多源遥感影像深度识别模型对抗攻击鲁棒性评估

孙浩[1]，徐延杰[1]，陈进[2]，雷琳[1]，计科峰[1]，匡纲要[1]

1. 国防科技大学 电子科学学院, 长沙 410073;
2. 北京市遥感信息研究所, 北京 100192

摘 要：基于深度神经网络的多源遥感影像目标识别系统已逐步在空天遥感情报侦察、无人作战自主环境认知、多模复合末制导等多个军事场景中广泛应用。然而，由于深度学习理论上的不完备性、深度神经网络结构设计工程上的强复用性、以及多源成像识别系统在复杂电磁环境中易受到各类干扰等多因素的影响，使得现有识别系统在对抗攻击鲁棒性方面评估不足，存在极大安全隐患。本文首先从深度学习理论不完备性和识别系统攻击样式两个方面分析了潜在安全风险，并重点介绍了深度识别模型对抗样本攻击基本原理和典型方法。其次，针对光学遥感影像和SAR遥感影像两类典型数据形式，从鲁棒正确识别率和对抗攻击可解释性两个方面开展多源遥感影像深度识别模型对抗攻击鲁棒性评估，覆盖了9类常见深度识别网络架构和7类典型对抗样本攻击方法，验证了现有深度识别模型对抗攻击鲁棒性普遍不足的问题，分析了对抗样本与正常样本的多隐层特征激活差异，为下一步设计对抗样本检测算法和提升模型对抗鲁棒性提供参考。

关键词：多源遥感影像目标识别，深度神经网络，对抗攻击，特征可视化，对抗鲁棒性评估

中图分类号：P2

引用格式：孙浩,徐延杰,陈进,雷琳,计科峰,匡纲要.2023.多源遥感影像深度识别模型对抗攻击鲁棒性评估.遥感学报,27(8):1951-1963

Sun H, Xu Y J, Chen J, Lei L, Ji K F and Kuang G Y. 2023. Adversarial robustness evaluation of multiple-source remote sensing image recognition based on deep neural networks. National Remote Sensing Bulletin, 27(8):1951-1963[DOI:10.11834/jrs.20210597]

1 引 言

随着成像方式的多样化以及数据获取能力的增强，空天地一体化对地观测网已经积累并将持续获取大量不同时—空—谱尺度的多源（多光谱/高光谱、雷达、多时相、多角度等）遥感影像，例如目前中国高分重大专项数据中心日生产数据超过25 TB，数据存储能力超过20 PB（童旭东，2016）；德国宇航中心（DLR）的多任务数据存储已经达到2 PB，欧洲航天局（ESA）的累积数据量在2020年达到20 PB（Kurte等，2017）。大规模多源遥感影像的持续获取和不断进化的深度学习技术为推动多源遥感影像智能解译及应用提供了良好的数据支撑和技术支撑（Cheng等，2020；Zhu等，2020）。基于深度神经网络的多源遥感影像目标识别方法近年来取得巨大进步，已经达到或超越人类的认知水平，并逐步在空天遥感情报侦察、无人作战自主环境认知、多模复合末制导等多个军事场景中广泛应用。

然而，由于深度学习理论本身的不完备性（Szegedy等，2014；Fawzi等，2017）、深度识别网络结构设计与优化方法的强复用性、以及多源成像识别系统在复杂电磁环境中易受到各类干扰等多因素的影响，使得现有识别系统在对抗攻击鲁棒性方面评估不足，给其在军事场景和安全敏感领域的深入应用和广泛部署带来巨大隐患。作为一种信息计算组件，多源遥感影像深度识别系统可能不会按预期程序工作或成为攻击者的目标（Berghoff等，2020；Sun等，2021；Wiyatno等，2019；Xu等，2021；Yuan等，2019）。对于意外故障和恶意对抗的鲁棒性是决定识别算法在现实物理世界中成功与否的关键因素。

收稿日期：2021-01-09；**预印本**：2021-11-09

基金项目：国家自然科学基金(编号：61971426)

第一作者简介：孙浩,研究方向为多源遥感影像协同解译、对抗机器学习等。E-mail：sunhao@nudt.edu.cn

2 遥感影像深度识别系统安全风险

2.1 多源遥感影像深度学习理论不完备性

深度神经网络模型具有处理不同模态信息的能力，其在处理不同模态信息时模型结构上的相似性及层次化的分析方法，为建立多源遥感影像特征表示模型提供了有力工具。深度神经网络已在多源遥感影像地物覆盖信息提取、目标检测、场景分类及图像检索等多个方面获得大量应用（Cheng等，2020；Zhu等，2020）。总结来看：基于深度神经网络的多源遥感数据分析的输入数据通常包括原始图像数据、多尺度图像块、对象分割结果、空间光谱滤波数据、空间地理信息等典型形式；深度网络模型包括卷积神经网络、全连接神经网络、递归神经网络、深度自编码器、深度置信网络、对抗生成网络等；网络架构包括单个识别网络的多尺度与小目标优化、多源与多谱段分组网络、多任务耦合分支网络、学生教师蒸馏网络、多网络集成等结构；并采用数据增广、参数迁移、自监督对比学习等策略解决大规模语义标注数据缺乏条件下的网络参数学习与优化。

然而，现有的多源遥感影像深度语义学习方法只有在训练数据和测试数据均来自同一特征空间且具有相同分布这一普遍假设下才有效。复杂动态对抗场景条件下，由于成像系统和成像过程引入的影像内容降质、地表时空变化造成的数据分布漂移、敌我伪装与干扰等多种因素影响，上述假设往往因为过于严格而难以成立，造成模型泛化能力大幅度下降。深度识别模型预测的不确定性和分布外数据泛化能力等问题在理论上还有待进一步深入研究。

多源遥感影像深度识别系统不能向操作人员解释其决策过程。许多安全敏感领域的高风险意味着深度识别系统必须透明，以取得决策者的信任并便于进行风险分析。然而现有的深度神经网络识别技术都是缺乏足够透明性的黑盒，模型的可解释性不足。此外，多源遥感影像深度识别网络架构设计及参数优化具有强通用性和高迁移性等特点，在工程上降低了恶意攻击的难度。

2.2 多源遥感影像深度识别系统潜在攻击面

面向高动态复杂电磁对抗军事应用场景，基于深度神经网络计算模型的多源遥感影像智能化识别系统面临着诸多安全性和可靠性的挑战（孙浩 等，2021）。根据数据处理过程和信息流向来看，多源遥感影像目标识别系统通常包括影像数据采集、数据预处理、深度识别模型训练与验证、深度识别模型推理与部署等关键步骤，每个步骤都可能面临潜在攻击，如图1所示。（1）采集阶段攻击：攻击者采用特定的方式对传感器的数据采集环节进行干扰，从系统输入源头发动攻击。例如光电成像系统中，攻击者可以通过布置特定的数字反光阵列，进行光路攻击；在雷达成像系统中，攻击者可以设置不同的干扰样式进行对抗压制干扰和欺骗干扰。（2）预处理阶段攻击：多平台传感器采集的大规模原始影像数据在进行目标识别和信息提取前，还需要进行一系列的校正和预处理操作，各种预处理过程中会产生新的安全风险。例如针对识别模型的重采样攻击，只有在特定空间分辨率时，攻击的信息才会被发现。（3）模型训练阶段攻击：典型攻击形式有数据投毒攻击和后门攻击。数据投毒攻击是指在模型预训练过程中添加恶意样本，或是在模型在线微调过程中设置混淆性场景，影响模型参数学习过程。后门攻击通常采用预训练模型投毒或开源框架代码修改等形式实现。（4）模型推理阶段攻击：针对数据建模在理论上存在的不完备性和模型学习过程中数据覆盖的不全面性进行攻击，设计针对性算法，通过估计模型的梯度参数和决策边界信息，迭代优化实现模型攻击。典型形式有逃避攻击、模仿攻击和模型逆向窃取攻击等。逃避攻击的代表性工作是针对深度神经网络的对抗样本生成技术。（5）模型部署阶段攻击：典型攻击形式有面向设备的硬件攻击、操作系统攻击和自动化复合攻击等。大量支持神经网络加速的开源软硬件平台的底层安全问题还没有得到充分的验证，给系统部署带来巨大安全风险。

