采用 GAN 的肺部疾病诊断模型黑盒可迁移性对抗攻击方法

王小银[1,2]，王丹[1]，孙家泽[1,2]，杨宜康[3]

（1. 西安邮电大学计算机学院，710121，西安；2. 西安邮电大学陕西省网络数据分析与智能处理重点实验室，710121，西安；3. 西安交通大学自动化科学与工程学院，710121，西安）

摘要：针对现有对抗攻击方法在黑盒场景下攻击成功率不高以及生成质量低等问题，提出了一种基于生成对抗网络(GAN)的肺部疾病诊断模型黑盒可迁移性对抗攻击方法。以肺部医学影像为基础，依托残差神经网络，在生成器中设计基于扩张卷积的残差块和金字塔分割注意力机制，以提高网络在更细粒度上的多尺度特征表达能力；设置带有辅助分类器的判别器对样本进行正确分类，并且添加攻击者实施对抗训练，以增强对抗样本的攻击能力和稳定 GAN 的训练。运用无数据黑盒对抗攻击框架训练替代模型，实现可迁移性对抗攻击，获得高黑盒攻击成功率。所提方法在目标攻击和无目标攻击任务下的对抗攻击成功率分别达到了 68.95%和 79.34%，与其他黑盒场景下基于 GAN 的对抗方法相比，迁移攻击成功率更高，且生成的对抗样本更接近真实样本。所提方法解决了传统基于 GAN 的攻击方法难以捕获肺部影像细节特征而导致无法获得更优的对抗性能的问题，对在实际应用场景下提高肺部疾病诊断模型的安全性和鲁棒性提供了参考方案。

关键词：肺部疾病诊断模型；黑盒对抗攻击；生成对抗网络；可迁移性

中图分类号：TP391.41 **文献标志码**：A

DOI：10.7652/xjtuxb202310019 **文章编号**：0253-987X(2023)10-0196-12

Black-Box Transferable Adversarial Attack Method Based on Generative Adversarial Networks for Lung Disease Diagnosis Models

WANG Xiaoyin[1,2], WANG Dan[1], SUN Jiaze[1,2], YANG Yikang[3]

(1. School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; 2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; 3. School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710121, China)

Abstract: A black-box transferable adversarial attack method based on GAN for lung disease diagnosis models was proposed to address the low success rate of attacks in black-box scenarios and low generation quality of existing adversarial attack methods. The method was built based on pulmonary medical images, with the residual neural network as the basic skeleton. In the generator, residual blocks based on dilated convolution and pyramidal segmentation attention mechanism were designed to improve the multi-scale feature representation capability of the network at finer granularity; discriminators with auxiliary classifiers were set up to correctly classify the samples, and the attackers were added to the discriminators for adversarial training to enhance

收稿日期：2023-03-22。 作者简介：王小银(1976—)，女，副教授，硕士生导师。 基金项目：陕西省重点研发计划资助项目(2023-YBGY-204，2023-YBGY-030)；西安市重点产业链核心技术攻关项目(人工智能领域)(2022JH-RGZN-0028)。
网络出版时间：2023-06-25 网络出版地址：https://kns.cnki.net/kcms2/detail/61.1069.T.20230625.1155.004.html

the adversarial sample attack capability and stabilize the training of GAN. The data-free black-box adversarial attack framework was also used to train alternative models to achieve transferable adversarial attack and obtain a more effective and higher black-box attack success rate. The method achieved adversarial success rates of 68.95% and 79.34% for targeted attacks and untargeted attacks respectively. Compared with other GAN-based attack methods in black-box scenarios, it presents a higher transferability attack success rate and the generated adversarial samples are closer to the real samples, solving the problem that traditional GAN-based attack methods cannot capture the detailed features of lung images and thus cannot obtain better adversarial performance. This method provides a reference for improving the security and robustness of lung disease diagnosis models in practical application scenarios.

Keywords: lung disease diagnosis model; black-box adversarial attack; generative adversarial network; transferability

受气候环境和生活习惯等因素影响，全球肺部疾病的发病率在逐年攀升，并且随着大流行疾病新冠肺炎的爆发，严重威胁了全球近 80 亿人的生命安全，因此准确识别肺部疾病极为重要。在传统医疗诊断中，医生通过对患者的表征及其相关的医学影像进行分析，但容易受到人为及环境因素的影响，因此仅通过人工技术难以精准快速地诊断病情。随着人工智能在医学领域的发展，基于深度学习的计算机辅助诊断技术在医疗诊断方面表现出了明显的优势和强大的生命力[1]，医生可以快速地获得计算机的辅助诊断结果作为参考，有利于更准确地做出最终的诊断与决策[2]。

深度神经网络(deep neural network，DNN)在医学图像分类中表现出了令人瞩目的性能，提高了对疾病的识别精度和工作效率。但是，由于其本身的不可解释性和脆弱性，被研究人员证实容易受到对抗攻击的影响，进而导致网络的安全性和鲁棒性受到威胁，因此其在疾病诊断中的实际应用仍值得商议[3]。

