基于动量迭代快速梯度符号的SAR-ATR深度神经网络黑盒攻击算法

万烜申 刘 伟* 牛朝阳 卢万杰

(中国人民解放军战略支援部队信息工程大学数据与目标工程学院 郑州 450000)

摘要：合成孔径雷达自动目标识别(SAR-ATR)领域缺乏有效的黑盒攻击算法，为此，该文结合动量迭代快速梯度符号(MI-FGSM)思想提出了一种基于迁移的黑盒攻击算法。首先结合SAR图像特性进行随机斑点噪声变换，缓解模型对斑点噪声的过拟合，提高算法的泛化性能；然后设计了能够快速寻找最优梯度下降方向的ABN寻优器，通过模型梯度快速收敛提升算法攻击有效性；最后引入拟双曲动量算子获得稳定的模型梯度下降方向，使梯度在快速收敛过程中避免陷入局部最优，进一步增强对抗样本的黑盒攻击成功率。通过仿真实验表明，与现有的对抗攻击算法相比，该文算法在MSTAR和FUSAR-Ship数据集上对主流的SAR-ATR深度神经网络的集成模型黑盒攻击成功率分别提高了3%～55%和6.0%～57.5%，而且生成的对抗样本具有高度的隐蔽性。

关键词：合成孔径雷达；目标识别；黑盒攻击；拟双曲动量算子；斑点噪声变换

中图分类号：TP391 **文献标识码**：A **文章编号**：2095-283X(2024)03-0714-16

DOI: 10.12000/JR23220

引用格式：万烜申, 刘伟, 牛朝阳, 等. 基于动量迭代快速梯度符号的SAR-ATR深度神经网络黑盒攻击算法[J]. 雷达学报(中英文), 2024, 13(3): 714–729. doi: 10.12000/JR23220.

Reference format: WAN Xuanshen, LIU Wei, NIU Chaoyang, *et al*. Black-box attack algorithm for SAR-ATR deep neural networks based on MI-FGSM[J]. *Journal of Radars*, 2024, 13(3): 714–729. doi: 10.12000/JR23220.

Black-box Attack Algorithm for SAR-ATR Deep Neural Networks Based on MI-FGSM

WAN Xuanshen LIU Wei* NIU Chaoyang LU Wanjie

(*PLA Strategic Support Force Information Engineering University*, *School of Data and Target Engineering*, *Zhengzhou* 450000, *China*)

Abstract: The field of Synthetic Aperture Radar Automatic Target Recognition (SAR-ATR) lacks effective black-box attack algorithms. Therefore, this research proposes a migration-based black-box attack algorithm by combining the idea of the Momentum Iterative Fast Gradient Sign Method (MI-FGSM). First, random speckle noise transformation is performed according to the characteristics of SAR images to alleviate model overfitting to the speckle noise and improve the generalization performance of the algorithm. Second, an AdaBelief-Nesterov optimizer is designed to rapidly find the optimal gradient descent direction, and the attack effectiveness of the algorithm is improved through a rapid convergence of the model gradient. Finally, a quasihyperbolic momentum operator is introduced to obtain a stable model gradient descent direction so that the gradient can avoid falling into a local optimum during the rapid convergence and to further enhance the success rate of black-box attacks on adversarial examples. Simulation experiments show that compared with existing adversarial attack algorithms, the proposed algorithm improves the ensemble model black-box attack success rate of mainstream SAR-ATR deep neural networks by 3%～55% and 6.0%～57.5% on the MSTAR and

收稿日期：2023-11-17；改回日期：2024-01-14；网络出版：2024-02-02

*通信作者： 刘伟 greatliuliu@163.com *Corresponding Author: LIU Wei, greatliuliu@163.com

基金项目：国家自然科学基金(42201472)

Foundation Item: The National Natural Science Foundation of China (42201472)

责任主编：李宁 Corresponding Editor: LI Ning

©The Author(s) 2024. This is an open access article under the CC-BY 4.0 License (https://creativecommons.org/licenses/by/4.0/)

FUSAR-Ship datasets, respectively; the generated adversarial examples are highly concealable.

Key words: Synthetic Aperture Radar (SAR); Target recognition; Black-box attack; Quasi-Hyperbolic Momentum (QHM) operator; Speckle noise transformation

1　引言

近年来，凭借深度神经网络(Deep Neural Networks, DNN)强大的特征提取能力，深度学习技术在合成孔径雷达自动目标识别(Synthetic Aperture Radar Automatic Target Recognition, SAR-ATR)领域取得了显著的成功[1–7]。然而，研究表明基于深度神经网络的SAR-ATR模型容易受到对抗样本的攻击[8–10]。对抗样本的概念由Szegedy等人[11]首次提出，通过在输入样本中添加精心设计的微小扰动，产生对抗样本，从而导致识别模型的错误分类，即实现对神经网络识别模型的攻击。针对SAR-ATR模型对抗攻击算法的研究能够拓展SAR目标识别的数据集，利用生成的对抗样本进行再训练即可提高SAR-ATR模型的鲁棒性。因此，SAR对抗攻击算法的研究对SAR-ATR模型的安全性具有重要意义。

目前针对SAR-ATR模型的对抗攻击算法处于起步阶段，学者主要将光学图像中的对抗攻击算法迁移到SAR图像。在光学图像领域中，已经提出了许多对抗攻击算法。根据对目标模型先验知识的掌握情况，这些对抗攻击算法一般可分为白盒攻击[12–16]和黑盒攻击。黑盒攻击大致可分为基于概率标签的黑盒攻击[17,18]、基于决策的黑盒攻击[19]和基于迁移的黑盒攻击[20–23]。前两种黑盒攻击算法需要大量查询神经网络，然而这在实际情况是难以实现的，因此基于迁移的黑盒攻击算法是学者研究的重点，其几乎都是通过基于梯度的对抗攻击算法实现的。在基于梯度的攻击方面，Goodfellow等人[12]提出快速梯度符号算法(Fast Gradient Sign Method, FGSM)，此算法通过寻找神经网络模型梯度变化最大的方向，并在此方向上添加微小扰动，获得对抗样本。Kurakin等人[13]提出一种迭代快速梯度符号算法(Iterative Fast Gradient Sign Method, I-FGSM)，此算法通过多次迭代添加较小的扰动，降低扰动被检测的概率，解决了FGSM攻击成功率低的问题。Dong等人[20]在I-FGSM的基础上首次引入动量的思想，提出了MI-FGSM (Momentum Iterative Fast Gradient Sign Method)，该攻击方法进一步提高了对抗样本的黑盒攻击能力，即迁移能力。Zhao等人[21]将Nesterov-Adam算法集成到I-FGSM中，提出了一种新的攻击算法NAM，此算法在成功发起黑盒攻击的同时，也提高了白盒攻击的有效性。Wang等人[22]提出了一种基于方差调整(Variance Tuning)的迭代快速梯度符号方法，简称为VMI-FGSM，此算法在每次迭代时减小梯度的方差，避免陷入局部最优，从而有效增强对抗样本的可迁移性。Xie等人[23]提出基于输入多样化(Diversity Input)的迭代快速梯度符号方法，简称为DI-FGSM，此算法在迭代攻击过程中对输入图片进行随机变换，增加输入图片的多样性，进一步提高对抗样本的黑盒攻击有效性。上述研究主要集中在光学图像领域，同时，对抗样本也存在于遥感图像领域。Czaja等人[24]通过实验首次验证了基于深度卷积神经网络的遥感图像识别存在对抗样本。Chen等人[25]对遥感图像场景分类的对抗样本问题进行了全面研究，实验结果表明对抗样本普遍存在。在SAR图像领域，学者早期主要利用光学图像中的对抗攻击算法验证SAR图像领域存在对抗样本，例如，Huang等人[8]验证了深度学习在SAR图像目标识别中存在安全性和鲁棒性问题。在迁移、复现光学图像对抗攻击算法的基础上，一些研究进一步加快SAR对抗样本的生成速度。Fast C&W算法[26]引入一个编码器网络，通过一步映射得到对抗样本，相比于C&W[16]算法，此算法有效提高了对抗样本的生成效率。Du等人[27]利用U-Net生成对抗网络(Generative Adversarial Network, GAN)构建对抗样本，实验结果表明，此算法提升了攻击成功率和计算效率。Zhou等人[28]提出了一种SAR通用对抗扰动(Universal Adversarial Perturbation, UAP)，从而大幅度缩短对抗样本的生成时间。考虑到SAR对抗攻击在真实场景的实现问题，Xia等人[29]尝试在信号域生成SAR对抗样本，通过所提出的欺骗干扰模型可以灵活地生成SAR对抗样本。实验结果表明，该方法能产生难以察觉的干扰，并能有效地攻击LeNet, VGGNet16, ResNet18和ResNet50这4种经典DNN模型。为了确保对抗样本的物理可行性，Peng等人[30]提出了一种基于参数化模型的SAR图像对抗样本生成方法，通过在8个DNN模型上进行实验，结果表明此算法具有较好的攻击性能。

目前针对SAR-ATR模型的对抗攻击算法大多为白盒攻击，即需要预先掌握敌方模型的类别和参数信息。但在实际情况的军事应用中，考虑到攻击对象SAR的非合作特性，其SAR-ATR模型对于攻击方来说是未知的，白盒攻击通常难以实施，因此

亟需发展无需获取敌方模型先验信息的SAR对抗攻击算法，即黑盒攻击算法。

为解决上述问题，本文提出了一种基于迁移的黑盒攻击算法(Transfer-based Black-box Attack Algorithm, TBAA)，有效实现了针对SAR-ATR模型的黑盒攻击。本文的贡献点如下：(1)在结合SAR图像特点方面，本文算法充分考虑SAR图像斑点噪声的特点，通过在每次迭代生成对抗样本期间不断利用Lee滤波算法滤除斑点噪声，并与服从截断指数分布的噪声相乘重构SAR图像，使得输入的SAR图像具有多样性，从而较好地缓解模型的过拟合现象；(2)为了稳定梯度更新方向和加快收敛速度，所提算法设计了梯度方向寻优器和梯度方向稳定算法，从而生成迁移性能强的对抗样本；(3)所提算法结合集成学习的思想，通过攻击集成模型实现对多个模型保持攻击的有效性，从而进一步提高黑盒攻击成功率。