3 多源遥感影像深度神经网络识别模型对抗样本攻击

3.1 深度识别模型对抗样本攻击

多源遥感影像深度识别系统潜在攻击最常见的一种形式是针对深度神经网络模型的对抗样本生成（Xu等，2021；Berghoff等，2020）。现有研

究表明：深度神经网络模型存在对抗脆弱性，对正常输入的样本图像，添加人眼无法察觉的微小扰动，可以造成模型以高置信度给出错误预测。假设$\boldsymbol{X}$表示输入的训练影像数据集合，$\boldsymbol{x}$表示其中的一个样本，通常是张量数据形式，例如多波段光学遥感影像或不同极化方式的微波遥感影像，$\boldsymbol{y}$表示$\boldsymbol{x}$对应的真实类别标记，$f(\boldsymbol{x};\ \boldsymbol{\theta})$表示参数为$\boldsymbol{\theta}$的深度神经网络。对抗样本通常定义为$\boldsymbol{x}'$，在欺骗目标模型$f$的同时保持对抗样本$\boldsymbol{x}'$和正常样本$\boldsymbol{x}$之间的差异在距离测度$d(\cdot)$下非常小，其中$d(\cdot)$通常采用$L_p$范数距离。对抗样本需同时满足式（1）：

$$d(\boldsymbol{x}',\boldsymbol{x})<\varepsilon \quad \text{且}\ \hat{\boldsymbol{y}}(\boldsymbol{x}')\neq\hat{\boldsymbol{y}}(\boldsymbol{x}) \tag{1}$$

式中，ε是一个很小的常数，用于限制扰动的幅度；$\hat{\boldsymbol{y}}(\cdot)$表示深度神经网络模型的预测标记，即$\hat{\boldsymbol{y}}(\boldsymbol{x})=\arg\max_c f(\boldsymbol{x};\ \boldsymbol{\theta})_{(c)}$，$c$是类别标记索引。计算机视觉领域光学影像标准数据集上，为了防止人类感知到扰动，通常限制扰动常数ε在区间［1/255，8/255］范围内。对红外遥感影像和微波遥感影像而言，扰动范围可以更大。图2可以看出，只对合成孔径雷达遥感影像添加了一个极小的扰动（为了显示效果，图中的对抗噪声进行了尺度放大）就轻松骗过了神经网络识别模型，将合成孔径雷达遥感影像车辆目标类别由BTR70识别为ZIL131。

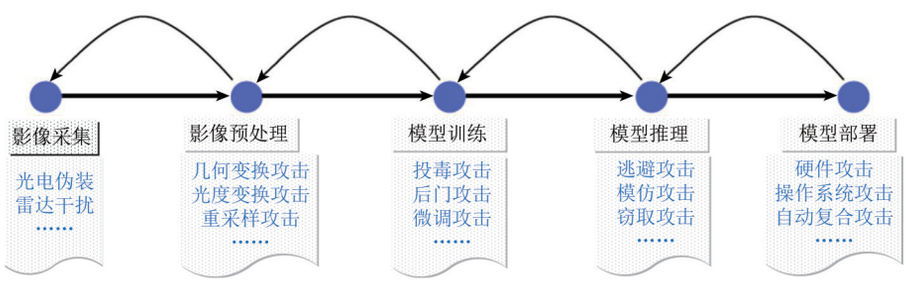


图1 多源遥感影像深度识别系统潜在攻击面

Fig. 1 Attack surface for deep learning based multiple source remote sensing images recognition

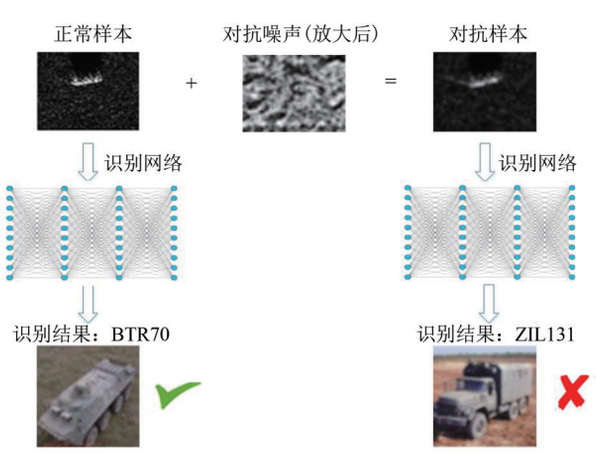


图2 遥感影像深度识别网络对抗样本示例

Fig. 2 Adversarial example for deep neural network based remote sensing image recognition

基于对抗样本的攻击通常可以分为白盒攻击和黑盒攻击（Wiyatno等，2019）。白盒攻击场景下，攻击者已知待攻击目标模型的网络结构、权重参数、训练过程或训练数据的相关信息；黑盒攻击场景下，攻击者仅能获取待攻击目标模型的外部输出信息，无法得到模型内部信息。对于军事应用和其他安全敏感领域来说，黑盒攻击更具危险性和现实性，因为多数情况下，攻击者很难获取模型的内部知识。评估深度神经网络模型对白盒攻击的鲁棒性是衡量模型在最坏情况下性能表现，同样具有重要意义。

基于对抗样本的攻击在信息链路对抗中属于模型推理阶段攻击，对抗样本攻击是逃避式攻击的代表性形式。以多源遥感影像军事侦察应用场景为例，应用流程上，首先是在数字域通过研究对抗样本生成技术，寻找具有强对抗性、高迁移性的扰动模式；然后在对应的物理域实现扰动模式，并将扰动模式合理地与待探测识别对象混合，实现智能化伪装，达到自动化反识别的目的。

3.2 对抗样本攻击基本原理与典型方法

对抗样本生成方法近几年发展迅速，研究成果非常丰富，但方法背后的基本原理变化不大。白盒攻击方法的基本原理可归类为基于梯度的优化、基于受限约束的优化、基于统计模型的对抗生成、基于敏感性分析的对抗生成等；黑盒攻击方法的基本原理可归类为梯度近似或决策边界近

似。此外诸如自监督学习、因果学习和注意力机制等新型学习机制也不断被引入到白盒优化或黑盒近似问题的求解过程。

（1）快速梯度符号攻击FGSM（Fast Gradient Sign Method）（Goodfellow等，2015）：基本原理是对于一个给定输入样本，快速寻找一个扰动方向，使得待攻击模型的训练损失函数增大，以实现减小分类置信度和增大类别间的混淆度。需要注意的是理论上无法保证增加训练函数损失一定会导致误分类。FGSM首先计算损失函数对于输入的梯度，然后对于梯度的符号向量乘以一个很小的常数ε产生扰动。

$$\boldsymbol{x}' = \boldsymbol{x} + \varepsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y})) \tag{2}$$

式中，$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y})$是损失函数对于输入$\boldsymbol{x}$的一阶导数。对于深度神经网络，可以通过后向传播算法计算。实际操作过程中，产生的对抗样本必须在输入空间范围内，需要进行数值重映射。

