SAR图像目标识别的可解释性问题探讨

郭炜炜[①] 张增辉*[②] 郁文贤[②] 孙效华[①]

[①](同济大学数字创新中心 上海 200092)

[②](上海交通大学智能探测与识别上海市重点实验室 上海 200240)

摘 要：合成孔径雷达(SAR)图像目标识别是实现微波视觉的关键技术之一。尽管深度学习技术已被成功应用于解决SAR图像目标识别问题，并显著超越了传统方法的性能，但其内部工作机理不透明、解释性不足，成为制约SAR图像目标识别技术可靠和可信应用的瓶颈。深度学习的可解释性问题是目前人工智能领域的研究热点与难点，对于理解和信任模型决策至关重要。该文首先总结了当前SAR图像目标识别技术的研究进展和所面临的挑战，对目前深度学习可解释性问题的研究进展进行了梳理。在此基础上，从模型理解、模型诊断和模型改进等方面对SAR图像目标识别的可解释性问题进行了探讨。最后，以可解释性研究为切入点，从领域知识结合、人机协同和交互式学习等方面进一步讨论了未来突破SAR图像目标识别技术瓶颈有可能的方向。

关键词：合成孔径雷达；自动目标识别；深度学习；可解释性；可解释机器学习

中图分类号：TN957.51 **文献标识码**：A **文章编号**：2095-283X(2020)03-0462-15

DOI: 10.12000/JR20059

引用格式：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.

Reference format: 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.

Perspective on Explainable SAR Target Recognition

GUO Weiwei[①] ZHANG Zenghui*[②] YU Wenxian[②] SUN Xiaohua[①]

[①](*Center of Digital Innovation*, *Tongji University*, *Shanghai* 200092, *China*)

[②](*Shanghai Key Lab of Intelligent Sensing and Recognition*,
Shanghai Jiaotong University, *Shanghai* 200240, *China*)

Abstract: SAR Automatic Target Recognition (ATR) is a key task in microwave remote sensing. Recently, Deep Neural Networks (DNNs) have shown promising results in SAR ATR. However, despite the success of DNNs, their underlying reasoning and decision mechanisms operate essentially like a black box and are unknown to users. This lack of transparency and explainability in SAR ATR pose a severe security risk and reduce the users' trust in and the verifiability of the decision-making process. To address these challenges, in this paper, we argue that research on the explainability and interpretability of SAR ATR is necessary to enable development of interpretable SAR ATR models and algorithms, and thereby, improve the validity and transparency of AI-based SAR ATR systems. First, we present recent developments in SAR ATR, note current practical challenges, and make a plea for research to improve the explainability and interpretability of SAR ATR. Second, we review and summarize recent research in and practical applications of explainable machine learning and deep learning. Further, we discuss aspects of explainable SAR ATR with respect to model understanding, model diagnosis, and model improvement toward a better understanding of the internal representations and decision mechanisms. Moreover, we emphasize the need to exploit interpretable SAR feature learning and recognition models that integrate SAR physical characteristics and domain knowledge.

收稿日期：2020-05-11；改回日期：2020-06-17；网络出版：2020-06-30

*通信作者： 张增辉 zenghui.zhang@sjtu.edu.cn *Corresponding Author: ZHANG Zenghui, zenghui.zhang@sjtu.edu.cn

基金项目：国家自然科学基金联合基金(U1830103)

Foundation Item: The National Natural Science Foundation of China(U1830103)

责任主编：邹焕新 Corresponding Editor: ZOU Huanxin

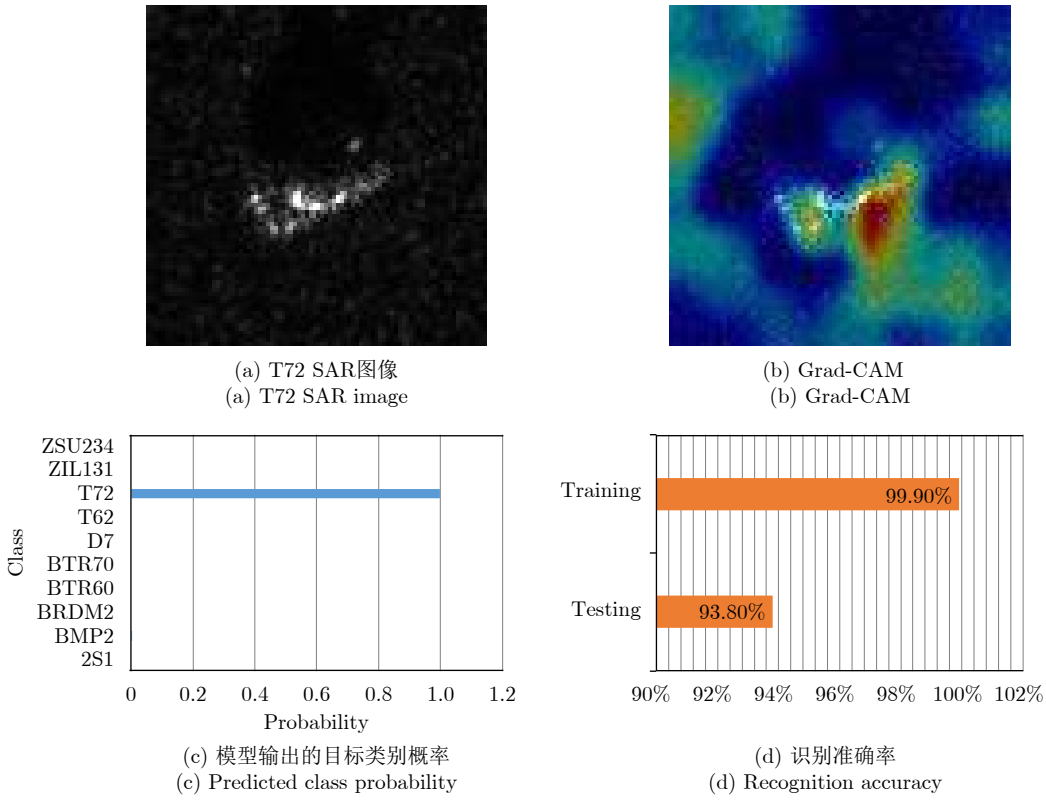
Finally, we draw our conclusion and suggest future work for SAR ATR that combines data and knowledge-driven methods, human–computer cooperation, and interactive deep learning.

Key words: SAR; Automatic Target Recognition (ATR); Deep learning; Explainability and interpretability; Explainable machine learning

1　引言

合成孔径雷达(SAR)是一种可实现高分辨率的微波主动成像雷达，具备全天时、全天候、大范围观测成像的能力，使其在国民经济和国防军事等领域的应用中具有独特的优势，甚至是极端气象条件下唯一可靠的观测数据来源。SAR图像自动目标识别(Automatic Target Recognition, ATR)是实现SAR图像智能解译的关键技术之一[1]，自上个世纪50年代SAR诞生以来至今持续获得大量的关注和研究[2]。特别是近年来随着深度学习技术的迅猛发展，深度神经网络也被应用于解决SAR图像目标检测和识别问题，并大幅超越了传统SAR图像目标检测识别技术[3−5]。尽管深度学习技术显著提升了SAR图像目标检测识别的性能，但主要依赖于大量标注数据的参数拟合能力，其内部过程犹如黑盒子，人们很难理解其背后的工作机理和决策逻辑，难以掌握系统决策行为的边界。如图1，笔者采用一个简单的具有5层卷积模块(Conv2d-ReLU-MaxPool2d)的卷积神经网络(Convolutional Neural Network, CNN)在MSTAR[2]测试集上的识别准确率可以达到93.80%(图1(d))，针对图1(a)输入样本能够正确判断其类别(图1(c))，但是基于Grad-CAM[6](Gradient-Class Activation Mapping)方法提取的决策显著性区域(图1(b))显示决策并不完全依赖于目标区域，还有部分背景区域对最终决策也有重要影响，其背后的决策合理性还需要结合SAR机理和特性进行分析和评估。

一方面，这样决策不透明和缺乏可解释性的SAR目标识别技术在军事目标侦察、精确打击等高风险应用中隐藏着一定的决策风险，在应用中难以取得用户的信任；另一方面，SAR图像是目标电磁散射特性的反映，难以被视觉所认知，深度神经网络从大量数据中自动挖掘的特征表示有可能蕴含一些新的知识，通过对这些特征的理解，可以启发人们反过来利用这些知识，进而提升SAR目标认知解译的能力；再次，深度神经网络工作机理复杂，且



(a) T72 SAR图像
(a) T72 SAR image

(b) Grad-CAM
(b) Grad-CAM

(c) 模型输出的目标类别概率
(c) Predicted class probability

(d) 识别准确率
(d) Recognition accuracy

图 1 一个简单CNN分类器在T72 SAR图像的梯度-类激活映射(Grad-CAM[6])

Fig. 1　The results of a CNN classifier and the Grad-CAM map[6]

具有一定的脆弱性[7]，需要通过理解深层网络模型背后的决策过程和依据，发现其中的缺陷，以便对模型和算法加以改进，提升SAR目标识别系统的鲁棒性；进一步地，SAR图像与光学图像特性存在着本质差异，其对成像参数高度敏感，很难获取完备的训练样本，因此在构建SAR图像目标识别的深层模型时需要考虑SAR图像数据的特点，结合SAR本身的物理、统计和语义等领域知识，建立可解释的SAR图像目标识别模型，从而增强SAR图像目标识别的可解释性、鲁棒性和在小样本上的泛化能力。