2 TBAA算法描述

本文基于动量迭代快速梯度符号算法(MI-FGSM)思想设计SAR-ATR模型的黑盒攻击算法，所提算法框架如图1所示。首先，结合SAR图像斑点噪声的特点，对SAR图像以概率p先后进行Lee滤波和随机斑点噪声变换(如图1中黄色框内所示)；然后设计了梯度方向寻优器(如图1中绿色框内所示)，利用其前瞻性和自适应调整学习率的优势进一步提高黑盒攻击成功率；进一步，利用模型增强的思路(如图1中灰色框内所示)，通过将n个模型的逻辑值输出Logits_$i(i=1,2,\cdots,n)$加权求和得到Logits，再通过标签和融合的Logits计算新的损失函数；最后，将损失函数计算得到的梯度通过拟双曲动量(Quasi-Hyperbolic Momentum, QHM)算子得到稳定的梯度更新方向(如图1中蓝色框内所示)，提升本文算法黑盒攻击性能。

2.1 算法设计思想

本文算法以MI-FGSM[20]的思想为基础进行设计。MI-FGSM为基于梯度的攻击算法，其思想是利用对抗样本的迁移性来攻击黑盒模型。其中，对抗样本的迁移性如图2所示。针对一个图像多分类任务，$f_1(\cdot)$, $f_2(\cdot)$和$f_s(\cdot)$分别为已经训练好的深度神经网络模型，δ表示添加的扰动，x和x^{adv}分别表示原始输入样本和添加扰动后的对抗样本。其中，$f_s(\cdot)$作为生成对抗样本的白盒模型，相反，$f_1(\cdot)$和$f_2(\cdot)$则为黑盒模型。右侧的柱状图可视化了分类结果，柱状体的高度越高则表示置信度越高，红色柱状体表示神经网络最终错误分类的类别，黄色柱状体表示正确的类别。从图中可以看出，对抗样本x^{adv}能够成功欺骗$f_s(\cdot)$，并且也能导致$f_1(\cdot)$和$f_2(\cdot)$输出错误的分类结果。因此，由$f_s(\cdot)$模型训练生成的对抗样本具有迁移性。

在基于迁移的黑盒攻击算法中，大部分为基于梯度的攻击算法，此类算法的思路与深度神经网络的训练过程是相似的。假设$f(\cdot)$为深度神经网络模型，损失函数为J，输入的SAR图像为x，y表示x的真实标签。深度神经网络模型的训练思路是沿着梯度下降的方向迭代更新模型参数θ，达到降低损失函数J的目的。因此，深度神经网络模型训练过程的参数更新公式为

$$\tilde{\theta}=\theta-\frac{\partial J}{\partial y}\cdot\frac{\partial y}{\partial \theta} \tag{1}$$

基于梯度的对抗样本的生成思路为沿着梯度上升的方向迭代更新输入SAR图像x，其目的是使得深度神经网络模型的损失函数J逐渐变大，从而导致模型输出错误的识别结果。具体为

$$x^{\mathrm{adv}}=x+\frac{\partial J}{\partial y}\cdot\frac{\partial y}{\partial x} \tag{2}$$

根据以上分析，基于梯度的攻击算法思路与深度神经网络的训练均是通过反向传播训练参数的过

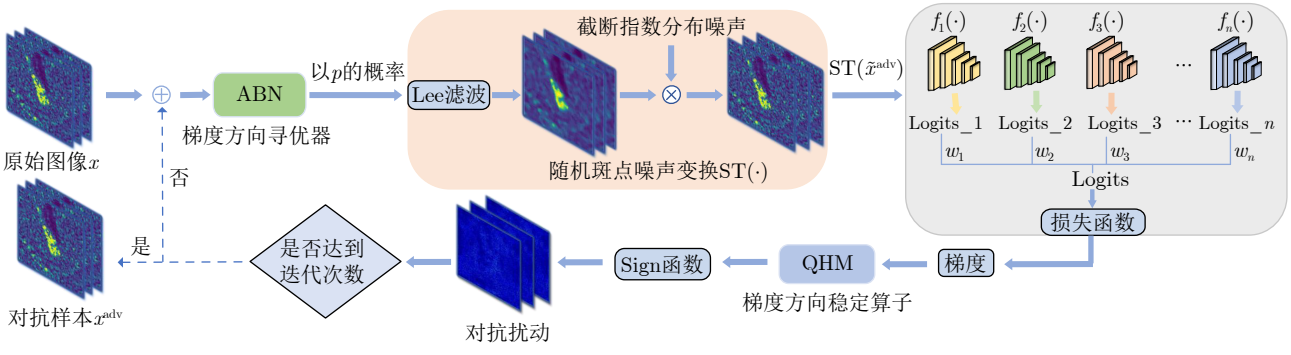


图 1 TBAA算法的原理图

Fig. 1 Schematic diagram of the TBAA algorithm

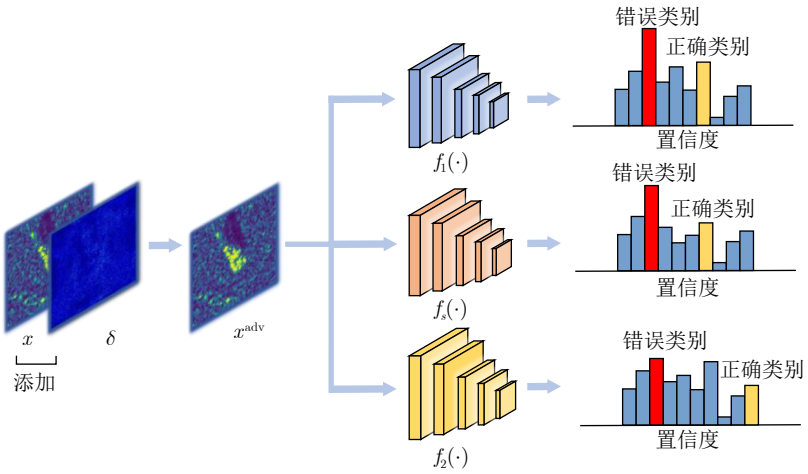


图 2 对抗样本的可迁移性

Fig. 2 Transferability of adversarial examples

程。在神经网络训练过程中，容易产生过拟合现象，即模型在训练集上识别精度高，但在测试集上识别精度低。基于梯度的攻击算法生成的对抗样本在白盒模型下表现出较好的攻击能力，但在黑盒条件下攻击性能较差，即对抗样本的迁移性不高。因此，可以将提高深度神经网络模型泛化性能的方法用于提高对抗样本的迁移性能上。结合更好的优化器能够有效提高深度神经网络模型的泛化性，如式(3)所示，μ为衰减因子，g_t为前t次迭代累积的梯度。MI-FGSM在梯度优化更新过程中引入动量项从而获得稳定更新的梯度方向，进而提高了对抗样本的迁移性，有效实现了黑盒攻击性能。

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{\text{adv}}, y)}{\left\|\nabla_x J(x_t^{\text{adv}}, y)\right\|_1} \tag{3}$$

同时，通过结合集成学习[31]的思想，如式(4)和式(5)所示，融合多个模型的全连接层输出Logits值，然后利用真实标签y和融合的Logits值构造新的损失函数。实现对多个模型保持攻击的有效性，进一步提高对抗样本的迁移性能，从而更易于攻击其他模型。

$$l(x) = \sum_{k=1}^{K} w_k l_k(x) \tag{4}$$

$$J(x, y) = -1_y \cdot \log(\text{softmax}(l(x))) \tag{5}$$

其中，K为模型数量，$l_k(x)$为第k个模型的Logits，w_k为集成系数，-1_y为标签的独热编码(One-Hot Encoding)。

2.2 随机斑点噪声变换

SAR图像斑点噪声为深度神经网络模型提供大量的高维特征，与SAR图像目标特征不同，该特征鲁棒性较差，当模型过拟合于斑点噪声特征时，会导致对抗样本的迁移性较差。

为了解决此问题，本文提出SAR图像随机斑点噪声变换(Speckle noise Transformation, ST)方法，有效提高输入训练样本的多样性，实现缓解模型对斑点噪声过拟合的目的，进而提高黑盒攻击成功率。此算法假设处理的SAR图像为单视SAR图像，拟定观察到的场景用乘性噪声模型建模，即实际SAR图像的每个分辨单元强度由一个反映该单元实际RCS的确定值和一个指数分布乘积而成[32]：

$$I = s \times n \tag{6}$$

其中，I表示观察到的强度，s表示SAR图像场景的RCS值，n表示服从截断指数分布的斑点噪声。

由于Lee滤波[33]被广泛应用于SAR图像的去噪，因此本文算法将观察到的SAR图像x以概率p首先对SAR图像进行Lee滤波，然后将Lee滤波后的SAR图像乘以服从截断指数分布的斑点噪声，具体实现公式为

$$\begin{aligned}\text{ST}(x) &= \text{Lee}(x) \cdot \frac{\mathrm{e}^{-z}}{1-\mathrm{e}^{-a}} \\ &= [\bar{x} + b(x-\bar{x})] \cdot \frac{\mathrm{e}^{-z}}{1-\mathrm{e}^{-a}},\ z>0\end{aligned} \tag{7}$$

其中，x表示SAR图像，$\text{Lee}(\cdot)$表示Lee滤波过程，$\bar{x}$表示SAR图像的均值，a为截断指数分布的参数，b代表权重系数。

