面向SAR图像目标分类的CNN模型可视化方法

李妙歌 陈 渤* 王东升 刘宏伟

(西安电子科技大学雷达信号处理全国重点实验室 西安 710071)

摘要： 卷积神经网络(CNN)在合成孔径雷达(SAR)图像目标分类任务中应用广泛。由于网络工作机理不透明，CNN模型难以满足高可靠性实际应用的要求。类激活映射方法常用于可视化CNN模型的决策区域，但现有方法主要基于通道级或空间级类激活权重，且在SAR图像数据集上的应用仍处于起步阶段。基于此，该文从神经元特征提取能力和网络决策依据两个层面出发，提出了一种面向SAR图像的CNN模型可视化方法。首先，基于神经元的激活值，对神经元在其感受野范围内的目标结构学习能力进行可视化，然后提出一种通道-空间混合的类激活映射方法，通过对SAR图像中的重要区域进行定位，为模型的决策过程提供依据。实验结果表明，该方法给出了模型在不同设置下的可解释性分析，有效拓展了卷积神经网络在SAR图像上的可视化应用。

关键词： 合成孔径雷达；可视化分析；卷积神经网络；类激活映射；神经元

中图分类号： TN957.51 **文献标识码：** A **文章编号：** 2095-283X(2024)02-0359-15

DOI: 10.12000/JR23107

引用格式： 李妙歌, 陈渤, 王东升, 等. 面向SAR图像目标分类的CNN模型可视化方法[J]. 雷达学报(中英文), 2024, 13(2): 359–373. doi: 10.12000/JR23107.

Reference format: LI Miaoge, CHEN Bo, WANG Dongsheng, *et al*. CNN model visualization method for SAR image target classification[J]. *Journal of Radars*, 2024, 13(2): 359–373. doi: 10.12000/JR23107.

CNN Model Visualization Method for SAR Image Target Classification

LI Miaoge CHEN Bo* WANG Dongsheng LIU Hongwei

(*National Key Laboratory of Radar Signal Processing*, *Xidian University*, *Xi'an* 710071, *China*)

Abstract: Convolutional Neural Network (CNN) is widely used for image target classifications in Synthetic Aperture Radar (SAR), but the lack of mechanism transparency prevents it from meeting the practical application requirements, such as high reliability and trustworthiness. The Class Activation Mapping (CAM) method is often used to visualize the decision region of the CNN model. However, existing methods are primarily based on either channel-level or space-level class activation weights, and their research progress is still in its infancy regarding more complex SAR image datasets. Based on this, this paper proposes a CNN model visualization method for SAR images, considering the feature extraction ability of neurons and their current network decisions. Initially, neuronal activation values are used to visualize the capability of neurons to learn a target structure in its corresponding receptive field. Further, a novel CAM-based method combined with channel-wise and spatial-wise weights is proposed, which can provide the foundation for the decision-making process of the trained CNN models by detecting the crucial areas in SAR images. Experimental results showed that this method provides interpretability analysis of the model under different settings and effectively expands the application of CNNs for SAR image visualization.

收稿日期：2023-06-16；改回日期：2023-09-24；网络出版：2023-10-20

*通信作者： 陈渤 bchen@mail.xidian.edu.cn *Corresponding Author: CHEN Bo, bchen@mail.xidian.edu.cn

基金项目：国家自然科学基金(U21B2006)，陕西省青年创新团队项目，中央高校基本科研业务费专项资金(QTZX23037, QTZX22160)，“111”计划(B18039)

Foundation Items: The National Natural Science Foundation of China (U21B2006), Shaanxi Youth Innovation Team Project, The Fundamental Research Funds for the Central Universities (QTZX23037, QTZX22160), The 111 Project (B18039)

责任主编：计科峰 Corresponding Editor: JI Kefeng

Key words: Synthetic Aperture Radar (SAR); Visualization; Convolutional Neural Network (CNN); Class Activation Mapping (CAM); Neurons

1 引言

合成孔径雷达(Synthetic Aperture Radar, SAR)作为一种微波成像雷达具备全天候全天时、高分辨率、大范围观测成像能力，不论在国防军事还是民用经济方面都发挥着重要的作用[1-4]。SAR图像目标识别作为一种关键的SAR图像智能解译技术受到了广泛的关注。近年来，随着深度学习领域的技术突破与蓬勃发展，越来越多的深度网络模型被用于实现SAR图像目标的检测与识别[5–8]。研究表明，以卷积神经网络(Convolutional Neural Network, CNN)为代表的深度神经网络模型能够有效超越传统的SAR图像目标检测与识别方法[9–11]。然而，深度网络模型因为缺乏可解释性和网络模型决策不透明性等缺陷，在实际SAR目标检测和识别任务中存在着一定的应用风险性和信任危机。一方面，SAR图像成像机理复杂且易受目标结构、入射角、材质、极化方式等特性的影响，人类视觉系统对其具有一定的认知局限性[12–14]。另一方面，深度神经网络模型的学习能力虽然很强，但其“黑盒”特性导致预测不确定性，在某些场景中有一定的脆弱性。目前对于深度模型的性能评估分析主要依赖于识别率，而国防军事、医疗诊断等实际应用领域要求模型所得预测结果应具备较高的可信度，因此仅依靠识别率这一指标可能会带来模型对问题描述不充分、模型出现错误等问题[15]。

深度网络模型的可解释性研究大致分为自解释(Self-explanatory)与事后解释(Post-hoc)两类。自解释通过构建决策树、线性回归核朴素贝叶斯等结构简单、可解释性强的模型或将物理、语义等领域知识的可解释性内置于具体的模型结构中。目前，一些最新研究进展将SAR图像的属性散射中心(Attributed Scattering Center, ASC)模型融入网络的特征学习过程中引导其学习含有物理意义的特征表示[16–18]。其中，Zhang等人[19]将ASC参数化表示并通过词向量转化为特征的方式进行了ASC模型与CNN特征图的融合学习；Feng等人[20]采用分部卷积与双向卷积循环神经网络进行ASC模型部分组件的学习，进而融合SAR图像全局特征强化模型对于目标的理解；Li等人[21]则提出了一种基于分量分析的多尺度识别网络。事后解释则主要聚焦于完成训练的深度网络模型，通过建立可视化等可解释方法进行模型的决策行为分析。CNN模型的可视化技术能够直观展示网络从数据中自主挖掘的神经元特征与决策特征，而其中蕴含的知识又可以启发人们对SAR图像的解译工作，从而协助模型可信度和可解释性的进一步提升，指导模型的评估、改进与更新[22]。针对SAR图像卷积识别网络的可视化，德国宇航中心Datch等人[23]对可解释SAR数据深度学习做出了初步研究。Su等人[24]将ResNet-101应用于OpenSARUrban数据集上，并用8种不同的可视化方法对网络预测进行解释。郭炜炜等人[25]就SAR图像目标识别网络的可解释性从模型理解、诊断与改进等方面进行了初步探讨。Belloni等人[26]基于可视化方法针对一组处于相同环境因素下的输入图像分析了CNN对于特定类别目标分类时的关键特征位置。Panati等人[27]使用局部可解释模型无关解释方法(Local Interpretable Model-agnostic Explanations, LIME)、沙普利加和解释(SHapley Additive exPlanations, SHAP)等可视化方法证实了7层CNN网络进行SAR图像识别时使用的是目标特征信息而非周围的杂波信息。然而上述方法面临着效率低与解释不稳定等缺陷。基于归因的积分梯度(integrated gradients)方法[28]一种属于事后可视化方法(Post-hoc Explanations)，利用输入图像在一幅基准图像上的相对梯度信息反映特征重要性。该方法虽然具有高分辨率但是其可视化结果中存在视觉可见噪声，且不具备类区分性。类激活映射(Class Activation Mapping, CAM)方法作为另一种典型的事后可视化方法，具有使用简单、通用性强和类判别性等优势。它以生成热力图的方式对CNN模型的决策区域进行可视化。其核心思想为针对网络特定层输出的特征图计算分配一组权重，通过权重衡量特征图对于网络最终决策的重要性程度，在此基础上对所有特征图进行加权求和生成热力图，最后与输入数据进行线性叠加完成模型的事后可视化。但不同于符合人类认知机理的自然光学图像，直接在SAR图像上使用此类方法往往会忽略SAR图像内部潜在的物理特性。

针对上述SAR图像解译难点与当前主流的卷积神经网络事后可视化技术，本文首先从神经元层面出发，提出了基于最大激活值的CNN模型神经元可视化方法，并在实验中载入已完成训练的网络模型展示神经元对输入SAR图像的特征提取效果。同时，受传统类激活映射方法启发提出一种新的类激活映射方法，结合神经元可视化模块，提出面向SAR图像的CNN模型可视化方法。该方法以离线