可解释性是人与决策模型之间的接口，旨在对模型的决策给出令人能够理解的清晰概括和指示，从而帮助人们理解模型从数据中学到了什么，针对每一个样本是如何决策的，决策是否合理和可靠等[8–10]。SAR的电磁成像机理与人类视觉系统和光学遥感的成像机理有着本质差异，导致对SAR图像的认知理解与解译应用非常困难。例如图2，SAR系统接收的是组成地物目标的每一个独立单元形成的散射能量，呈现在SAR图像上的地物目标是散射单元构成的集合体，多表现为离散的点、线组合。SAR系统独特的成像方式会造成相干斑、结构缺失、几何畸变(透视收缩、叠掩)、阴影等现象，导致SAR图像在视觉特性上与光学图像有着明显差异，表现为“所见非所知”的特点，同时SAR图像对观测参数敏感、获取样本困难，导致SAR图像目标识别仍是一个世界性难题。本文在总结当前SAR图像目标识别技术及其存在问题的基础上，结合当前机器学习、深度学习可解释性的研究进展，从模型理解、模型诊断和模型改进等方面对SAR图像目标识别的可解释性问题进行了探讨，以突破当前SAR目标识别的技术瓶颈和应用限制。最后，本文还从领域知识的引入与结合、人机协同、交互式学习等方面对SAR目标识别未来可能的研究工作进行了讨论，以期推动SAR目标识别技术的进一步发展。

2 SAR图像目标识别研究进展与挑战

2.1 SAR图像目标识别研究进展

SAR图像目标解译一般采用“检测->鉴别->识别”的处理流程[11]。SAR图像目标检测和鉴别的主要目的是定位目标在图像中的位置和区域，为进一步的目标识别奠定基础，杜兰教授等人在文献[4]中对目前SAR目标检测及鉴别的研究工作进行了很好的总结。SAR目标识别的目的是确定目标的类别，甚至细粒度的型号等信息。它实际上是一个模式识别问题，通常采用“特征提取+模式分类”的经典模式识别框架，其中特征提取是关键。传统SAR目标识别技术主要是基于图像处理、统计分析等方法手工设计对识别有效的特征表示[12,13]。典型的SAR图像目标识别特征包括原始图像、Garbor纹理特征、散射点分布特征、阴影形状特征等[14–16]；而分类器设计方面，从早期的相关滤波到支持矢量机(Support Vector Machine, SVM)、基于稀疏表示的分类器、Adaboosting集成分类器等都有被应用于SAR图像目标识别[17–20]。SAR图像目标识别的另一类方法是基于散射中心模型匹配的方法，主要思想是将未知目标的散射中心特征与目标模型库中的散射中心模板或者电磁计算预测的特征进行匹配识别，主要涉及目标散射中心参数化建模、参数估计和匹配相似度计算[21–24]，例如Potter等人[21]提出了属性散射中心模型用于SAR目标识别，计科峰教授等人[22]研究了图像域的属性散射中心参数估计方法。基于模型的方法主要困难在于：一是难于建立目标，特别是非合作目标的模型库，而SAR目标图像易受目标、传感器、环境等操作条件的影响，模型数量往往呈几何级数增长，制约了该类方法在实际中的应用。总的来说，传统SAR目标识别方法主要是基于图像的统计、物理特性进行手工建模，该框架可解释性强，识别的特征和模型具有明确的统计或物理含义，但是手工建模难以适应SAR图像的复杂多变，从而在实际应用中很难取得很高的性能。

近年来，随着计算能力的显著提升、数据规模的大幅扩大以及机器学习算法的不断改进，从数据中自动进行特征学习日益成为模式识别的主要范式。在绝大部分底层图像处理任务(例如图像去噪、超分辨等[25,26])及高层图像理解任务(图像分类、物体检测、语义分割[27–29])中，深度学习方法

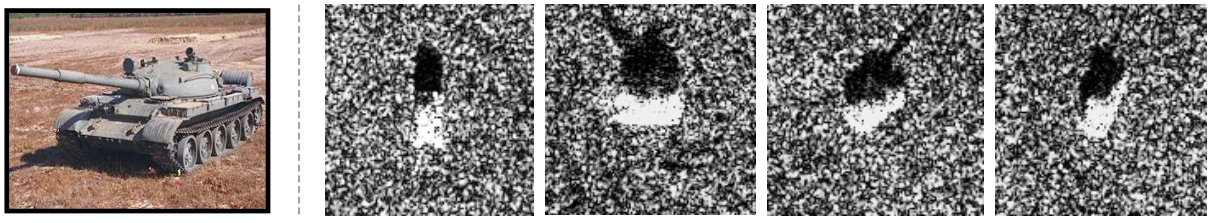


图 2 MSTAR T62光学图像与不同方位角下的SAR图像

Fig. 2 The Optical image and SAR image samples of T62 tank at different azimuth angles in MSTAR dataset

尤其是基于卷积神经网络(Convolutional Neural Networks, CNNs)的方法已成为目前表现最好的方法，在SAR图像目标检测和识别中也同样显示出巨大的优势[3,30–32]。例如，Zhao等人[3]提出了基于多尺度网络融合的SAR舰船目标检测方法，提升了对SAR图像中小目标的检测能力，并进一步根据SAR特性提出了一种基于脉冲余弦变换(Pulse cosine transformation)的视觉关注算法，其利用频域信息来进一步地进行舰船鉴别，去除虚警，提升了复杂场景下的SAR目标检测能力[33]；陈慧元等人[34]设计了一种由目标预筛选全卷积网络(Fully Convolutional Networks for Prescreening, P-FCN)和目标精细检测全卷积网络(Detection Fully Convolutional Network for Detection, D-FCN)两个全卷积网络级联而成的目标检测框架，在保持检测精度的前提下显著提升了大场景SAR图像的目标检测效率。

对于SAR目标识别问题，国内外学者设计和改进了不同的网络结构和学习算法来提高SAR图像目标识别性能，文献[5,31,32]较好地总结了当前基于深度学习的SAR目标分类识别技术。例如，Chen等人[30]提出了所谓的AConvNets，其将全连接层去掉形成全卷积网络，降低了网络训练中的过拟合风险，在MSTAR数据集上取得了目前最好的性能；Wagner等人[35]提出了将图像强度和梯度信息多通道特征融合的方法，提升了SAR图像分类性能；并将CNN与传统SVM分类器结合，将CNN作为特征提取器提取深度特征后采用SVM作为分类器。在应用深度神经网络解决SAR目标识别问题所面临的主要困难是没有足够训练数据，目前常用的MSTAR数据包含10类目标，也仅有5631个样本，其中训练样本2813个，测试数据2818个[2]。通常采用数据扩充、迁移学习、元学习等策略来解决小样本目标识别问题[31]。例如，Wagner[36]采用弹性变形和仿射变换生成扩充数据，Huang等人[37]研究了自然图像ImageNet、不同源SAR图像之间深层特征的迁移性。由于SAR特殊的成像原理，非直观性强，人工标注极易出错，导致学习能力和泛化能力急剧下降。针对这种含噪声标签的SAR图像分类问题，赵娟萍等人[38]提出了一种基于概率转移模型的CNN方法。在传统CNN模型基础上，可潜在地校正错误标记，增强了含噪标记下CNN分类模型的鲁棒性。

总的来说，目前基于深度学习的SAR目标识别方法主要是借鉴光学图像中的神经网络模型和框架，侧重于对网络结构和学习算法进行有针对性的改进，来提升SAR目标检测识别的性能，在MSTAR等类别确定、数量有限、标注充分的特定数据集上性能已趋于饱和。但是在面对SAR图像与光学图像的本质差异性、SAR图像的多参数敏感性以及小样本等问题时，深度学习方法在SAR目标识别任务上仍面临着不小的挑战，其严重依赖大量标注数据，在机理分析、知识利用、可解释性、逻辑推理等方面还有很大局限性，单纯通过改进通用算法来提升识别性能存在着“天花板”。

2.2 SAR图像目标识别存在的问题与挑战

不同于光学图像，SAR图像作为目标电磁散射情况的反映，与人的视觉认知有着很大差别，它所蕴含的目标信息难以被直观理解；同时，SAR图像还伴随有固有相干斑噪声，几何畸变(如叠掩、前视收缩)等现象，对观测参数也更加敏感，使得对SAR图像的目标稳健特征描述和分类识别更加困难；并且，目标识别通常需要完备的训练集，但是SAR图像的获取成本较高，真值标记不易获得，导致SAR目标识别还面临小样本的困境。传统的SAR目标检测识别方法可解释性强，但是手工建模的泛化能力和鲁棒性严重不足，可挖掘的潜力有限，而深度学习能在特定数据集上取得较好结果，但是对样本数量和网络规模具有较高要求，特别是对于非合作目标，很难建立起完备的样本集，在对大场景、多尺度、密集排布、地物干扰等复杂场景下的目标检测识别有待进一步研究，且目前的深度学习方法主要依赖于图像域数据，很容易对图像噪声过拟合；同时，深度模型可解释性差，在先验知识的引入和利用方面仍存在较大局限性，限制了其性能的进一步提升。