2013 年，Szegedy 等[4]证明了在原样本上添加轻微扰动会使模型以高置信度被错误分类，因此引发了人们对 DNN 泛化能力的质疑，限制了深度学习在关键安全环境中的应用[5]。目前，研究人员已经提出了一系列不同的方法来进行对抗攻击。Goodfellow 等[6]提出了快速梯度符号法(fast gradient sign method，FGSM)，该方法在梯度变化的最大方向生成对抗扰动并添加到原图像中，使模型被错误分类。在 FGSM 基础上衍生了投影梯度下降法(project gradient descent，PGD)[7]，该算法目前是公认的最强白盒攻击方法。此外，Carlini 等[8]提出了一种基于优化的攻击方法(Carlini&Wagner，C&W)，利用不同范数来限制扰动无法被察觉。Xiao 等[9]受生成对抗网络(generative adversarial networks，GAN)的启发，利用生成模型提出了生成对抗样本网络(AdvGAN)，该方法生成的对抗样本不仅更容易混淆目标网络，人眼看起来也更加逼真。与基于梯度和优化的方法相比，生成模型减少了对抗样本的生成时间，并且也不需要获取更多的背景知识，适合攻击难度较高的半白盒和黑盒攻击。然而，现有的生成方法在医学图像上具有两个明显的缺点：①与自然图像不同，医学图像具有复杂的生物纹理从而导致对抗扰动的敏感区域更多，因此仅使用原始的 GAN 网络无法提取图像中更深层次的特征信息，不能完全捕捉肺部图像特征的语义信息，生成能力受限；②大多数攻击算法只能在白盒场景中执行有效攻击，并且需要对模型的体系结构和参数进行无限次访问，在设置更实际的黑盒对抗场景时无法获得高攻击成功率[10]。

为了解决上述问题，本文提出了一种基于 GAN 的黑盒可迁移性对抗攻击方法，以有效地生成对抗样本。针对肺部影像构建深度学习目标模型，利用生成对抗网络生成对抗样本，并运用无数据黑盒对抗攻击框架实现可迁移性对抗攻击，获得高黑盒攻击成功率，提高对抗样本的可迁移性和攻击性能。与其他黑盒攻击场景下基于 GAN 的对抗方法进行实验对比，结果表明本文方法生成的对抗样本攻击效果更好，生成的图像也更接近于真实图像。

1　GAN 网络对抗攻击

1.1　生成对抗网络

GAN 是一种有效的深度生成模型，由生成器 G 和判别器 D 组成，如图 1 所示。

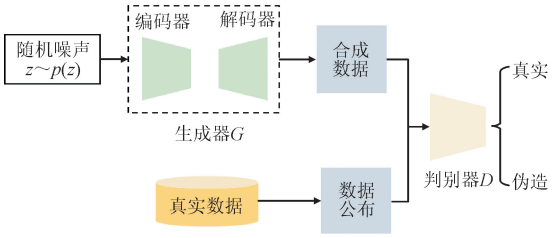

图 1 GAN 网络架构

Fig. 1 The network architecture diagram of GAN

生成器 G 用于学习真实图像的数据分布 $x\sim p(x)$，合成以假乱真的图像 $G(z)$。判别器 D 负责判断输入图像的真假，为生成器提供指导信息。GAN 损失表示为

$$\min_G \max_D L_{\mathrm{GAN}} = E_{x\sim p(x)}[\log D(x)] + E_{z\sim p(z)}[\log(1-D(G(z)))] \tag{1}$$

在训练过程中，生成器 G 和判别器 D 通过交替优化的方式互相学习，以提高自身的生成能力和判别能力。直至判别器不足以区分真实图像和伪造图像，即生成器能够生成最为逼真的假样本时，博弈双方达到纳什平衡状态。

1.2 对抗攻击

尽管 DNN 诊断模型具有高性能，但研究发现其容易受到对抗攻击的影响。根据攻击场景，将对抗攻击分为白盒攻击和黑盒攻击。在白盒场景中执行攻击，可以访问目标模型的网络结构、训练参数等。但是，由于隐私和安全性，这种攻击场景在现实中通常不可用[11]。在执行更为实际的黑盒攻击时，既可以通过查询目标模型直接实施攻击，也可以利用对抗样本的可迁移性训练替代模型生成对抗样本来欺骗目标模型。由于无法直接获取目标模型的训练数据，Truong 等[12]提出了在无数据黑盒场景中训练替代模型：生成模型负责合成输入图像，并在合成图像上训练替代模型来模仿目标模型。在训练过程中，替代模型和生成模型分别尝试最小化和最大化替代模型和目标模型之间的匹配率，直至替代模型收敛。然而，准确量化替代模型和生成模型之间的分歧非常困难，而且不稳定的训练过程会导致模型难以收敛甚至崩溃，继而无法在实践中达到理想的纳什均衡点。

因此，本文利用一个更优的黑盒攻击框架：改变生成模型和替代模型之间的博弈，让双方不用在最小化-最大化中直接竞争，使生成模型和替代模型具有相对独立的优化过程，并且通过平衡数据分布和促进数据多样性来缓解替代模型在训练过程中出现的模型崩溃问题，使替代模型可以更快更稳定地收敛到目标模型，实现可迁移性对抗攻击。