的形式从神经元、同异类目标、模型识别准确率、模型初始化等多个角度对CNN模型展开分析。实验结果显示所提面向SAR图像的CNN模型可视化方法能够清晰可视化模型的神经元特征识别重点与核心决策区域，进一步增强了SAR图像目标识别模型的解释性和鲁棒性。

2　面向SAR图像的CNN模型可视化方法

所提面向SAR图像的CNN模型可视化方法如图1所示，主要由两个模块所组成：神经元可视化模块和类激活映射模块。给定输入SAR图像，模块1可视化展示了感兴趣神经元感受野内的目标区域，直观表征了CNN模型中神经元对目标特征的捕捉能力；模块2以生成类激活热力图的形式展示了输入图像对CNN模型当前决策影响最大的区域，同时还探究对比了相似类别与非相似类别下模型的决策重要性区域的响应差异。该方法以离线的形式在不修改网络结构及参数的前提下通过两个模块实现了目标特征识别重点与输出决策的可视化。

2.1　基于最大激活值的神经元可视化方法

CNN中的感受野(Receptive Field)表征了网络内部神经元对原始输入数据的感受范围大小。神经元之所以无法感知输入数据中的所有信息，是因为网络结构中交替使用了卷积层和池化层，使得层与层之间为局部相连，每个神经元只能观测输入数据的部分区域。随着网络层数的递增，神经元的感受野越大、所接触到的输入数据范围也就越大，这意味着神经元可能捕捉到更为全局、语义层次更高的特征。相比之下，浅层神经元的感受野范围较小，所包含的特征更加趋向于局部与细节特征。因此借助感受野大小，可以大致判断每一层的抽象层次。

基于最大激活值进行神经元的可视化时，首先需要定位网络中待观测神经元的最大激活值位置，设网络第l层中第k个神经元对应特征图$\boldsymbol{h}_k^{(l)}(x)$的最大激活值为$x_k^{*(l)}$，则有

$$x_k^{*(l)} = \underset{x}{\operatorname{argmax}}\left(\boldsymbol{h}_k^{(l)}(x)\right) \tag{1}$$

然后将$x_k^{*(l)}$的坐标标记为起始点并对神经元的感受野参数进行计算。具体地，CNN第l层待观测最大激活值神经元的感受野大小可由相邻浅层$l-1$层的感受野大小$\mathrm{RF}^{(l-1)}$和网络参数推导得到：

$$\mathrm{RF}^{(l)} = \mathrm{RF}^{(l-1)} \times k^{(l)} - \left(\mathrm{RF}^{(l-1)} - \prod_{i=0}^{l-1} s^{(i)}\right) \times \left(k^{(l)} - 1\right)，即：$$

$$\mathrm{RF}^{(l)} = \mathrm{RF}^{(l-1)} + \left[\left(k^{(l)} - 1\right) \times \prod_{i=0}^{l-1} s^{(i)}\right] \tag{2}$$

其中，$k^{(l)}$为第l层待观测神经元的卷积核大小，$s^{(i)}$为第l层待观测神经元的步长大小。获得感受野大小后还需求解待观测神经元在输入图像上的感受野中心坐标：

$$\mathrm{RF}_x^{(l)} = \mathrm{RF}_x^{(l-1)} + \left(\frac{k^{(l)}-1}{2} - p^{(l)}\right) \times \prod_{i=0}^{l-1} s^{(i)}$$

$$\mathrm{RF}_y^{(l)} = \mathrm{RF}_y^{(l-1)} + \left(\frac{k^{(l)}-1}{2} - p^{(l)}\right) \times \prod_{i=0}^{l-1} s^{(i)} \tag{3}$$

设输入二维图像有x, y两个轴，式(3)中$\mathrm{RF}_x^{(l)}$和$\mathrm{RF}_y^{(l)}$分别表示感受野中心在x轴和y轴上的坐标，$p^{(l)}$表示第l层的padding大小。最后结合感受野大小与中心坐标参数，自起始点出发由深层至浅层逐层递推即可展示待观测神经元关注的目标区域及捕捉到的目标特征，完成神经元的可视化。Luo等人[29]研究表明上述理论感受野往往会小于实际感受野，且实际感受野往往受到网络训练结果、卷积层数等影响。

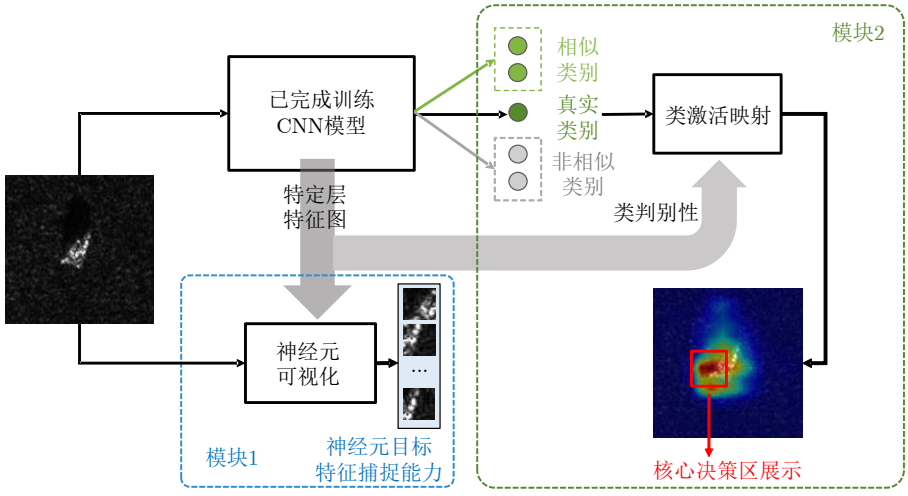


图 1 面向SAR图像的CNN模型可视化方法框图

Fig. 1 CNN model visualization method for SAR images

因此实际有效感受野难以准确计算。此处，本文将直接采用理论感受野方法进行神经元可视化计算。

2.2 CS-CAM (Channel-wise and Spatial-wise weighted Class Activation Mapping)可视化方法

类激活映射方法的原型(Class Activation Mapping, CAM)最早由Zhou等人[30]提出，CAM认为CNN随着层数的递增，神经元特征图内与决策不相关的信息会越来越少，因此高层所提取的目标信息在更加抽象的同时蕴含的语义信息也会更加丰富，其中最高层卷积层输出的特征图含有最抽象的目标级语义信息，且每个神经元激活响应的目标位置也有所区别。因此，CAM首先用全局平均池化层来取代原始网络中除softmax层外的所有全连接层，将每个通道的特征图压缩为一个数，进而将所有通道特征图数据压缩为一个包含全局信息的特征向量。其次，重新训练新的网络参数用以学习原网络最高层卷积层输出特征图对特定类别的类激活权重，最后基于该权重对特征图进行加权求和得到类激活热力图以高亮的形式指示输入数据中与指定类别最相关的区域。然而，CAM在实际应用中受限于特定的CNN结构，同时网络结构的修改与重新训练将大大增加时间开销，因而在此基础上计算机视觉领域又相继衍生出了一系列适用于所有CNN网络、通用性更强的类激活图可视化方法[31–36]。

如表1所示，目前基于CAM的方法间的核心区别在于如何计算目标层特征图的类激活权重，设第i层卷积层的特征图个数为K_i，表1中$\boldsymbol{h}_j^{(L)}(x,y)$为CNN最高层卷积层(第$L$层)的第$j(j=1,2,\cdots,K_L)$个特征图，$(x,y)$为特征图中的空间坐标位置，针对特定类别$c$，$S^c$为网络输出的类别分数，$S_j^c(j=1,2,\cdots,K_L)$为将特征图依次置零后经前向传播计算所得的输出类别分数，ΔS_j^c为第j个特征图对输入图像进行掩码前后在类别c上网络预测分数的变化，$a_{j_\text{channel}}^c$为第j个特征图的通道级权重值，$a_j^c(x,y)_{_\text{spatial}}$为第$j$个特征图在$(x,y)$位置处的空间级权重值。$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$，$\boldsymbol{U}$和$\boldsymbol{V}$均为酉矩阵且$\boldsymbol{U}$中每个特征向量称为$\boldsymbol{X}$的左奇异向量，$\boldsymbol{V}$中每个特征向量称为$\boldsymbol{X}$的右奇异向量，$\boldsymbol{\Sigma}$为奇异值矩阵，$\boldsymbol{v}_1$为$\boldsymbol{V}$的第1个特征向量。其中Grad-CAM[31], Grad-CAM++[32]基于网络当前预测值的回传梯度；Ablation-CAM[33]基于特征图的消融；Score-CAM[34]基于模型对于特征图的全局置信度分数；Eigen-CAM[35], Eigengrad-CAM基于SVD分解分别针对每一个特征图计算了通道级类激活权重，而Layer-CAM[36]则计算了像素空间级的类激活权重。笔者同时从特征图的通道级权重与像素空间级权重出发，提出了通道-空间混合类激活权重的类激活映射方法(Channel-wise and Spatial-wise weighted Class Activation Mapping, CS-CAM)。其流程图如图2所示。