（2）投影梯度下降攻击PGD（Project Gradient Descent）（Madry等，2018）：FGSM方法的扩展，给定输入扰动的约束，多次迭代，每次一小步，并且迭代后把扰动重映射到原始数值区间。

$$\boldsymbol{x}'_{i+1} = \prod_{B(\boldsymbol{x},\varepsilon)} \left(\boldsymbol{x}'_i - \alpha \cdot \mathrm{sign}\left(\nabla_{\boldsymbol{x}'_i} L(\boldsymbol{x}'_i, \boldsymbol{y}) \right) \right) \tag{3}$$

式中，i是迭代次数索引，$0 < \alpha < \varepsilon$是迭代扰动幅度，$\prod_{B(\boldsymbol{x},\ \varepsilon)}$表示约束扰动范围的重映射函数，PGD在约束范围内随机选择一个初始化样本点进行迭代。

（3）C&W攻击（Carlini and Wagner Attacks）（Carlini和Wagner，2017）：基于受限约束优化的攻击方案，通过求解经验损失函数寻找潜在对抗样本。

$$L_{CW}(\boldsymbol{x}', t) = \max\left(\max_{i \neq t} \left\{ Z(\boldsymbol{x}')_{(i)} - Z(\boldsymbol{x}')_{(t)}, \kappa \right\} \right) \tag{4}$$

式中，$Z(\boldsymbol{x}')_{(i)}$表示表示对抗样本输入神经网络后logit输出，t表示目标类别，κ表示对抗样本的最小期望置信边界，用于约束对抗扰动幅度。

（4）深度欺骗攻击DF（Deep Fool Algorithm）（Moosavi-Dezfooli等，2016）：估计输入样本到多类分类器的最近决策边界距离，该距离既可以作为模型鲁棒性的度量，也可以用于最小扰动方向。对于非线性多类神经网络，算法通过线性化每个类别的决策边界，迭代扰动输入样本，在多个网络架构攻击中都取得了较好的攻击效果。

（5）弹性网络攻击EAD（Elastic-Net Attack to Deep Neural Network）（Chen等，2018）：基于约束条件下优化搜索的弹性白盒攻击，使用C&W攻击目标函数，与之前采用L_1或L_2范数正则化项不同，加入弹性网络正则化项，同时利用L_1和L_2范数进行正则化。

（6）多类型决策估计攻击HSJA（Hop Skip Jump Attack）（Chen等，2020）：一种高效的决策近似黑盒攻击方法，仅依赖对模型决策的访问，估计决策边界处的梯度方向，并设计一系列攻击算法。迭代型算法，每个迭代步骤中包括3项内容：梯度方向估计、几何级数步长搜索和二分法边界搜索。

（7）自监督扰动攻击SSP（Self Supervised Perturbation）（Naseer等，2020）：引入自监督学习的新型攻击方法，核心思想是特征层的失真造成决策错误，无需标记信息，具有跨任务、跨模型的强迁移特性。

3.3 对抗攻击特征扰动可解释性

深度神经网络的可解释性研究方法可分为可视化技术、模型蒸馏技术和内部机制（Ras等，2020）。本文关注基于可视化技术的深度神经网络对抗扰动分析。可视化方法思想是分析神经网络模型特征与决策间的关联度。可视化方法最常见的解释形式是显著性图，展示了输入特征影响模型输出的重要程度。可视化方法主要分为两大类：基于反向传播的可视化和基于遮挡结果分析的可视化。（1）基于反向传播的可视化方法：基于网络训练过程中从输出回传到输入的梯度信息鉴别输入特征的显著性。常用的处理策略包括激活最大化、反卷积、类别激活图、逐层相关性传播等。（2）基于遮挡结果分析的可视化方法：通过对输入图像进行特定方式的遮挡或者加噪，观察输出结果的变化，进而评估输入特征的显著性，如图3所示。

采用基于反向传播的梯度加权类别激活图Grad-CAM（Gradient-weighted Class Activation Map）（Selvaraju等，2017）方法对正常样本和对抗样本在不同模型、不同深度时的特征激活情况进行可视化分析，以期得到更多对抗样本欺骗网络决策的直观理解。类别激活图CAM（Class Activation Map）（Zhou等，2016）通过在卷积神经网络架构中采用全局平均池化GAP（Global Average Pooling）实现，通常的配置是GAP（Conv）层→FC层→

softmax层，全连接层FC有C个节点，每个节点代表一个类别。CAM方法将来自卷积层Conv的激活$\boldsymbol{A}$与全连接层的权重$\boldsymbol{w}_{k,\,c}$结合构建相关性得分图，其中卷积层包括K个卷积滤波器。

$$map_c = \sum_k^K \boldsymbol{w}_{k,c}\boldsymbol{A}_k \tag{5}$$

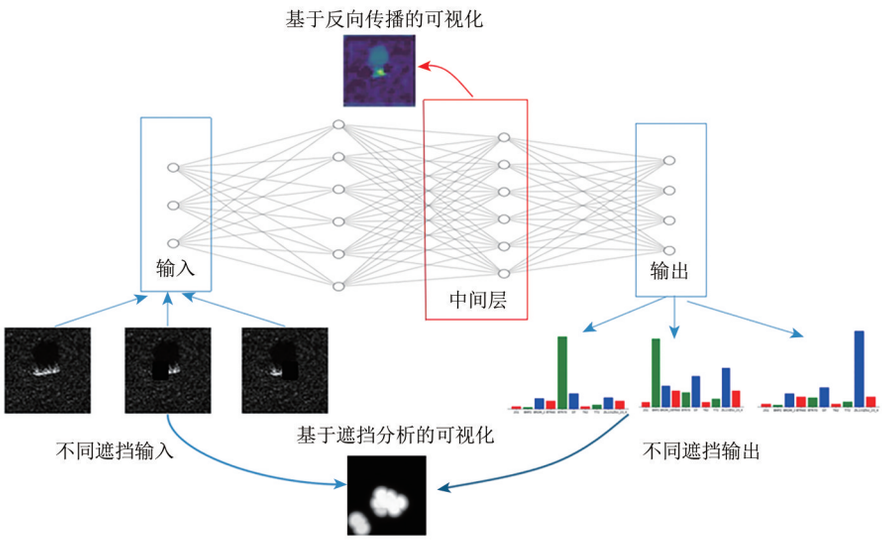


图3 深度神经网络特征可视化技术

Fig. 3 Visualization techniques of deep neural network based feature analysis

Grad-CAM采用网络输出对于最后一层卷积层的梯度实现类别激活图，仅要求网络预测的激活函数可微。对于神经网络最后一个卷积层的每幅特征图$\boldsymbol{A}_k$，计算类别c的梯度得分y_c并进行平均化得到特征图$\boldsymbol{A}_k$的重要性$\alpha_{k,\,c}$。

$$\alpha_{k,c} = \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial y_c}{\partial \boldsymbol{A}_{k,i,j}} \tag{6}$$

$$map_c = \mathrm{ReLU}\left(\sum_k^K \alpha_{k,c}\boldsymbol{A}_k\right) \tag{7}$$

式中，$\boldsymbol{A}_{k,\,i,\,j}$是$m \times n$大小的特征图$\boldsymbol{A}_k$中位置坐标为$(i,\ j)$的神经元。