2 肺部疾病诊断模型黑盒对抗攻击

本文提出了一种基于 GAN 的肺部疾病诊断模型黑盒可迁移性对抗攻击方法 BA-GAN。构建肺部疾病诊断目标模型，利用无数据黑盒对抗攻击框架(data-free black-box adversarial attack，DBAA)获得替代模型，实现黑盒可迁移性对抗攻击。利用 AI-GAN(attack-inspired GAN)方法生成对抗样本，其中：使用残差神经网络(ResNet)作为生成器的基本骨架，并设计带有扩张卷积的残差块[13]和轻量高效的金字塔分割注意力机制(pyramid split attention，PSA)[14]对生成器进行改进优化，以提高网络的特征表达能力；设置带有辅助分类器的判别器对生成的样本进行分类，并且添加攻击者对判别器实施对抗训练，增强对抗样本的攻击能力和稳定 GAN 的训练。整体对抗攻击流程如图 2 所示。

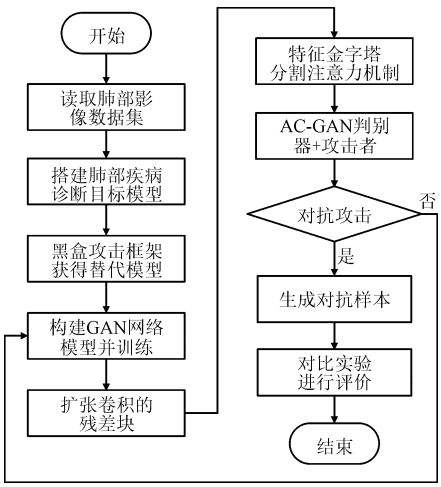

图 2 肺部疾病诊断模型对抗攻击流程

Fig. 2 The flow chart of adversarial attack

2.1 黑盒对抗攻击框架

构建无数据黑盒攻击框架，具体包括数据合成和替代模型蒸馏两个阶段。在对抗攻击时使用模型蒸馏技术，对未知的诊断模型进行参数学习和模型复制，实现对未知网络模型的高可靠攻击，增加对攻击对象的迁移性与通用性[15]。无数据黑盒对抗攻击框架如图 3 所示。

在第一阶段，生成器 G 合成分布接近于目标训练所需的数据 X。为了缓解替代模型在训练过程中容易崩溃的问题以及提高替代模型的准确率，引入最

大化信息熵和随机标签平滑策略，生成损失表示为

$$L_{ce} = \mathrm{CE}(S(X), \hat{Y}) \tag{2}$$

$$L_G = L_{ce} + \alpha L_H \tag{3}$$

式中：CE 是交叉熵损失函数；$S(X)$ 是向替代模型输入生成器合成的数据；$\hat{Y}$ 是随机平滑标签；α 是调整正则化值的超参数；L_H 是信息熵损失。

在第二阶段，替代模型用合成的数据进行训练，从而有效地模仿目标模型。对替代模型和目标模型采用蒸馏技术，蒸馏损失表示为

$$L_d = \mathrm{CE}(T(X), S(X)) \tag{4}$$

式中：$T(X)$ 表示目标模型的输出；$S(X)$ 表示蒸馏替代模型的输出。

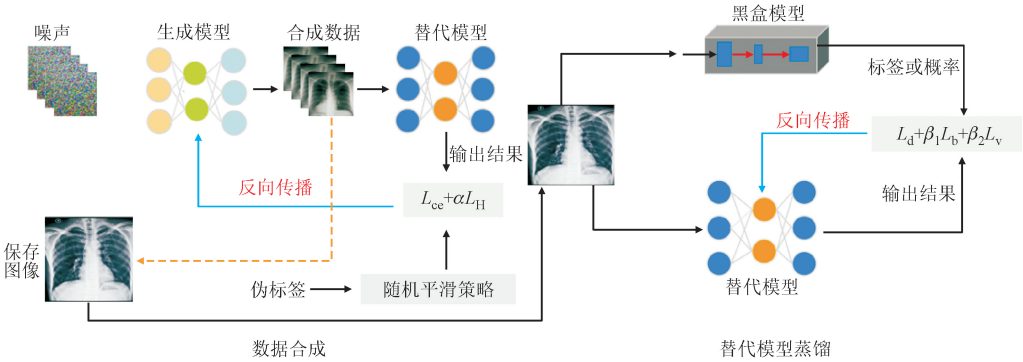

图 3　无数据黑盒对抗攻击框架

Fig. 3　Data-free black box adversarial attack framework

为了获得高攻击成功率，使替代模型和目标模型具有高度一致的决策边界来促进替代模型的训练，需要在蒸馏过程中对两种类型数据多加关注[16]。第一类是目标模型和替代模型边界之间具有决策分歧的数据，给予这些数据更多权重有助于弥合两个决策边界之间的差距，由此引入边界支持损失 L_b

$$\begin{cases} L_b = \mathrm{CE}(T(X), S(X)) \\ \arg\max T(X) \neq \arg\max S(X) \end{cases} \tag{5}$$

