研究与开发

基于掩模提取的SAR图像对抗样本生成方法

章坚武[1]，能豪[1]，李杰[1]，钱建华[2]，方银锋[1]
（1. 杭州电子科技大学，浙江 杭州 310018；
2. 中国联通（浙江）产业互联网有限公司，浙江 杭州 311199）

摘　要：合成孔径雷达（synthetic aperture radar，SAR）图像的对抗样本生成在当前已经有很多方法，但仍存在对抗样本扰动量较大、训练不稳定以及对抗样本的质量无法保证等问题。针对上述问题，提出了一种SAR图像对抗样本生成模型，该模型基于AdvGAN模型架构，首先根据SAR图像的特点设计了一种由增强Lee滤波器和最大类间方差法（OTSU）自适应阈值分割等模块组成的掩模提取模块，这种方法产生的扰动量更小，与原始样本的结构相似性（structural similarity，SSIM）值达到0.997以上。其次将改进的相对均值生成对抗网络（relativistic average generative adversarial network，RaGAN）损失引入AdvGAN中，使用相对均值判别器，让判别器在训练中同时依赖于真实数据和生成的数据，提高了训练的稳定性与攻击效果。在MSTAR数据集上与相关方法进行了实验对比，实验表明，此方法生成的SAR图像对抗样本在攻击防御模型时的攻击成功率较传统方法提高了10%～15%。

关键词：对抗样本；生成对抗网络；合成孔径雷达；半白盒攻击；掩模提取
中图分类号：TN957
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024081

Adversarial example generation method for SAR images based on mask extraction

ZHANG Jianwu[1], NAI Hao[1], LI Jie[1], QIAN Jianhua[2], FANG Yinfeng[1]
1. Hangzhou Dianzi University, Hangzhou 310018, China
2. China Unicom (Zhejiang) Industrial Internet Co., Ltd., Hangzhou 311199, China

Abstract: There are many ways to generate adversarial samples for synthetic aperture radar (SAR) images at present, but some problems such as large amount of perturbation of adversarial samples, unstable training, and unguaranteed quality of adversarial samples still exist. To solve the above problems, a SAR image adversarial sample generation model was proposed. The model was based on the AdvGAN model architecture. Firstly, according to the characteristics of the SAR images, an adaptive threshold segmentation method based on the en-

收稿日期：2023−09−30；**修回日期：**2024−01−10
通信作者：章坚武，jwzhang@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.IEC\NSFC\181300）；浙江省自然科学基金重点项目（No.LZ23F010001）
Foundation Items: The National Natural Science Foundation of China (No.IEC\NSFC\181300), The Natural Science Foundation of Zhejiang Province (No.LZ23F010001)

hanced Lee filter OTSU was designed. The mask extraction module composed of equal modules, this method produced a smaller amount of disturbance, and the structural similarity (SSIM) with the original sample reached that more than 0.997. Secondly, the improved relativistic average GAN (RaGAN) loss was introduced into AdvGAN, and the relative mean discriminator was used to make the discriminator rely on both real data and generated data during training, which improved the stability of training and the attack effect. Experiments were compared with related methods on the MSTAR dataset. Experiments show that the attack success rate of SAR image adversarial samples generated by this method is increased by 10%～15% than that of traditional methods when attacking defense models.

Key words: adversarial sample, generative adversarial network, synthetic aperture radar, semi-white box attack, mask extraction

0　引言

合成孔径雷达（synthetic aperture radar，SAR）可以在各种恶劣天气环境下成像，它在地理调查、气候变化研究、环境监测、军事信息处理等应用中发挥着重要作用[1]。近年来，深度学习在SAR目标识别领域取得了显著成功[2-4]。SAR目标识别模型在准确性方面比传统识别方法表现得更好，这得益于深度卷积神经网络的非线性映射带来的强大特征提取能力。虽然现代深度学习网络的精度越来越高，然而由于其层次非线性特性，它们极其容易受到对抗攻击的影响[5]。

对抗样本是指在清晰的图像中添加了难以察觉的扰动，导致目标识别模型对这些样本进行了错误的分类[6]。研究表明，对抗样本容易降低深度学习网络的目标识别能力，导致模型识别失效[7]。所以研究SAR图像对抗样本的生成有着重要意义。目前已经有很多对抗样本的生成方法，如迭代快速梯度符号法（I-FGSM）[8]、DeepFool迭代计算法[9]以及C&W[10]等攻击算法。这些攻击算法虽然在白盒攻击中已经具有了很高的攻击成功率，但是在对抗样本的生成中需要目标模型的所有信息，这在很多情况下是不现实的。而使用生成对抗网络（generative adversarial network，GAN）的方式来生成对抗样本，只需要在训练时用到目标模型的信息，训练完成后则不再需要目标模型，并且可以快速生成大量对抗样本，如AdvGAN[11]、RobGAN[12]、基于CycleGAN的攻击[13]以及基于WGAN的攻击[14]等。

虽然使用GAN的方式已经比较流行，但在SAR图像的对抗样本生成中存在着一些不足：GAN的训练不够稳定，生成的对抗样本质量无法保证；在SAR图像的对抗样本生成中，没有利用SAR图像的特点，并且生成的扰动是一种全局扰动，对抗样本容易被察觉，在这种对抗样本攻击下，通过一些相应的图像预处理操作来去除背景区域的信息，即可轻松抵御。为了解决上述问题，本文提出了一种SAR图像对抗样本生成模型，主要贡献如下。