相对于SAR系统数据获取能力的显著增强，对目标“辨得明、认得准”的SAR目标认知解译能力仍严重滞后，还面临以下挑战性问题，如图3所示：

(1)“看得懂”问题(机理层面)：SAR成像机理导致的目标结构性缺失、电磁散射机制的多样性和复杂性、特性反演困难等问题使得对SAR图像的理解与人类对于可见光影像的认知存在巨大鸿沟，如何实现从稳健的特征提取到逼近视觉认知的模型再到目标和场景的检测识别存在诸多困难。

(2)“认得准”问题(算法层面)：SAR图像存在固有的相干斑噪声以及在平台、目标和环境耦合作用下目标几何和电磁特征的多参数敏感性问题；对于非合作目标，难以建立起多参数完备的样本库，面临小样本识别的困境；对于高分辨SAR图像在提供目标丰富细节信息的同时，也存在目标观测尺度变大(从目标内部精细结构到大中小目标、目标群

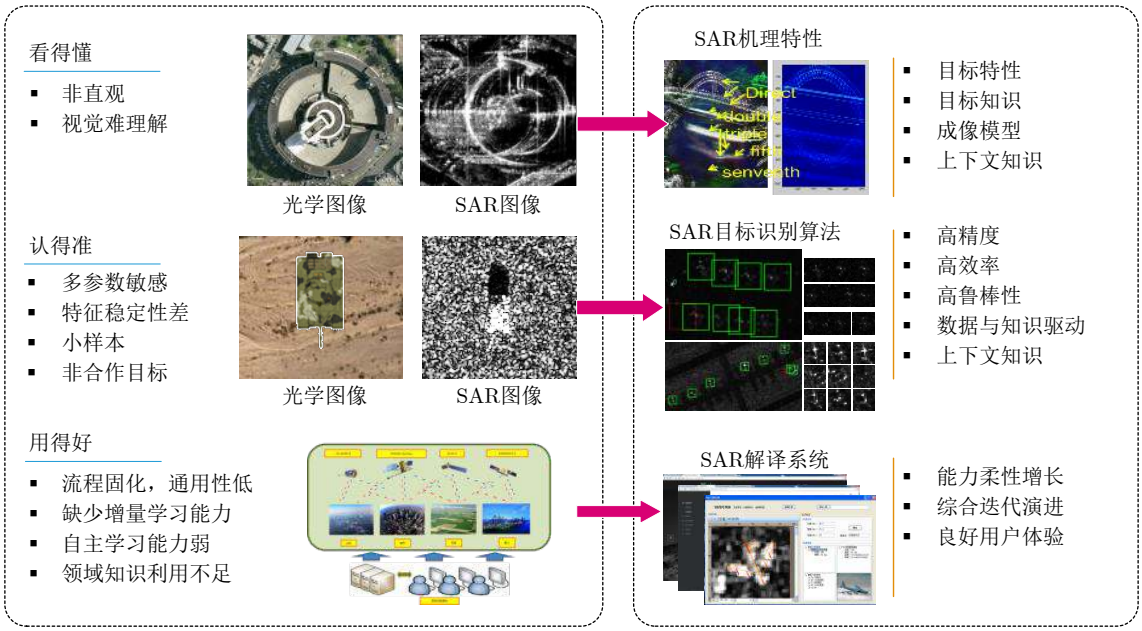


图 3 SAR目标识别面临的挑战

Fig. 3 Challenges of SAR ATR

密级排布)、目标表现变化大，目标与背景互相干扰耦合。如何设计高精度、高效率和高鲁棒性的SAR图像目标检测识别等核心算法，支持SAR图像情报分析与挖掘，仍存在较大挑战。

(3)“用得好”问题(系统层面)：SAR目标和场景解译流程固化、通用性差、核心算法与辅助工具集成度低；系统缺少增量学习、迁移和持续学习的能力，对知识与经验的利用不足，还不能实现识别能力的迭代增长。

3 深度学习的可解释性

随着深度学习技术的进步和在诸多领域的大量应用，其可解释性问题日益受到政府、学术界和工业界的广泛重视，例如美国国防部高级研究计划署(Defense Advanced Research Projects Agency, DARPA) 启动了一项名为可解释性人工智能(EXplainable Artificial Intelligence, XAI)的大型项目[39]，我国也在《新一代人工智能发展规划》中明确将“实现具备高可解释性、强泛化能力的人工智能”作为未来我国人工智能发展的重要突破口。深度学习技术的可解释性问题源于其“黑盒”性质，其工作机理、决策过程和决策逻辑对用户的不透明，会存在安全隐患，特别是在医疗诊断、金融投资、国防军事等高风险领域，深度神经网络的可解释性对于理解和信任模型的决策至关重要。同时由于深度神经网络机理复杂，在应用过程中主要依靠经验调参，迫切需要打开深层网络的“黑盒子”，才能针对具体任务需求和数据特点对神经网络进行有针对性地改进。如图4，从目的上来说，可解释性旨在帮助人们理解机器学习模型是如何学习的，它从数据中学到了什么；针对每一个输入样本，它为什么会做出如此决策以及它所做的决策是否可靠等；从方法来说，可解释性方法是挖掘模型决策背后的信息并给出令人理解的指示。文献[8–10,40–42]对当前机器学习，特别是深度学习可解释性的研究进行了较好地总结。

传统机器学习模型大多具有可解释性，例如决策树(Decision tree)、线性模型(Linear model)、广义加性模型(Generalized Additive Mode, GAM)、稀疏表示(Sparse representation)模型等，其解释性主要体现在能够给出特征对决策的重要性度量，但是传统方法的解释性需要输入特征本身就具有一定物理或者语义含义，而且模型准确度不够高，可解释性和模型性能存在一定的矛盾。目前对深度学习的可解释性研究大致有两个方面：一是模型的可解释性，即对一个已经训练好的神经网络模型，通过建立可解释方法或者代理模型，从整体上理解深

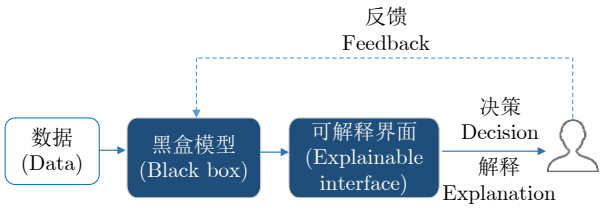


图 4 可解释性学习

Fig. 4 Explainable machine learning

层模型的决策行为以及针对每一个测试样本的局部决策依据；另一个是可解释的模型，即基于一定的物理或者语义等领域知识设计和构建自身具有一定可解释性的神经网络模型。

3.1 深度模型的可解释性方法

根据解释对象的不同，深度模型的可解释性方法，主要分为针对模型的全局性解释方法(Explain model)和针对单个样本的局部解释方法(Explain sample)；根据解释方法是否依赖具体模型内部参数，又可分为模型依赖(Model-specific)的解释方法和模型无关的解释方法(Model-agnostic)。表1列出一些典型的可解释性方法。

表 1 典型的可解释性方法

Tab. 1　Typical methods for explainablitiy

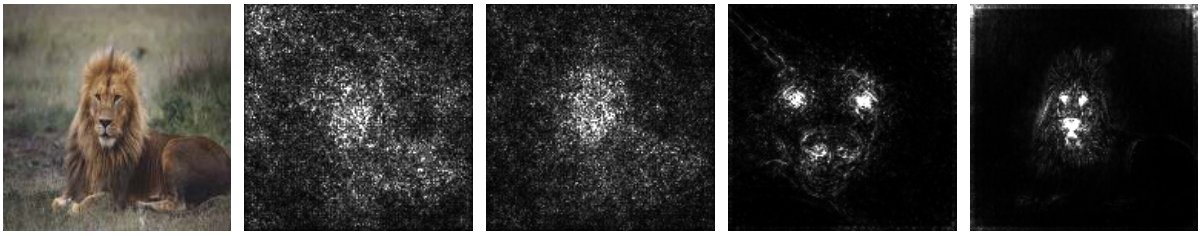
解释的对象	模型依赖(Model-specific)	模型无关(Model-agnostic)
解释模型 Explain model	■激活最大化方法AM[43,44] ■概念激活矢量TCAV[45]	■知识蒸馏(Knowledge distilling)[46] ■特征置换(Permutation)[47]
解释样本 Explain sample	■基于梯度的方法Grad[48], GuidedBP[49], IntegratedGrad[50], SmoothGrad[51] ■特征扰动分析Perturbation[52] ■层次相关传播LRP[53] ■类激活映射CAM[54], Grad-CAM[6]	■基于局部代理模型的方法，如LIME[55] ■基于实例的方法，如Influence function[56], Critic样本方法[57] ■基于Shapley值的方法[58]

针对模型的全局性解释方法，主要是从整体上理解模型从数据中学到的内容及其行为逻辑。为了理解神经网络从数据中学到了什么，通常采用特征可视化方法，但是直接对神经网络的权重进行可视化对于用户来说仍然过于抽象。为此，许多研究者探索如何在输入空间实现对任意隐含神经元计算内容的可视化，以此捕捉神经网络中内部神经元计算内容的特定含义[8]。激活最大化方法(Activation Maximization, AM)是一类典型方法，即寻找最大化激活给定的隐藏单元或者重构满足一定条件的输入模式[43]，其可形式化为如下的最优化问题