另一类是目标模型和替代模型决策边界更加接近的数据，对这类数据给予更多权重可确保在替代模型上对抗样本可以继续朝着目标模型的方向移动，由此引入对抗样本支持损失 L_v

$$\begin{cases} L_v = \mathrm{CE}(T(X), S(X)) \\ \arg\max T(\hat{X}) = \arg\max S(\hat{X}) \end{cases} \tag{6}$$

在模型蒸馏过程中要使得目标函数最小化来确保替代模型的输出结果逐渐逼近于目标模型，用 β_1 和 β_2 控制损失占比，目标函数表示为

$$L = L_d + \beta_1 L_b + \beta_2 L_v \tag{7}$$

通过优化合成图像的蒸馏目标，得到特征非常接近黑盒目标模型的替代模型，然后对经过蒸馏提炼的网络使用白盒攻击的方式进行攻击，生成攻击性能最优的对抗样本。相比于基于查询的黑盒攻击方式，该方法能在查询预算有限的情况下显著提高黑盒攻击方式下的对抗成功率，并且利用替代模型制作对抗样本消耗的计算资源更少，在实践中更为现实。

2.2　基于 AI-GAN 的对抗样本生成方法

构造基于 GAN 的对抗样本生成模型，由生成器 G、双头判别器 D、攻击者、辅助分类器和攻击模型组成。采用 AI-GAN 方法生成对抗样本的整体架构，如图 4 所示。

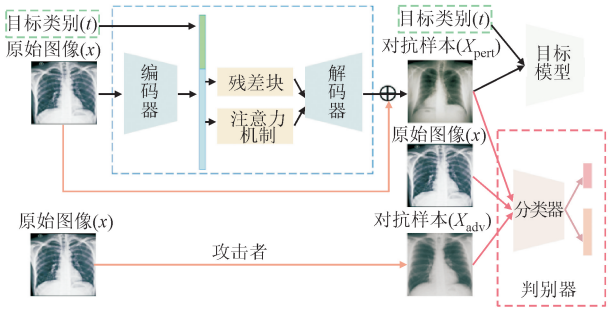

图 4　AI-GAN 整体架构

Fig. 4　The overall architecture diagram of AI-GAN

生成器 G 和判别器 D 均以端到端的方式进行训练。生成器 G 将原始肺部图像 x 和目标类别 t 作为输入以生成对抗扰动 $G(x,t)$，叠加到原始图像 x 上获得对抗样本 X_{pert}，并在 L_2 范数的约束下尽可能减小扰动，从而保证生成图像的真实性。为了增强对抗样本的攻击能力，添加了攻击者进行对抗训练。在判别器中设置了辅助分类器来提高攻击效率，因

此判别器既可以区分图像的真伪性，也可以对样本进行正确分类。一旦 GAN 网络训练结束后，就可以设置不同的目标类别对模型执行半白盒和黑盒攻击，并且能在保持图像质量的前提下显著提高对抗样本的攻击性能。

2.2.1 判别器网络结构

基于 GAN 的生成方法不能像基于梯度的方法一样直接在分类决策边界上运行，并且随着图像大小的增加，攻击性能会急剧下降，这是由于分类空间和生成空间存在的潜在差异引起的。因此，本文的判别器采用 AC-GAN 的判别器[17]，在原始判别器中引入了分类器来关联生成器，将生成器中的高级特征空间与分类器的分类特征空间对齐，此时生成空间中的扰动可以映射到原始图像的分类空间以保证对抗攻击发生在模型的决策边界上，提高攻击效率。

将攻击者添加到训练过程中，增强了判别器的健壮性和鲁棒性，从而有助于稳定和加速整个训练。最终，判别器的损失函数由 3 部分组成：用于区分真实/扰动图像产生的损失 L_S，攻击者和生成器生成的对抗样本产生的分类损失 L_a 和 L_g，分别定义为

$$L_S = E[\log P(X_{real})] + E[\log P(X_{pert})] \quad (8)$$

$$L_a = E[\log P(C = y \mid X_{adv})] \quad (9)$$

$$L_g = E[\log P(C = y \mid X_{pert})] \quad (10)$$

式中 y 代表真实的标签。最大化损失函数 $L_s + L_a + L_g$ 促使生成图像无限逼近于真实图像，从而保证对抗样本的质量。

2.2.2 生成器网络结构

生成器采用 ResNet-50 作为基本模型，利用编码-解码结构进行特征提取，并且设计基于扩张卷积的残差块和金字塔分割注意力机制以提高特征表达能力，在解决深度神经网络退化问题时还能有效地增大卷积核的感受野。在医学图像中，由于病变区域或模型感知的潜在疾病表达区域改变在攻击中起决定性作用，所以将针对自然图像的攻击方法应用到医学图像特征中缺乏特异性，因此引入金字塔分割注意力机制，使扰动更多地集中发生在病变区域，生成有针对性的对抗扰动。生成器和判别器的网络结构如图 5 所示。