（1）基于AdvGAN的架构，设计了一个掩模提取模块，该模块由增强Lee滤波器、最大类间方差法（OTSU）自适应阈值分割等模块组成，它可以从原始图像中提取出掩模参数。将此模块加入AdvGAN训练，可以生成只针对目标区域、扰动量更小的对抗样本。

（2）SAR图像对抗样本生成模型主要使用了GAN损失、对抗攻击损失和扰动限制损失，在标准GAN损失的基础上引入改进的相对均值GAN（relativistic average GAN，RaGAN）损失，使用了一个相对均值判别器，用来计算生成的数据比真实数据“更真实”的概率，代替了原来判别器只是计算生成的数据是真实数据的概率，判别器

同时依赖于真实数据和生成的数据。与原来相比，这种损失明显更稳定并且可以生成更高质量的数据样本。

（3）在MSTAR数据集上与AdvGAN、基于CycleGAN以及基于WGAN的攻击方法进行了实验对比，实验表明，相比于传统的基于GAN的攻击方法，本文方法生成的SAR图像对抗样本有着更小的扰动量以及更高的攻击成功率。

1 相关知识

1.1 SAR图像特点

SAR图像不同于一般的光学、红外传感器获取的图像，SAR图像通过雷达回波信号进行成像，有着很强的穿透效果，可以穿透云层进行成像，但其分辨率远低于光学图像。SAR图像只记录了一个通道中的回波信息，并且以二进制复数的形式记录下来，其灰度信息反映了目标的电磁散射特性，而光学图像一般由3个通道组成，更加方便了后续的识别和分类。由于SAR的成像原理，SAR图像中存在很明显的相干噪声，使其可读性进一步变差，光学图像和SAR图像的差异见表1。

表1 光学图像和SAR图像的差异

特性	光学图像	SAR图像
波长	纳米级	厘米级
应用	广泛应用于各领域	军用
数据形式	RGB三通道	单通道复数
分辨率	高	低
样本数量	多	少
成像原理	能量聚焦积累	相位相干积累

1.2 AdvGAN

尽管目前有许多基于某种设定好的策略迭代优化对抗扰动的对抗攻击算法，但这些方法仍然存在生成对抗样本速度较慢的问题，并且生成对抗样本需要用到目标模型的所有信息。为了解决这一问题，Xiao等[11]提出了一种针对特定被攻击模型的AdvGAN方法，它能够快速生成逼真的对抗样本并保证较高的攻击成功率，并且只需要在训练时用到目标模型的信息。该方法是基于生成对抗网络的，生成器可以专门为输入图像量身定制对抗扰动。此外，为了使得生成的样本尽可能真实，AdvGAN使用判别器来监督生成扰动。训练完成后，仅需要使用生成器即可快速批量生成对抗样本而无须再去访问被攻击网络。AdvGAN能够快速地帮助人们生成真实、强大、有效的对抗样本。但是，AdvGAN在进行攻击时生成的扰动是一种全局扰动，在SAR图像的对抗样本生成中，由于背景区域较大，容易被各种防御模型检测出，并且由于AdvGAN是结合GAN的方式，训练容易不稳定，产生的对抗样本质量无法保证。

2 对抗样本生成

2.1 总体思路

本文在AdvGAN原始的网络架构基础上，重新设计了网络模型和参数，并且加入掩模提取模块。此架构包含3种网络结构，分别是生成器网络、判别器网络和目标网络。生成器中定义了3种损失函数来约束GAN的训练，用来提高生成样本的感知相似性、攻击成功率以及限制扰动大小。

由于生成器网络输出的扰动图像是一种全局扰动，这种全局扰动扰动范围较大，使对抗样本容易被察觉。为了降低扰动量，本文在AdvGAN的基础上加入掩模提取模块，使其在训练时只生成特定范围的扰动信息，并且在AdvGAN的训练下，对抗样本也能够保持较强的欺骗能力。

2.2 攻击模型总体架构

攻击模型总体架构如图1所示。生成器使用自编码器结构，其中编码器使用3个二维卷积层，解码器使用3个二维反卷积层，每个卷积层后紧跟IN层进行归一化，IN层可以加速模型收敛，

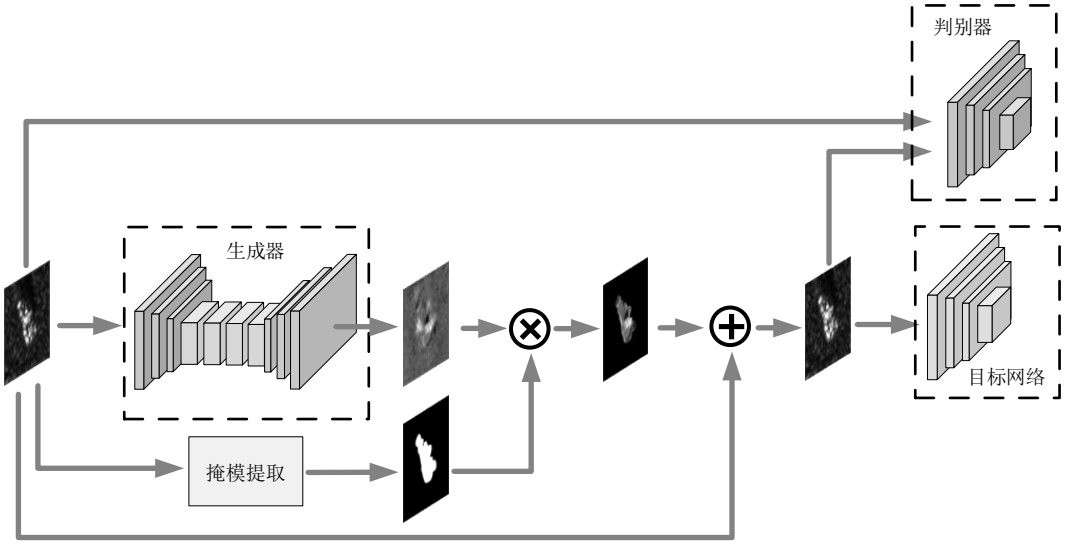


图1　攻击模型总体架构

并且保持每个图像样本之间的独立。编码器和解码器之间增加4个残差块，提高网络性能。根据经验，生成器最后一层使用Tanh进行激活，其他每层输出后使用ReLU进行激活。