4 深度识别模型对抗攻击鲁棒性实验评估

4.1 数据集与实验配置

为了验证多源遥感影像深度神经网络识别模型的对抗脆弱性，本文实验采用微波遥感影像识别领域广泛使用的MSTAR（Moving and Stationary Target Acquisition and Recognition）数据集（Ross等，1998）和光学遥感影像场景分类领域广泛使用的UC-Merced数据集（Yang和Newsam，2010）。MSTAR数据集典型目标的SAR影像切片与光学影像切片实例如图4所示（Blasch，2020）。本文实验中我们采用地面军事车辆MSTAR-10数据子集，目标类别包括2S1、BMP2、BRDM_2、BTR60、BTR70、D7、T62、T72、ZIL131、ZSU_23_4。分辨率为0.3 m×0.3 m，共10类，每类样本数目不同，空间大小不同。按照文献通用做法，俯仰角17°作为训练集，19°作为测试集。所有样本上采样变换后，中心裁剪为224×224像素。UC-Merced数据集覆盖21类光学遥感场景，每个类别包括100幅光学遥感影像样本，样本图像空间大小为256×256像素。

图5为UC-Merced数据集中21个类别的代表性样本示例。本文实验中UC-Merced数据集每类样本的训练集和测试集的比例为0.8∶0.2。

为了更加全面评估不同深度识别模型的对抗攻击鲁棒性，实验中选取了9种广泛使用的AlexNet、ResNet18、VGG16、Densenet201、Inceptionv3、GoogleNet、NASNet1_0、Mobilenet_v2、SqueezeNet架构，覆盖了手工设计深度网络（AlexNet、ResNet18、VGG16、Densenet201、Inceptionv3、GoogleNet）、自动搜索深度网络（NASNet1_0）和轻量深度网络（Mobilenet_v2、SqueezeNet）3类典型深度神经网络识别架构。为了避免不同识别模型实现方法和训练方法带来的不一致性，我们选择使用PyTorch标准库中实现的模型架构，采用同样的学习策略进行模型参数优化。

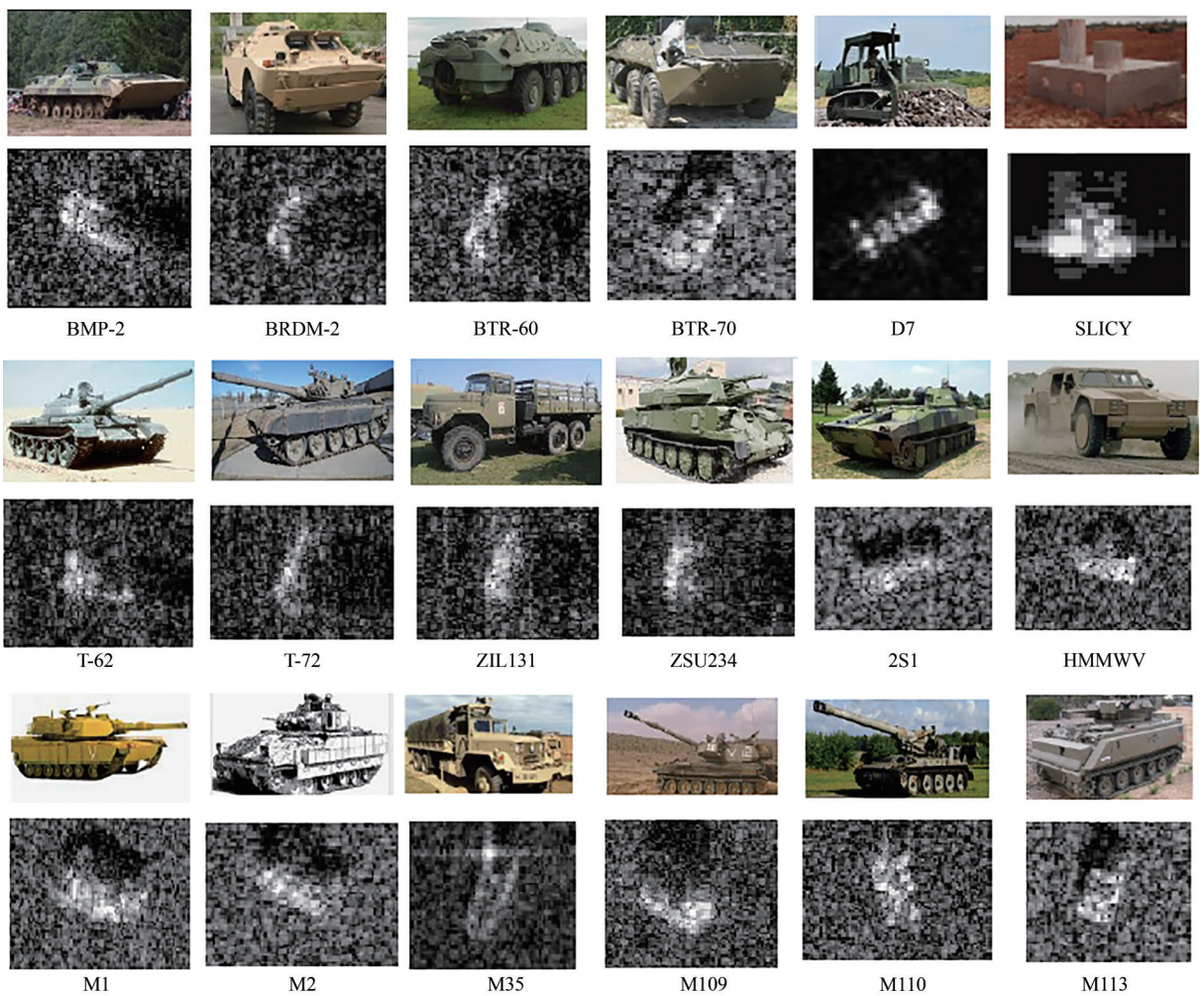


图4 MSTAR 数据集遥感影像样例

Fig. 4 MSTAR data set exemplar imagery

图5 UC-Merced 数据集遥感影像样例

Fig. 5 UC-Merced data set exemplar imagery

对抗攻击方法包括FGSM、PGD、DF、C&W、EAD等5类白盒攻击方法和HSJA、SSP两类黑盒攻击方法。实验中FGSM攻击噪声范数选用L_∞，扰动强度约束$\varepsilon = 0.3$；PGD攻击噪声范数选用L_∞，

扰动强度约束$\varepsilon = 0.3$，最大迭代次数设定为100；C&W攻击学习率0.01，最大迭代次数设定为100；DF攻击扰动步长设定为10^{-6}，最大迭代次数100；EAD攻击噪声范数选用L_2，学习率0.01，最大迭代次数100；HSJA攻击和SSP攻击噪声范数选用L_2，强度约束$\varepsilon = 0.3$，最大迭代次数100。

区分两类典型攻击场景，开展非定向对抗攻击和定向对抗攻击两组实验。在非定向对抗攻击实验中，主要实验目的是验证多类别平均意义下的多源遥感影像深度识别模型对抗攻击鲁棒性，并重点对比分析光学遥感影像和SAR遥感影像深度识别模型对抗扰动之间的相似性与差异性。在定向对抗攻击实验中，主要实验目的是分析深度识别模型特定类别的对抗鲁棒性，进一步细化分析不同类别间存在的对抗攻击鲁棒性不平衡问题。此外，我们还结合多隐层特征激活特性，尝试理解对抗样本攻击的作用机理，从深度识别模型的可解释性角度研究对抗攻击鲁棒性。