图2中，$\odot$表示哈达马积，$\otimes$表示加权线性和，两个全连接层完全相同。CS-CAM首先将选定第i层内的每个特征图上采样到输入图像大小，并与输入图像对应元素相乘得到掩码输入：

$$\boldsymbol{M}_j^{(i)}=\mathrm{Up}\left(\boldsymbol{h}_j^{(i)}\right)\odot\boldsymbol{I} \tag{4}$$

表 1 不同类激活映射方法类激活权重计算表

Tab. 1 Class activation weights for different CAM-based methods

类激活方法名称	类激活权重计算公式
Grad-CAM[31]	$a_{j\_\text{channel}}^c=\dfrac{1}{Z}\sum\limits_x\sum\limits_y\dfrac{\partial S^c}{\partial\boldsymbol{h}_j^{(L)}(x,y)}$
Grad-CAM++[32]	$a_{j\_\text{channel}}^c=\sum\limits_x\sum\limits_y\dfrac{\dfrac{\partial^2 S^c}{\left(\partial\boldsymbol{h}_j^{(L)}(x,y)\right)^2}}{2\dfrac{\partial^2 S^c}{\left(\partial\boldsymbol{h}_j^{(L)}(x,y)\right)^2}+\sum\limits_x\sum\limits_y\boldsymbol{h}_j^{(L)}(x,y)\left\{\dfrac{\partial^3 S^c}{\left(\partial\boldsymbol{h}_j^{(L)}(x,y)\right)^3}\right\}}\cdot\mathrm{ReLU}\left(\dfrac{\partial S^c}{\partial\boldsymbol{h}_j^{(L)}(x,y)}\right)$
Ablation-CAM[33]	$a_{j\_\text{channel}}^c=\dfrac{S^c-S_j^c}{S^c}$
Score-CAM[34]	$a_{j\_\text{channel}}^c=\dfrac{\exp\left(\Delta S_j^c\right)}{\sum\limits_j\exp\left(\Delta S_j^c\right)}$
Eigen-CAM[35]	$\boldsymbol{h}^{(L)}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}},a_{j\_\text{channel}}^c=\boldsymbol{v}_1$
Eigengrad-CAM	$\boldsymbol{h}^{(L)}\cdot\dfrac{1}{Z}\sum\limits_x\sum\limits_y\dfrac{\partial S^c}{\partial\boldsymbol{h}_j^{(L)}(x,y)}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}},a_{j\_\text{channel}}^c=\boldsymbol{v}_1$
Layer-CAM[36]	$a_j^c(x,y)_{\_\text{spatial}}=\dfrac{\partial S^c}{\partial\boldsymbol{h}_j^{(L)}(x,y)}$

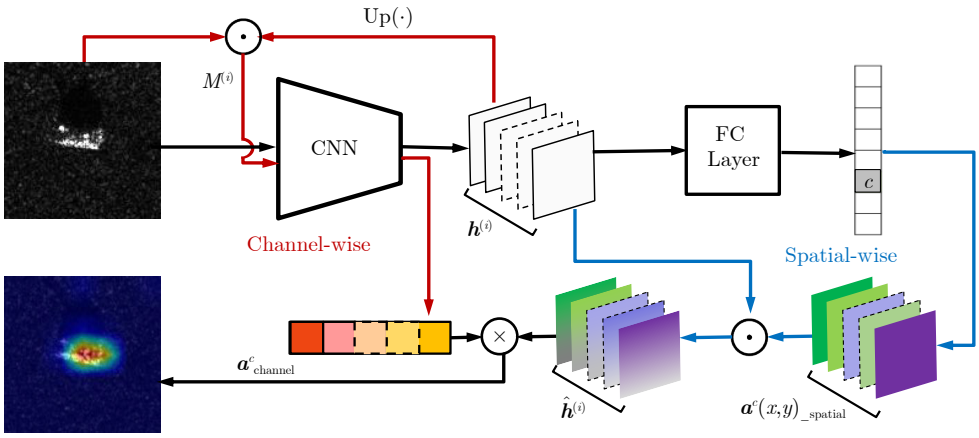


图 2 CS-CAM算法流程图

Fig. 2 Pipeline of CS-CAM

式(4)中，Up(·)表示上采样函数，$\boldsymbol{M}_j^{(i)}$表示经第j个特征图掩码后的输入，$\boldsymbol{I}$表示输入SAR图像，然后计算掩码前后网络在类别c上预测分数的变化ΔS_j^c并做归一化处理得到特征图的通道级类激活权重$a_{j_\text{channel}}^c$：

$$\Delta S_j^c = f\left(\boldsymbol{M}_j^{(i)}\right) - f(\boldsymbol{I}) \tag{5}$$

$$a_{j\_\text{channel}}^c = \frac{\exp\left(\Delta S_j^c\right)}{\sum_j \exp\left(\Delta S_j^c\right)} \tag{6}$$

式(5)中，$f(\cdot)$表示所用的CNN模型。此外，基于梯度信息计算特征图的空间级类激活权重$a_j^c(x,y)_{_\text{spatial}}$：

$$a_j^c(x,y)_{\_\text{spatial}} = \text{ReLU}\left(\frac{\partial S_j^c}{\partial \boldsymbol{h}_j^{(i)}(x,y)}\right) \tag{7}$$

最后，为经空间位置加权后的特征图$\hat{\boldsymbol{h}}_j^{(i)}(x,y)$施加通道级权重并通过ReLU(·)保留其中对网络预测起正向作用的部分，即得到CS-CAM的类激活热力图表达式：

$$\hat{\boldsymbol{h}}_j^{(i)}(x,y) = a_j^c(x,y)_{\_\text{spatial}} \times \boldsymbol{h}_j^{(i)}(x,y) \tag{8}$$

$$L_{\text{CS-CAM}}^c = \text{ReLU}\left(\sum_j a_{j\_\text{channel}}^c \times \hat{\boldsymbol{h}}_j^{(i)}(x,y)\right) \tag{9}$$

观察式(8)和式(9)可知，所提方法同时考虑了特征图的空间级和通道级类激活权重，前者根据当前决策为特征图内每个空间位置的像素给予了一个类激活权重，有助于网络可视化过程中(尤其对于低层)梯度回传所带来的背景噪声的缓解与消除，而后者不依赖于梯度信息，通过网络预测分数的变化情况获得不同通道下特征图的权重，将感兴趣网络层内与当前类别相关性更强的特征图进一步增强的同时抑制了弱相关通道的特征图的影响。

3　实验结果分析

3.1　实验数据与实验设置

本文采用MSTAR数据集[37]进行实验，该数据集为当前SAR目标识别方法测试与评价的代表数据集之一。它包含了10类军事目标车辆，各类目标的SAR图像由X波段的机载雷达采集得到，距离分辨率为0.3 m。根据俯仰角可将采集数据划分为训练集与测试集，其中17°俯仰角数据为训练集，15°俯仰角数据为测试集。具体数据分布如表2所示，其中训练集样本共计2746个，测试集样本共计2426个。

实验采用VGG-16[38]进行SAR图像识别，识别率如表3所示。其中，RI (Random Initialization)表示随机初始化网络参数后直接进行训练，PT (Pre-Trained)表示载入基于光学数据集ImageNet的预训练模型参数作为初始化，随后不冻结网络各层参数(除根据输入类别数调整模型全连接层参数)进行微调。

表 2　MSTAR数据集

Tab. 2　MSTAR dataset

目标名称	目标类型	训练集(17°俯仰角)	测试集(15°俯仰角)
2S1	自行榴弹炮	299	274
BMP2	步兵战车	232	196
BRDM2	装甲侦察车	298	274
BTR60	装甲侦察车	233	196
BTR70	装甲侦察车	256	195
T62	坦克	299	273
T72	坦克	232	196
ZIL131	军用卡车	299	274
ZSU234	自行高炮	299	274
D7	推土机	299	274

表 3 MSTAR数据集模型训练结果(%)
Tab. 3 Model training results (%)