2.3 梯度方向寻优器

为了有效提高黑盒攻击成功率，实现快速寻找梯度最优下降方向的目的，本文将AdaBelief[34]和Nesterov[35]算法进行有效结合，设计了梯度方向寻优器，简称ABN。如式(8)和式(9)所示，首先计算梯度g_t的指数移动平均值m_t和$(g_t - m_t)^2$的移动平

均值s_t。当g_t与m_t的差值大时，证明当前预测的梯度偏离前一时刻的梯度，导致学习率下降；然而当预测的梯度与前一时刻的梯度相差较小时，学习率上升。因此，能够实现快速收敛的目的。同时，在式(10)中，通过预先向前走一步，并利用下一时刻的梯度代替当前时刻的梯度。因此具有较好的前瞻性，从而能够有效跳出局部最优。

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) g_t \tag{8}$$

$$s_t = \beta_2 \cdot s_{t-1} + (1-\beta_2)(g_t - m_t)^2 + \varepsilon \tag{9}$$

$$g_t^* = \nabla_{x_t^{\mathrm{adv}}} \left(x_t^{\mathrm{adv}} + \frac{\alpha m_t}{\sqrt{s_t} + \varepsilon} \right) \tag{10}$$

其中，下标t表示迭代至第t步，α表示学习率参数，ε为极小值，用于防止分母为零。

2.4 梯度方向稳定算子

为了稳定梯度的更新方向，有效跳出局部最优，实现进一步提高对抗样本黑盒攻击性能的目的。本文引入拟双曲动量算子[36](简称QHM)，增强梯度方向下降方向的稳定性。该算子优点具体如下：第一，此算法是动量梯度算法的简易变式，计算方面不涉及二次求导，具有计算量小且计算流程简单的特点；第二，此算法能够充分结合历史的梯度动量来修正梯度下降的方向，有效避免陷入局部最优值。该算子的计算公式为

$$g_{t+1} = \beta \cdot g_t + (1-\beta) \cdot \frac{g_t^*}{\|g_t^*\|_1} \tag{11}$$

$$\tilde{g}_{t+1} = (1-v) \cdot \frac{g_t^*}{\|g_t^*\|_1} + v g_{t+1} \tag{12}$$

其中，g_t表示迭代到t时刻累计的梯度，g_t^*表示通过ABN寻优器得到的梯度，v和β为动量系数。

2.5 算法伪代码

TBAA计算公式如算法1所示。具体来说，TBAA首先在步骤1和步骤2中对各个参数进行初始化。步骤4至步骤8为梯度方向寻优器ABN具体计算过程，其中使用m_t累计前t次迭代的梯度，衰减因子为β_1，s_t累计前t次迭代的梯度与m_t之间差值的平方，其衰减因子为β_2，设置稳定系数ζ防止步骤8中公式分母为零。步骤9将随机斑点噪声变换得到的ST($\tilde{x}_t^{\mathrm{adv}}$)输入至模型中，再通过模型的损失函数计算得到梯度$g_t^*$，利用梯度方向稳定算子QHM对步骤10得到的梯度进行更新得到$\tilde{g}_{t+1}$，最终在步骤13生成对抗样本。其中，在步骤9中，当$K=1$时，为单模型攻击；当$K>1$时，为集成模型攻击。

算法 1 基于迁移的SAR-ATR黑盒攻击算法

Alg. 1 SAR-ATR Transfer-based Black-box Attack Algorithm (TBAA)

输入：干净样本x，K个深度神经网络模型$f_1, f_2, \cdots, f_K$，对应的网络模型逻辑值$l_1, l_2, \cdots, l_K$以及相应的网络模型集成权重$w_1, w_2, \cdots, w_K$，扰动量大小ε，步长α，迭代次数T，系数v, β, β_1和β_2

输出：对抗样本x^{adv}

步骤1 $\alpha \leftarrow \varepsilon/T, g_0 \leftarrow 0, m_0 \leftarrow 0, n_0 \leftarrow 0$

步骤2 $g_0 \leftarrow 0, m_0 \leftarrow 0, s_0 \leftarrow 0, x_0^{\mathrm{adv}} \leftarrow x$

步骤3 **For** $t = 0$ to $T-1$ **do**

步骤4 Update m_t by $m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) g_t$

步骤5 Update $\hat{m}_t = \dfrac{m_t}{1-\beta_1^t}$

步骤6 Update $s_t = \beta_2 \cdot s_{t-1} + (1-\beta_2)(\hat{g}_t - m_t)^2$

步骤7 Update $\hat{s}_t = \dfrac{s_t + \zeta}{1-\beta_2^t}$

步骤8 $\tilde{x}_t^{\mathrm{adv}} = x_t^{\mathrm{adv}} + \dfrac{\alpha}{\sqrt{\hat{s}_t + \zeta}} \hat{m}_t$

步骤9 $l(\tilde{x}_t^{\mathrm{adv}}) = \sum_{k=1}^{K} w_k l_k \left(\mathrm{ST}(\tilde{x}_t^{\mathrm{adv}}; p) \right)$

步骤10 Update g_t^* by $g_t^* = \nabla_{x_t^{\mathrm{adv}}} J \left(\mathrm{ST}(\tilde{x}_t^{\mathrm{adv}}; p), y \right)$

步骤11 Update g_{t+1} by $g_{t+1} = \beta g_t + (1-\beta) \cdot \dfrac{g_t^*}{\|g_t^*\|_1}$

步骤12 Update $\tilde{g}_{t+1}$ by $\tilde{g}_{t+1} = (1-v) g_{t+1} + v \cdot \dfrac{g_t^*}{\|g_t^*\|_1}$

步骤13 $x_{t+1}^{\mathrm{adv}} = \mathrm{Clip}_x^{\varepsilon} \left\{ x_t^{\mathrm{adv}} + \alpha \cdot \mathrm{sign}(\tilde{g}_{t+1}) \right\}$

步骤14 **End for**

步骤15 **Return** $x_t^{\mathrm{adv}} = x_{t+1}^{\mathrm{adv}}$

3 实验结果与分析

3.1 实验设置

3.1.1 实验数据集

为了验证本文算法的有效性，本文采用两个SAR数据集，分别是移动和静止目标采集与识别(The Moving and Stationary Target Acquisition and Recognition, MSTAR)[37]和FUSAR-Ship数据集[38]。

MSTAR由美国国防高级研究计划局(DAPRA)和空军研究实验室(AFRL)提供。此数据集利用高分辨率的聚束式合成孔径雷达获取，分辨率为0.3 m×0.3 m。目前，MSTAR数据集广泛应用于国内外SAR-ATR性能评估研究，图像尺寸为128像素×128像素，包括10类军事车辆目标，如2S1, BMP2, BRDM2, BTR60, BTR70, D7, T62, T72, ZIL131和ZSU23/4，其SAR图像如图3所示。此数据集可分为标准操作条件(Standard Operating Condition, SOC)和扩展工作条件(Extended Operating Condition, EOC)。本文算法采用SOC数据集中的10类SAR目标，SOC数据集一般将17°俯仰角的数据作为训练集，将15°俯仰角的数据作为测试集，其具体目标类别和数量分布如表1所示。

FUSAR-Ship数据集由复旦大学电磁波信息科学重点实验室提供。此数据集取自高分三号卫星遥感图像，分辨率为1.124 m×1.728 m，极化模式包含DH和DV，覆盖了各种海、陆、海岸、河流和岛屿场景。FUSAR-Ship数据集适用于复杂海面的船只检测与识别工作，一共包含5000多张不同类别船舶图像，所有图像的尺寸为512像素×512像素。本文选取此数据集中的4类子目标进行实验测试。具体而言，包括BulkCarrier, CargoShip, Fishing和Tanker，其SAR图像如图4所示，训练集和测试集的划分情况如表2所示。

3.1.2 实验网络

本文实验选取10种应用广泛的深度神经网络AlexNet[39], VGGNet16[40], ResNet18[41], ResNet50[41], InceptionV3[42], A-ConvNet[5], MobileNet[43], SqueezeNet[44], PVTv2[45]和MobileViTv2[46]作为SAR-ATR模型。在预处理阶段对SAR图像进行随机翻转、旋转、亮度变化等数据增强操作，在训练阶段，本文通过从训练数据集中统一采样10%的数据来形成验证数据集，本文将学习率设置为0.001，将训练轮数设置为50，将批量大小设置为64，并使用Adam优化器[47]。以上SAR-ATR模型在MSTAR和FUSAR-Ship测试集识别精度如表3所示，其中，FUSAR-Ship数据集仅在上述SAR-ATR模型中的5种模型进行训练。实验使用Windows 10操作系统，PyTorch深度学习开发框架，Python作为开发语言。实验采用的CPU为Intel酷睿i9-11900H，GPU为 NVIDIA GeForce RTX 3080 Laptop GPU。

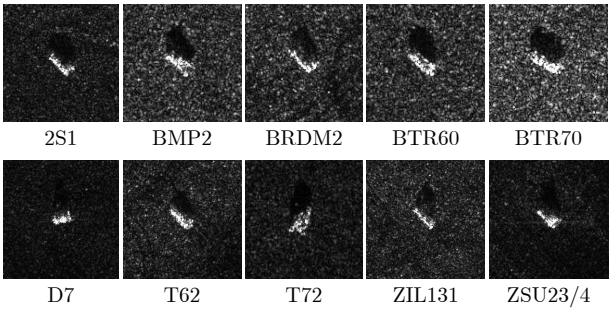


图 3 MSTAR数据集的SAR图像

Fig. 3 SAR images of the MSTAR dataset

表 1　MSTAR数据中SOC下的SAR图像类别与样本数量

Tab. 1　SAR image categories and number of samples under SOC in MSTAR dataset

目标类别	训练集		测试集	
	俯仰角(°)	数量	俯仰角(°)	数量
2S1	17	299	15	274
BRDM2	17	298	15	274
BTR60	17	233	15	195
D7	17	299	15	274
T62	17	299	15	273
ZIL131	17	299	15	274
BMP2	17	233	15	195
ZSU23/4	17	299	15	274
T72	17	232	15	196
BTR70	17	233	15	196