4.2 多源遥感影像深度识别模型非定向对抗攻击

表1给出了在MSTAR数据集上不同深度神经网络模型在原始无攻击测试数据集、添加7种非定向攻击对抗扰动后的测试数据集上的识别率统计信息。不同测试数据条件下正确识别率前3名进行了加粗显示。在原始无攻击测试数据集上，正确识别率较高的深度网络有InceptionV3、ResNet18、VGG16，3者都属于手工设计大计算量网络架构。在存在对抗攻击的测试数据集上，所有深度网络模型的正确识别率都出现了30%—70%左右的大幅度下降，凸显了深度识别模型的对抗脆弱性。DenseNet201在7组添加了不同对抗扰动的测试数据上取得了6组前3的优异表现，对不同攻击方法的综合鲁棒性最优。7类攻击方法中，5类白盒攻击方法的成功率优于2类黑盒攻击方法；基于梯度的PGD白盒攻击方法成功率最高；基于自监督扰动的SSP黑盒攻击方法由于没有利用梯度和决策边界信息，对抗攻击成功率最低。实验结果表明：对抗攻击的成功率与深度识别模型的梯度信息或决策边界信息紧密相关，梯度混淆和识别输出信息查询保护是提升深度识别模型对抗鲁棒性的有效途径。

表1 MSTAR数据集对抗攻击统计结果

Table 1 Adversarial attack results on MSTAR dataset

/%

攻击方法	AlexNet	VGG16	ResNet18	InceptionV3	DenseNet201	GoogleNet	MNASNet1_0	SqueezeNet	MobileNetV2
无攻击	71.62	**95.59**	**96.87**	**98.60**	81.61	92.62	87.67	92.33	75.05
FGSM	15.79	**20.99**	1.48	9.81	**23.16**	13.16	7.89	**27.18**	16.84
PGD	0.53	2.11	0.53	0.53	**8.95**	7.89	**17.89**	4.21	**8.95**
DF	7.89	1.05	5.79	**21.58**	**16.32**	7.37	8.42	3.16	**15.26**
C&W	**37.37**	5.79	7.89	**21.05**	**21.05**	7.89	8.95	5.79	17.37
EAD	**18.95**	1.58	4.74	0.93	8.95	3.27	**9.47**	5.79	**11.68**
HSJA	**18.42**	3.16	17.89	16.84	**28.95**	14.21	8.95	14.21	**18.42**
SSP	23.83	24.30	14.95	24.87	**31.78**	**28.04**	11.21	**57.01**	16.36

注：每行正确率前3名加粗标识。

表2给出了在UC-Merced数据集上不同深度识别模型在原始无攻击测试数据集、添加7种非定向攻击对抗扰动后的测试数据集上的识别率统计信息。不同测试数据条件下正确识别率的前3名进行了加粗显示。在原始无攻击测试数据集上，正确识别率较高的深度网络有InceptionV3、DenseNet201、MobileNetV2、ResNet18等。在存在对抗攻击的测试数据集上，所有深度模型的正确识别率都出现了20%—90%左右的大幅度下降。手工设计网络AlexNet、InceptionV3、DenseNet201在多组攻击测试数据集上取得了不错的效果，对多种类型的对抗攻击综合鲁棒性较好；特别是AlexNet在7组攻击测试数据上有5组正确识别率进入前3，反映出深度识别模型的结构复杂度与对抗鲁棒性是不同的属性，两者之间的关系还需要进一步深入研究。

综合表1和表2的统计信息，可以看到：(1) 覆盖手工设计深度网络、自动搜索深度网络和轻量

化深度网络的3大类9种典型深度卷积神经识别网络模型，无论是在无攻击测试数据上还是在添加不同对抗扰动的测试数据上，光学遥感影像多类别平均正确识别率均优于SAR遥感影像。一方面由于现有的主流网络架构都是基于光学影像应用启发，另一方面也反映出SAR遥感影像解译更加困难。（2）现有主流识别网络模型都普遍存在对抗攻击鲁棒性较差的问题。现有深度识别方法直接部署于实际系统时将面临极大安全隐患，急需开展对抗环境下深度识别模型的鲁棒正确识别率评估研究工作。手工设计网络在抵御对抗攻击方面，在光学遥感影像和SAR遥感影像数据集上都普遍优于自动搜索网络和轻量化网络，反映出专家知识和领域经验可以提升识别模型的对抗鲁棒性。边缘智能和端侧智能常用的轻量识别网络在抵御对抗攻击方面效果较差。（3）各类攻击方法中，白盒攻击方法由于获取的先验信息更多，所以攻击效果普遍优于黑盒攻击方法。在视觉领域被用于对抗攻击基准的PGD攻击方法，在多源遥感影像数据集上同样成功率较高，反映出梯度信息控制的重要性。（4）深度识别模型的多类别平均正确识别率与对抗攻击鲁棒性是两个不同的属性，在光学遥感影像和SAR遥感影像数据集上没有发现正确识别率和鲁棒正确识别率两者之间的明显规律，需要设计新的评估体系评价深度识别模型的在对抗场景中的可用性。

表2 UC-Merced数据集对抗攻击统计结果

Table 2 Adversarial attack results on UC-Merced dataset

/%

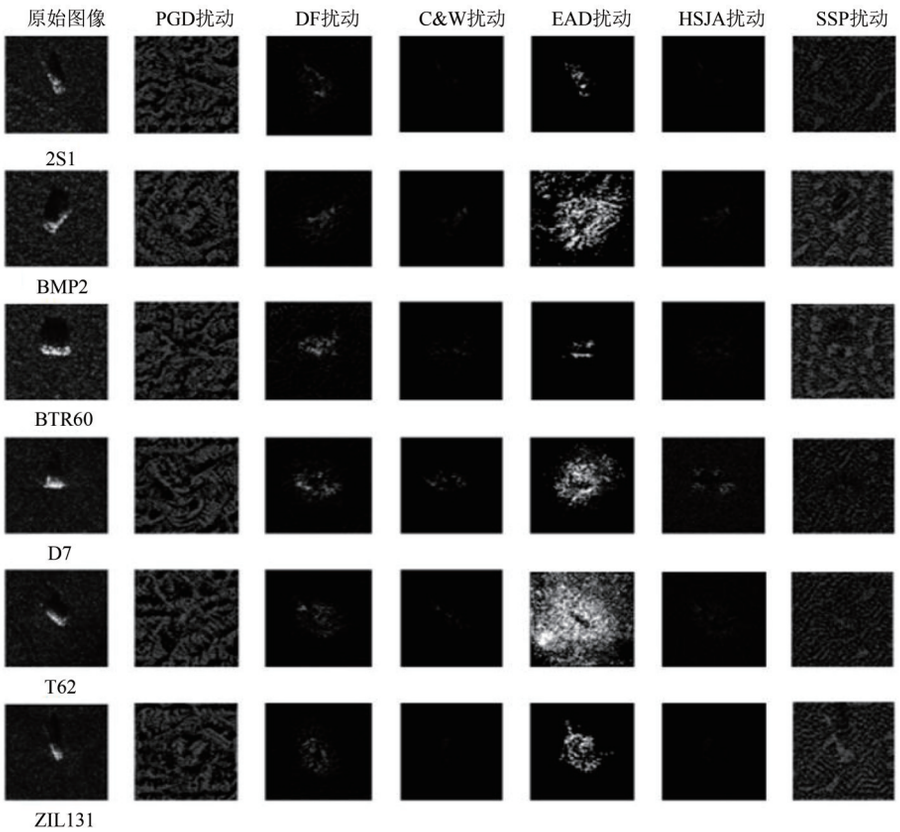
攻击方法	AlexNet	VGG16	ResNet18	InceptionV3	DenseNet201	GoogleNet	MNASNet1_0	SqueezeNet	MobileNetV2
无攻击	91.43	96.43	96.90	**98.33**	**98.31**	94.76	92.14	93.57	**97.14**
FGSM	**77.38**	47.86	40.48	45.71	**67.14**	54.05	40.00	**70.24**	44.52
PGD	**7.14**	6.19	4.29	2.38	5.95	**6.67**	**8.81**	6.43	5.95
DF	35.95	25.95	24.52	**35.96**	**56.43**	**41.67**	36.19	35.95	29.05
C&W	**5.95**	2.38	1.90	**4.29**	1.43	3.81	**5.00**	3.81	1.67
EAD	2.62	1.19	2.37	**19.05**	9.29	**16.43**	**11.90**	1.43	1.90
HSJA	**44.76**	7.38	6.43	6.19	**14.29**	13.33	9.52	**31.19**	2.62
SSP	**76.19**	8.57	29.76	**51.19**	48.81	46.67	32.14	**55.00**	13.10