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} l(\boldsymbol{x}),\ \ l(\boldsymbol{x}) = l_{\text{inv}}(\boldsymbol{x};\varphi_0) + \lambda R(\boldsymbol{x}),\\ \boldsymbol{x} \in \mathbb{R}_+^{W\times H\times C} \tag{1}$$

其中，$\boldsymbol{x}$为待求解的输入模式图像，$l(\boldsymbol{x})$是目标函数，$l_{\text{inv}}(\boldsymbol{x},\varphi_0)$是与特定神经元特征响应$\varphi_0$有关的损失函数，$R(\boldsymbol{x})$为表示图像先验的正则化项，$\lambda$为正则化系数。Nguyen等人[44]利用生成对抗网络来对图像的先验分布进行建模，并与激活最大化相结合来产生更真实、更具有可解释性的模式图像。Kim等人[45]提出概念激活矢量测试的方法(Testing with Concept Activation Vectors, TCAV)以此来捕捉神经网络内部节点对某一类别的敏感性。为了从整体上理解模型的行为逻辑，另一种方法是用一个可解释的代理模型来近似理解黑盒模型的决策机制，例如Frosst等人[46]基于知识蒸馏(Distilling)原理提出使用决策树作为学生模型来提取深度神经网络模型的决策规则，它是模型无关的。针对基于CNN的图像分类任务，Zhang等人[42,59]提出了基于And-Or图模型来解释CNN卷积层特征内在的图像知识层次，进而提取决策树规则来揭示卷积层中哪些滤波器会参与预测以及这些滤波器对预测结果的贡献程度。

针对样本的局部解释方法主要是针对每一个特定输入样本，通过分析和提取输入样本的每一维特征对模型最终决策的贡献程度，即提取特征的决策重要性，并通过可视化手段进行呈现，使用户能从语义和视觉上直观理解模型对输入样本的决策逻辑和依据。典型方法是基于反向传播的方法，它是模型依赖的，核心思想是对于一个给定输入样本图像，利用神经网络的反向传播机制将对决策的重要性信号从模型的输出层逐层传播到模型的输入层，以推导输入样本的特征重要性，生成与之对应的决策显著性热力图(Heatmap)，对输入图像中对决策的重要部分进行标注和显示，例如图5所示的Grad[48], GuidedBP[49], IntegratedGrad[50]和SmoothGrad[51]等基于梯度的系列方法，但是基于梯度信息只能用于定位重要特征，而无法量化特征对决策结果的重要程度。Du等人[60]提出了基于敏感性分析的方法，Bach等人[53]提出了层级相关性传播(Layer-wise Relevance Propagation, LRP)方法，Zeiler等人[61]提出了利用反卷积操作(DeConv)将高层激活反向传播到模型的输入层以辨识输入图片中负责激活的重要部分。这些方法的主要区别在于如何在输出层建立与决策相关的度量，以及决策相关信号从输出层到输入层的传播机制。最近，Samek等人[62]基于扰动分析定量地比较了敏感性分析方法、DeConv 和LRP方法，表明LRP方法能够较准确地定位输入中对决策起重要作用的区域。Zhou等人[54]提出了类激活映射(Class Activation



(a) Input image (b) Gradient (c) IntegratedGradient (d) GuidedBP (e) SmoothGradient

图 5 基于梯度系列方法的决策显著性[1]

Fig. 5 Decision saliency of the Gradient-based methods

Mapping, CAM)方法，其利用全局平均池化(Global Average Pooling, GAP)层来替代传统CNN 模型中除Softmax层以外的所有全连接层，并通过将输出层的权重投影到卷积特征图来定位图像中的重要区域。Selvaraju[6]则将基于梯度的方法与CAM方法结合，提出了梯度加权类激活映射方法(Grad-CAM)。但是CAM和Grad-CAM方法只能对决策重要区域进行粗粒度的定位，无法像基于梯度的方法提供像素级别的细粒度解释。

另一类针对样本的局部解释方法是模型无关的方法，主要有基于局部代理模型和基于样例的方法。比较典型的基于局部代理模型方法是Ribeiro等人[55]提出的LIME(Local Interpretable Model-agnostic Explanations)模型，如图6所示其在一个样本邻域内，用一个线性模型来近似原非线性神经网络模型，线性模型的权重可作为输入特征局部重要性的指示器。但是LIME方法需要针对每一个样本重新训练线性模型，存在解释效率的问题。其实，人们在做复杂决策的时候，有时候并不是通过仔细分析和计算，而是基于相似的经验进行类比得出结论。基于样例的解释是通过选择有代表性的或者关键样本，来解释模型的决策行为。比较典型的方法是Liang Percy等人[56]提出的基于影响力函数(Influence function)的方法来选择对一个分类器决策起到重要作用的样本，并以此来评估决策的合理性。基于影响力函数的方法还可以用来构建对抗样本，评估训练集与测试集分布一致性以及发现训练集中的标记错误样本等。Kim[57]提出基于MMD(Maximum Mean Discrepancy)的方法来同时选择数据集的原型样本和所谓的Critic样本，进一步提升解释性。Lundberg S M等人[58]提出基于博弈论Shapley值的特征重要性评估方法SHAP (SHapley Additive exPlanations)，理论上可以获得唯一可能的一致的、局部精确的加性特征解释，结合DeepLIFT[63]和Shapley方法可应用于图像分类任务的解释。

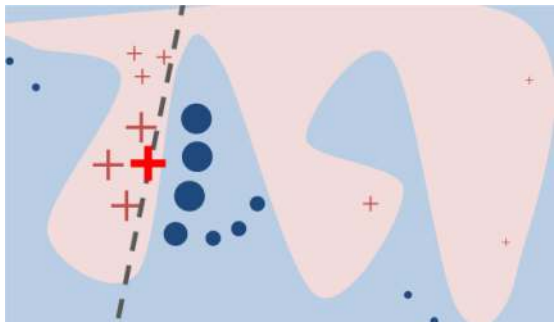


图 6 LIME示意图[55]

Fig. 6 Illustration of LIME[55]

3.2 可解释的深度模型

以上无论是对模型的全局解释方法还是针对单个样本的局部解释方法，都是对已训练好的模型进行后验解释(Post-hoc)，实际上只是对原始模型的一种近似理解和间接解释，与模型真实决策行为有可能存在不一致性，从而导致错误的解释。一方面错误的理解在实际应用中反而会适得其反，另一方面后验理解的方法并不能完全预测原始模型的行为，导致系统的不可控。Rudin[64]认为模型的性能和可解释性并不完全矛盾，目前在可解释的研究中亟须对可解释的神经网络进行研究，而不应局限于神经网络可解释性的研究。可解释的神经网络是指网络的结构和中间层具有明确的物理或者语义含义。目前，构建本身具有内在可解释性的神经网络大致有如下方法：

一是基于注意力模型(Attention)。注意力模型源于人脑的注意力机制，其数学本质是一种对数据的加权策略，注意力矩阵体现了模型在决策过程中的感兴趣区域，因而具有良好的可解释性。例如Xu等人[65]将注意力机制应用于看图说话(Image caption)任务中以产生对图片的描述，其利用带注意力机制的循环神经网络(Recurrent Neural Network, RNN)生成图片描述。通过可视化注意力权重矩阵，人们可以清楚地了解到模型在生成每一个单词时所对应的感兴趣图片区域。

[1] https://pair-code.github.io/saliency/

二是浅层统计模型的深度化。相较于大多数深度神经网络模型，统计学习模型具有完备的理论基础、可解释性强和易于优化等诸多优点。因此，研究人员考虑基于统计学习的模型来构建神经网络，主要有两个策略：一种是将一些统计机器学习方法，特别是优化算法展开成一个循环神经网络RNN，这样就能同时兼具传统方法的可解释性强和深度学习性能优的优点，例如LeCun Yann教授等人[66]将稀疏编码方法中经典的ISTA(Iterative Shrinkage and Thresholding Algorithm)展开成一个循环神经网络结构，提出了LISTA(Leaned ISTA)模型，Zheng等人[67]将条件随机场(Conditional random field)的平均场优化算法(Mean field)展开成一个循环神经网络结构，应用于图像语义分割；另一种是基于统计模型来设计目标函数，例如PENG等人[68]通过改写K-Means聚类算法的目标函数，形成了一个具有自注意力机制的神经网络结构。

三是基于物理模型。真实物质世界其实是遵循着一定的物理规律，机器学习模型也要遵循这样的物理规则，因此可以尝试对物理模型进行建模，例如Zhu等人[69]提出的去雾模型根据雾形成的物理过程建立端到端的深层网络模型，这样神经网络的每一个模块都具有明确的物理含义。Karpatne等人[70]提出了基于物理模型引导的神经网络学习模型，其核心思想是在输入空间根据物理模型的预测数据来增广输入，在输出空间根据物理模型来设计相应的正则化损失函数来进行网络学习。