生成器的损失函数由攻击目标模型的交叉熵损失 L_t 和判别损失 L_D 组成，定义为

$$L_t = E[\log P(C = t \mid X_{pert})] \quad (11)$$

$$L_D = E[\log P(C = t \mid X_{pert})] \quad (12)$$

式中 t 是目标攻击的类别。最大化损失函数 $L_t + L_D - L_S$ 使对抗样本在攻击过程中的结果更接近于期望值。

在原始图像输入和生成器输出之间引入金字塔分割注意力机制，该注意力模块能够充分提取多尺度特征图的空间信息，实现跨维度通道的注意力特征交互。PSA 模块主要通过 4 个步骤实现，如图 6 所示。

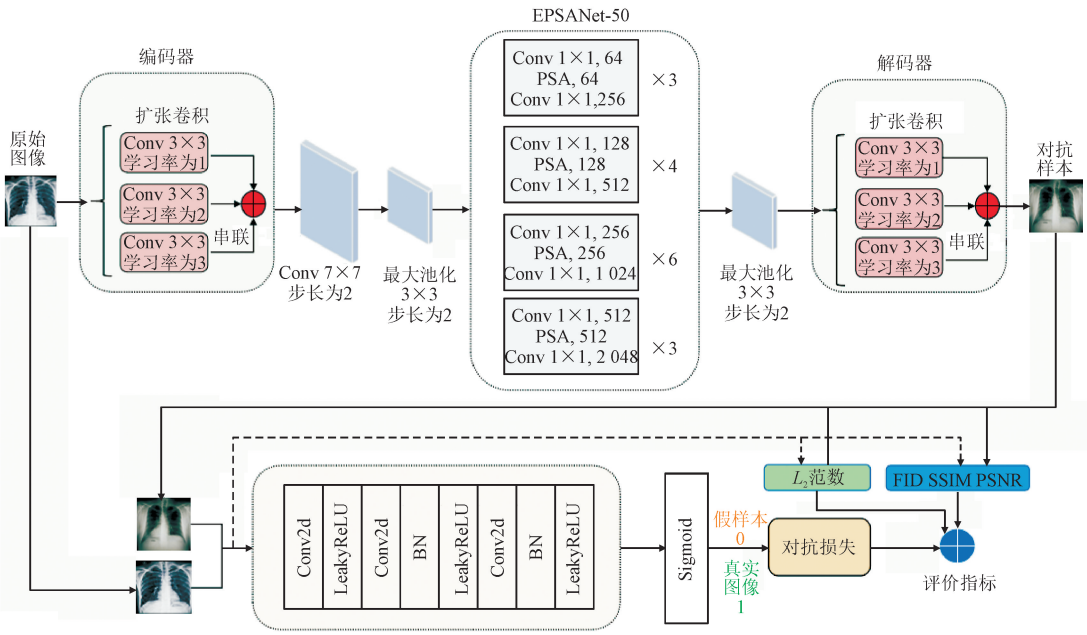

图 5 生成器和判别器网络结构

Fig. 5 The network structure diagram of generator and discriminator

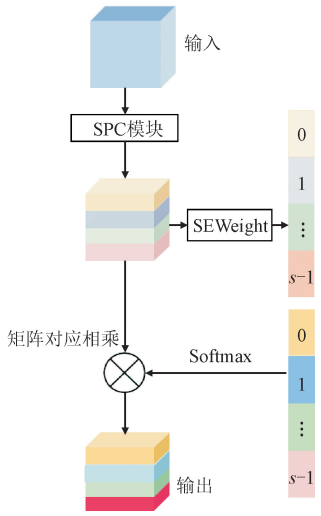

图 6　金字塔分割注意力机制

Fig. 6　The pyramid split attention mechanism

(1)利用 SPC 模块将通道切分为 S 组，然后针对每个通道特征图上的空间信息进行多尺度特征提取，获得多尺度特征图

$$\boldsymbol{F}=\mathrm{Cat}([\boldsymbol{F}_0,\boldsymbol{F}_1,\cdots,\boldsymbol{F}_{S-1}]) \tag{13}$$

(2)采用 SEWeight 模块提取不同尺度特征图的通道注意力[18]，得到每个尺度上的通道注意力向量

$$\boldsymbol{Z}_i=\mathrm{SEWeight}(\boldsymbol{F}_i),\ i=0,1,2,\cdots,S-1 \tag{14}$$

为了不破坏原始通道注意力向量的跨维度融合，以串联的方式获得整个多尺度通道注意力向量

$$\boldsymbol{Z}=\boldsymbol{Z}_0\oplus\boldsymbol{Z}_1\oplus\cdots\oplus\boldsymbol{Z}_{S-1} \tag{15}$$

(3)使用 Softmax 函数对多尺度通道注意力向量进行特征重新标定，得到新的多尺度通道交互注意力权重

$$\boldsymbol{t}_i=\mathrm{Softmax}(\boldsymbol{Z}_i)=\frac{\exp(\boldsymbol{Z}_i)}{\sum_{i=0}^{S-1}\exp(\boldsymbol{Z}_i)} \tag{16}$$

(4)对重新校准的权重和相应的特征图按像素进行点乘操作，输出一个多尺度特征信息加权之后的特征图

$$\boldsymbol{O}=\mathrm{Cat}([\boldsymbol{Y}_0,\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_{S-1}]) \tag{17}$$