注：每行正确率前3名加粗标识。

图6为多源遥感影像深度识别模型对抗扰动的可视化结果，分别从MSTAR数据集和UC-Merced数据集中选择6幅测试样本影像，基于训练好的VGG16深度识别网络，采用不同对抗攻击方法为测试样本生成扰动。由于对抗扰动的绝对值很小，对于人类视觉不可区分，因此为了便于实验分析，我们对生成的扰动图像进行了放大显示。可以看到，无论是在光学遥感影像还是SAR遥感影像上，大多数攻击方法针对不同类别的对抗攻击扰动都在视觉上呈现出“噪声”特性。对抗扰动空间分布模式的在图像内容上的主要特性表现为纹理特性，这个现象与深度神经网络识别模型决策的“纹理偏好”具有关联性。图6（a）为MSTAR数据集上的对抗扰动示例，可以看到，对抗扰动的分布区域和强度具有明显的规律性，例如DF扰动和EAD扰动主要分布在图像中显著目标区域，PGD扰动则遍布整个图像区域。

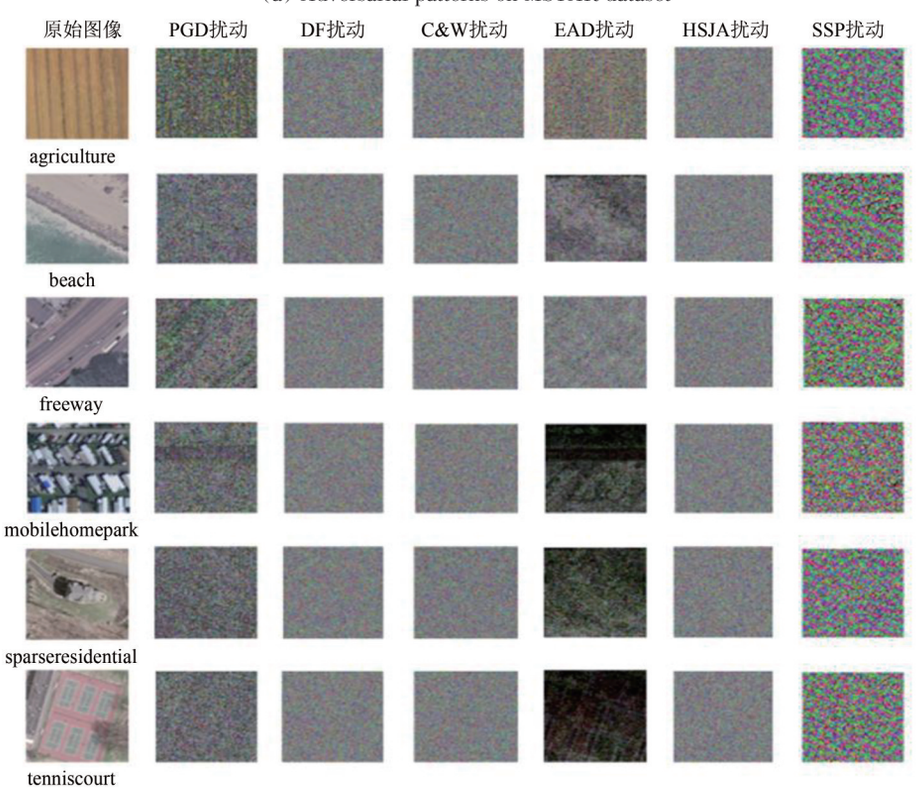
4.3 多源遥感影像深度识别模型定向对抗攻击

由于各个类别样本数据量和类别区分难度的不同，多源遥感影像深度识别模型对每个类别的正确识别率并不相同，因此需要进一步细化分析面向特定类别的对抗攻击鲁棒性，诊断深度识别模型的定向攻击脆弱性。图7（a）为无攻击场景MSTAR数据集上VGG16模型的多类别识别混淆矩阵，图7（b）为无攻击场景UC-Merced数据集上VGG16模型的多类别识别混淆矩阵，其中类别名称使用数字序号代替。图7中观察得到：VGG16模型多类别正确识别率存在差异，部分类别间发生混淆的概率更大意味着更容易成为定向攻击的目标。MSTAR数据集中2S1、BMP2（第1、2类）正确识别率较低，易被误识别为ZIL131、T72（第9、8类）；UC-Merced数据集中buildings、dense residential、intersection（第5、7、12类）正确识别率较低，易被错误识别为dense residential、medium residential、mobile home park（第7、13、14类）。