判别器由3个二维卷积层和全连接层组成，每个卷积层后紧跟BN层进行归一化，避免训练时发生梯度消失。每个卷积层的输出使用Leaky-ReLU进行激活，避免了梯度方向的锯齿化问题。

将原始样本输入生成器网络中得到扰动图像，扰动图像与掩模提取模块生成的掩模参数相乘，得到提取后的扰动图像，然后将原始图像与扰动图像相加即可生成对抗样本。对抗样本和原始样本同时输入判别器中，判别器会输出损失L_{D}和$L_{\mathrm{G_GAN}}$，这两个损失统称为GAN损失，GAN损失约束整个网络朝着感知相似的方向训练。生成的对抗样本也会输入目标网络中，判断是否成功愚弄目标网络，并输出损失L_{adv}，L_{adv}约束整个网络朝着攻击成功的方向训练。在AdvGAN的原始网络架构中，还包括一个扰动限制损失L_{hinge}，用来限制扰动的大小以及稳定GAN的训练。$L_{\mathrm{G\_GAN}}$、L_{adv}和L_{hinge}组成了生成器损失L_{G}，L_{G}和L_{D}交替训练，最终达到纳什平衡状态。具体算法伪代码如下。

算法1　生成器和判别器训练算法

输入　真实样本X，真实样本类别t，真实样本数量n，生成器$G(x)$，判别器$D(x)$，待攻击模型f，训练次数epochs，批次大小bs，权重值α、β、γ，每次训练中生成器的训练次数n_G，判别器的训练次数n_D

输出　训练完成的生成器网络权重值和判别器网络权重值

for i in 1，…，epochs do：
　for j in 1，…，n/bs do：
　　取出一个批次的原始样本
　　for k in 1，…，n_D do：
　　X输入$G(x)$，得到扰动pertubation
　　X输入掩模提取模块，得到掩模参数M
　　pertubation = M · pertubation
　　对抗样本adv_image = pertubation + X
　　$D(x)$梯度清零
　　输入X、adv_image计算L_{D}
　　计算并更新$D(x)$梯度信息
　　for k in 1，…，n_G do：

$G(x)$梯度清零

输入pertubation计算L_{hinge}

输入t、adv_image和f，计算L_{adv}

输入X、adv_image计算$L_{\text{G_GAN}}$

$L_G=\alpha\times L_{\text{G\_GAN}}+\beta\times L_{\text{hinge}}+\gamma\times L_{\text{adv}}$

计算并更新$G(x)$梯度信息

2.3 掩模提取模块

本文设计了一种基于增强Lee滤波器以及自适应阈值分割的方法来提取掩模参数。掩模参数可以控制对抗样本的扰动范围，分别用0和1来表示，0表示被屏蔽的区域，1表示待扰动的目标区域，这一参数对对抗样本的生成十分重要。为了提取SAR图像的掩模参数，需要去除SAR图像的相干斑噪声，并对图像进行分割。掩模提取模块流程如图2所示，其中①②表示对SAR图像首先采用引入直方图均衡的Lee滤波器，增强SAR图像目标区域的特征信息，降低噪声的干扰。③④表示根据SAR图像直方图的分布特征，通过自适应阈值法对SAR图像进行分割。⑤⑥表示对于第一次分割后存在的一些不平滑成分，采用平滑滤波的方法来处理，⑦⑧表示再次通过自适应阈值分割模块，最终得到分割后的图像。最后根据分割结果来设置目标区域的掩模参数。

使用引入直方图均衡算法的Lee滤波器，该滤波器改正了原来增强Lee滤波器使SAR图像的细节、结构信息破坏严重，无法很好地保持边缘信息的缺点[15]。引入直方图均衡后既提高了Lee滤波去噪效果又不损失图像边缘细节信息。灰度直方图是统计图像灰度分布的重要手段，修正图像灰度直方图可以提高图像对比度。直方图均衡化是将原始的灰度直方图映射到均匀分布的灰度直方图，映射函数为以灰度级累积的分布函数，分布函数如式（1）所示：

$$S_k=T(r_k)=\sum_{j=0}^{k}\frac{n_j}{n}=\sum_{j=0}^{k}p_r(r_j) \tag{1}$$

其中，$0\leqslant r_j\leqslant 1$，$k=0$，1，2，3，…，$L$-1，$r_j$表示映射前图像归一化灰度级，$T(r_k)$为映射关系函数，$S_k$表示映射后图像归一化灰度级，$n_j$表示在第$k$灰度级中像素点存在的个数，$n$为像素点的总数，$p_r(r_j)$表示图像在变换前第$k$级灰度值概率。

改进的Lee滤波算法分别使用局域变差系数C_l和同域变差系数C_v作为度量的依据，$C_{\max}$是噪声相对标准差系数的最大值，将信号分为3类不同的区域分别进行处理。统计像素点个数，快速