网络模型	训练集识别率	测试集识别率
VGG-16 (RI)	98.7	94.1
VGG-16 (PT)	99.7	98.0

3.2 神经元可视化分析

针对已完成训练的VGG-16 (PT)模型，输入一张SAR图像对网络中第6层至第22层(间隔4层取值)部分神经元进行可视化展示。图3中红框表示对应神经元的感受野范围，图3(a)和图3(b)各图左下角的子图为感受野范围内目标切片放大结果。分析可知，低层神经元的感受野较小，神经元关注的是目标的部分结构特征，以图3(a)中第6层神经元为例，0号神经元捕捉的主要为T72坦克的末端边缘结构、而1号神经元则主要捕捉到了T72坦克的炮筒结构。对比图3(a)—图3(f)可知，随着网络深度的增加，高层的神经元感受野逐渐变大，从第22层开始，部分神经元(如第8号、第95号神经元)的感受野已经可以完整地观测到整个目标。网络完成分类任务时，不同神经元往往起着不同的作用，其对于输入图像的激活响应大小也具有差异，其中激活值较大的神经元往往对网络最后的分类决策影响更大。如图4所示，选取两张T72目标图像展示了网络中第12层内最大激活值TOP-9的神经元可视化结果。图4中神经元在两个输入SAR图像上均可以捕捉到一定的目标信息，如输入1中109号神经元和输入2中224号神经元关注于坦克的炮筒结构。同时，对于同一类目标的不同输入样本，网络中激活响应值TOP-9的神经元有重叠部分，如25, 54, 55, 68, 123, 190号神经元。此外，本文还进一步统计了网络第12层TOP-9的神经元在同、异类目标条件下的分布情况(考虑到测试集内目标总数的差异性，本文采用归一化后的神经元个数统计值)。图5内各子图的神经元分布存在局部峰值，说明同类目标的不同图像输入网络后处于高激活状态的关键神经元具有一致性，因此进行神经元可视化时应当重点关注具有强激活值的关键神经元。对比图5(a)与图5(b)可知，异类目标的关键神经元分布存在部分相同的神经元，说明网络提取到了目标共有的一些基本特征，同时异类目标也激活了各自独有的神经元，这使得最终提取的特征具有可分性，有助于CNN模型的正确识别。

3.3 类激活映射方法对比分析

类激活映射方法的选择对CNN模型的解释效果至关重要，选取合适的方法才能更好地对模型展

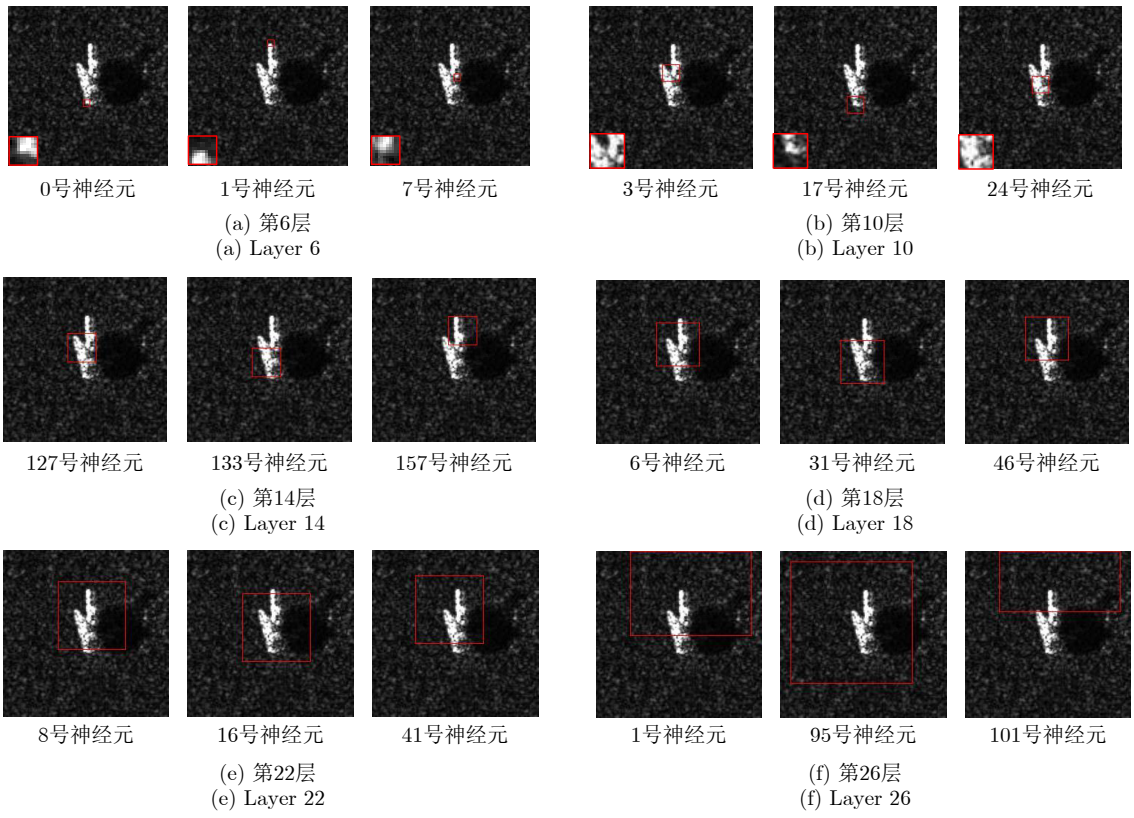


图 3 SAR图像神经元可视化结果图

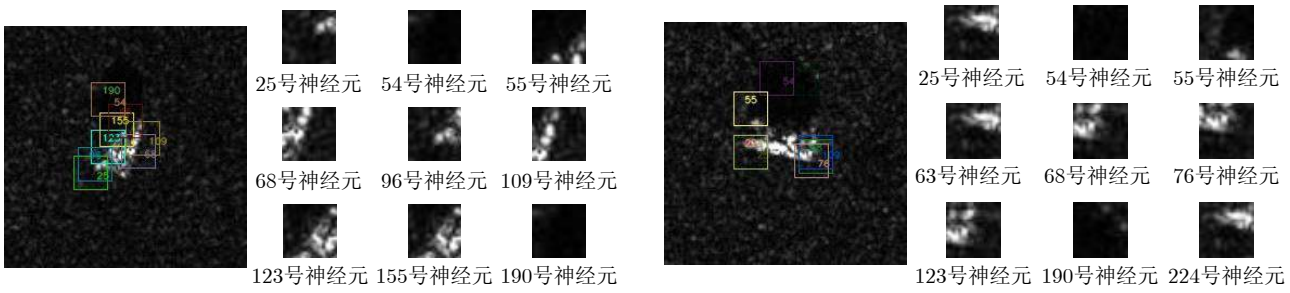
Fig. 3 Visualization of neurons at different network layers

开解释性更强、更合理的可视化研究。下面采用定性分析、定量分析和充分性评估3种形式针对表1中传统方法和所提CS-CAM方法的可视化效果进行比较。

3.3.1 类激活映射方法定性分析

如图6所示，展示了8种类激活映射方法针对目标真实类别所得的类激活热力图，热力图中偏红区域为高亮区，且颜色越红代表激活程度越高，反之，偏蓝区域表示低激活或未激活区。可以得出，虽然不同方法关注的核心决策区域有差异，但目标均处于热力图高亮区。基于自然光学图像的卷积识别网络可视化中，部分热力图中可能存在高亮区包含目标周围的弱相关背景区域的情况(如对于汽车这一目标类别，热力图高亮区可能包含周围的车道区域)。同理，CNN模型进行SAR目标识别时，模型可能利用背景信息进行识别辅助，具体表现为热力图中部分背景区域被激活表示。

理论上，类激活映射方法适用于CNN网络中的任意层，本文还实验对比了各方法在不同深度的卷积层下生成的热力图。首先如图7所示选取两张T72目标图像输入VGG-16 (PT)模型，基于网络中各阶段(Stage1—Stage5)内最后一层卷积层得到可视化结果如图8所示。