3.1.3 基线设置

在实验中，本文将所提算法与MI-FGSM[20], NAM[21], VMI-FGSM[22], DI-FGSM[23], Attack-Unet-GAN[27]和Fast C&W[26]进行对比分析，其中，MI-FGSM, NAM, VMI-FGSM和DI-FGSM为目前应用较为广泛的基于迁移的黑盒攻击算法，Attack-Unet-GAN和Fast C&W为主流的SAR-ATR攻击算法。

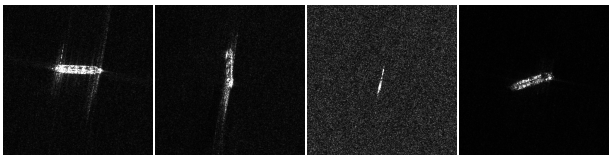


图 4 FUSAR-Ship数据集的SAR图像

Fig. 4 SAR images of the FUSAR-Ship dataset

表 2　FUSAR-Ship数据集中SAR图像类别与样本数量

Tab. 2　SAR image categories and number of samples in FUSAR-Ship dataset

目标类别	训练集数量	测试集数量
BulkCarrier	97	25
CargoShip	126	32
Fishing	75	19
Tanker	36	10

表 3　模型识别精度

Tab. 3　Model recognition accuracy

模型	MSTAR ACC (%)	FUSAR-Ship ACC (%)
AlexNet	95.1	69.47
VGG16	95.6	70.23
ResNet18	96.6	68.10
ResNet50	97.7	—
InceptionV3	99.1	—
A-ConvNet	99.8	—
MobileNet	97.8	—
SqueezeNet	95.4	72.25
PVTv2	98.8	—
MobileViTv2	99.4	72.70

3.1.4 超参数设置

在本文实验中，对于MI-FGSM, NAM, VMI-FGSM, DI-FGSM和TBAA，将最大扰动值ε设置为0.06，迭代次数T设置为10。对于Attack-Unet-GAN和Fast C&W，按照文献[27]和文献[26]的参数设置。对于TBAA，本文遵循以下默认的设置：衰减因子$\beta=0.999$，$\beta_1=0.99$，$\beta_2=0.999$，滑动平均系数$v=0.7$，稳定性参数$\zeta=10\mathrm{E}{-8}$，概率$p=0.5$。

3.1.5 评估指标

本实验从攻击有效性和攻击隐蔽性两个方面对实验结果进行评估。

在攻击有效性方面，实验使用攻击成功率[20]作为评价指标，如式(13)所示：

$$\text{攻击成功率}=\frac{\text{错误分类的样本数量}}{\text{正确分类的样本数量}} \tag{13}$$

其中，正确分类的样本数量为SAR-ATR模型分类正确的样本数量，错误分类的样本数量表示在添加扰动之后输入SAR-ATR导致分类错误的样本数量。

在攻击隐蔽性方面，本文使用平均结构相似度(Average Structural Similarity, ASS)[48]，同时平均结构相似度越高，则攻击的隐蔽性越好，其具体计算公式为

$$\begin{aligned}\mathrm{ASS}(x,x^{\mathrm{adv}})&=\frac{1}{M}\sum_{i=1}^{M}\mathrm{SSIM}(x,x^{\mathrm{adv}})\\&=\frac{1}{M}\sum_{i=1}^{M}\frac{(2\mu_{x_i}\mu_{x_i^{\mathrm{adv}}}+C_1)(2\sigma_{x_ix_i^{\mathrm{adv}}}+C_2)}{(\mu_{x_i}^2+\mu_{x_i^{\mathrm{adv}}}^2+C_1)(\sigma_{x_i}^2+\sigma_{x_i^{\mathrm{adv}}}^2+C_2)}\end{aligned} \tag{14}$$

其中，M为样本数量，x^{adv}表示对抗样本，μ_{x_i}，$\mu_{x_i^{\mathrm{adv}}}$和σ_{x_i}，$\sigma_{x_i^{\mathrm{adv}}}$分别为对应图像的均值和标准差，$\sigma_{x_ix_i^{\mathrm{adv}}}$表示协方差，$C_1$和$C_2$是用于保持度量稳定的常数。

3.2 单模型攻击有效性分析

在本节中，为了验证本文所提算法的攻击性能，分别在MSTAR数据集和FUSAR-Ship数据集上对单个神经网络模型进行对抗攻击。MSTAR数据集和FUSAR-Ship数据集的单模型攻击成功率分别如表4和表5所示，其中标$*$数值表示白盒攻击成功率，其余数值表示黑盒攻击成功率。

通过分析表4可得，在MSTAR数据集上，TBAA算法在所有的黑盒模型上均优于其他基线攻击算法，同时在所有的白盒模型上保持较高的成功率。以在InceptionV3上生成的对抗样本为例，7种基线攻击算法的白盒攻击成功率均达到了100%，同时这7种算法在MobileNet上的攻击成功率分别为31.0%, 33.9%, 34.0%, 33.5%, 25.0%, 12.0%和49.5%。

通过分析表5可得，在FUSAR-Ship数据集上，以在VGGNet16上生成的对抗样本为例，7种基线攻击算法的白盒攻击成功率均在98%以上，7种对比算法在MobileViTv2模型上的攻击成功率分别为46.0%, 50.0%, 52.4%, 52.6%, 26.0%, 24.0%和56.0%。

显然，TBAA的黑盒攻击成功率在所有基线对比算法中是最高的。究其原因，结合2.1节的分析，本文认为，MI-FGSM,NAM和VMI-FGSM算法为通过结合优化算法提升对抗样本的黑盒攻击能力，DI-FGSM算法通过多样化的输入缓解模型对对抗

表 4 MSTAR数据集单模型攻击成功率(%)

Tab. 4 Single model attack success rate on the MSTAR dataset (%)

代理模型	攻击算法	受害者模型									
		AlexNet	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	SqueezeNet	PVTv2	MobileViTv2
AlexNet	MI-FGSM	100*	10.9	12.0	9.0	5.0	28.0	35.0	18.9	14.0	19.6
	NAM	100*	12.0	13.0	10.0	6.9	36.0	37.0	22.9	20.7	22.0
	VMI-FGSM	100*	19.5	19.5	17.0	6.0	29.5	39.5	27.0	40.5	21.5
	DI-FGSM	100*	21.0	26.5	16.0	7.5	29.5	43.5	32.5	32.0	20.5
	Attack-Unet-GAN	98.69*	7.0	8.0	7.5	4.0	20.5	32.5	14.5	12.5	9.5
	Fast C&W	100*	4.5	7.0	6.0	3.0	17.5	19.5	12.5	8.0	3.5
	TBAA	100*	23.5	29.0	20.4	14.5	64.0	53.4	33.9	56.0	32.8
VGGNet16	MI-FGSM	61.0	100*	58.0	56.0	40.0	55.0	41.0	43.0	26.0	30.0
	NAM	60.0	100*	61.0	59.0	42.0	61.0	45.0	47.0	31.0	35.0
	VMI-FGSM	62.5	100*	59.5	58.5	42.5	57.5	41.0	46.5	38.5	38.5
	DI-FGSM	63.5	100*	60.5	67.5	46.5	59.5	42.5	48.0	37.5	38.5
	Attack-Unet-GAN	53.0	100*	40.5	32.5	24.5	32.5	38.5	39.0	23.0	24.5
	Fast C&W	44.5	100*	31.0	37.5	24.0	31.5	22.0	24.5	13.5	14.5
	TBAA	69.5	100*	72.0	78.5	56.9	74.0	56.5	63.5	48.0	48.5