式中 $\boldsymbol{Y}_i=\boldsymbol{t}_i\boldsymbol{F}_i$，$\boldsymbol{Y}_i$ 的多尺度信息表示能力更加丰富。

使用 PSA 模块替换 ResNet-50 残差块中的 3×3 卷积，从而获得高效的金字塔分割注意模块(EPSA)。EPSA 可以提高模型在更细粒度上的多尺度特征表示能力，捕捉远程通道之间的相互依赖关系，提升网络性能。

3　实验结果与分析

在黑盒场景设置下，将本文方法 BA-GAN 与其他 4 种基于 GAN 的对抗方法在肺部疾病诊断模型上的攻击效果进行比较，通过对抗攻击成功率、可迁移性以及样本生成质量指标对攻击性能进行评估。设置消融实验来验证生成模型中各模块的必要性和有效性。

3.1　数据集

使用新冠肺炎数据集[19]和 Chest X-ray14 数据集，图像均是从公开可用的数据集和已发表的文章中收集的，Chest X-ray14 数据集中包含 14 种病理：肺不张、心脏肥大、渗出、浸润、肿块、肺结节、普通肺炎、气胸、肺实变、水肿、肺气肿、纤维变性、胸膜增厚和肺疝。

3.2　构建肺部疾病诊断目标模型

对新冠肺炎数据集和 Chest X-ray14 数据集按照 7∶1∶2 的比例划分为训练集、验证集和测试集。为充分解决训练数据的不均衡问题，使用随机翻转、平移以及不同缩放比例进行不同程度的数据增广，经过数据增广后各单类疾病基本到达平衡。

构建肺部疾病诊断目标模型 CheXNet，以 DenseNet169 网络为基本骨架，利用 3×3 小卷积替换 7×7 大卷积减少模型参数量，并通过密集连接充分提取肺部图像中的纹理边缘信息，将全连接层的输出维度修改为 15，使用 Sigmoid 非线性激活函数实现对模型最终的分类输出，完成肺部 X-Ray 图像的多种疾病智能诊断。改进 CheXNet 模型架构如图 7 所示。

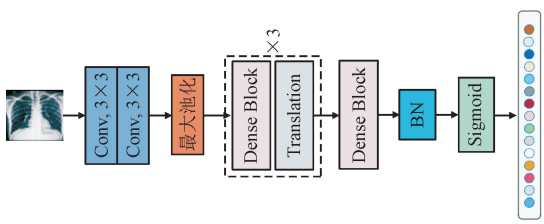

图 7　改进 CheXNet 模型架构

Fig. 7　The model architecture of improved CheXNet

模型在训练时，将图像统一采样为 224×224 的分辨率，设置 Batch_size 为 32，并根据 ImageNet 数据集上预训练的权重初始化模型，采用 Momentum 和 SGD 算法进行优化迭代，初始学习率为 10^{-5}，每次验证损失在趋于稳定一个周期后，学习率衰减为原来的 1/10，直至目标模型达到收敛状态后保存目标模型。

肺部疾病目标模型的诊断正确率如表 1 所示。可以看出,改进 CheXNet 模型在各种疾病上的诊断准确度明显超过了对比模型,能实现对多种肺部疾病的精准诊断,可以作为下一阶段对抗攻击的目标模型。

表 1 肺部疾病目标模型的诊断正确率

Table 1 The diagnostic accuracy of lung disease target model

肺部疾病类型	诊断正确率/%			
	Wang 等[20]	Yang 等[21]	CheXNet[22]	改进 CheXNet
新冠肺炎	67.61	73.21	78.94	82.39
肺不张	71.64	77.72	80.94	83.65
心脏肥大	80.73	90.46	92.48	92.43
渗出	78.47	85.19	86.18	89.09
浸润	60.92	69.35	73.45	72.57
肿块	70.64	79.82	86.46	87.12
肺结节	67.15	71.37	78.08	79.74
普通肺炎	63.72	71.35	76.84	78.25
气胸	80.96	84.51	88.87	87.64
肺实变	70.08	78.82	79.01	81.38
水肿	83.53	88.21	88.72	90.05
肺气肿	81.25	82.29	93.14	93.63
纤维化	76.69	76.67	80.62	82.91
胸膜增厚	70.48	76.45	80.63	81.07
肺疝	76.75	91.46	91.37	92.23

3.3 肺部疾病诊断模型对抗攻击

3.3.1 对比实验及结果分析

使用改进 CheXNet 模型作为目标模型,训练 DenseNet121 网络作为替代模型,使用 CycleGAN 的生成器模型合成数据[23]。为了对比 DBAA 在黑盒攻击场景下的效果,与 3 种先进的黑盒攻击方法 JPBA[24]、DaST[25]和 DFME 进行比较,并设置无目标对抗攻击和有针对性的目标攻击两种方式,结果如表 2 所示。可以看出:在相同的对抗样本生成方法下,本文提出的黑盒对抗攻击框架 DBAA 的平均对抗攻击成功率高于其他 3 种黑盒攻击方法;在相同的黑盒攻击方式下,本文提出的对抗样本生成方法 AI-GAN 的对抗攻击成功率高于其他 4 种基于 GAN 的对抗样本生成方法。