四是知识的嵌入与融合。深度神经网络具有强大的表示学习能力，但是忽略了一定的先验知识，通过引入语义概念和语义关联等高层信息，引导模型进行特征学习和推理，不仅可以增强特征的表示能力，还能使模型具有更好的解释性。例如Chen等人[71]基于图网络模型(Graph Neural Network, GNN)进行类别与属性域关联信息的嵌入(Embedding)，并引导网络学习具有特定语义的特征。在这类方法中，如何进行领域知识的表达和嵌入是其中的难点。

3.3 可解释性的评估

可解释的评估对于可解释性研究至关重要，但是目前可解释性还没有一个严格的定义，而且由于解释性与人的认知密切相关，导致目前对可解释性的评估还没有一个科学的评价准则。目前一般有主观和客观两种方式：客观评估方面，Chu等人[72]提出利用输入样本与其邻近样本的解释结果的余弦相似性来评估解释结果的一致性，Samek等人[62]基于扰动分析(Perturbation)提出了AOPC指标(Area of MoRF Perturbation Curve)来评估解释性热力图定位到决策重要区域的准确性，Bau等人[73]从目标、部件、场景、纹理、材料和颜色等方面来评估CNN特征表示的语义可解释性，将CNN得到的响应模式与人工真值标注区域的IoU作为评价准则。主观方面主要是进行用户测试，通过终端用户的评价来进行评估。

总得来说，面向深度学习的可解释性研究还处于初步探索阶段，研究人员针对不同任务需求，从不同角度对深度学习的可解释性问题进行了研究，但还没形成一个完整的科学体系。目前的可解释性方法在因果关系、知识融合和推理、解释性评价、智能人机交互等方面还存在着很大局限性，如何设计兼具解释性强和高性能的深度神经网络是其中一个重要问题。

4 可解释的SAR图像目标识别展望

针对SAR数据解译的可解释性问题，德国宇航局(DLR)的Datcu M教授及其团队[74]在面向SAR数据的可解释性人工智能方面开始进行了初步的探索，Huang等人[75]从SAR复数数据的时频谱中学习SAR目标的散射特征，并与图像域特征融合进行SAR图像分类。Zhao等人[76]提出了对比度正则化卷积神经网络从极化复散射数据中学习目标的散射特征。通过引入注意力模块可以增强神经网络的可解释性，Chen等人[77]提出基于空间注意力和通道注意力的双通道机制的网络模型，实现SAR图像机场区域提取。Li等人[78]提出基于空间像素级注意力机制的SAR目标识别方法，引导网络能够自聚焦于目标区域，消除背景杂波的影响。但总得来说，SAR图像目标识别的可解释性问题目前还没有引起足够关注和重视，亟须研究和解决。为此，如图7所示，本文从模型理解、模型诊断和模型改进等3个方面来探讨SAR图像目标识别的可解释性问题。

4.1 模型理解

由于SAR图像与光学图像的显著差异，从而需要提取深层网络针对SAR图像的特征模式，分析其内在的物理与语义含义，回答深度神经网络从SAR数据中学到了什么的问题，以及对特定样本是如何做决策的。对于模型的整体理解，可以考虑采用激活最大化原理，同时考虑SAR图像散射点的稀疏分布特性，设计散射增强的SAR图像深层特征反演方法，从而发现SAR图像深层特征的整体特征模式。例如考虑如下的特征反演模型

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} l(\boldsymbol{x}),\ \ l(\boldsymbol{x}) = -\langle \varphi(\boldsymbol{x}), \varphi_0 \rangle + \lambda R(\boldsymbol{x}),\quad \boldsymbol{x} \in \mathbb{R}_+^{W\times H\times C} \tag{2}$$

其中，$\boldsymbol{x}$为待求解的SAR特征模式图像，$R(\boldsymbol{x})$为表

示图像先验的正则化项，λ为正则化系数，$\varphi(\cdot)$表示神经网络函数。当$\varphi_0=\boldsymbol{e}_i$, $\boldsymbol{e}_i$是第i元素为1其余为0的指示向量，则式(2)表示最大化特征单元响应$[\varphi(\boldsymbol{x})]_i$。考虑到SAR图像散射点的稀疏分布特性[79]，采用如下正则化

$$R(\boldsymbol{x}) = \|\boldsymbol{x}\|_p^p + \|\boldsymbol{D}\boldsymbol{x}\|_p^p \tag{3}$$

其中，$\|\cdot\|_p$表示l_p范数，$\boldsymbol{D}$表示1阶微分算子。这样，可以采用随机子梯度(Stochastic sub-gradient)来最小化目标函数式(2)。

对于每一个样本的决策机理分析，可如图8基于SAR目标识别的样本决策重要性分析基于层次相关传播、敏感性分析等方法提取决策显著性区域，定位图像中的决策重要性区域，并进一步基于SAR散射中心模型提取决策显著性区域的散射中心参数，进而分析模型决策背后的物理含义。

4.2 模型诊断

深度神经网络因其强大的非线性参数拟合能力能够提升SAR图像目标识别的性能，特别是在类别确定、数据量有限、完全标注的情况下，但是这种性能的提升及其背后模型的决策逻辑在多大程度上具有合理性是存疑的。特别是在SAR训练数据不足、图像特性受观测参数影响变化大、存在固有斑点噪声等情况下，加之神经网络本身易受对抗样本攻击，因此可通过建立可解释的代理模型的方式来近似原神经网络模型的决策行为，以此发现模型所蕴含的新知识以及可能存在的缺陷和漏洞。

可考虑模型损失函数为$L(\boldsymbol{Z},\theta)=\ln p(\boldsymbol{Z}|\theta)$，这样可得到Fisher信息矩阵为

$$\boldsymbol{H} = \mathbb{E}_{p(\boldsymbol{Z})}\left[\frac{\partial \ln p(\boldsymbol{Z}|\theta)}{\partial\theta}\frac{\partial \ln p(\boldsymbol{Z}|\theta)^{\mathrm{T}}}{\partial\theta}\right] \tag{4}$$

其中，$\boldsymbol{Z}=[\boldsymbol{z}_i]_{i=1}^n$。定义基于Fisher信息矩阵的核函数为

$$k(\boldsymbol{z}_i,\boldsymbol{z}_j) = \|\boldsymbol{z}_i-\boldsymbol{z}_j\|_{\mathcal{H}} = \boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{p}_j \tag{5}$$

式中，$\boldsymbol{p}=\partial \ln p(\boldsymbol{z}|\theta)/\partial\theta|_{\theta=\hat{\theta}}$。这样基于Fisher核函数，可以使用高斯过程来近似原始神经网络的决策，并且可进一步地通过高斯过程选择对深层网络决策起到重要作用的原型(Prototype)样本，实现基于原型或者样例的解释。

高斯过程赋予决策函数f以高斯过程先验，即$f\sim\mathcal{GP}(0,k)$，其中k为核函数，给定训练集$\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}|_{i=1}^n, y_i=f(\boldsymbol{x}_i)$，这样$f$的后验概率均值为

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x},\boldsymbol{X})^{\mathrm{T}}\boldsymbol{K}^{-1}\tilde{\boldsymbol{y}} \tag{6}$$

方差为

$$\mathrm{var}(f) = k(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x},\boldsymbol{X})^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}) \tag{7}$$

其中，$\boldsymbol{k}(\boldsymbol{x},\boldsymbol{X})=[k(\boldsymbol{x},\boldsymbol{x}_1),k(\boldsymbol{x},\boldsymbol{x}_2),\cdots,k(\boldsymbol{x},\boldsymbol{x}_n)]^{\mathrm{T}}$, $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]^{\mathrm{T}}$, $\tilde{\boldsymbol{y}}=[\tilde{y}_1,\tilde{y}_2,\cdots,\tilde{y}_n]^{\mathrm{T}}$。借鉴模型蒸馏原理，这里并不直接采用原始训练数据的标记，而是采用神经网络的输出作为高斯过程的标记值。进一步，根据最小化后验估计方差来选择对模型决策起到重要作用且有代表性的原型(Prototype)