（a）MSTAR数据对抗扰动可视化

（a）Adversarial patterns on MSTAR dataset

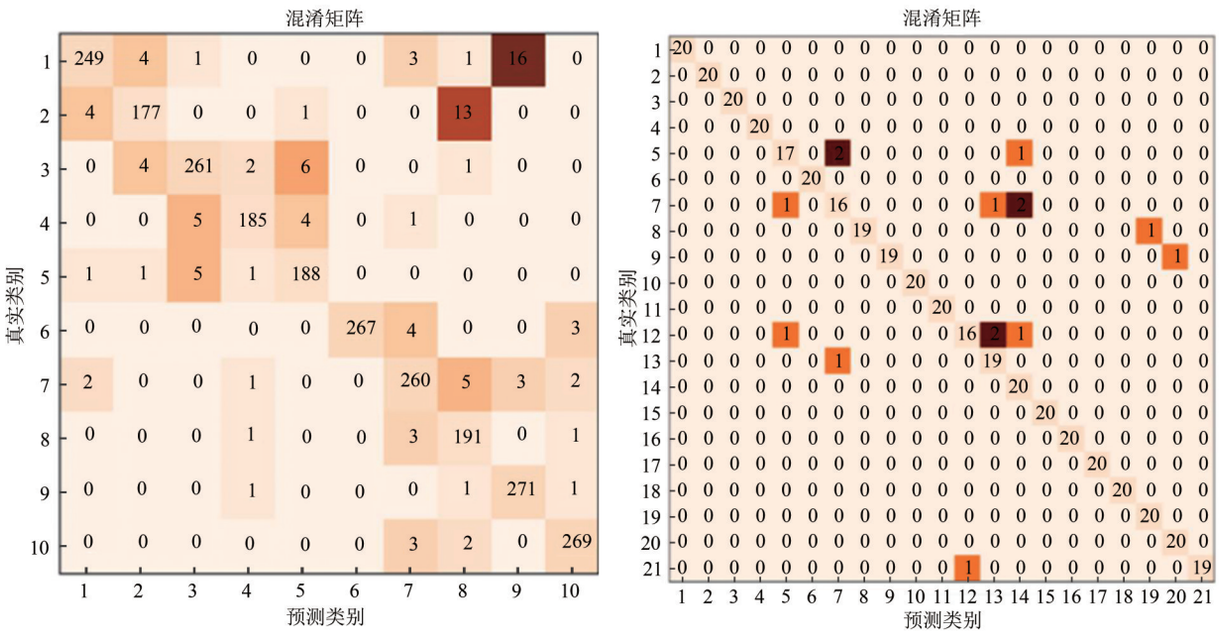


（b）UC-Merced数据对抗扰动可视化

（b）Adversarial patterns on UC-Merced dataset

图6 对抗扰动可视化

Fig. 6 Visualization of adversarial perturbation



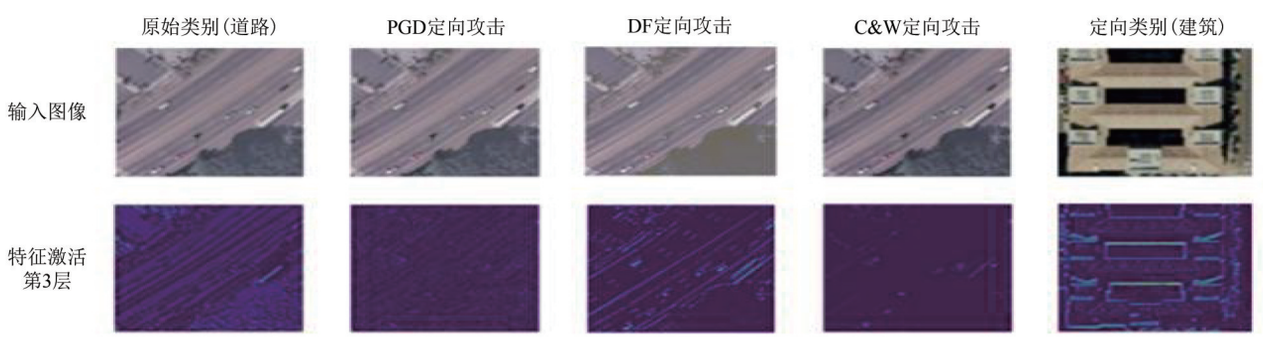
(a) MSTAR数据集无攻击识别混淆矩阵
(a) Confusion matrix of VGG16 on MSTAR dataset

(b) UC-Merced数据集无攻击识别混淆矩阵
(b) Confusion matrix of VGG16 on UC-Merced dataset

图7 无对抗攻击时VGG16模型识别正确率混淆矩阵(颜色越深代表错误分类样本数越多)

Fig. 7 Confusion matrix of VGG16 recognition accuracy (dark colors represent more classification errors)

图8为多源遥感影像深度识别模型定向对抗攻击实验结果，在UC-Merced数据集上，我们将场景目标类别高速公路类别样本图像定向攻击为建筑物类别。实验中以VGG16模型为例，采用PGD、DF和C&W方法的定向形式进行扰动生成。为了观察定向攻击对抗样本图像和目标类别无攻击干净样本图像的模型激活异同，理解定向攻击的基本原理，采用Grad-CAM方法对定向攻击样本和无攻击样本进行显著性可视化分析。实验中对VGG16识别模型的第3、8、15、22、29层进行激活可视化，各层的激活情况的变化反应了对抗样本对于网络决策的影响。从第3层输出可以看到，PGD对抗样本使影像内各元素界限变得模糊，对抗样本激活图中可见高速公路与树木的界限被明显弱化；C&W对抗样本则是隐匿了大量信息，使激活部分大大减少；DF攻击将树木遮挡道路这一信息进行了放大。对比第8、15、22、29层输出可以看到原始影像激活以线性特征为主，激活区域较为连贯，而对抗样本的激活逐渐趋向于突出激活个别点，与建筑物的激活情况逐层相似。定向攻击样本的特征激活特性与目标类别正常样本的特征激活特性在深度识别模型的低层特征中存在明显差异，在中层特征中区分逐渐弱化，在高层特征中激活具有强相似性。下一步研究中，可以利用这种规律，设计多隐层特征激活诊断提升深度识别模型定向对抗攻击的鲁棒性。



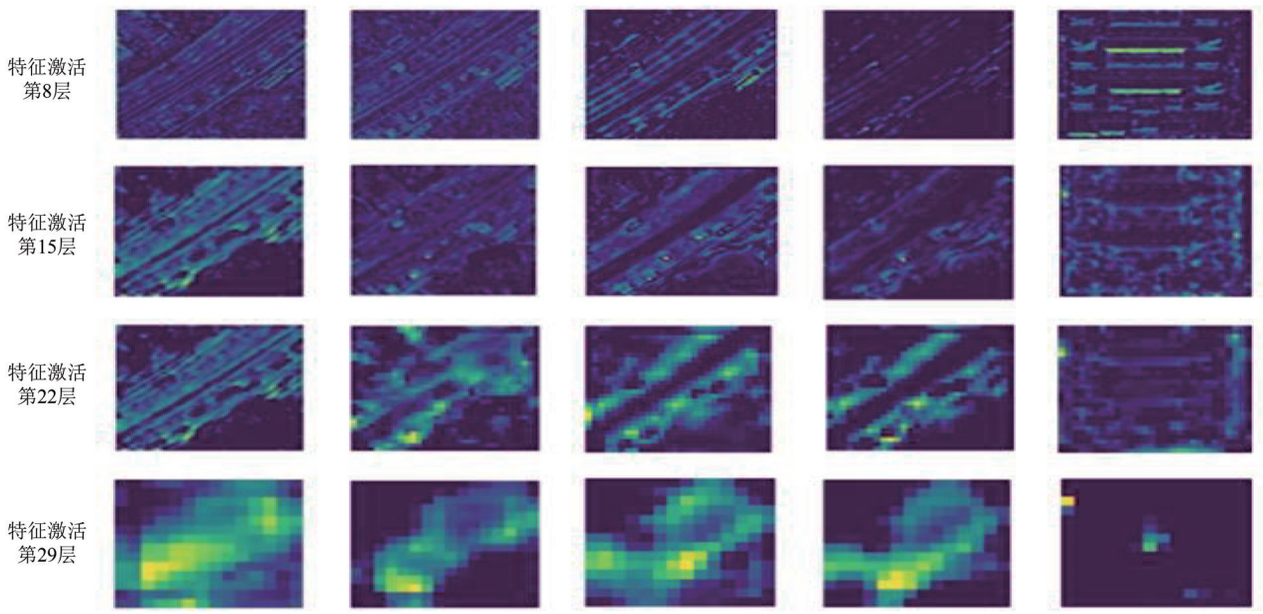


图8 定向攻击特征激活分析

Fig. 8 Feature activation of targeted adversarial attack

5 结 论

深度神经网络模型具有处理不同模态信息的能力，其在处理不同模态信息时模型结构上的相似性及层次化的分析方法，为建立多源遥感影像特征表示与语义识别模型提供了有力工具。然而，由于深度学习理论本身的不完备性、深度识别网络结构设计与优化方法的强复用性、以及多源成像识别系统在复杂电磁环境中易受到各类干扰等多因素的影响，使得现有识别系统在对抗鲁棒性方面评估不足，给其在军事场景和安全敏感领域的深入应用和广泛部署带来巨大隐患。本文首先分析了多源遥感影像深度神经网络识别系统在理论上和应用中的潜在安全风险；其次重点介绍了面向深度模型推理阶段的对抗样本攻击形式，并在讨论基本原理和可解释性的基础上，针对光学遥感影像和SAR遥感影像开展了深度神经网络识别模型对抗攻击实验，从对抗攻击正确识别率和对抗扰动可视化两个方面评估对抗攻击。