续表 4

代理模型	攻击算法	受害者模型									
		AlexNet	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	SqueezeNet	PVTv2	MobileViTv2
ResNet18	MI-FGSM	13.0	9.9	100*	20.9	13.9	39.0	26.0	15.0	14.0	5.0
	NAM	15.0	9.0	100*	20.9	16.0	38.0	31.0	17.0	21.0	5.3
	VMI-FGSM	17.0	16.5	100*	25.0	15.8	45.5	31.5	25.0	32.5	10.5
	DI-FGSM	18.0	14.0	100*	21.0	19.0	41.0	29.5	30.0	23.5	8.6
	Attack-Unet-GAN	12.5	6.5	100*	11.5	5.0	19.5	18.5	11.5	11.0	3.0
	Fast C&W	10.0	4.0	100*	6.0	3.0	9.0	11.5	12.0	13.5	4.0
	TBAA	29.0	19.0	100*	25.0	35.5	64.0	42.5	30.0	54.0	24.0
ResNet50	MI-FGSM	8.0	12.0	10.5	100*	21.0	16.0	22.9	10.0	12.0	9.0
	NAM	10.0	14.0	14.0	100*	22.0	27.0	24.0	13.0	17.0	14.9
	VMI-FGSM	19.5	19.0	14.5	100*	22.0	33.0	33.5	21.0	28.5	13.5
	DI-FGSM	19.5	19.0	22.5	100*	23.5	26.5	23.0	22.5	28.0	11.5
	Attack-Unet-GAN	6.5	10.5	6.5	100*	7.0	15.0	18.0	8.0	8.0	7.0
	Fast C&W	5.0	4.5	7.5	100*	13.0	7.5	10.5	7.5	10.5	6.0
	TBAA	24.5	19.9	27.5	100*	27.4	48.5	44.9	25.5	43.9	22.0
InceptionV3	MI-FGSM	29.0	31.4	65.5	38.0	100*	65.0	31.0	39.0	12.0	28.0
	NAM	36.0	35.0	67.9	42.0	100*	66.9	33.9	41.9	18.0	29.5
	VMI-FGSM	33.0	31.5	52.5	39.0	100*	68.5	34.0	43.0	30.0	29.5
	DI-FGSM	34.0	34.5	56.0	41.0	100*	66.0	33.5	41.5	23.5	28.5
	Attack-Unet-GAN	20.6	24.5	53.0	31.0	100*	32.5	25.0	26.5	9.0	24.5
	Fast C&W	11.0	16.5	30.0	28.0	100*	20.0	12.0	15.0	10.5	16.5
	TBAA	41.0	50.0	73.5	52.0	100*	76.5	49.5	47.0	45.9	43.9
A-ConvNet	MI-FGSM	19.9	15.5	29.5	20.9	11.5	100*	29.0	15.0	21.9	9.0
	NAM	23.5	17.5	35.5	24.5	18.9	100*	32.5	18.0	24.0	13.0
	VMI-FGSM	25.5	19.0	37.0	25.5	19.5	100*	36.5	31.0	31.0	12.5
	DI-FGSM	28.0	17.5	37.0	23.0	21.5	100*	36.5	29.5	26.5	10.5
	Attack-Unet-GAN	10.8	5.6	9.0	13.0	7.0	98.0*	11.6	11.0	14.7	8.0
	Fast C&W	11.5	4.0	8.5	5.0	3.0	97.5*	10.5	12.5	13.5	4.0
	TBAA	29.5	21.9	40.5	30.5	24.5	100*	38.0	32.9	36.0	24.0
MobileNet	MI-FGSM	16.0	15.1	10.0	15.0	15.6	18.0	100*	18.9	8.0	9.0
	NAM	18.0	14.9	12.0	18.9	18.9	25.0	100*	26.9	9.5	10.5
	VMI-FGSM	21.0	18.0	12.0	18.0	21.5	23.0	100*	23.5	22.0	14.0
	DI-FGSM	19.0	17.5	10.5	17.5	18.0	19.5	100*	20.5	22.0	14.5
	Attack-Unet-GAN	9.0	3.5	7.5	7.8	2.5	12.5	100*	11.0	7.3	5.0
	Fast C&W	10.0	4.0	6.0	5.0	3.0	7.0	100*	10.0	6.5	4.0
	TBAA	24.0	20.5	18.9	26.0	25.4	32.9	100*	30.0	24.0	25.0
SqueezeNet	MI-FGSM	19.5	9.5	20.5	18.0	6.0	40.5	31.4	100*	18.0	18.0
	NAM	18.5	10.3	20.9	19.5	6.5	40.5	32.9	100*	24.0	21.0
	VMI-FGSM	26.5	15.5	28.5	25.5	11.0	42.5	32.0	100*	28.5	19.5
	DI-FGSM	21.0	11.5	30.5	22.5	12.0	41.0	31.5	100*	23.0	21.5
	Attack-Unet-GAN	13.0	8.0	16.5	17.0	4.5	17.5	17.0	100*	12.5	14.5
	Fast C&W	10.0	4.5	7.0	5.5	3.0	18.0	10.0	100*	13.5	14.0
	TBAA	28.0	18.5	32.5	31.0	12.5	53.5	38.5	100*	41.9	39.0

续表 4

代理模型	攻击算法	受害者模型									
		AlexNet	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	SqueezeNet	PVTv2	MobileViTv2
PVTv2	MI-FGSM	10.0	7.3	9.0	12.0	15.5	6.0	18.0	7.8	100*	11.3
	NAM	13.0	3.5	10.7	13.5	21.5	10.4	19.9	9.0	100*	18.5
	VMI-FGSM	12.0	12.0	9.5	20.0	22.5	16.0	23.0	11.0	100*	19.5
	DI-FGSM	11.0	13.5	11.0	15.0	23.5	12.0	23.5	12.6	100*	13.0
	Attack-Unet-GAN	8.5	5.0	7.5	7.9	12.5	3.5	11.6	4.5	100*	9.0
	Fast C&W	10.0	4.0	6.5	4.5	13.0	5.5	9.0	3.7	100*	4.0
	TBAA	26.0	23.0	22.0	27.0	35.0	28.9	37.0	25.0	100*	32.0
MobileViTv2	MI-FGSM	14.0	16.0	19.0	18.3	7.9	43.8	30.0	18.0	52.0	100*
	NAM	21.4	24.0	26.2	20.7	11.6	47.8	33.9	25.4	58.0	100*
	VMI-FGSM	21.0	25.0	29.0	21.5	11.5	45.0	35.0	27.0	56.0	99.5*
	DI-FGSM	22.5	23.1	29.0	24.0	10.5	46.3	38.0	27.5	58.5	98.0*
	Attack-Unet-GAN	11.0	6.5	14.0	11.5	5.5	29.0	15.5	15.0	46.0	100*
	Fast C&W	11.0	4.0	8.5	5.5	3.0	17.5	10.5	11.5	45.0	99.5*
	TBAA	40.0	31.9	33.9	35.9	20.3	66.0	48.0	45.9	65.9	100*

注：标红字体为最优值，标蓝字体为次优值。*表示白盒攻击成功率，其余数值表示黑盒攻击成功率。

表 5 FUSAR-Ship数据集单模型攻击成功率(%)

Tab. 5 Single model attack success rate on the FUSAR-Ship dataset (%)

代理模型	攻击算法	受害者模型				
		AlexNet	VGGNet16	ResNet18	SqueezeNet	MobileViTv2
AlexNet	MI-FGSM	100*	38.00	40.00	33.90	68.00
	NAM	100*	48.00	62.00	45.90	76.00
	VMI-FGSM	100*	47.10	63.60	42.60	70.00
	DI-FGSM	98.41*	47.40	63.56	44.93	74.94
	Attack-Unet-GAN	100*	23.80	33.20	15.60	30.00
	Fast C&W	99.96*	18.70	29.10	12.40	24.00
	TBAA	100*	60.00	84.00	56.00	80.00
VGGNet16	MI-FGSM	28.00	100*	24.00	40.00	46.00
	NAM	33.90	100*	30.00	38.00	50.00
	VMI-FGSM	33.90	98.62*	28.10	42.80	52.40
	DI-FGSM	37.60	98.76*	36.60	42.40	52.60
	Attack-Unet-GAN	13.50	100*	20.40	24.00	26.00
	Fast C&W	9.30	99.96*	19.60	22.90	24.00
	TBAA	43.90	100*	48.00	62.00	56.00
ResNet18	MI-FGSM	6.00	7.90	100*	15.90	40.00
	NAM	7.90	9.90	100*	21.90	50.00
	VMI-FGSM	9.30	10.60	100*	28.60	53.80
	DI-FGSM	13.50	12.80	99.96*	29.91	54.60
	Attack-Unet-GAN	4.50	5.60	100*	10.40	17.80
	Fast C&W	5.30	6.20	99.98*	6.70	12.90
	TBAA	38.00	18.00	100*	50.00	60.00
SqueezeNet	MI-FGSM	16.00	9.90	28.00	100*	45.90
	NAM	21.90	14.00	43.90	100*	56.00
	VMI-FGSM	25.10	19.30	44.20	99.86*	53.30
	DI-FGSM	25.07	16.32	45.39	99.59*	59.40
	Attack-Unet-GAN	14.50	7.90	22.00	100*	28.00
	Fast C&W	10.10	6.30	20.10	98.69*	26.90
	TBAA	39.90	31.90	65.90	100*	64.00

续表 5

代理模型	攻击算法	受害者模型				
		AlexNet	VGGNet16	ResNet18	SqueezeNet	MobileViTv2
MobileViTv2	MI-FGSM	4.00	7.90	42.00	21.90	100*
	NAM	9.00	16.00	48.00	26.00	100*
	VMI-FGSM	12.80	23.50	45.90	24.10	100*
	DI-FGSM	13.20	18.60	47.00	26.40	99.52*
	Attack-Unet-GAN	2.90	5.30	25.60	18.00	100*
	Fast C&W	2.60	4.20	21.60	14.60	100*
	TBAA	24.00	16.00	67.90	43.90	100*

注：标红字体为最优值，标蓝字体为次优值。*表示白盒攻击成功率，其余数值表示黑盒攻击成功率。

样本的过拟合，从而有效攻击黑盒模型，Attack-Unet-GAN和Fast C&W在生成对抗样本的过程中容易对白盒模型产生过拟合的情况，然而本文所提算法TBAA在利用随机斑点噪声变换得到多样化输入的同时结合更加优秀的优化算法，因此能够生成具有最强迁移性能的对抗样本。

3.3 集成模型攻击有效性分析

3.3.1 集成模型攻击实验结果

虽然在3.2节中，本文提出的攻击算法在单模型黑盒攻击方面，能够有效提升对抗样本的迁移性，但还可以通过集成模型的方法生成迁移性更强的对抗样本，本文通过对多个网络逻辑值的集成来进行攻击。在本节实验中，分别利用MI-FGSM, NAM,VMI-FGSM, DI-FGSM, Attack-Unet-GAN, FastC&W以及TBAA针对多个正常训练的SAR-ATR模型的集成来生成对抗样本，并在所有的网络上进行测试。

表6给出7种对抗攻击算法在黑盒条件下对集成模型的攻击成功率。MSTAR数据集上的“AlexNet”一列中，本节通过将除AlexNet外其他9个模型进行集成生成对抗样本，即为黑盒条件下。同理，FUSAR-Ship数据集上的“AlexNet”一列表示在除AlexNet外其他4个模型进行集成生成对抗样本。在具有挑战性的黑盒模型下，本文算法TBAA总是在所有的网络上生成比其他基线算法具有更好迁移性的对抗样本。例如，通过将VGGNet16设为黑盒模型，在MSTAR数据集上，7种攻击算法的攻击成功率分别为39.0%, 41.5%, 43.5%, 44.3%, 30.5%, 26.8%, 62.0%；在FUSAR-Ship数据集上，7种攻击算法的攻击成功率分别为40.9%, 50.5%, 53.2%, 56.0%, 25.0%, 22.5%和62.0%。与基线算法相比，本文算法在MSTAR数据集上的黑盒攻击成功率提高了3%～55%；在FUSAR-Ship数据集上的黑盒攻击成功率提高了6.0%～57.5%。

