2017.10.011.

[29] 匡纲要, 高贵, 蒋咏梅, 等. 合成孔径雷达: 目标检测理论、算法及应用[M]. 长沙: 国防科技大学出版社, 2007: 45–50.
KUANG Gangyao, GAO Gui, JIANG Yongmei, *et al*. Synthetic Aperture Radar Target: Detection Theory Algorithms and Applications[M]. Changsha: National University of Defense Technology Press, 2007: 45–50.

[30] ANASTASSOPOULOS, LAMPROPOULOS G A, DROSOPOULOS A, *et al*. High resolution radar clutter statistics[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 1999, 35(1): 43–60. doi: 10.1109/7.745679.

[31] KURUOGLU E E and ZERUBIA J. Modeling SAR images with a generalization of the Rayleigh distribution[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 527–533. doi: 10.1109/TIP.2003.818017.

[32] BELLONI C, BALLERI A, AOUF N, *et al*. Explainability of deep SAR ATR through feature analysis[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2021, 57(1): 659–673. doi: 10.1109/taes.2020.3031435.

[33] RICE J A. Mathematical Statistics and Data Analysis[M]. 3rd ed. Belmont: Cengage Learning, 2006: 71–99.

[34] BERTSEKAS D P and TSITSIKLIS J N. Introduction to Probability[M]. Cambridge: Massachusetts Institute of Technology, 2000: 6–48.

[35] SIMONYAN K, VEDALDI A, and ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps[C]. 2nd International Conference on Learning Representations, Banff, Canada, 2014.

[36] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[C]. IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1026–1034. doi: 10.1109/iccv.2015.123.

[37] FONG R C and VEDALDI A. Interpretable explanations of black boxes by meaningful perturbation[C]. IEEE International Conference on Computer Vision, Venice, Italy, 2017: 3449–3457. doi: 10.1109/ICCV.2017.371.

[38] ANCONA M, OZTIRELI C, and GROSS M. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation[C]. In International Conference on Machine Learning. PMLR, 2019: 272–281. doi: 10.48550/arXiv.1903.10992.

[39] DIEMUNSCH J R and WISSINGER J. Moving and stationary target acquisition and recognition (MSTAR) model-based automatic target recognition: Search technology for a robust ATR[C]. SPIE 3370, Algorithms for synthetic aperture radar Imagery V, Orlando, USA, 1998: 481–492. doi: 10.1117/12.321851.

[40] HUANG Lanqing, LIU Bin, LI Boying, *et al*. OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, 11(1): 195–208. doi: 10.1109/jstars.2017.2755672.

[41] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[42] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/cvpr.2016.90.

[43] HEILIGERS M and HUIZING A. On the importance of visual explanation and segmentation for SAR ATR using deep learning[C]. 2018 IEEE Radar Conference (RadarConf18), Oklahoma City, USA, 2018: 394–399. doi: 10.1109/radar.2018.8378591.

[44] DEVRIES T and TAYLOR G W. Learning confidence for out-of-distribution detection in neural networks[EB/OL]. https://arxiv.org/abs/1802.04865, 2018.

[45] LI Weijie, YANG Wei, LIU Li, *et al*. Discovering and explaining the noncausality of deep learning in SAR ATR[J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 4004605. doi: 10.1109/lgrs.2023.3266493.

作 者 简 介

崔宗勇，博士，副教授，研究方向为SAR图像处理、目标识别、深度学习等。

杨致远，硕士生，研究方向为SAR目标的可解释性等。

蒋　阳，硕士生，研究方向为SAR目标分类、深度学习可解释性等。

曹宗杰，博士，教授，研究方向为SAR目标检测识别、图像处理、人工智能等。

杨建宇，博士，教授，博士生导师，研究方向为雷达前视成像、实孔径超分辨成像、双多基合成孔径雷达成像等。

(责任编辑：高山流水)