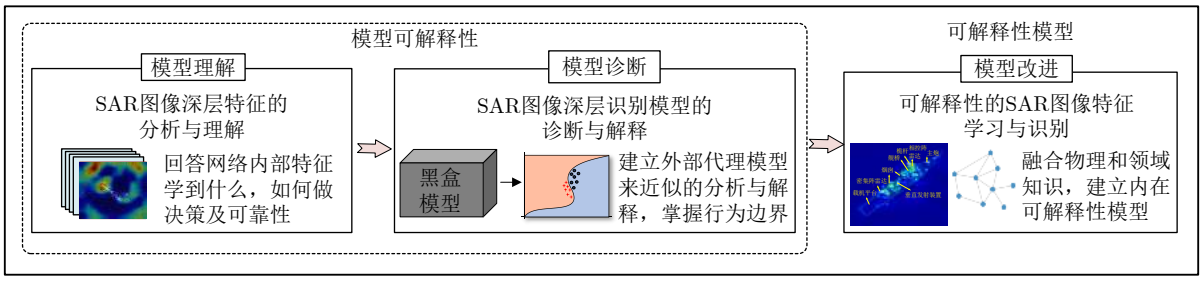


图 7 SAR目标识别可解释性研究

Fig. 7 Explainable SAR automatic target recognition

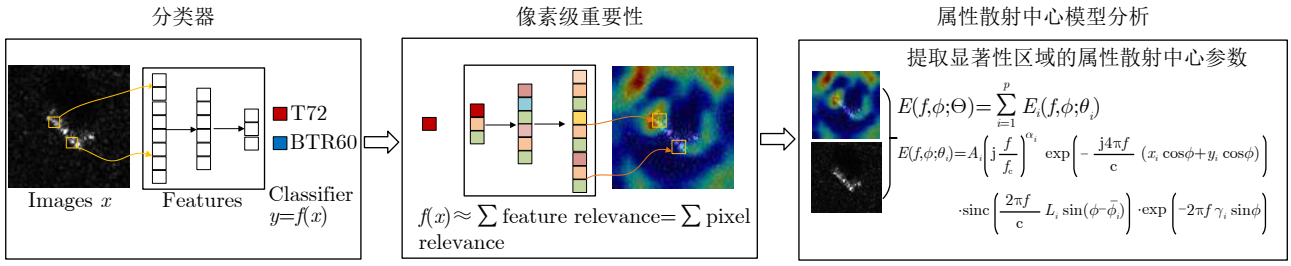


图 8 SAR目标识别的样本决策重要性分析

Fig. 8 Decision importance analysis for the SAR target recognition model

样本，对神经网络的结果进行解释和诊断。令原型样本集为$S_m=\{\boldsymbol{x}_1,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m\}$，原型样本的选择通过最小化如下的方差函数[80]

$$\begin{aligned}\mathrm{var}\left(Z(S_m)\right)=&\frac{1}{|\mathcal{D}|^2}\sum_{\boldsymbol{x}_i\in\mathcal{D}}\sum_{\boldsymbol{x}_j\in\mathcal{D}}k(\boldsymbol{x_i},\boldsymbol{x_j})\\&-\frac{1}{|\mathcal{D}|^2}\sum_{\boldsymbol{x}_i\in\mathcal{D}}\sum_{\boldsymbol{x}_j\in\mathcal{D}}\boldsymbol{k}(\boldsymbol{x_i},\boldsymbol{X}_{S_m})^{\mathrm{T}}\\&\cdot\boldsymbol{K}_{S_mS_m}^{-1}\boldsymbol{k}(\boldsymbol{X}_{S_m},\boldsymbol{x_j})\end{aligned}\tag{8}$$

式中，第1项相对于S_m是常数。为了最小化式(8)，基于序贯贝叶斯采样的贪心算法来对原始数据进行采样得到原型样本集$S_j=\{\boldsymbol{x}_1,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_j\}$，即

$$i_{j+1}^{*}=\underset{i\in\mathcal{D}\backslash S_j,S=S\cup\{i\}}{\arg\max}\ \boldsymbol{z}_S^{\mathrm{T}}\boldsymbol{K}_{SS}^{-1}\boldsymbol{z}_S\tag{9}$$

这样通过基于高斯过程回归模型的原型样本选择，可以理解模型的决策行为，对模型进行诊断，检测训练集中是否含有异常训练样本等。

4.3 模型改进

模型改进主要是指建立可解释的SAR图像目标识别模型。目前，SAR目标识别模型大多是借鉴光学图像中的模型和学习方法，这些模型虽具有一定的通用性，但主要是依赖大量标注数据拟合大量参数进行预测，忽略了SAR本身的特性和先验知识，对数据量和标注要求高，限制了模型性能的进一步提升。因此，需要考虑如何结合SAR特有的物理和语义知识，自然嵌入到神经网络模型中进行特征的学习和推理，建立兼具可解释性强和高性能的SAR图像特征学习和识别模型，同时也能降低对大量训练数据的依赖。例如图9，可结合SAR图像的属性散射中心模型，将SAR图像成像的物理机理融入到SAR特征学习过程中，通过属性散射中心模型引导模型学习更具物理意义的特征表示，进而增强模型本身的可解释性。

理论和实验表明，在高频区目标总的电磁散射可以看成由有限个局部散射源叠加而成，这些局部散射源称为散射中心。属性散射中心模型是基于几何绕射理论和物理光学理论提出的描述高频区复杂目标散射特性的参数模型[21]。假设目标的电磁散射响应可以认为是p个独立的散射中心叠加而成，具体形式为

$$E(f,\phi;\Theta)=\sum_{i=1}^{p}E_i(f,\phi;\theta_i)\tag{10}$$

$$\begin{aligned}E_i(f,\phi;\theta_i)=&A_i\left(\mathrm{j}\frac{f}{f_{\mathrm{c}}}\right)^{\alpha_i}\\&\cdot\exp\left(-\frac{\mathrm{j}4\pi f}{\mathrm{c}}\left(x_i\cos\phi+y_i\cos\phi\right)\right)\\&\cdot\mathrm{sinc}\left(\frac{2\pi f}{\mathrm{c}}L_i\sin(\phi-\bar{\phi}_i)\right)\\&\cdot\exp\left(-2\pi f\gamma_i\sin\phi\right)\end{aligned}\tag{11}$$

其中，$E(f,\phi;\Theta)$是目标总的散射场，其中$\Theta=\{\theta_i\}=[A_i,\alpha_i,x_i,y_i,L_i,\bar{\phi}_i,\gamma_i]$为目标散射中心的属性集。$x_i$,$y_i$为散射中心方位向、距离向的位置，$A_i$为幅度，$\alpha_i$为频率依赖因子，$L_i$为散射中心长度，$\bar{\phi}_i$为散射中心方位角，$\gamma_i$为散射中心对方位角的方向依赖性。当$L_i$=0,$\bar{\phi}_i=0$时，表示散射中心是局部散射中心，当$L_i\neq0,\gamma_i=0$时表示该散射中心是分布式散射中心。$E(f,\phi;\Theta)$是按照频率$f$和方位角$\phi$等间隔采样的极坐标格式，而通常目标识别是在图像域进行的，因而需要先将其转换到欧式坐标系下，再进行二维逆傅里叶变换2D-IFT得到SAR图像。注意到模型参数间量级相差较大，在参数估计时会存在收敛问题，因此可对式(11)进行规整化处理，得到规整化的属性散射中心模型[22]。

由此，本文设计一个深度网络结构从图像中直接回归目标的属性散射中心参数，物理知识引导的SAR特征学习网络如图9所示。首先，图像经过一个卷积神经神经网络特征提取层，然后经过属性散射参数回归层，得到5个特征图，每个特征图分别对应于5个属性散射中心参数$A,\alpha,\gamma,L,\bar{\phi}$，对应的位置参数$x_p,y_p$为特征图中每一个点对应于原输入图像中的坐标，可以根据卷积神经网络的感受野大

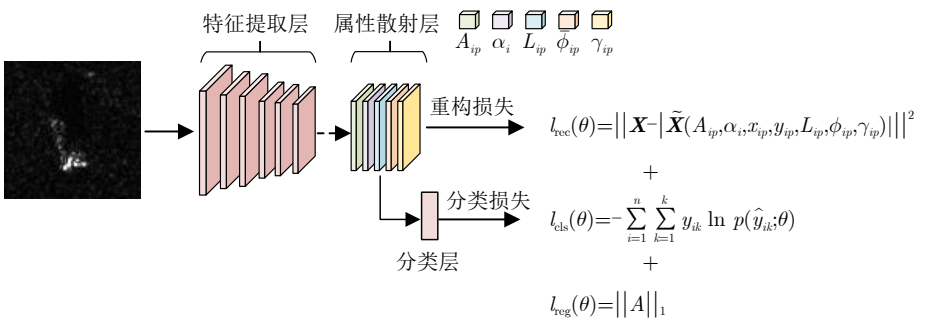


图 9 物理知识引导的SAR特征学习网络

Fig. 9 Physical model guided feature learning for SAR images

小和降采样系数确定。同时，考虑到散射中心分布的稀疏性，即对应于散射幅度特征图A是稀疏的。由此，得到网络训练的损失函数为

$$L(X,y;\theta)=l_{\mathrm{cls}}(y,\hat{y};\theta)+\lambda_1 l_{\mathrm{rec}}(X,\tilde{X};\theta)+\lambda_2 l_{\mathrm{reg}}(\theta) \tag{12}$$

其中，$l_{\mathrm{cls}}(y,\hat{y};\theta)$为交叉熵分类损失，$l_{\mathrm{rec}}(X,\tilde{X};\theta)$为基于属性散射中心模型的重构损失，$l_{\mathrm{reg}}(\theta)$为稀疏正则化损失。

$$l_{\mathrm{rec}}(\theta)=\left\|\boldsymbol{X}-\left|\tilde{\boldsymbol{X}}(A_{ip},\alpha_i,x_{ip},y_{ip},L_{ip},\bar{\phi}_{ip},\gamma_{ip})\right|\right\|^2 \tag{13}$$

和$l_{\mathrm{reg}}(\theta)=\|A\|_1$

由式(13)，通过引入属性散射模型重构损失，可以引导模型学习具有物理意义的特征，增强了模型的解释性。在式(12)中，去掉分类损失函数项，则可以通过散射模型的引导进行无监督的学习；如果数据中只含有部分标记的训练数据，利用式(12)则可以进行半监督的学习。这样通过物理模型引导学习的模型，不仅具有较强的解释性，而且能够降低对标记训练样本的依赖。