此外，在MSTAR数据集上，与表4针对VGG-Net16单模型攻击中的最高黑盒攻击成功率50%

表 6　集成模型攻击成功率(%)

Tab. 6　Ensemble model attack success rate (%)

数据集	攻击算法	AlexNet	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	SqueezeNet	PVTv2	MobileViTv2
MSTAR	MI-FGSM	62.9	39.0	52.0	65.0	42.0	67.9	50.0	51.5	68.0	46.0
	NAM	63.1	41.5	68.2	70.5	45.0	75.7	53.2	54.0	75.6	51.4
	VMI-FGSM	66.4	43.5	72.5	65.8	46.5	74.6	52.0	53.0	76.3	56.0
	DI-FGSM	69.0	44.3	74.0	70.0	50.3	76.0	55.0	55.8	70.0	51.0
	Attack-Unet-GAN	53.6	30.5	47.0	35.0	30.0	35.0	41.0	43.0	52.3	31.0
	Fast C&W	46.0	26.8	35.0	38.0	28.0	33.0	28.5	30.0	51.0	24.0
	TBAA	72.0	62.0	86.0	88.0	70.0	88.0	70.0	66.0	92.0	78.0
FUSAR-Ship	MI-FGSM	31.9	40.9	48.0	—	—	—	—	45.9	—	71.9
	NAM	34.5	50.5	68.5	—	—	—	—	48.5	—	78.0
	VMI-FGSM	35.8	53.2	67.0	—	—	—	—	51.3	—	76.5
	DI-FGSM	36.0	56.0	68.0	—	—	—	—	50.0	—	78.0
	Attack-Unet-GAN	16.0	25.0	38.0	—	—	—	—	28.5	—	38.4
	Fast C&W	12.5	22.5	34.2	—	—	—	—	26.0	—	32.0
	TBAA	70.0	62.0	86.0	—	—	—	—	64.0	—	88.0

注：标红数字为最优值，标蓝数字为次优值。

相比，本节集成攻击能够有效提高对抗样本的迁移性能。

3.3.2 参数影响分析

本节在MSTAR数据集上进一步分析不同参数对TBAA的影响。

概率p：首先，本文研究在白盒和黑盒模型下，概率p对攻击成功率的影响。概率p的取值范围在0到1之间。图5展示了TBAA算法分别在白盒和黑盒模型下的攻击成功率。从图中可分析得，随着p的增大，算法的白盒攻击成功率逐渐降低，黑盒成功率逐渐升高。究其原因，本文认为随着p的增大，增加了输入样本的多样性，能够有效缓解过拟合的问题，从而提高黑盒攻击有效性。因此，图5所示的变化趋势能够为实际中构建有效的对抗攻击提供有用的建议，例如，可以选择一个合适的概率p值，在满足白盒攻击成功率大于等于90%的条件下，最大限度提高黑盒攻击成功率。

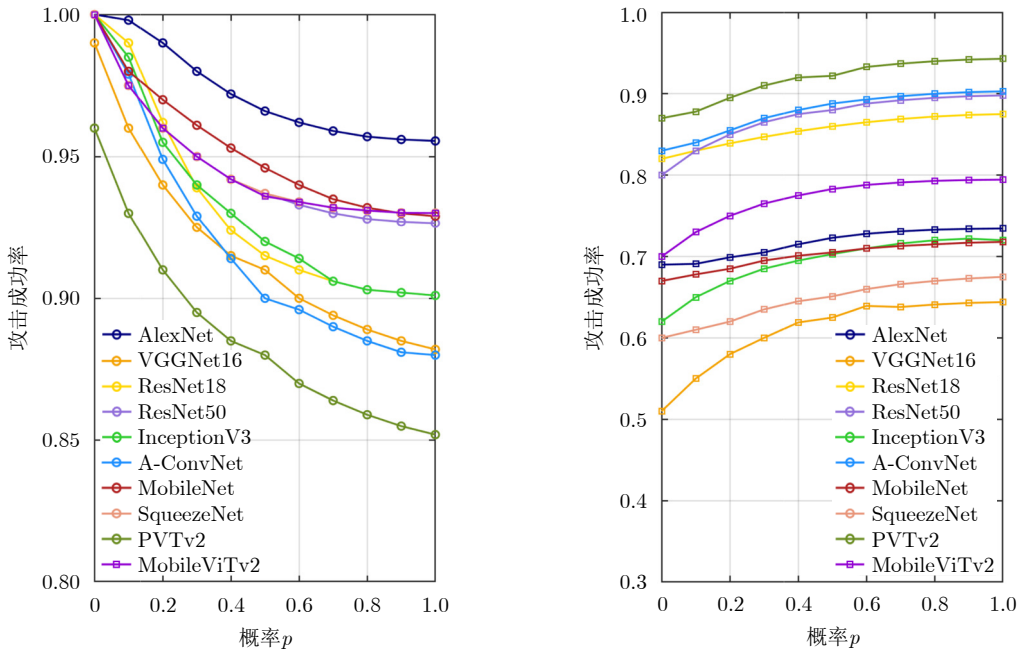
最大扰动量ε：接下来，重点研究在黑盒模型下，最大扰动量ε对攻击成功率的影响。在实验中设置最大扰动量ε从0.02变化到0.10。图6(a)展示了在不同网络模型下的攻击成功率，从中可得，随着最大扰动量ε的增加，TBAA的攻击成功率升高。

迭代次数T：最后，本文研究在黑盒模型下，迭代次数T对攻击成功率的影响。实验设置迭代次数从2以2的步长大小变化到12，具体实验结果如图6(b)所示。从中可得出，当迭代次数较小时，攻击成功率上升幅度相对较大；当迭代次数大于等于8时，攻击成功率上升速度变缓。以此类推，当迭代次数不断增大时，攻击成功率将趋于稳定。

3.3.3 消融实验

本节主要分析所提算法中各个模块对攻击有效性的提升。实验以MI-FGSM为基准算法，通过逐步添加本文各个模块来生成对抗样本，具体对比方法设置如表7所示。本节在MSTAR数据集和FUSAR-Ship数据集上将表7中的4种对比算法在集成模型条件下进行对比分析，具体实验结果如表8所示。

通过对比AN-QHMI-FGSM和MI-FGSM算法在两个数据集上的攻击成功率可得，梯度方向稳定算子QHM通过稳定梯度更新的方向帮助算法找到全局最优解，从而提高对抗样本的迁移性能；当梯度更新方向不稳定时，优化过程可能会出现震荡或停滞的情况，导致收敛速度较慢或无法达到较好的解。因此梯度方向稳定算子QHM有助于提高算法的黑盒攻击能力。ABN-QHMI-FGSM算法在AN-QHMI-FGSM集成了梯度方向寻优器ABN，实验结果表明ABN-QHMI-FGSM的黑盒攻击成功率优于AN-QHMI-FGSM，其原因在于ABN具有前瞻性和自适应性，能够结合历史和未来的梯度信息更新梯度方向，有效避免算法陷入局部最优解，进一步对抗样本的迁移性能。同时，TBAA算法在ABN-

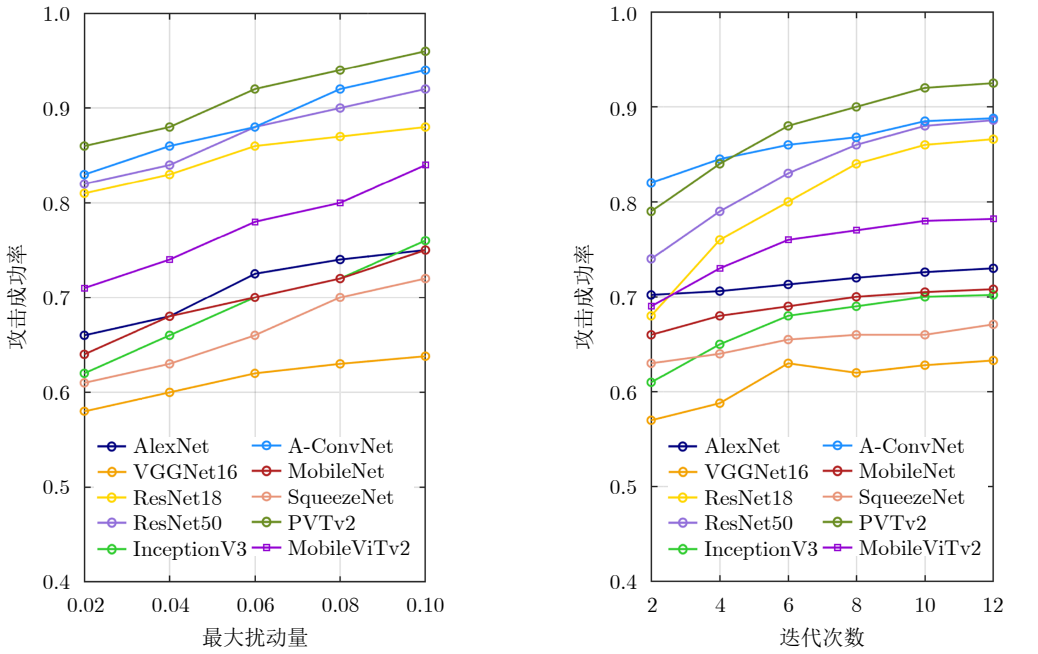


(a) 白盒模型
(a) White-box model

(b) 黑盒模型
(b) Black-box model

图 5 攻击成功率随概率p变化折线图

Fig. 5 The attack success rate changes with probability p



(a) 攻击成功率随最大扰动量变化折线图
(a) Line chart of the attack success rate changing with the maximum perturbation amount