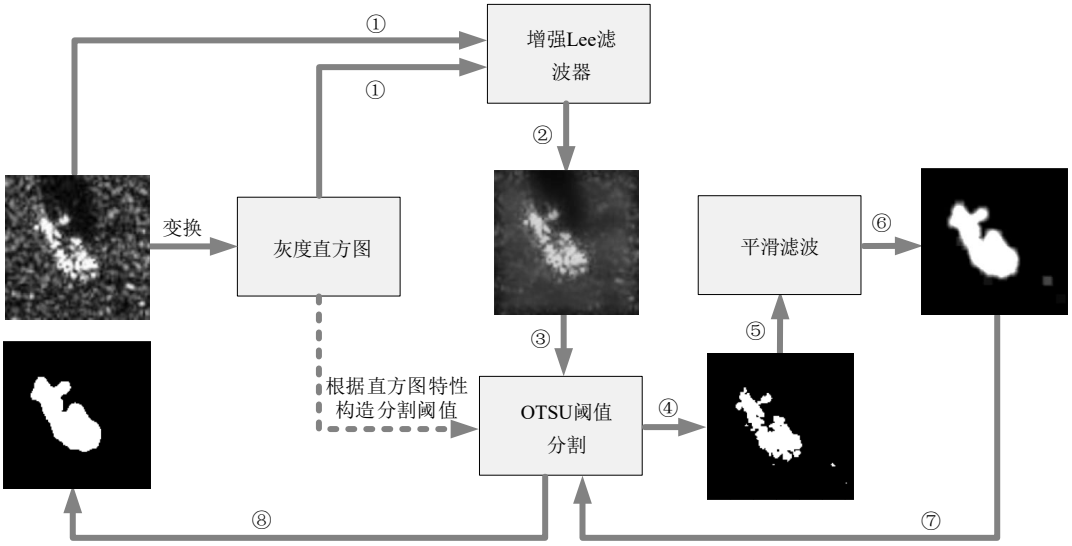


图2 掩模提取模块流程

确定终止元素的位置。具体表达式如式（2）所示[15]：

$$\hat{R}=\begin{cases} I_{\text{med}} & ,C_I<C_v \\ I_{\text{med}}+W\left(I-I_{\text{med}}\right) & ,C_v\leqslant C_I\leqslant C_{\max} \\ I & ,C_I>C_{\max} \end{cases} \tag{2}$$

其中：

$$W=\exp\left[\frac{-K\left(C_I-C_v\right)}{C_{\max}-C_I}\right] \tag{3}$$

其中，$\hat{R}$表示图像去噪后的图像值，I_{med}表示区域像素的均值，I表示原始图像的像素值，W表示权重系数，K为阻尼因子。经过滤波处理后，目标与背景区域之间的差异性有效增强，接下来就可以进行自适应阈值分割。本文使用OTSU进行阈值分割，OTSU是一种采用最大类间方差的自动确定阈值的方法，这是一种基于全局的二值化算法，它可以将图像分为前景和背景两部分。在OTSU算法中，选取最佳阈值时，目标和背景之间的差别应该是最大的，衡量这种差别的标准就是最大类间方差。当类间方差越大时，就表明图像中目标和背景之间的差别越大；如果部分目标被错误地归入背景或者部分背景被错误地归入目标，就会导致两部分之间的差异变小。最佳阈值表达式如式（4）所示[16]：

$$T_d=\underset{0\leqslant t\leqslant L-1}{\operatorname{argmax}}\left[P_a\left(w_a-w_0\right)^2+P_b\left(w_b-w_0\right)^2\right] \tag{4}$$

其中：

$$P_a=\sum_{i=1}^{T}P_i \tag{5}$$

$$P_b=\sum_{i=T+1}^{L}P_i \tag{6}$$

$$w_a=\sum_{i=1}^{T}i\frac{P_i}{P_a} \tag{7}$$

$$w_b=\sum_{i=T+1}^{L}i\frac{P_i}{P_b} \tag{8}$$

$$w_0=\sum_{i=1}^{L}iP_i \tag{9}$$

$$p_i=\frac{N_i}{N} \tag{10}$$

这里w_a和w_b表示类别a和类别b的数据权重，w_0表示预设的权重，用来和其他权重比较，图像像素总数为N，灰度级为L，N_i为灰度级为i的像素个数，p_i为灰度级i的概率。

根据SAR图像的分割结果，可以得到相应的掩模参数。如式（11）所示，将目标区域的值设置为1，背景区域的值设置为0，从而可以获得掩模参数M。其中X表示SAR图像的原始样本数据，T_d为式（4）计算得到的最佳分割阈值。

$$M=\begin{cases} 1, X>T_d \\ 0, X\leqslant T_d \end{cases} \tag{11}$$

2.4　损失函数

攻击模型在SAR图像的对抗样本生成中，生成器使用了3种损失函数，下面分别说明3种损失函数以及总损失函数。

（1）GAN损失L_{D}和$L_{\text{G_GAN}}$

将相对均值判别器加入攻击网络的损失函数中，在标准GAN中的判别器是用来评估输入图像真实的概率，而相对均值判别器则是预测真实图像比生成的图像相对更加真实的概率[17]，标准判别器和相对均值判别器的区别如图3所示。