表 2 5 种对抗方法的无目标和目标平均对抗攻击成功率

Table 2 Average adversarial attack success rate of untarget and target attack of five adversarial methods

攻击方式	黑盒攻击方法	平均对抗攻击成功率/%				
		AdvGAN	AdvGAN++	Natural-GAN	Rob-GAN	AI-GAN
目标攻击	JPBA	14.17	17.25	24.07	32.63	41.75
	DaST	15.04	19.68	26.45	35.49	42.80
	DFME	21.18	28.79	35.68	46.81	60.04
	DBAA	26.64	34.42	43.19	52.34	**68.95**
无目标攻击	JPBA	22.86	27.78	36.07	43.80	53.06
	DaST	20.59	24.29	35.39	45.56	52.57
	DFME	29.73	35.22	46.82	57.39	68.32
	DBAA	33.54	42.03	53.67	64.82	**79.34**

注:表中加粗数据为最优值。

表 3 展示了常见的肺部疾病在 BA-GAN 方法下的黑盒攻击成功率。可以看出，本文对抗攻击方法 BA-GAN 能显著地提高黑盒攻击方式下的对抗成功率，能有效地欺骗肺部疾病诊断模型进而产生错误的诊断结果。

表 3　常见肺部疾病在 BA-GAN 方法下的对抗攻击成功率

Table 3　Success rate of adversarial attacks on common lung diseases using the BA-GAN method

肺部疾病类型	对抗攻击成功率/%	
	目标攻击	无目标攻击
新冠肺炎	70.13	81.29
渗出	64.16	77.43
肺结节	67.89	76.90
气胸	68.84	79.58
普通肺炎	69.25	80.47
水肿	68.75	79.25
肺气肿	69.42	80.36
肺实变	68.37	79.51

BA-GAN 不仅增强了对抗攻击性能，也提高了对抗样本的可迁移性。可迁移性是衡量替代模型生成的对抗样本迁移到未知内部信息的黑盒目标模型上的能力。表 4 展示了黑盒攻击场景下 4 种基于 GAN 的攻击方法以及 BA-GAN 生成的对抗样本的非目标迁移攻击成功率和目标迁移攻击成功率。可以看出，BA-GAN 生成的对抗样本具有一定的迁移能力，可以显著转移到其他未知的目标模型上，即生成的对抗样本能以黑盒形式攻击其他未知的网络模型并取得高攻击成功率。

表 4　对抗样本的可迁移性

Table 4　Transferability of adversarial samples

攻击方法	可迁移率/%	
	目标攻击	无目标攻击
AdvGAN	15.71	20.65
AdvGAN++	19.34	25.79
Natural-GAN	20.56	27.08
Rob-GAN	21.09	29.30
BA-GAN	**23.40**	**32.52**

注：表中加粗数据为最优值。

AdvGAN、AdvGAN++、Natural-GAN、Rob-GAN 和 BA-GAN 的对抗攻击效果对比如图 8 所示。可以看出，随着迭代轮数的增加，BA-GAN 攻击模型的收敛速度更快，且设置在黑盒场景 DBAA 下的无目标对抗攻击对抗成功率达到了 79.34%，优于其他 4 种基于 GAN 的对抗攻击方法，因此 BA-GAN 的对抗攻击效果更佳。

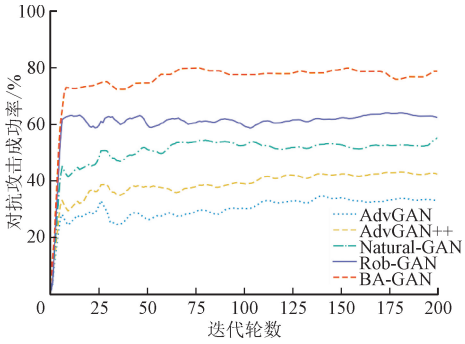

图 8　5 种方法对抗攻击效果对比

Fig. 8　The comparison diagram of adversarial attack

图 9 可视化了原始肺部图像、不同方法生成的对抗样本以及改变的像素点，使用红色标注与原图相比发生改变的像素点，用绿色标注未改变的像素点。计算结果显示，AdvGAN、AdvGAN++、Natural-GAN、Rob-GAN 和 BA-GAN 生成的对抗样本与原图相比像素数改变了 55.01%、49.88%、35.69%、42.72%和 25.65%。由实验结果可知，本文 BA-GAN 方法能以更少地改变攻击目标中的像素来达到成功攻击的目的，并且生成的对抗样本更真实自然。