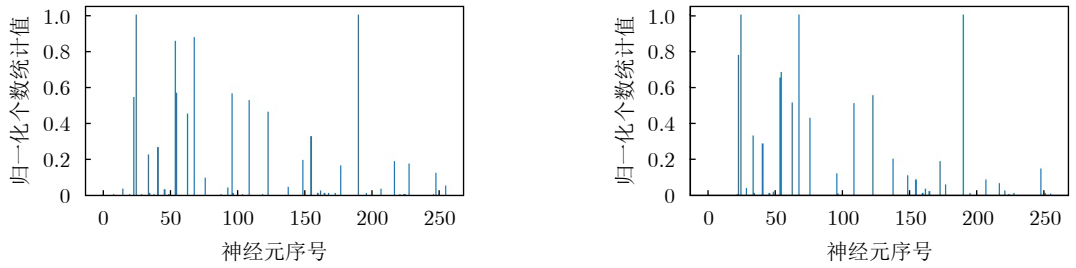


(a) 输入图像1神经元可视化
(a) Neurons visualization of input image 1

(b) 输入图像2神经元可视化
(b) Neurons visualization of input image 2

图 4 第12层TOP-9神经元可视化结果图

Fig. 4 Visualization of layer 12 TOP-9 neurons



(a) T72　　(b) ZSU234

图 5 第12层TOP-9神经元归一化个数统计图

Fig. 5 Layer 12 TOP-9 neuron normalization statistics chart

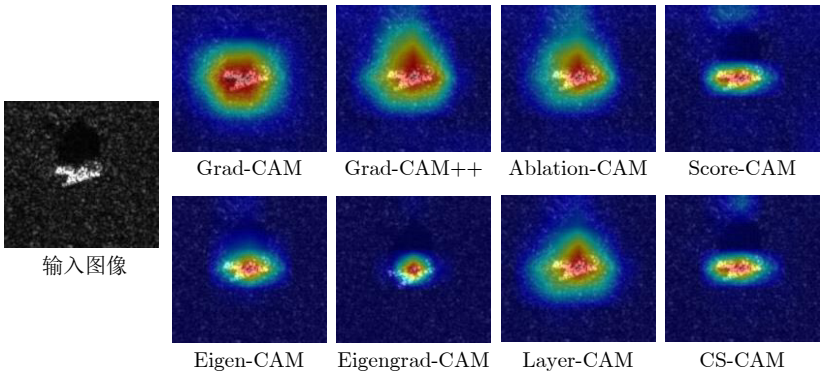


图 6 各类激活映射方法可视化结果图

Fig. 6 CAM-based methods' visualization of the same input image

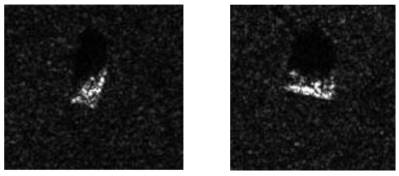


图 7 输入图像

Fig. 7 Input images

观察可知，浅层热力图主要高亮表示了目标更细粒度的细节且大多呈现点状分布形式，如图8(b)—图8(h)中Stage1捕捉了目标的部分结构，这与浅层对应的特征图有更大的空间分辨率密不可分。同时，Grad-CAM和Grad-CAM++生成的浅层热力图存在目标区域低响应、背景区域高亮的情况，如Grad-CAM基于Stage1, Stage2生成的热力图含有大量噪声，这一方面可能与梯度带噪相关，即梯度可能会由于激活函数中平坦的零梯度区域而趋于消失，使输出对应于输入或中间网络层激活的梯度在视觉上呈现含噪；另一方面，图9统计了网络各阶段最后一个卷积层内各通道特征图的方差变化，结果表明Stage5的特征图方差较小(多个通道趋于0)，表征了最高层网络层中用全局权重代表特征图中每个空间位置权重的合理性。而在浅层，由于方差较

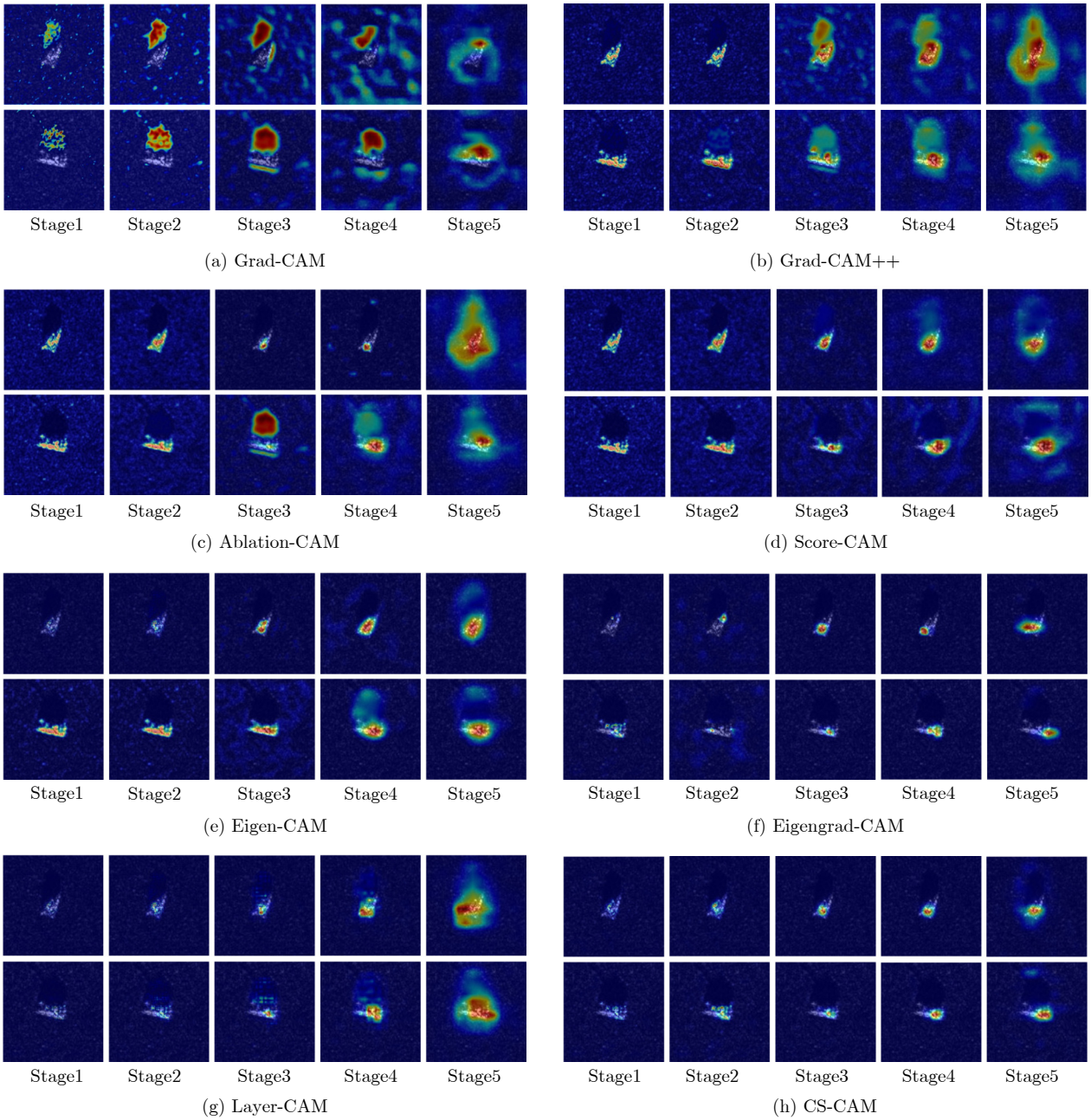


图 8 各类激活映射方法针对CNN不同层的可视化结果

Fig. 8 Visualization of CAM-based methods in each stage of CNN

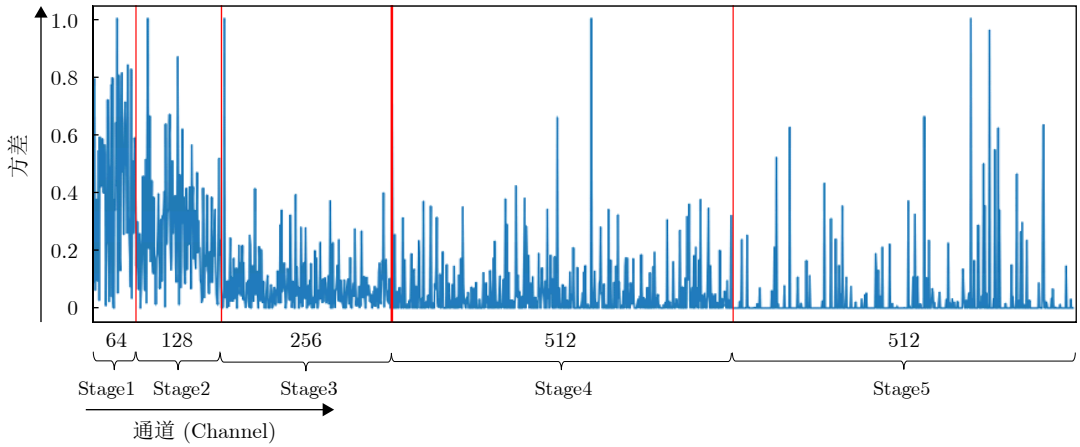


图 9 VGG-16网络各Stage最后一个卷积层内各通道特征图的方差变化图

Fig. 9 The variance statistics of the feature maps for each channel in the last convolutional layer of each Stage in VGG-16