在图3中，x_r为真实样本，x_f为假样本（生成器生成的样本），σ为激活函数，$C(x)$是鉴别器没有经过最后层激活函数的输出。RaGAN的损失函数如式（12）、式（13）所示：

$$L_{\text{D}}^{\text{RaGAN}}=E_x\left[f_1\left(D\left(x_r\right)-E_xD\left(x_f\right)\right)\right]+E_x\left[f_2\left(D\left(x_f\right)-E_xD\left(x_r\right)\right)\right] \tag{12}$$

$$L_{\text{G\_GAN}}^{\text{RaGAN}}=E_x\left[g_1\left(D\left(x_r\right)-E_xD\left(x_f\right)\right)\right]+E_x\left[g_2\left(D\left(x_f\right)-E_xD\left(x_r\right)\right)\right] \tag{13}$$

其中，$f_1(y)=g_2(y)=-y$，$f_2(y)=g_1(y)=y$。经实验，当加入矩阵维度相同的全1矩阵，并且取平方后，效果更佳，改造后的GAN损失如式（14）、式（15）所示：

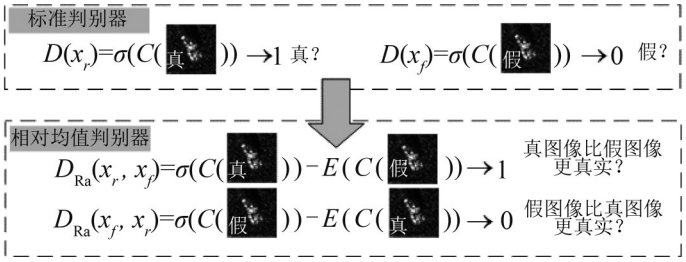


图3　标准判别器和相对均值判别器的区别

$$L_{\mathrm{D}}^{\mathrm{RaGAN\_r}}=E_x\left[\left(D(x_r)-E_xD(x_f)-1\right)^2\right]+E_x\left[\left(D(x_f)-E_xD(x_r)+1\right)^2\right] \tag{14}$$

$$L_{\mathrm{G\_GAN}}^{\mathrm{RaGAN\_r}}=E_x\left[\left(D(x_r)-E_xD(x_f)+1\right)^2\right]+E_x\left[\left(D(x_f)-E_xD(x_r)-1\right)^2\right] \tag{15}$$

（2）对抗攻击损失L_{adv}

本文中使用目标攻击中的C&W损失来捕获对抗攻击损失[18]，使用C&W中的L_2范数攻击,并且将扰动信息与上文得到的掩模参数相乘，损失函数如式（16）所示：

$$L_{\mathrm{adv}}=\|\delta\cdot M\|_2^2+c\cdot f(X+\delta\cdot M) \tag{16}$$

其中：

$$\delta=\frac{1}{2}(\tanh(w)+1)-X \tag{17}$$

$$f(X+\delta\cdot M)=\max\left(\max\left\{Z(X+\delta\cdot M)_i: i\neq t\right\}-Z(X+\delta\cdot M)_t, -k\right) \tag{18}$$

其中，δ表示生成器生成的扰动信息，M为上文得到的掩模参数，X为原始样本，c为设置的超参数，根据C&W之前的研究，设置c=100。由于$X+\delta\in[0,1]^n$，所以设置了δ来限制其大小，w为δ的替代值。$f(\cdot)$是攻击的关键，用来减少原始类别t和其他类别i的最大差异。这个损失函数会使生成的对抗样本朝着愚弄目标网络的方向训练。

（3）扰动限制损失L_{hinge}

本文使用了基于L_2范数的扰动限制损失，用来限制扰动的大小以及稳定GAN的训练，如式（19）所示：

$$L_{\mathrm{hinge}}=E_x\max\left(0, \|G(x)\|_2-c_0\right) \tag{19}$$

c_0为扰动限制因子，可以通过改变c_0的值来限制扰动的大小。

（4）生成器总损失L_{G}

$$L_{\mathrm{G}}=\alpha\times L_{\mathrm{G\_GAN}}^{\mathrm{RaGAN\_r}}+\beta\times L_{\mathrm{hinge}}+\gamma\times L_{\mathrm{adv}} \tag{20}$$

联合上面的3种损失函数，即可得到生成器的总损失L_{G}。其中，α、β、γ为权重值，并且应满足$\alpha+\beta+\gamma=1$。L_{G}与$L_{\mathrm{D}}^{\mathrm{RaGAN_}r}$交替训练，最终达到纳什平衡状态。

3　实验及分析

3.1　实验环境和数据集

实验平台为Pytorch深度学习平台，工作站为NVIDIA GeForce GTX 1080 Ti单卡的环境。为验证本文提出的模型在生成对抗样本问题上的可行性，本文在由美国国防高级研究计划署和空军研究室提供的MSTAR数据集上进行实验分析，数据集样本如图4所示。该数据集共有10类目标，分别为2S1（自行榴弹炮）、BMP2（步兵战

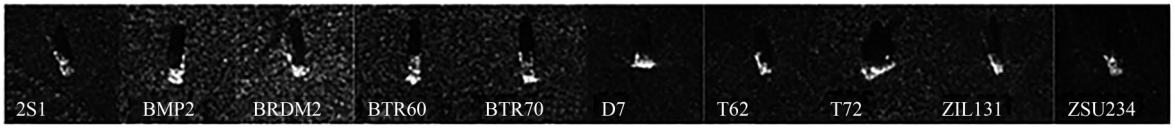


图4　MSTAR数据集样本

车）、BRDM2（装甲侦察车）、BTR60（装甲运输车）、BTR70（装甲运输车）、D7（推土机）、T62（坦克）、T72（坦克）、ZIL131（货运卡车）、ZSU234（自行高炮）。图像大小均为 128×128（单位为像素），训练集有 2 536 个图像，成像侧视角为 15°；测试集有 2 636 个图像，成像侧视角为 17°。为了重点研究目标区域，将所有的样本中心裁剪至 64×64 大小，再统一放大到 128×128，减小背景区域的占比。