现有的深度神经网络识别模型存在巨大的安全隐患，对抗鲁棒性普遍不足，模型准确率与模型对抗攻击鲁棒性之间的关系还有待进一步深入研究；在机器学习、计算机视觉和自然语言处理等领域对抗攻击与防御已经开展了大量研究，在多源遥感影像解译领域需要引发重点关注。复杂电磁环境下新一代人工智能多源遥感影像识别系统需要融合来自不同模态传感器的数据，开展基于领域知识图谱嵌入、多粒度知识深度迁移、鲁棒对抗样本防御的多模型融合新方法研究，充分挖掘先验领域知识，建立模型驱动的多任务深度学习模型，提升学习模型的可解释性和透明性，提升识别系统在对抗场景中的准确性和安全性。

参考文献(References)

Berghoff C, Neu M and Von Twickel A. 2020. Vulnerabilities of connectionist AI applications: evaluation and defence. arXiv preprint arXiv:2003.08837

Blasch E. 2020. Self-proficiency assessment for ATR systems//Proceedings of SPIE 11393, Algorithms for Synthetic Aperture Radar Imagery XXVII. [s.l.]: SPIE: 113930T [DOI: 10.1117/12.2563259]

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//Proceedings of 2017 IEEE Symposium on Security and Privacy. San Jose: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Chen J B, Jordan M I and Wainwright M J. 2020. HopSkipJumpAttack: a query-efficient decision-based attack//Proceedings of 2020 IEEE Symposium on Security and Privacy. San Francisco: IEEE: 1277-1294 [DOI: 10.1109/SP40000.2020.00045]

Chen P Y, Sharma Y, Zhang H, Yi J F and Hsieh C J. 2018. EAD: elastic-net attacks to deep neural networks via adversarial examples. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI: 2

Cheng G, Xie X X, Han J W, Guo L and Xia G S. 2020. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13:

3735-3756 [DOI: 10.1109/JSTARS.2020.3005403]

Fawzi A, Moosavi-Dezfooli S M and Frossard P. 2017. The robustness of deep networks: a geometrical perspective. IEEE Signal Processing Magazine, 34(6): 50-62 [DOI: 10.1109/MSP.2017.2740965]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR

Kurte K R, Durbha S S, King R L, Younan N H and Vatsavai R. 2017. Semantics-enabled framework for spatial image information mining of linked earth observation data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(1): 29-44 [DOI: 10.1109/JSTARS.2016.2547992]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards deep learning models resistant to adversarial attacks. Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Naseer M, Khan S, Hayat M, Khan F S and Porikli F. 2020. A self-supervised approach for adversarial robustness//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 259-268 [DOI: 10.1109/CVPR42600.2020.00034]

Ras G, Xie N, Van Gerven M and Doran D. 2020. Explainable deep learning: a field guide for the uninitiated. arXiv preprint arXiv: 2004.14545

Ross T D, Worrell S W, Velten V J, Mossing J C and Bryant M L. 1998. Standard SAR ATR evaluation experiments using the MSTAR public release data set//Proceedings of the SPIE 3370, Algorithms for Synthetic Aperture Radar Imagery V. Orlando: SPIE, 1998. 566-573 [DOI: 10.1117/12.321859]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Sun H, Chen J, Lei L, Ji K F and Kuang G Y. 2021. Adversarial robustness of deep convolutional neural network based image recognition models: a review. Journal of Radars, 10(4): 571-594 (孙浩, 陈进, 雷琳, 计科峰, 匡纲要. 2021. 深度卷积神经网络图像识别模型对抗鲁棒性技术综述. 雷达学报, 10(4): 571-594) [DOI: 10.12000/JR21048]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J and Fergus R. 2014. Intriguing properties of neural networks. Proceedings of the 2nd International Conference on Learning Representations. Banff: ICLR

Tong X D. 2016. Development of China high-resolution earth observation system. Journal of Remote Sensing, 20(5): 775-780 (童旭东. 2016. 中国高分辨率对地观测系统重大专项建设进展. 遥感学报, 20(5): 775-780) [DOI: 10.11834/JRS.20166302]

Wiyatno R R, Xu A Q, Dia O and De Berker A. 2019. Adversarial examples in modern machine learning: a review. arXiv preprint arXiv:1911.05268

Xu Y H, Du B and Zhang L P. 2021. Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: attacks and defenses. IEEE Transactions on Geoscience and Remote Sensing, 59(2): 1604-1617 [DOI: 10.1109/TGRS.2020.2999962]

Yang Y and Newsam S. 2010. Bag-of-visual-words and spatial extensions for land-use classification//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose: ACM: 270-279 [DOI: 10.1145/1869790.1869829]

Yuan X Y, He P, Zhu Q L and Li X L. 2019. Adversarial examples: attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9): 2805-2824 [DOI: 10.1109/TNNLS.2018.2886017]

Zhou B L, Khosla A, Lapedriza A, Oliva A and Torralba A. 2016. Learning deep features for discriminative localization//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 2921-2929 [DOI: 10.1109/CVPR.2016.319]

Zhu X X, Montazeri S, Ali M, Hua Y S, Wang Y Y, Mou L C, Shi Y L, Xu F and Bamler R. 2020. Deep learning meets SAR. arXiv preprint arXiv:2006.10027

Adversarial robustness evaluation of multiple-source remote sensing image recognition based on deep neural networks

SUN Hao[1], XU Yanjie[1], CHEN Jin[2], LEI Lin[1], JI Kefeng[1], KUANG Gangyao[1]

1. College of Electronic Science, National University of Defense Technology, Changsha 410073, China;
2. Beijing Institute of Remote Sensing Information, Beijing 100192, China

Abstract: Deep-neural-network-based multiple-source remote sensing image recognition systems have been widely used in many military

scenarios, such as in aerospace intelligence reconnaissance, unmanned aerial vehicles for autonomous environmental cognition, and multimode automatic target recognition systems. Deep learning models rely on the assumption that the training and testing data are from the same distribution. However, these models show poor performance under common corruption or adversarial attacks. In the remote sensing community, the adversarial robustness of deep-neural-network-based recognition models have not received much attention, thence increasing the risk for many security-sensitive applications.

This article evaluates the adversarial robustness of deep-neural-network-based recognition models for multiple-source remote sensing images. First, we discuss the incompleteness of deep learning theory and reveal the presence of great security risks. The independent identical distribution assumption is often violated, and the system performance cannot be guaranteed under adversarial scenarios. The whole process chain of a deep-neural-network-based image recognition system is then analyzed for its vulnerabilities. Second, we introduce several representative algorithms for adversarial example generation under both the white- and black-box settings. The gradient-propagation-based visualization method is also proposed for analyzing adversarial attacks.

We perform a detailed evaluation of nine deep neural networks across two publicly available remote sensing image datasets. Both optical remote sensing and SAR remote sensing images are used in our experiments. For each model, we generate seven perturbations, ranging from gradient-based optimization to unsupervised feature distortion, for each testing image. In all cases, we observe a significant reduction in average classification accuracy between the original clean data and their adversarial images. Apart from adversarial average recognition accuracy, feature attribution techniques have also been adopted to analyze the feature diffusion effect of adversarial attacks, hence contributing to the present understanding of the vulnerability of deep learning models.

Experimental results demonstrate that all deep neural networks have suffered great losses in classification accuracy when the testing images are adversarial examples. Understanding such adversarial phenomena improves our understanding of the inner workings of deep learning models. Additional efforts are needed to enhance the adversarial robustness of deep learning models.

Key words：multiple source remote sensing images, deep neural networks, adversarial attack, feature visualization, adversarial robustness evaluation

Supported by National Natural Science Foundation of China (No. 61971426)