(b) 攻击成功率随迭代次数变化折线图
(b) Line chart of attack success rate changing with the number of iterations

图 6 黑盒模型下攻击成功率变化折线图

Fig. 6 Line chart of attack success rate change under black-box model

QHMI-FGSM的基础上结合了随机斑点噪声变换模块，在深度学习中通常利用数据增强提升模型的泛化性，在本文算法利用随机斑点噪声变换增加输入的多样性，缓解白盒模型对对抗样本的过拟合，最终有效增强对抗样本的迁移性。

3.4 攻击隐蔽性评估

在实际对抗攻击的应用中，SAR对抗样本面临着两个挑战，其一是对SAR目标识别网络的欺骗，其二是其对人工检查的欺骗。其中完成对SAR目标识别网络的欺骗仅需要使其判断错误即可，而对于人工检查，SAR对抗样本需要与原始图像保持高度相似性。本节为了验证SAR对抗样本的隐蔽性，利用3.1.5节的平均结构相似度对原始SAR图像与SAR对抗样本图像的差异进行评估。在本节中，主要评估在MSTAR数据集上，集成模型攻击生成的对抗样本的攻击隐蔽性。实验结果如表9所示，与基线算法相比，本文算法在所有DNN模型的平均

表 7 消融实验方法设置

Tab. 7 Ablation experiment method setup

攻击算法	QHM	ABN	ST
MI-FGSM	—	—	—
AN-QHMI-FGSM	√	—	—
ABN-QHMI-FGSM	√	√	—
TBAA	√	√	√

表 8 消融实验攻击成功率(%)

Tab. 8 Ablation experiment attack success rate (%)

数据集	攻击算法	AlexNet	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	SqueezeNet	PVTv2	MobileViTv2
MSTAR	MI-FGSM	62.9	39.0	52.0	65.0	42.0	67.9	50.0	51.5	68.0	46
	AN-QHMI-FGSM	65.7	48.0	75.0	78.0	56.0	82.0	56.0	57.2	82.0	58.0
	ABN-QHMI-FGSM	69.3	51.6	82.0	81.0	63.0	85.2	68.3	60.8	88.3	69.5
	TBAA	72.0	62.0	86.0	88.0	70.0	88.0	70.0	66.0	92.0	78.0
FUSAR-Ship	MI-FGSM	31.9	40.9	38.0	—	—	—	—	45.9	—	71.9
	AN-QHMI-FGSM	36.9	52.0	76.0	—	—	—	—	50.0	—	81.6
	ABN-QHMI-FGSM	43.9	59.9	81.5	—	—	—	—	52.0	—	85.0
	TBAA	70.0	62.0	86.0	—	—	—	—	64.0	—	88.0

注：标红数字为最优值，标蓝数字为次优值。

ASS最高。分析原因，本文设计的ABN寻优器具有快速收敛和前瞻性的特点，同时QHM算子能够稳定梯度的下降方向。因此，所提算法可以有效找到攻击性最强的梯度下降方向；并且本文算法利用L_∞范数和Clip(·)函数限制每次迭代扰动的最大范围，实现仅通过小幅度改变SAR图像像素值，即在保持图像结构相似性的同时，提高对抗样本的黑盒攻击成功率。

图7展示了由TBAA算法生成的对抗样本、相对应的原始干净样本以及对抗扰动图像，从图中可以看出原始图像和对应生成的对抗样本之间的差别很小，即对抗扰动是人眼几乎不可见的。

3.5 攻击效率评估

由于本文算法基于MI-FGSM算法的思想，设计了ABN寻优器、QHM算子以及随机斑点噪声变换，在一定程度上增加了算法的复杂度。因此本节探究在黑盒条件下，不同基线攻击算法在单模型和多模型集成攻击时生成对抗样本的运算时间，并将运算时间的长短作为衡量攻击算法效率的指标。

如表10所示，以AlexNet为例，在MSTAR数据集上分别进行单模型黑盒攻击运算时间测试和集成模型黑盒攻击时间测试。由表中数据可得，Attack-Unet-GAN和Fast C&W生成对抗样本的运算时间最短且较为接近，因为Attack-Unet-GAN和Fast C&W均为通过训练好的U-Net生成器模型一步映射得到对抗样本，生成对抗样本的速度远远快于基于梯度的对抗攻击算法。在其余5种基于梯度的对抗攻击算法中，MI-FGSM的运算时间最短，TBAA的运算时间最长。在单模型黑盒攻击中，TBAA与MI-FGSM的运算时间差值最大约为0.0659 s，最短约为0.0459 s。因此，本文的攻击算法虽然在一定程度上增加了运算时间，生成对抗样本的效率有所下降，但仍在可接受范围之内，没有导致增加大量的额外运算时间。

4 结语

本文结合MI-FGSM的思想，考虑到神经网络的训练过程与基于梯度的对抗攻击算法是相似的，两者均是采用梯度迭代算法对参数进行更新。基于此，提升神经网络模型泛化性能的算法同样也可以用于提升对抗样本的黑盒攻击性能。因此，本文算法结合SAR图像的特点，利用随机斑点噪声变换有效缓解模型的过拟合情况；为了加快收敛速度，设计了梯度方向寻优器ABN，实现提高黑盒攻击成功率的目的；利用QHM算子能够在深度神经网络

表 9 MSTAR数据集在集成模型攻击下原始SAR图像和SAR对抗样本的平均结构相似度

Tab. 9 ASS of original SAR images and SAR adversarial examples under ensemble model attack on MSTAR dataset

攻击算法	AlexNet	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	SqueezeNet	PVTv2	MobileViTv2	Mean
MI-FGSM	0.951	0.959	0.968	0.976	0.970	0.962	0.969	0.960	0.963	0.960	0.9638
NAM	0.962	0.965	0.971	0.978	0.973	0.967	0.973	0.966	0.968	0.962	0.9685
VMI-FGSM	0.965	0.961	0.972	0.976	0.975	0.969	0.977	0.967	0.969	0.965	0.9696
DI-FGSM	0.960	0.970	0.974	0.974	0.976	0.971	0.979	0.974	0.970	0.963	0.9711
Attack-Unet-GAN	0.968	0.975	0.975	0.978	0.978	0.974	0.982	0.975	0.972	0.968	0.9745
Fast C&W	0.969	0.974	0.976	0.979	0.979	0.974	0.980	0.975	0.971	0.967	0.9744
TBAA	0.969	0.975	0.979	0.981	0.978	0.975	0.981	0.973	0.972	0.967	0.9750

注：标红数字为最优值，标蓝数字为次优值。

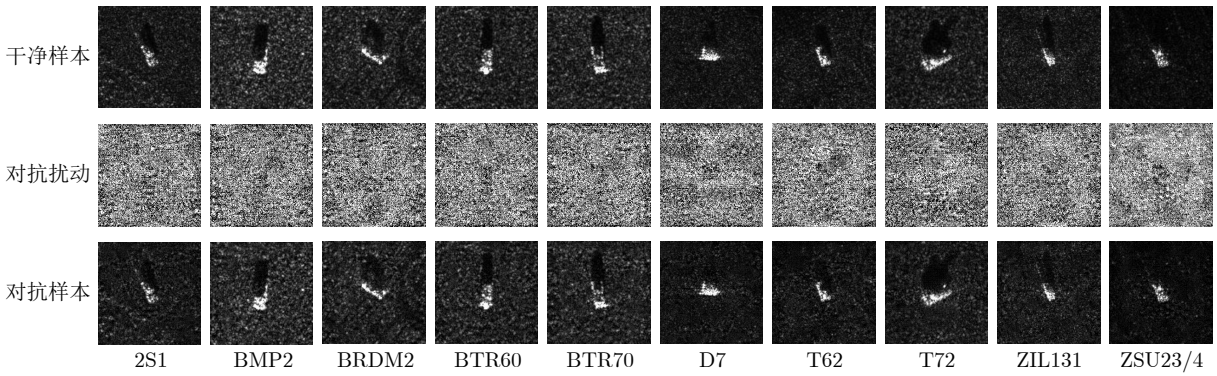


图 7 TBAA干净样本、对抗扰动以及对抗样本展示

Fig. 7 TBAA clean examples, adversarial perturbations and adversarial examples display

表 10　对抗样本生成效率(s)

Tab. 10　Adversarial examples generation efficiency (s)

攻击方法	VGGNet16	ResNet18	ResNet50	InceptionV3	A-ConvNet	MobileNet	Squeezenet	Ensemble
MI-FGSM	0.2970	0.2404	0.3621	0.5217	0.1797	0.3258	0.2550	1.8953
NAM	0.3014	0.2410	0.3623	0.5303	0.1822	0.3248	0.2257	1.8973
VMI-FGSM	0.2980	0.2498	0.3625	0.5289	0.1826	0.3289	0.2274	1.9766
DI-FGSM	0.2984	0.2485	0.3623	0.5280	0.1823	0.3283	0.2294	1.9795
Attack-Unet-GAN	0.0052	0.0052	0.0052	0.0052	0.0052	0.0052	0.0052	0.0052
Fast C&W	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
TBAA	0.3588	0.3046	0.4159	0.5676	0.2456	0.3876	0.2824	2.1357

注：标红数字为最大值，标蓝数字为最小值。

训练中稳定梯度下降方向的优势，进一步提升黑盒攻击的有效性。同时，本文结合集成学习的思想，利用集成模型的方法实现对多个模型保持攻击的有效性，从而更加容易成功攻击其他模型。实验结果表明，本文算法能够在不增加大量额外运行时间的情况下，有效提升黑盒攻击能力和隐蔽性。随着SAR智能识别技术的发展，考虑到在实际对抗场景下难以获取敌方SAR的波位参数，因此在未来的研究中将进一步探索不依赖于视角和成像参数的通用对抗攻击算法。

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] XU Yan and SCOOT K A. Sea ice and open water classification of SAR imagery using CNN-based transfer learning[C]. 2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 2017: 3262–3265. doi: 10.1109/IGARSS.2017.8127693.