3.2　实验参数

在 Ra-AdvGAN 的训练中，设置训练数 epochs=200，批次大小 bs=64，损失函数优化过程中使用自适应动量估计（Adam）优化器，学习率初始化为 0.001，并且每 50 轮衰减 10 倍速度。扰动限制因子 c_0=0.1，生成器损失权重分别取 α=0.3，β=0.1，γ=0.6。每次训练中生成器的训练次数 n_G=3，判别器的训练次数 n_D=1，即生成器每更新 3 次，判别器更新 1 次。

3.3　评价标准

为了观察 RaGAN 引入后的实验效果，在此部分实验中，使用没有添加掩模提取模块的原始 AdvGAN 模型来进行实验。攻击成功率是指在训练趋于稳定后的 100 次训练中，攻击成功率的平均值。训练的稳定性影响生成样本的质量，即训练越稳定，生成样本的质量也会越高。生成样本的质量可以用图像相似性（structural similarity，SSIM）算法来评估，SSIM 着重关注原始样本与对抗样本之间的直接结构差异，SSIM 越接近于 1，表示对抗样本与原始样本越难以区分。SSIM 计算表达式如式（21）所示。

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{21}$$

其中，μ_x 和 μ_y 分别表示图像 x 和 y 的均值，σ_x^2 和 σ_y^2 分别表示图像 x 和 y 的方差，σ_{xy} 表示图像 x 和 y 的协方差。

3.4　针对损失函数的实验结果

使用目前较为主流的分类模型 AlexNet、VggNet、ResNet 以及 DenseNet 作为目标模型，Ra_AdvGAN 的对抗样本攻击成功率和标准差见表 2，其中 AdvGAN 表示传统模型，Ra_AdvGAN 表示引入本文损失后的模型。从表 2 中可以看出，在引入改进的 RaGAN 后，每种目标模型的攻击成功率均有提高，并且 SSIM 值更高，说明对抗样本的攻击性能更好，并且生成的对抗样本与原始样本的相似性更高。

表 2　Ra_AdvGAN 的对抗样本攻击成功率和标准差

目标模型	攻击模型	攻击成功率	SSIM
AlexNet	AdvGAN	92.80%	0.980 6
	Ra_AdvGAN	**96.30%**	**0.986 9**
VggNet16	AdvGAN	94.59%	0.939 3
	Ra_AdvGAN	**98.15%**	**0.948 8**
ResNet18	AdvGAN	92.15%	0.974 2
	Ra_AdvGAN	**97.82%**	**0.976 6**
ResNet50	AdvGAN	92.17%	0.962 4
	Ra_AdvGAN	**96.89%**	**0.980 6**
DenseNet121	AdvGAN	95.27%	0.744 0
	Ra_AdvGAN	**98.60%**	**0.792 3**

3.5　针对掩模提取的实验结果

SAR 图像滤波前后分析如图 5 所示，从图 5（b）和图 5（e）中可以看出，滤波后的图像像素灰度取值的三维显示更加清晰。并且从图 5（c）和图 5（f）中可以看出，经过 Lee 滤波的增强后，以低灰度区域为主，这些低灰度区域对应的是滤波图像中的背景区域，其分布大致呈现高斯分布特性。

经过增强 Lee 滤波器的处理后，目标与背景区域之间的差异性得到了有效增强，满足了使用 OTSU 进行阈值分割的条件，它可以将图像分为目标和背景两部分，完全适用于滤波后 SAR 图像的分割。10 类 MSTAR 图像掩模提取过程如图 6 所示，从图 5 中可以看出，每一类 SAR 图像经过掩模提取模块后都成功提取出目标区域的掩模参数。