大，采用全局权重替代不同空间位置权重可能使得Grad-CAM, Grad-CAM++和Score-CAM无法生成可信度较高的热力图。这进一步证明了CS-CAM与Layer-CAM由于采用了空间级的类激活权重计算方法，使其在浅层的可视化效果更优。观测网络中间层Stage3，热力图捕捉到的目标特征范围更广，相比其他方法，CS-CAM, Score-CAM, Eigen-CAM和Layer-CAM可以更好地反映网络学习到的目标整体轮廓特征。最后观测网络高层Stage4, Stage5，各方法所生成的热力图呈现为区域级形式，除了Grad-CAM外，其他方法均能较准确地对目标位置进行定位。因此在选用类激活映射方法进行CNN的可视化时，浅层网络层可采用效果较好的CS-CAM, Layer-CAM和Eigen-CAM，而对于其他类激活映射方法采用网络高层生成热力图会更加合理。

3.3.2 类激活映射方法定量分析

为更好地指导CNN模型可视化时类激活映射方法的选取，采用文献[31]提出的平均上升率(Average Increase)、平均下降率(Average Drop)和文献[39]提出的失真度(Infidelity)和最大灵敏度(Max-sensitivity)4个量化指标，以数值形式清晰地说明各算法在MSTAR数据集整体上的综合表现，具体定义如下。

(1) 平均上升率。

用$\boldsymbol{I}$表示输入SAR图像，$\boldsymbol{L}^c_{\text{CAM-based}}$表示所用类激活映射方法针对类别$c$所得的类激活热力图，将$\boldsymbol{L}^c_{\text{CAM-based}}$作为位置掩码信息对输入图像进行掩码操作所得的解释图(Explanation Map)$\boldsymbol{E}^c$为

$$\boldsymbol{E}^c = \boldsymbol{L}^c_{\text{CAM-based}} \odot \boldsymbol{I} \tag{10}$$

其中，$\odot$表示哈达马积。观察掩码前后图像$\boldsymbol{I}$和$\boldsymbol{E}^c$输入网络后在目标类别c上的置信度分数变化衡量所得热力图对模型的解释程度。Average Increase统计了$\boldsymbol{E}^c$输入网络后输出置信度分数有所提升的样本百分比：

$$\text{Average Increase} = \sum_{i=1}^{N} \text{sign}\left(0, O_i^c - S_i^c\right) \times \frac{100}{N} \tag{11}$$

式(11)中，S_i^c表示输入原始SAR图像网络输出的置信度分数，O_i^c表示输入解释图$\boldsymbol{E}^c$网络输出的置信度分数，N为样本总数，$\text{sign}(\cdot)$函数的表达式为

$$\text{sign}(x) = \begin{cases} 1, & x>0 \\ 0, & x \le 0 \end{cases} \tag{12}$$

Average Increase反映了热力图中高亮的特征对于网络预测结果的提升能力，该指标数值越大说明所选方法的解释性越好。

(2)平均下降率。

在大多数些情况下，O^c会小于S^c，当解释图$\boldsymbol{E}^c$遮挡了对网络分类结果贡献程度较小的区域时，能够保留输入图像中更多的关键特征，使置信度分数的降低率较低，相应计算公式为

$$\text{Average Drop} = \sum_{i=1}^{N} \frac{\max\left(0, S_i^c - O_i^c\right)}{S_i^c} \times \frac{100}{N} \tag{13}$$

指标Average Drop的数值越小说明所选的类激活映射方法解释性越好。

(3) 失真度。

Infidelity指标反映了可视化方法的准确性，它首先对输入图像引入扰动$\boldsymbol{P}$，然后通过计算引入扰动$\boldsymbol{P}$的热力图同引入扰动前后输出$f(\boldsymbol{I})$和$f(\boldsymbol{I}-\boldsymbol{P})$变化的均方误差来衡量类激活映射方法表征网络模型输出变化情况的能力，其计算公式为

$$\begin{aligned}&\text{Infidelity}(\boldsymbol{L}^c_{\text{CAM-based}}, f, \boldsymbol{I})\\&= \text{E}_{\boldsymbol{P}\sim\mu_P}\left[\left(\boldsymbol{P}^{\text{T}}\boldsymbol{L}^c_{\text{CAM-based}} - (f(\boldsymbol{I}) - f(\boldsymbol{I}-\boldsymbol{P}))\right)^2\right]\end{aligned} \tag{14}$$

式(14)中所施加的扰动$\boldsymbol{P}$服从分布μ_P，$f(\cdot)$表示所用的CNN模型。该指标的数值越小说明所选的类激活映射方法的准确性越好。

(4) 最大灵敏度。

Max-sensitivity评估了可视化方法的鲁棒性，该指标通过蒙特卡罗采样法计算引入领域半径为r的扰动后所得类激活热力图的最大变化，其计算公式为

$$\begin{aligned}&\text{Max-sensitivity}(\boldsymbol{L}^c_{\text{CAM-based}}, f, \boldsymbol{I}, r)\\&= \max_{\|\boldsymbol{I}'-\boldsymbol{I}\|\le r}\|\boldsymbol{L}^c_{\text{CAM-based}}(\boldsymbol{I}') - \boldsymbol{L}^c_{\text{CAM-based}}(\boldsymbol{I})\|\end{aligned} \tag{15}$$

该指标的数值越小说明所选的类激活映射方法的鲁棒性越好。

表4给出了各方法在定量分析指标上的结果，分析可知，CS-CAM在Average Increase和Average Drop两项评价指标上的综合性能优于其他方法，说明所提方法能够更有效地定位输入图像中对网络决策贡献值最高的目标区域。CS-CAM和Eigen-CAM在Infidelity指标上表现较好，说明两者能够更准确地反映CNN模型在输入数据受到扰动后产生的输出变化情况。CS-CAM, Ablation-CAM与Layer-CAM在Max-sensitivity指标上表现较好，说明这3种方法的鲁棒性更好。综合前面的定性分析结果与以上数值指标分析，所提CS-CAM方法具备较好的CNN模型可视化性能，生成的类激活热力图含噪较小且对模型的核心决策区域定位准确。

3.3.3 充分性评估

Sanity Check由Adebayo等人[40]提出，是一种解释方法充分性评估的测试，常用于验证类激活映射方法是否适用于当前任务。实验采用网络参数级联随机化(Cascading Randomization)与独立随机化(Independent Randomization)两种测试方式对所提CS-CAM可视化方法进行测试。其中，前者自顶向下，以级联的方式依次随机初始化CNN模型参数，探究可视化方法在原始网络模型与未经训练的相同架构网络模型上输出的热力图差异；后者则独立地对各层模型参数进行随机初始化，探究可视化方法是否对某一层参数具有依赖性。

若通过Sanity Check测试，则两种热力图会存在明显差异，证明该类激活方法对CNN模型的可视化依赖于模型学习的参数，忠于模型的当前任务，有助于模型有效性的验证与进一步调试。本文将VGG-16网络参数从Logit层出发级联或独立随机至Conv21层，CS-CAM测试结果如图10所示。可以看到，相比图10(a)中原始网络生成的热力图，图10(b)和图10(c)中可视化结果会随着网络参数的破坏而变差，说明CS-CAM对模型参数具有敏感性，能够反映模型质量，通过了Sanity Check。

表 4 定量分析指标结果表

Tab. 4 Quantitative analysis results

方法	评估指标			
	Average Increase (%)	Average Drop (%)	Infidelity	Max-sensitivity
Grad-CAM	13.6	86.4	2.425	0.752
Grad-CAM++	13.6	64.0	2.987	0.612
Ablation-CAM	4.5	62.1	0.698	0.447
Layer-CAM	13.6	63.1	0.609	0.450
Eigengrad-CAM	13.6	77.8	0.619	0.459
Eigen-CAM	13.6	60.3	0.452	0.452
Score-CAM	13.6	58.6	0.818	0.486
CS-CAM	13.6	57.3	0.537	0.442