[2] ZHANG Yue, SUN Xian, SUN Hao, *et al*. High resolution SAR image classification with deeper convolutional neural network[C]. International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018: 2374–2377. doi: 10.1109/IGARSS.2018.8518829.

[3] SHAO Jiaqi, QU Changwen, and LI Jianwei. A performance analysis of convolutional neural network models in SAR target recognition[C]. 2017 SAR in Big Data Era: Models, Methods and Applications, Beijing, China, 2017: 1–6. doi: 10.1109/BIGSARDATA.2017.8124917.

[4] ZHANG Ming, AN Jubai, YU Dahua, *et al*. Convolutional neural network with attention mechanism for SAR automatic target recognition[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 4004205. doi: 10.1109/LGRS.2020.3031593.

[5] 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.

[6] 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.

[7] 吕艺璇, 王智睿, 王佩瑾, 等. 基于散射信息和元学习的SAR图像飞机目标识别[J]. 雷达学报, 2022, 11(4): 652–665. doi: 10.12000/JR22044.

LYU Yixuan, WANG Zhirui, WANG Peijin, *et al*. Scattering information and meta-learning based SAR images interpretation for aircraft target recognition[J]. *Journal of Radars*, 2022, 11(4): 652–665. doi: 10.12000/JR22044.

[8] HUANG Teng, ZHANG Qixiang, LIU Jiabao, *et al*. Adversarial attacks on deep-learning-based SAR image target recognition[J]. *Journal of Network and Computer Applications*, 2020, 162: 102632. doi: 10.1016/j.jnca.2020.102632.

[9] 孙浩, 陈进, 雷琳, 等. 深度卷积神经网络图像识别模型对抗鲁棒性技术综述[J]. 雷达学报, 2021, 10(4): 571–594. doi: 10.12000/JR21048.

SUN Hao, CHEN Jin, LEI Lin, *et al*. Adversarial robustness of deep convolutional neural network-based image recognition models: A review[J]. *Journal of Radars*, 2021, 10(4): 571–594. doi: 10.12000/JR21048.

[10] 高勋章, 张志伟, 刘梅, 等. 雷达像智能识别对抗研究进展[J]. 雷达学报, 2023, 12(4): 696–712. doi: 10.12000/JR23098.

GAO Xunzhang, ZHANG Zhiwei, LIU Mei, *et al*. Intelligent radar image recognition countermeasures: A review[J]. *Journal of Radars*, 2023, 12(4): 696–712. doi: 10.12000/JR23098.

[11] SZEGEDY C, ZAREMBA W, SUTSKEVER I, *et al*. Intriguing properties of neural networks[C]. The 2nd

International Conference on Learning Representations, Banff, Canada, 2014.

[12] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[C]. The 3rd International Conference on Learning Representations, San Diego, CA, USA, 2015: 1050.

[13] KURAKIN A, GOODFELLOW L J, and BENGIO S. Adversarial examples in the physical world[C]. The 5th International Conference on Learning Representations, Toulon, France, 2017: 99–112.

[14] PAPERNOT N, MCDANIEL P, JHA S, *et al*. The limitations of deep learning in adversarial settings[C]. 2016 IEEE European Symposium on Security and Privacy, Saarbruecken, Germany, 2016: 372–387. doi: 10.1109/EuroSP.2016.36.

[15] BRENDEL W, RAUBER J, and BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]. The 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

[16] CARLINI N and WAGNER D. Towards evaluating the robustness of neural networks[C]. 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 2017: 39–57. doi: 10.1109/SP.2017.49.

[17] SU Jiawei, VARGAS D V, and SAKURAI K. One pixel attack for fooling deep neural networks[J]. *IEEE Transactions on Evolutionary Computation*, 2019, 23(5): 828–841. doi: 10.1109/TEVC.2019.2890858.

[18] CHEN Pinyu, ZHANG Huan, SHARMA Y, *et al*. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]. The 10th ACM Workshop on Artificial Intelligence and Security, Dallas, USA, 2017: 15–26. doi: 10.1145/3128572.3140448.

[19] CHEN Jianbo, JORDAN M I, and WAINWRIGHT M J. HopSkipJumpAttack: A query-efficient decision-based attack[C]. 2020 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 2020: 1277–1294. doi: 10.1109/SP40000.2020.00045.

[20] DONG Yinpeng, LIAO Fengzhou, PANG Tianyu, *et al*. Boosting adversarial attacks with momentum[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 9185–9193. doi: 10.1109/CVPR.2018.00957.

[21] ZHAO Haojun, LIN Yun, GAO Song, *et al*. Evaluating and improving adversarial attacks on DNN-based modulation recognition[C]. GLOBECOM 2020–2020 IEEE Global Communications Conference, Taipei, China, 2020: 1–5. doi: 10.1109/GLOBECOM42002.2020.9322088.

[22] WANG Xiaosen and HE Kun. Enhancing the transferability of adversarial attacks through variance tuning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2021: 1924–1933. doi: 10.1109/CVPR46437.2021.00196.

[23] XIE Cihang, ZHANG Zhishuai, ZHOU Yuyin, *et al*. Improving transferability of adversarial examples with input diversity[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019: 2725–2734. doi: 10.1109/CVPR.2019.00284.

[24] CZAJA W, FENDLEY N, PEKALA M J, *et al*. Adversarial examples in remote sensing[C]. The 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, USA, 2018: 408–411. doi: 10.1145/3274895.3274904.

[25] CHEN Li, XU Zewei, LI Qi, *et al*. An empirical study of adversarial examples on remote sensing image scene classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(9): 7419–7433. doi: 10.1109/TGRS.2021.3051641.

[26] DU Chuan, HUO Chaoying, ZHANG Lei, *et al*. Fast C&W: A fast adversarial attack algorithm to fool SAR target recognition with deep convolutional neural networks[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 4010005. doi: 10.1109/LGRS.2021.3058011.

[27] DU Chuan and ZHANG Lei. Adversarial attack for SAR target recognition based on UNet-generative adversarial network[J]. *Remote Sensing*, 2021, 13(21): 4358. doi: 10.3390/rs13214358.

[28] ZHOU Junfan, SUN Hao, and KUANG Gangyao. Template-based universal adversarial perturbation for SAR target classification[C]. The 8th China High Resolution Earth Observation Conference, Singapore, Singapore, 2023: 351–360. doi: 10.1007/978-981-19-8202-6_32.

[29] XIA Weijie, LIU Zhe, and LI Yi. SAR-PeGA: A generation method of adversarial examples for SAR image target recognition network[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2023, 59(2): 1910–1920. doi: 10.1109/TAES.2022.3206261.

[30] PENG Bowen, PENG Bo, ZHOU Jie, *et al*. Scattering model guided adversarial examples for SAR target recognition: Attack and defense[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5236217. doi: 10.1109/TGRS.2022.3213305.

[31] HANSEN L K and SALAMON P. Neural network ensembles[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, 12(10): 993–1001. doi: 10.1109/34.58871.

[32] DING Jun, CHEN Bo, LIU Hongwei, *et al*. Convolutional neural network with data augmentation for SAR target recognition[J]. *IEEE Geoscience and Remote Sensing*

Letters, 2016, 13(3): 364–368. doi: 10.1109/LGRS.2015.2513754.

[33] LEE J S. Digital image enhancement and noise filtering by use of local statistics[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1980, PAMI-2(2): 165–168. doi: 10.1109/TPAMI.1980.4766994.

[34] ZHUANG Juntang, TANG T, DING Yifan, *et al*. AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients[C]. The 34th International Conference on Neural Information Processing Systems, 2020: 795–806.

[35] NESTEROV Y. A method for unconstrained convex minimization problem with the rate of convergence[J]. *Mathematics*, 1983, 269: 543–547.

[36] MA J and YARATS D. Quasi-hyperbolic momentum and Adam for deep learning[C]. The 7th International Conference on Learning Representations, New Orleans, LA, USA, 2019: 1–38.

[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: 228–242. doi: 10.1117/12.242059.

[38] HOU Xiyue, AO Wei, SONG Qian, *et al*. FUSAR-Ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition[J]. *Science China Information Sciences*, 2020, 63(4): 140303. doi: 10.1007/s11432-019-2772-5.

[39] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. The 25th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012: 1106–1114.

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

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

[42] SZEGEDY C, VANHOUCKE V, IOFFE S, *et al*. Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 2818–2826. doi: 10.1109/CVPR.2016.308.

[43] HOWARD A G, ZHU Menglong, CHEN Bo, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. https://arxiv.org/abs/1704.04861, 2017.

[44] IANDOLA F N, HAN Song, MOSKEWICZ M W, *et al*. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[EB/OL]. https://arxiv.org/abs/1602.07360, 2016.

[45] WANG Wenhai, XIE Enze, LI Xiang, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 2021: 548–558. doi: 10.1109/ICCV48922.2021.00061.

[46] MEHTA S and RASTEGARI M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer[C]. The Tenth International Conference on Learning Representations, 2022.

[47] KINGMA D P and BA J. Adam: A method for stochastic optimization[C]. The 3rd International Conference on Learning Representations, San Diego, CA, USA, 2015: 1–15.

[48] WANG Zhou, BOVIK A C, SHEIKH H R, *et al*. Image quality assessment: From error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612. doi: 10.1109/TIP.2003.819861.

作 者 简 介

万炬申，硕士生，主要研究方向为SAR智能解译与对抗。

刘　伟，副教授，博士，主要研究方向为智能信息处理、遥感图像分析。

牛朝阳，副教授，博士，主要研究方向为SAR信息处理与对抗。

卢万杰，讲师，博士，主要研究方向为智能信息处理、遥感图像分析。

(责任编辑：于青)