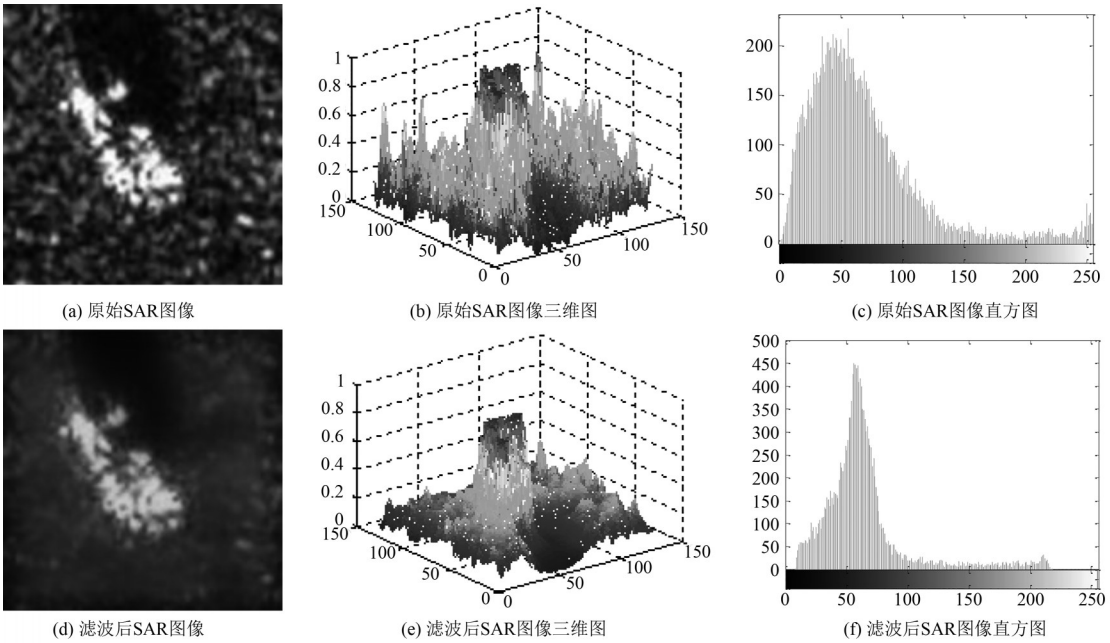


图5　SAR图像滤波前后分析

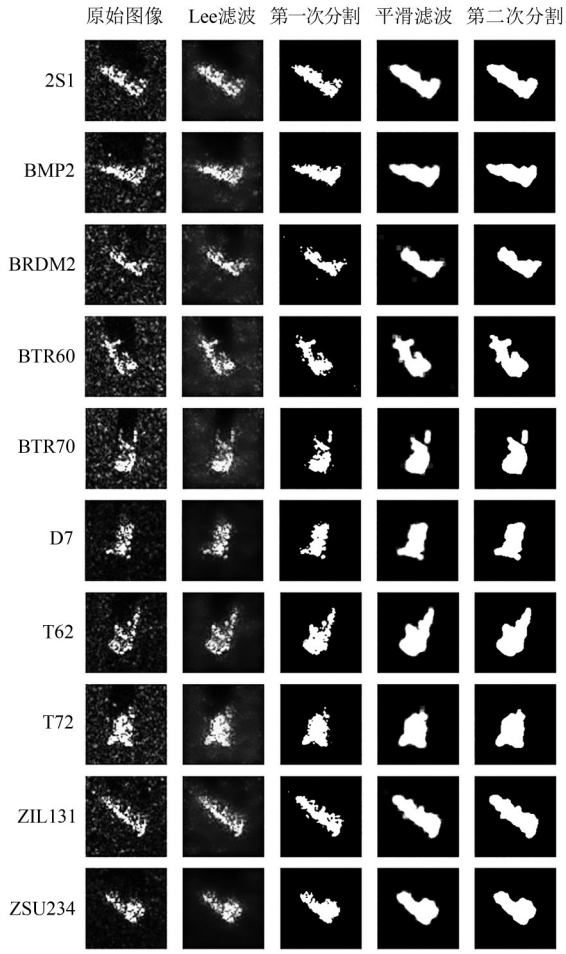


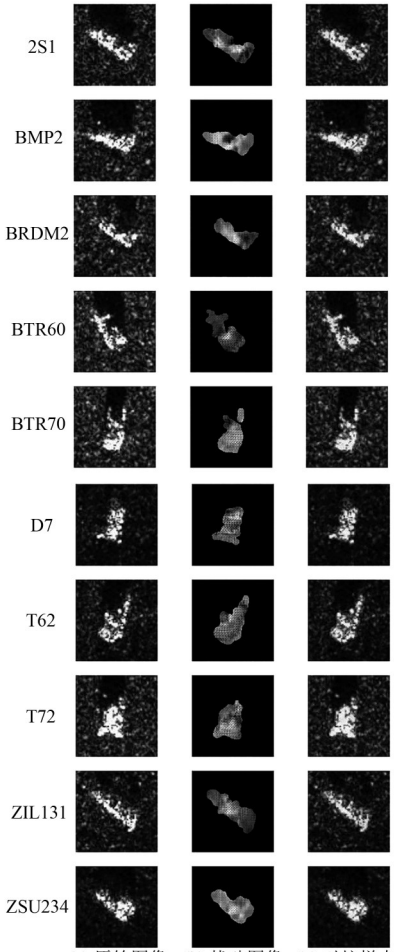
图6　10类MSTAR图像掩模提取过程

将掩模提取模块引入Ra_AdvGAN中，进行对抗攻击实验，使用3个目标模型作对比，其中防御模型是指在ResNet50基础上，加入去噪等预处理操作的模型，由于防御模型只在测试攻击成功率中使用，不参与对抗样本的生成，所以其SSIM值不参与实验对比。本文方法与其他基于GAN的攻击方法对比结果见表3。

从表3中可以看出，在Ra_AdvGAN的基础上加入掩模提取后，由于扰动产生的范围大幅减小，虽然攻击成功率相比于其他攻击方法有所降低，但SSIM指标与其他方法相比更接近1，说明此方法产生的对抗样本与原始样本有高度的相似，扰动更加具有针对性。本文提出的方法在对防御模型进行攻击时，可以看出攻击成功率与针对无防御模型（VggNet16、ResNet50）进行攻击的攻击成功率相比，并没有降低太多，而其他攻击方法却有大幅降低，说明对具有一定防御能力的模型有较好的攻击效果。MSTAR对抗样本展示如图7所示，可以看出原始样本和对抗样本肉眼无法区分，生成的样本质量较高。

表3　本文方法与其他基于GAN的攻击方法对比

攻击模型	目标模型	攻击成功率	SSIM
AdvGAN	VggNet16	94.59%	0.939 3
	ResNet50	92.17%	0.962 4
	防御模型	67.05%	—
基于CycleGAN	VggNet16	94.51%	0.965 4
	ResNet50	93.01%	0.963 7
	防御模型	68.97%	—
基于WGAN	VggNet16	95.39%	0.968 8
	ResNet50	95.44%	0.957 9
	防御模型	74.44%	—
Ra_AdvGAN	VggNet16	**98.15**%	0.976 6
	ResNet50	**96.89**%	0.980 6
	防御模型	74.17%	—
基于Ra_AdvGAN和掩模提取	VggNet16	86.21%	**0.998 0**
	ResNet50	87.19%	**0.997 6**
	防御模型	**83.11**%	—