总得来说，可解释性问题是当前基于深度学习的SAR目标识别研究中还没有引起足够关注但亟待研究的一个关键问题，但目前还鲜有这方面的研究工作。本文对面向SAR图像目标识别的可解释问题，从模型理解、模型诊断和模型改进等方面进行了初步的探讨，提供了一些可能的研究思路，以启发SAR领域的研究人员进一步探索，以突破SAR图像目标认知解译的技术瓶颈。

5 总结和进一步工作

本文系统地总结和分析了当前SAR图像目标识别的研究进展以及在技术和应用中存在的重要挑战，对当前机器学习、深度学习的可解释性研究和主要方法进行了梳理和总结。深度学习的可解释性问题是当前人工智能领域的研究热点和难点，是实现可靠、可信和透明的人工智能系统的重要基础。目前面向深度学习的可解释性研究还处于初步探索阶段，研究人员针对不同任务需求、从不同角度对深度学习的可解释性进行了研究，但还没形成一个科学完整的体系。对基于深度学习的SAR图像目标识别的可解释性问题的研究，目前还鲜有这方面的研究工作。为此，本文从SAR目标认知解译所面临的挑战和技术瓶颈出发，强调开展SAR图像目标识别的可解释性问题的研究，从模型理解、模型诊断和模型改进等3个方面对SAR图像目标识别可解释问题进行了探讨，以启发研究人员对此问题开展进一步地探索和研究。这对于剖析SAR目标认知机理，提升SAR图像目标识别能力具有重要意义。在未来SAR目标识别的研究工作中，还可以以可解释性为切入点，从以下几个方面进一步发展SAR图像目标识别技术：

(1) 与SAR领域知识的有机融合。单纯依靠数据暴力的深度学习方法并不能完全解决SAR目标认知解译问题。目前，SAR图像目标解译仍严重依赖于判读专家的经验知识和推理，如何构建和表达这些先验知识，并嵌入到深度模型的学习和推理过程中，发展解释性强、泛化性好、鲁棒性高的SAR目标识别模型和方法是一个重要方向。

(2) 人机智能协同。由于SAR图像的特殊性，仅仅依赖图像信号并没有很大的提升空间，判读专家的经验和知识在SAR图像解译中仍不可或缺，因此需要将人的作用引入到模型学习和推理的环路中，充分发挥判读专家快速聚焦、语义关联、知识推理和计算机快速计算与自动化处理的两方面优势，实现有效的“人在环路”的人机协同计算。可解释性作为人与模型的接口，在人机智能协同中发挥着重要作用，通过建立可解释的智能交互SAR目标识别系统能够为突破当前SAR目标识别的技术和应用瓶颈提供一条切实可行的途径。

(3) 交互式学习。SAR图像视觉认知困难，图像特性不稳定，获取和标注样本难，良好的SAR目标识别系统应具备从小样本数据学习以及在与人的交互过程中持续学习的能力。通过结合主动学习、增量学习、知识图谱等技术，在人-模型-数据的动态互动中进行渐进式模型训练和知识更新，从而实现系统能力的迭代增长。

参 考 文 献

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336–359. doi: 10.1007/s11263-019-01228-7.

[7] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[C]. 2015 International Conference on Learning Representations, San Diego, USA, 2015.

[8] 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.

[9] 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.

[10] GUIDOTTI R, MONREALE A, RUGGIERI S, *et al*. A survey of methods for explaining black box models[J]. *ACM Computing Surveys*, 2018, 51(5): 93. doi: 10.1145/3236009.

[11] NOVAK L M, OWIRKA G J, and NETISHEN C M. Performance of a high-resolution polarimetric SAR automatic target recognition system[J]. *The Lincoln Laboratory Journal*, 1993, 6(1): 11–23.

[12] GAO Gui. Statistical modeling of SAR images: A survey[J]. *Sensors*, 2010, 10(1): 775–795. doi: 10.3390/s100100775.

[13] 高贵. SAR图像统计建模研究综述[J]. 信号处理, 2009, 25(8): 1270–1278. doi: 10.3969/j.issn.1003-0530.2009.08.019.
GAO Gui. Review on the statistical modeling of SAR images[J]. *Signal Processing*, 2009, 25(8): 1270–1278. doi: 10.3969/j.issn.1003-0530.2009.08.019.

[14] 郭炜炜. SAR图像目标分割与特征提取[D]. [硕士论文], 国防科学技术大学, 2007: 28–35.
GUO Weiwei. SAR image target segmentation and feature extraction[D]. [Master dissertation], National University of Defense Technology, 2007: 28–35.

[15] HUAN Ruohong and YANG Ruliang. SAR target recognition based on MRF and gabor wavelet feature extraction[C]. 2008 IEEE International Geoscience and Remote Sensing Symposium, Boston, USA, 2008: II-907–II-910. doi: 10.1109/igarss.2008.4779142.

[16] PAPSON S and NARAYANAN R M. Classification via the shadow region in SAR imagery[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2012, 48(2): 969–980. doi: 10.1109/taes.2012.6178042.

[17] CASASENT D and CHANG W T. Correlation synthetic discriminant functions[J]. *Applied Optics*, 1986, 25(14): 2343–2350. doi: 10.1364/ao.25.002343.

[18] ZHAO Q and PRINCIPE J C. Support vector machines for SAR automatic target recognition[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2001, 37(2): 643–654. doi: 10.1109/7.937475.

[19] SUN Yijun, LIU Zhipeng, TODOROVIC S, *et al*. Adaptive boosting for SAR automatic target recognition[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2007, 43(1): 112–125. doi: 10.1109/taes.2007.357120.

[20] SUN Yongguang, DU Lan, WANG Yan, *et al*. SAR automatic target recognition based on dictionary learning and joint dynamic sparse representation[J]. *IEEE Geoscience and Remote Sensing Letters*, 2016, 13(12): 1777–1781. doi: 10.1109/lgrs.2016.2608578.

[21] POTTER L C and MOSES R L. Attributed scattering centers for SAR ATR[J]. *IEEE Transactions on Image Processing*, 1997, 6(1): 79–91. doi: 10.1109/83.552098.

[22] 计科峰, 匡纲要, 粟毅, 等. 基于SAR图像的目标散射中心特征提取方法研究[J]. 国防科技大学学报, 2003, 25(1): 45–50. doi: 10.3969/j.issn.1001-2486.2003.01.010.
JI Kefeng, KUANG Gangyao, SU Yi, *et al*. Research on the extracting method of the scattering center feature from SAR imagery[J]. *Journal of National University of Defense Technology*, 2003, 25(1): 45–50. doi: 10.3969/j.issn.1001-2486.2003.01.010.

[23] 丁柏圆, 文贡坚, 余连生, 等. 属性散射中心匹配及其在SAR目标识别中的应用[J]. 雷达学报, 2017, 6(2): 157–166. doi: 10.12000/JR16104.
DING Baiyuan, WEN Gongjian, YU Liansheng, *et al*. Matching of attributed scattering center and its application to synthetic aperture radar automatic target recognition[J]. *Journal of Radars*, 2017, 6(2): 157–166. doi: 10.12000/JR16104.

[24] JONES III G and BHANU B. Recognizing articulated objects in SAR images[J]. *Pattern Recognition*, 2001, 34(2): 469–485. doi: 10.1016/s0031-3203(99)00218-6.

[25] MAO Xiaojiao, SHEN Chunhua, and YANG Yubin. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections[C]. The 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 2810–2818.

[26] DONG Chao, LOY C C, HE Kaiming, *et al*. Image super-resolution using deep convolutional networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(2): 295–307. doi: 10.1109/tpami.2015.2439281.

[27] LIU Li, OUYANG Wanli, WANG Xiaogang, *et al*. Deep learning for generic object detection: A survey[J]. *International Journal of Computer Vision*, 2020, 128(2): 261–318. doi: 10.1007/s11263-019-01247-4.

[28] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 770–778.

[29] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 834–848. doi: 10.1109/tpami.2017.2699184.

[30] 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.

[31] 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.

[32] 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 and Information Technology*, 2020, 42(1): 119–131. doi: 10.11999/JEIT180899.

[33] ZHAO Juanping, ZHANG Zenghui, YU Wenxian, *et al*. A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images[J]. *IEEE Access*, 2018, 6: 50693–50708. doi: 10.1109/access.2018.2869289.

[34] 陈慧元, 刘泽宇, 郭炜炜, 等. 基于级联卷积神经网络的大场景遥感图像舰船目标快速检测方法[J]. 雷达学报, 2019, 8(3): 413–424. doi: 10.12000/JR19041.

CHEN Huiyuan, LIU Zeyu, GUO Weiwei, *et al*. Fast detection of ship targets for large-scale remote sensing image based on a cascade convolutional neural network[J]. *Journal of Radars*, 2019, 8(3): 413–424. doi: 10.12000/JR19041.

[35] WAGNER S. Combination of convolutional feature extraction and support vector machines for radar ATR[C]. The 17th International Conference on Information Fusion (FUSION), Salamanca, Spain, 2014: 1–6.

[36] WAGNER S A. SAR ATR by a combination of convolutional neural network and support vector machines[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2016, 52(6): 2861–2872. doi: 10.1109/taes.2016.160061.