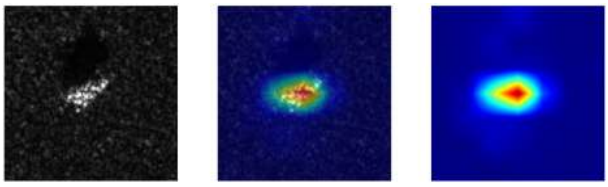
3.4 网络性能分析

3.4.1 模型准确率对比分析

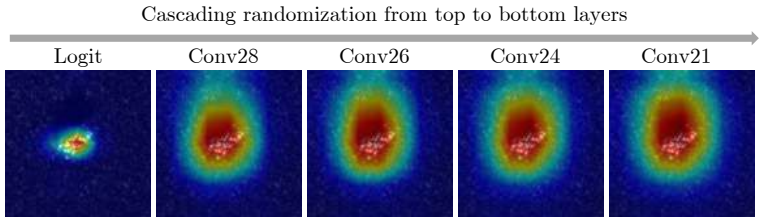
以VGG-16 (PT)为代表，在MSTAR数据集上训练3种准确率的模型参数，用于探究不同准确率下CNN模型的核心决策区的变化情况。如图11所示，从左至右对应网络准确率依次为18%, 85%, 98%。观察可得，CS-CAM生成的热力图的质量与模型准确率呈正相关关系，即随准确率的提升，热力图中高亮区域的面积以目标为中心缩小，说明网络对目标特征的学习能力在逐步提升，对目标位置的定位也更加准确。

3.4.2 初始化方式对比分析

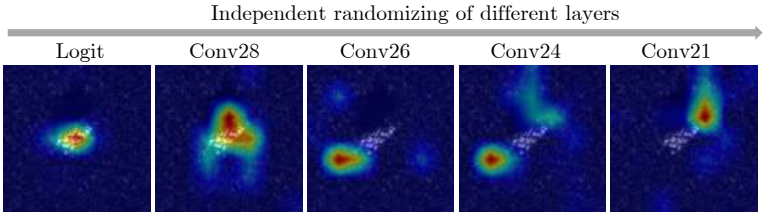
注意到表3中基于预训练权重微调的网络(PT)准确率相比随机初始化训练的网络(RI)准确率有一定提升。为作进一步探究，如图12所示采用CS-CAM分别可视化了加载随机权重(RI)、自然图像(ImageNet)和遥感图像(高分三号飞机)预训练权重3种初始化方式对于同一输入图像的热力图。从可视化结果可知，总体来说，本文所提方法对于3种初始化方式具有较强的鲁棒性，可视化结果具有一定的相似性。相比于随机初始化训练网络，两种预训练网络对于目标位置信息的捕捉能力更强且对于边缘背景的响应更小，该结果指示SAR图像的特征提取机制并不唯一，光学图像明确语义的纹理特征同样适用于如SAR图像等不同源数据[15,41]。此外，相比于随机初始化训练网络，基于遥感图像的预训练网络



(a) CS-CAM类激活热力图
(a) Heat maps of CS-CAM



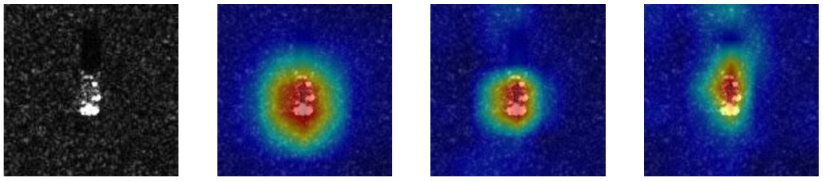
(b) 级联随机化
(b) Cascading randomization



(c) 独立随机化
(c) Independent randomization

图 10 CS-CAM Sanity Check结果图

Fig. 10 Sanity Check results of CS-CAM



(a) 输入图像
(a) Input image

(b) 准确率=18%
(b) Accuracy=18%

(c) 准确率=85%
(c) Accuracy=85%

(d) 准确率=98%
(d) Accuracy=98%

图 11 不同网络准确率下类激活热力图

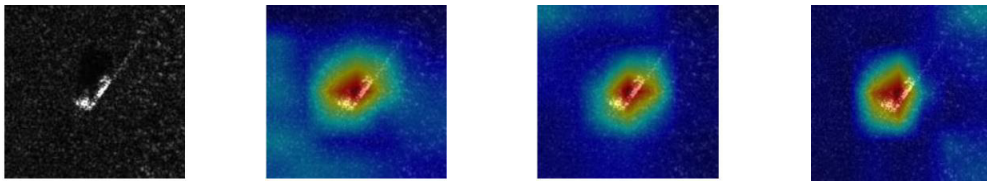
Fig. 11 Heat maps with different network accuracy

利用了遥感图片的先验信息，过滤了大多弱相关的背景噪声，有利于模型提取更准确的目标特征。

3.4.3 类判别性分析

本节利用类激活映射方法针对模型同类目标与异类目标的识别重点进行决策可视化分析。对于同类目标，输入5张T72类别的SAR图像，所得CS-CAM可视化结果如图13所示。可以看到，对于网络最终决策结果起到显著性影响的区域主要为SAR图像中的目标区域，而非背景区域，符合人类主观认识。

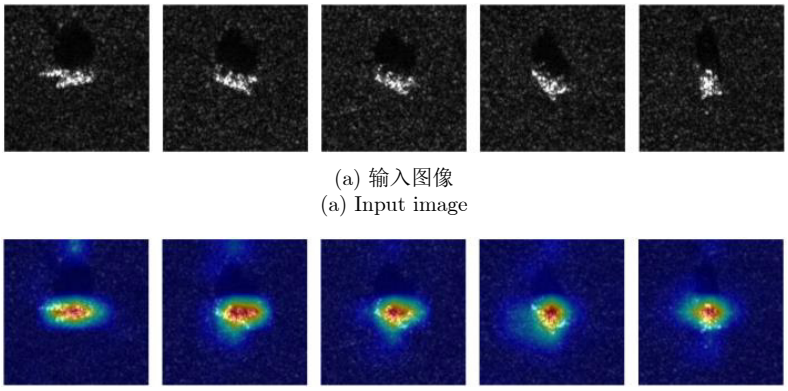
对于异类目标，可以分为相似类别与非相似类别两种情况，如图14所示，可视化分析了两种情况下CNN模型的决策差异性。图14中第1列为输入图像和CS-CAM针对真实类别所得的热力图，第2列为CS-CAM针对BRDM2和BTR70两个相似类别所得的热力图，第3列为针对D7和ZSU234两个非相似类别所得的热力图。对比可知，针对同属于装甲侦察车的相似类别，网络能够提取到相似的目标信息从而将输入图像中的目标区域进行高亮激活；而针对非相似的自行高炮、推土机两个类别，热力图表现为对背景区域呈现高亮状态(ZSU234)，甚至是完全不响应(D7)。以上结果说明所用网络模型有效学习到了类判别性，即将输入图片判别为不同类别时模型捕捉到的特征具有差异性。



(a) 输入图像 (a) Input image　(b) 随机初始化训练 (b) RI　(c) 基于自然图像预训练权重微调 (c) PT (ImageNet)　(d) 基于遥感图像预训练权重微调 (d) PT (Remote sensing image)

图 12 不同初始化方式下类激活热力图

Fig. 12 Heat maps under different initialization modes

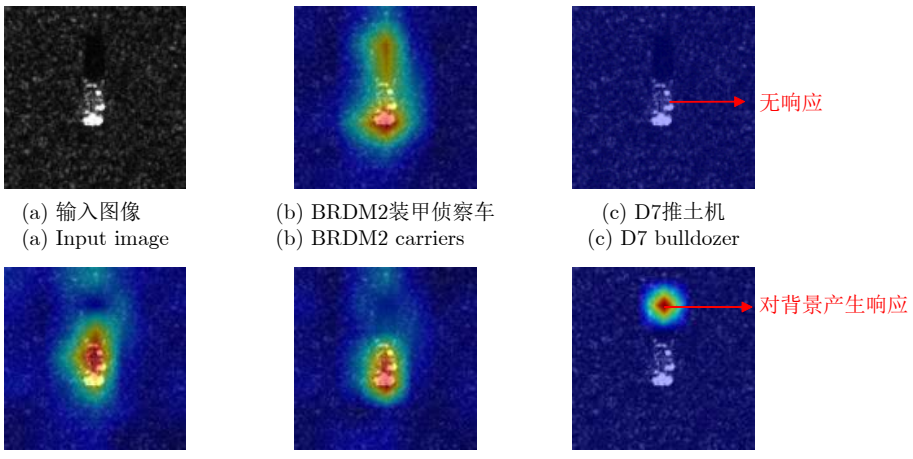


(a) 输入图像 (a) Input image

(b) CS-CAM类激活热力图 (b) Heat maps of CS-CAM

图 13 同类输入类激活热力图

Fig. 13 Heat maps of inputs with same class



(a) 输入图像 (a) Input image　(b) BRDM2装甲侦察车 (b) BRDM2 carriers　(c) D7推土机 (c) D7 bulldozer