(a) 原始图像　(b) 扰动图像　(c) 对抗样本

图7　MSTAR对抗样本展示

4　结束语

本文提出了一种SAR图像对抗样本生成模型，相比于其他相关方法，该模型生成的对抗样本与原始样本高度相似，并且在攻击具有一定防御能力的模型时，攻击成功率高于其他攻击方法。该模型利用SAR图像的特点，采用增强Lee滤波器和OTSU自适应阈值分割组成的掩模提取模块，产生局部扰动。同时，引入改进的RaGAN损失到AdvGAN中，使用相对均值判别器提高了训练的稳定性与攻击效果。因此，该模型可以被应用于SAR图像的对抗样本生成以及评估目标识别模型的鲁棒性等方面，具有很好的实际应用价值。但由于扰动范围变小，攻击成功率没有足够高，如何提高攻击成功率值得进一步探索，同时如何针对这种攻击模型进行防御也是下一步研究的方向。

参考文献：

[1] XIA W J, LIU Z, LI Y. 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.

[2] 何涛, 李大亮, 曹兰英. 基于深度学习的机载SAR典型目标识别算法[J]. 现代雷达, 2022, 44(12): 87-92.
HE T, LI D L, CAO L Y. An algorithm of typical target recognition for airborne SAR based on deep learning[J]. Modern Radar, 2022, 44(12): 87-92.

[3] SHI J. SAR target recognition method of MSTAR data set based on multi-feature fusion[C]//Proceedings of the 2022 International Conference on Big Data, Information and Computer Network (BDICN). Piscataway: IEEE Press, 2022: 626-632.

[4] FANG C, SONG Y M, GUAN F H, et al. A robust complex-valued deep neural network for target recognition of UAV SAR imagery[J]. IEEE Journal on Miniaturization for Air and Space Systems, 2023, 4(2): 175-185.

[5] 王志波, 王雪, 马菁菁, 等. 面向计算机视觉系统的对抗样本攻击综述[J]. 计算机学报, 2023, 46(2): 436-468.
WANG Z B, WANG X, MA J J, et al. Survey on adversarial example attack for computer vision systems[J]. Chinese Journal of Computers, 2023, 46(2): 436-468.

[6] 赵宏, 常有康, 王伟杰. 深度神经网络的对抗攻击及防御方法综述[J]. 计算机科学, 2022, 49(S2): 662-672.
ZHAO H, CHANG Y K, WANG W J. Summary of anti-attack and defense methods of deep neural networks[J]. Computer Science, 2022, 49(S2): 662-672.

[7] WANG X M, LI J, KUANG X H, et al. The security of machine learning in an adversarial setting: a survey[J]. Journal of Parallel and Distributed Computing, 2019(130): 12-23.

[8] HUANG T, ZHANG Q X, LIU J B, et al. Adversarial attacks on deep-learning-based SAR image target recognition[J]. Journal of Network and Computer Applications, 2020(162): 102632.

[9] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2574-2582.

[10] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[11] XIAO C W, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of the Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2018: 3905-3911.

[12] LIU X Q, HSIEH C J. Rob-GAN: generator, discriminator, and adversarial attacker[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 11226-11235.

[13] 张高志, 刘新平, 邵明文. 用于白盒目标攻击的GAN对抗样本生成[J]. 模式识别与人工智能, 2020, 33(9): 830-838.
ZHANG G Z, LIU X P, SHAO M W. Generating adversarial example with GAN for white-box target attacks[J]. Pattern Recognition and Artificial Intelligence, 2020, 33(9): 830-838.

[14] 田宇. 基于GAN的图像对抗样本生成方法研究[D]. 北京:北京邮电大学, 2020.
TIAN Y. Research of method on generating image adversarial samples based on GAN[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.

[15] 付睢宁, 路泽忠, 王舜瑶. 一种改进的Lee滤波SAR图像去噪算法[J]. 计算机与数字工程, 2019, 47(8): 2018-2021.
FU S N, LU Z Z, WANG S Y. An improved lee filter SAR image denoising algorithm[J]. Computer & Digital Engineering, 2019, 47(8): 2018-2021.

[16] 谢鹏鹤. 图像阈值分割算法研究[D]. 湘潭: 湘潭大学, 2012.
XIE P H. The study on the image thresholding segmentation algorithm[D]. Xiangtan: Xiangtan University, 2012.

[17] JOLICOEUR-MARTINEAU A. The relativistic discriminator: a key element missing from standard GAN[EB]. 2018: arXiv: 1807.00734.

[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[作者简介]

章坚武（1961- ），男，杭州电子科技大学教授、博士生导师，中国电子学会、中国通信学会高级会员，浙江省通信学会常务理事，主要研究方向为无线通信与移动通信、通信网络与信息安全。

能豪（1999- ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为无线通信与信息安全。

李杰（1976- ），女，杭州电子科技大学信息工程学院副教授，主要研究方向为无线通信与移动通信。

钱建华（1971- ），男，中国联通（浙江）产业互联网有限公司总经理、高级工程师，主要研究方向为数据通信。

方银锋（1986- ），男，杭州电子科技大学校副教授、硕士生导师，主要从事人机接口设计、生物信号处理与分析、模式识别与智能系统等领域的研究。