[37] 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.

[38] 赵娟萍, 郭炜炜, 柳彬, 等. 基于概率转移卷积神经网络的含噪标记SAR图像分类[J]. 雷达学报, 2017, 6(5): 514–523. doi: 10.12000/JR16140.

ZHAO Juanping, GUO Weiwei, LIU Bin, *et al*. Convolutional neural network-based sar image classification with noisy labels[J]. *Journal of Radars*, 2017, 6(5): 514–523. doi: 10.12000/JR16140.

[39] GUNNING D. EXplainable Artificial Intelligence (XAI)[R]. DARPA/I2O, 2017.

[40] ADADI A and BERRADA M. Peeking inside the black-box: A survey on EXplainable Artificial Intelligence (XAI)[J]. *IEEE Access*, 2018, 6: 52138–52160. doi: 10.1109/access.2018.2870052.

[41] LIPTON Z C. The mythos of model interpretability[J]. *Communications of the ACM*, 2018, 61(10): 36–43. doi: 10.1145/3233231.

[42] ZHANG Quanshi and ZHU Songchun. Visual interpretability for deep learning: A survey[J]. *Frontiers of Information Technology & Electronic Engineering*, 2018, 19(1): 27–39. doi: 10.1631/fitee.1700808.

[43] MAHENDRAN A and VEDALDI A. Visualizing deep convolutional neural networks using natural pre-images[J]. *International Journal of Computer Vision*, 2016, 120(3): 233–255. doi: 10.1007/s11263-016-0911-8.

[44] NGUYEN A, CLUNE J, BENGIO Y, *et al*. Plug & play generative networks: Conditional iterative generation of images in latent space[J]. arXiv: 1612.00005, 2016.

[45] KIM B, WATTENBERG M, GILMER J, *et al*. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV)[J]. arXiv: 1711.11279, 2017.

[46] FROSST N and HINTON G. Distilling a neural network

into a soft decision tree[J]. arXiv: 1711.09784, 2017.

[47] ALTMANN A, TOLOŞI L, SANDER O, *et al*. Permutation importance: A corrected feature importance measure[J]. *Bioinformatics*, 2010, 26(10): 1340–1347. doi: 10.1093/bioinformatics/btq134.

[48] SIMONYAN K, VEDALDI A, and ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv: 1312.6034, 2013.

[49] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, *et al*. Striving for simplicity: The all convolutional net[J]. arXiv: 1412.6806, 2014.

[50] SUNDARARAJAN M, TALY A, and YAN Qiqi. Gradients of counterfactuals[J]. arXiv: 1611.02639, 2016.

[51] SMILKOV D, THORAT N, KIM B, *et al*. SmoothGrad: Removing noise by adding noise[J]. arXiv: 1706.03825, 2017.

[52] FONG R, PATRICK M, and VEDALDI A. Understanding deep networks via extremal perturbations and smooth masks[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019: 2950–2958. doi: 10.1109/iccv.2019.00304.

[53] BACH S, BINDER A, MONTAVON G, *et al*. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. *PLoS One*, 2015, 10(7): e0130140. doi: 10.1371/journal.pone.0130140.

[54] 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.

[55] RIBEIRO M, SINGH S, and GUESTRIN C. "Why should I trust you?": Explaining the predictions of any classifier[C]. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, USA, 2016: 97–101. doi: 10.18653/v1/n16-3020.

[56] KOH P W and LIANG P. Understanding black-box predictions via influence functions[C]. The 34th International Conference on Machine Learning, Sydney, Australia, 2017: 1885–1894.

[57] KIM B, KHANNA R, and KOYEJO O. Examples are not enough, learn to criticize! Criticism for Interpretability[C]. The 30th Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 2280–2288.

[58] LUNDBERG S M and LEE S I. A unified approach to interpreting model predictions[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 4768–4777.

[59] ZHANG Quanshi, YANG Yu, MA Haotian, *et al*. Interpreting CNNs via decision trees[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 6254–6263. doi: 10.1109/cvpr.2019.00642.

[60] DU Mengnan, LIU Ninghao, SONG Qingquan, *et al*. Towards explanation of DNN-based prediction with guided feature inversion[C]. The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 2018: 1358–1367.

[61] ZEILER M D and FERGUS R. Visualizing and understanding convolutional networks[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 818–833.

[62] SAMEK W, BINDER A, MONTAVON G, *et al*. Evaluating the visualization of what a deep neural network has learned[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28(11): 2660–2673. doi: 10.1109/tnnls.2016.2599820.

[63] NAM W J, GUR S, CHOI J, *et al*. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks[C]. The 34th Conference on Artificial Intelligence (AAAI), New York, USA, 2020: 2501–2508.

[64] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. *Nature Machine Intelligence*, 2019, 1(5): 206–215. doi: 10.1038/s42256-019-0048-x.

[65] XU K, BA J L, KIROS R, *et al*. Show, attend and tell: Neural image caption generation with visual attention[C]. The 32nd International Conference on Machine Learning(ICML), Lille, France, 2015: 2048–2057.

[66] GREGOR K and LECUN Y. Learning fast approximations of sparse coding[C]. The 27th International Conference on Machine Learning, Haifa, Israel, 2010: 399–406.

[67] ZHENG Shuai, JAYASUMANA S, ROMERA-PAREDES B, *et al*. Conditional random fields as recurrent neural networks[C]. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015: 1529–1537. doi: 10.1109/iccv.2015.179.

[68] PENG Xi, TSANG I W, ZHOU J T, *et al*. K-meansNet: When k-means meets differentiable programming[J]. arxiv: 1808.07292, 2018.

[69] ZHU Hongyuan, PENG Xi, Chandrasekhar V, *et al*. DehazeGAN: When image dehazing meets differential programming[C]. The 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 1234–1240.

[70] KARPATNE A, WATKINS W, READ J, *et al*. Physics-guided Neural Networks (PGNN): An application in lake temperature modeling[J]. arxiv: 1710.11431, 2017.

[71] CHEN Tianshui, XU Muxin, HUI Xiaolu, *et al*. Learning semantic- specific graph representation for multi-label image recognition[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 522–531.

[72] CHU Lingyang, HU Xia, HU Juhua, *et al*. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution[C]. The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 2018: 1244–1253.

[73] BAU D, ZHOU Bolei, KHOSL A, *et al*. Network dissection: Quantifying interpretability of deep visual representations[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 3319–3327. doi: 10.1109/cvpr.2017.354.

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

[75] 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.

[76] 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.

[77] CHEN Lifu, TAN Siyu, PAN Zhouhao, *et al*. A new framework for automatic airports extraction from SAR images using multi-level dual attention mechanism[J]. *Remote Sensing*, 2020, 12(3): 560. doi: 10.3390/rs12030560.

[78] LI Chen, DU Lan, DENG Sheng, *et al*. Point-wise discriminative auto-encoder with application on robust radar automatic target recognition[J]. *Signal Processing*, 2020, 169: 107385. doi: 10.1016/j.sigpro.2019.107385.

[79] CETIN M, KARL W C, and CASTANON D A. Feature enhancement and ATR performance using nonquadratic optimization-based SAR imaging[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2003, 39(4): 1375–1395. doi: 10.1109/taes.2003.1261134.

[80] KHANNA R, KIM B, GHOSH J, *et al*. Interpreting black box predictions using fisher kernels[C]. The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), Okinawa, Japan, 2019: 3382–3390.

作 者 简 介

郭炜炜(1983–)，男，江苏南通人，博士，分别于2005，2007，2014年获得国防科技大学信息工程学士，信息与通信专业硕士和博士学位。2008年—2010年在英国Queen Mary，University of London联合培养，2014年12月至2018年6月在上海交通大学电子信息与电气工程学院从事博士后研究工作，2018年12月至今为同济大学设计创意学院助理教授。研究方向为遥感图像理解、模式识别与机器学习、人机交互等。

E-mail: weiweiguo@tongji.edu.cn

张增辉(1980–)，男，山东金乡人，博士，分别于2001年、2003年和2008年在国防科技大学获得应用数学、计算数学、信息与通信工程专业学士、硕士和博士学位。2008年6月至2013年7月，为国防科技大学数学与系统科学系讲师；2014年2月至今，为上海交通大学电子信息与电气工程学院副研究员。研究方向为SAR图像解译、雷达信号处理等。

E-mail: zenghui.zhang@sjtu.edu.cn

郁文贤(1964–)，男，上海松江人，博士，教授，博士生导师，上海交通大学讲席教授，教育部长江学者特聘教授，上海市领军人才。现为上海交通大学信息技术与电气工程研究院院长，北斗导航与位置服务上海市重点实验室主任，智能探测与识别上海市高校重点实验室主任。研究方向为遥感信息处理、多源融合导航定位、目标检测识别等。

E-mail: wxyu@sjtu.edu.cn

孙效华(1972–)，女，河南安阳人，麻省理工学院设计与计算专业硕士与博士，教授，博士生导师，同济大学设计创意学院副院长。曾在MIT CECI、MIT媒体实验室、FXPAL、IBM研究院、美国克拉克森大学等机构从事研究与教学。研究方向为人机智能交互与共融、人-机器人交互HRI、可视分析等。

E-mail: xsun@tongji.edu.cn