(d) BTR60装甲侦察车 (真实类别) (d) BTR60 carriers (Ground truth)　(e) BTR70装甲侦察车 (e) BTR70 carriers　(f) ZSU234自行高炮 (f) ZSU234 air defense unit

图 14 异类输入激活热力图

Fig. 14 Heat maps of inputs with different classes

4 结语

本文围绕SAR图像卷积识别网络的可视化方法展开研究，提出一种新的混合通道、空间类激活权重的类激活映射方法，通过定性、定量实验与充分性评估结果证实CS-CAM能够更准确地反映模型的决策行为和依据，进而定位输入SAR图像中的重要区域。同时，针对已完成训练的CNN模型提出基于最大激活值的神经元可视化方法，展示了网络中特定层关键神经元的目标识别重点。综合上述各模块，所提面向SAR图像的CNN模型可视化方法，以离线的形式通过神经元感受野与类激活热力图分析的方式实现了模型的可视化研究，有效地提升了模型的可解释性与可信度，一定程度上缓解了模型

解译难的问题，有望为SAR图像卷积识别网络的设计与优化提供全新的视角和知识。

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] PALLOTTA L, CLEMENTE C, DE MAIO A D, *et al*. Detecting covariance symmetries in polarimetric SAR images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(1): 80–95. doi: 10.1109/TGRS.2016.2595626.

[2] WANG Zhen, WANG Shuang, XU Caijin, *et al*. SAR images super-resolution via cartoon-texture image decomposition and jointly optimized regressors[C]. 2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, USA, 2017: 1668–1671. doi: 10.1109/IGARSS.2017.8127294.

[3] LI Weike, ZOU Bin, and ZHANG Lamei. Ship detection in a large scene SAR image using image uniformity description factor[C]. 2017 SAR in Big Data Era: Models, Methods and Applications, Beijing, China, 2017: 1–5. doi: 10.1109/BIGSARDATA.2017.8124933.

[4] YUAN Ye, WU Yanxia, FU Yan, *et al*. An advanced SAR image despeckling method by bernoulli-sampling-based self-supervised deep learning[J]. *Remote Sensing*, 2021, 13(18): 3636. doi: 10.3390/rs13183636.

[5] SHU Yuanjun, LI Wei, YANG Menglong, *et al*. Patch-based change detection method for SAR images with label updating strategy[J]. *Remote Sensing*, 2021, 13(7): 1236. doi: 10.3390/rs13071236.

[6] CHEN Sizhe, WANG Haipeng, XU Feng, *et al*. Target classification using the deep convolutional networks for SAR images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, 54(8): 4806–4817. doi: 10.1109/TGRS.2016.2551720.

[7] 潘宗序, 安全智, 张冰尘. 基于深度学习的雷达图像目标识别研究进展[J]. 中国科学: 信息科学, 2019, 49(12): 1626–1639. doi: 10.1360/ssi-2019-0093.

PAN Zongxu, AN Quanzhi, and ZHANG Bingchen. Progress of deep learning-based target recognition in radar images[J]. *SCIENTIA SINICA Informationis*, 2019, 49(12): 1626–1639. doi: 10.1360/ssi-2019-0093.

[8] 贺丰收, 何友, 刘准钆, 等. 卷积神经网络在雷达自动目标识别中的研究进展[J]. 电子与信息学报, 2020, 42(1): 119–131. doi: 10.11999/JEIT180899.

HE Fengshou, HE You, LIU Zhunga, *et al*. Research and development on applications of convolutional neural networks of radar automatic target recognition[J]. *Journal of Electronics & Information Technology*, 2020, 42(1): 119–131. doi: 10.11999/JEIT180899.

[9] 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.

[10] 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.

[11] 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.

[12] 黄钟泠, 姚西文, 韩军伟. 面向SAR图像解译的物理可解释深度学习技术进展与探讨[J]. 雷达学报, 2022, 11(1): 107–125. doi: 10.12000/JR21165.

HUANG Zhongling, YAO Xiwen, and HAN Junwei. Progress and perspective on physically explainable deep learning for synthetic aperture radar image interpretation[J]. *Journal of Radars*, 2022, 11(1): 107–125. doi: 10.12000/JR21165.

[13] DATCU M, HUANG Zhongling, ANGHEL A, *et al*. Explainable, physics-aware, trustworthy artificial intelligence: A paradigm shift for synthetic aperture radar[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2023, 11(1): 8–25. doi: 10.1109/MGRS.2023.3237465.

[14] LI Yi and DU Lan. Design of the physically interpretable sar target recognition network combined with electromagnetic scattering characteristics[C]. 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022: 4988–4991. doi: 10.1109/IGARSS46834.2022.9883598.

[15] 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.

[16] 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.

[17] LI Chen, DU Lan, LI Yi, *et al*. A novel SAR target recognition method combining electromagnetic scattering information and GCN[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 4508705. doi: 10.1109/LGRS.2022.3178234.

[18] LIU Zhunga, WANG Longfei, WEN Zaidao, *et al*. Multilevel scattering center and deep feature fusion learning framework for SAR target recognition[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5227914. doi: 10.1109/TGRS.2022.3174703.

[19] ZHANG Jinsong, XING Mengdao, and XIE Yiyuan. FEC: A feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(3): 2174–2187. doi: 10.1109/TGRS.2020.3003264.

[20] FENG Sijia, JI Kefeng, ZHANG Linbin, *et al*. SAR target classification based on integration of ASC parts model and deep learning algorithm[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 10213–10225. doi: 10.1109/JSTARS.2021.3116979.

[21] 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.

[22] FENG Sijia, JI Kefeng, WANG Fulai, *et al*. PAN: Part attention network integrating electromagnetic characteristics for interpretable SAR vehicle target recognition[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 1–17. doi: 10.1109/TGRS.2023.3256399.

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

[24] SU Shenghan, CUI Ziteng, GUO Weiwei, *et al*. Explainable analysis of deep learning methods for SAR image classification[C]. 2022 – 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, 2570–2573. doi: 10.1109/IGARSS46834.2022.9883815.

[25] 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.

[26] 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.

[27] 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, Gdansk, Poland, 2022: 457–462. doi: 10.23919/IRS54158.2022.9904989.

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

[29] LUO Wenjie, LI Yujia, URTASUN R, *et al*. Understanding the effective receptive field in deep convolutional neural networks[C]. 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 4898–4906.

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

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

[32] 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, Lake Tahoe, USA, 2018: 839–847. doi: 10.1109/WACV.2018.00097.

[33] DESAI S and RAMASWAMY H G. Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization[C]. 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass, USA, 2020: 972–980. doi: 10.1109/WACV45572.2020.9093360.

[34] 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, Seattle, USA, 2020: 111–119. doi: 10.1109/CVPRW50498.2020.00020.

[35] MUHAMMAD M B and YEASIN M. Eigen-CAM: Class activation map using principal components[C]. 2020 International Joint Conference on Neural Networks, Glasgow, UK, 2020: 1–7. doi: 10.1109/IJCNN48605.2020.9206626.

[36] JIANG Pengtao, ZHANG Changbin, HOU Qibin, *et al*.

LayerCAM: Exploring hierarchical class activation maps for localization[J]. *IEEE Transactions on Image Processing*, 2021, 30: 5875–5888. doi: 10.1109/TIP.2021.3089943.

[37] 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. doi: 10.1117/12.242059.

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

[39] YEH C K, HSIEH C Y, SUGGALA A S, *et al*. On the (in)fidelity and sensitivity of explanations[C]. 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019.

[40] ADEBAYO J, GILMER J, MUELLY M, *et al*. Sanity checks for saliency maps[C]. 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 9525–9536.

[41] HUANG Zhongling, PAN Zongxu, and LEI Bin. What, where, and how to transfer in SAR target recognition based on deep CNNs[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(4): 2324–2336. doi: 10.1109/TGRS.2019.2947634.

作 者 简 介

李妙歌，硕士生，主要研究方向为雷达目标识别、SAR图像解译、机器学习与人工智能等。

陈　渤，博士，教授，博士生导师，主要研究方向为机器学习、统计信号处理、雷达目标识别与检测、深度学习网络、大规模数据处理等。

王东升，博士生，主要研究方向为贝叶斯概率统计、生成模型、机器学习等。

刘宏伟，博士，教授，博士生导师，主要研究方向为雷达目标识别、认知雷达、网络化协同探测、雷达智能化探测等。

(责任编辑：高山流水)