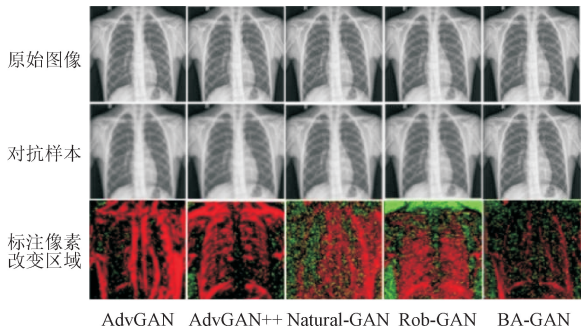

图 9　5 种方法对抗样本可视化对比

Fig. 9　Visualization of adversarial samples

为了评估生成的对抗样本图像质量，使用 FID (Frechet inception distance score)分数[26]、结构相似度(structural similarity，SSIM)、L_2 范数以及峰值信噪比(peak signal to noise ratio，PSNR)对上述 5 种对抗攻击方法进行实验对比。

FID 分数是一种评估生成图像质量的指标，用来计算原始图像和生成图像之间的相似度。在无目标攻击任务下，5 种对抗方法的 FID 分数如图 10 所

示。可以看出，BA-GAN 的 FID 分数中位数最小、上四分位点和下四分位点相比于其他 4 种方法也较小，其次是 Natural-GAN，然后是 Rob-GAN、AdvGAN++和 AdvGAN。根据小提琴图的概率密度和分布状态可知，该方法生成的对抗样本与原始图像计算的 FID 分数值最小，表明生成的对抗样本更接近于原始图像，具有更清晰的纹理细节，人眼看起来更加自然。

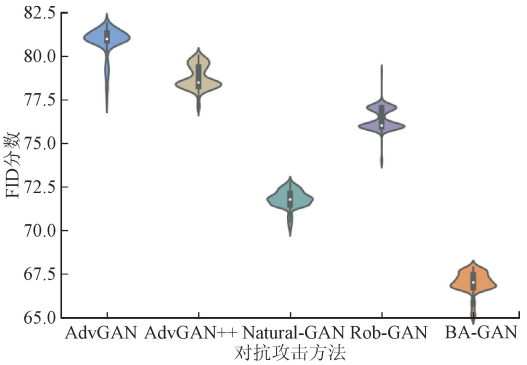

图 10 对抗攻击方法的 FID 分数对比

Fig. 10 FID scores comparison diagram of adversarial methods

结构相似度是一种从亮度、结构和对比度 3 个方面综合评价图像的指标，5 种对抗方法的结构相似度指数如图 11 所示。可以看出，BA-GAN 和 Natural-GAN 的结构相似度随着迭代轮数的增加显著优于其他方法，比只添加了攻击者的 Rob-GAN、使用原始 GAN 网络的 AdvGAN 和 AdvGAN++方法在生成方面能力更强，表明通过优化生成器的网络结构，可以进一步提高网络性能。

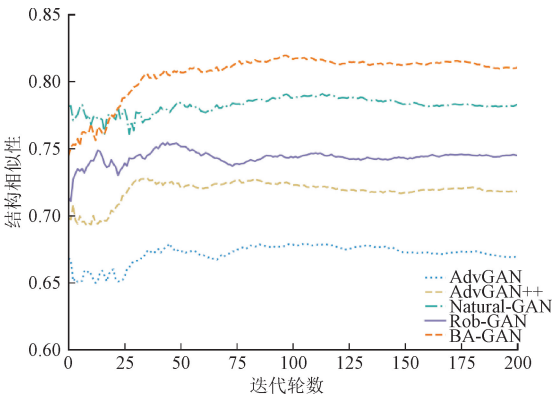

图 11 对抗攻击方法的结构相似度对比

Fig. 11 SSIM comparison diagram of adversarial methods

L_2 范数是为了提高模型的抗过拟合能力而被加入到损失函数中的扰动约束，能定量判断攻击结果是否符合视觉的感知判断。5 种对抗方法的 L_2 范数变化趋势如图 12 所示。可以看出，在攻击程度趋于收敛时，对抗扰动幅度也在减小。BA-GAN 方法计算出的扰动量明显小于其他 4 方法，生成的对抗样本与原始图像更加接近，质量更高，对抗攻击效果也更加可信。

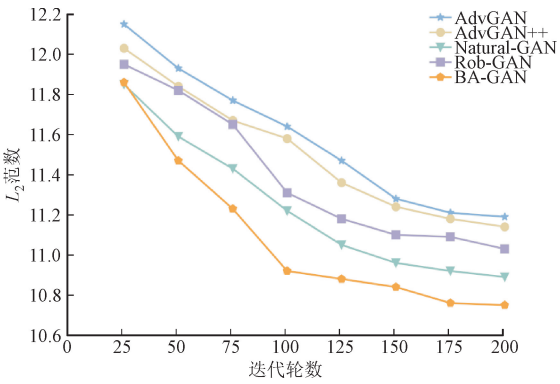

图 12 对抗攻击方法的 L_2 约束对比

Fig. 12 L_2-norms comparison diagram of adversarial methods

峰值信噪比是一种基于误差敏感的图像质量评价指标，图 13 展示了 BA-GAN 与其他 4 种对抗攻击方法的峰值信噪比。可以看出，Natural-GAN 和 BA-GAN 的峰值信噪比高于其他 3 种方法，说明通过修改生成器的结构，充分提取图像空间中的特征关系可以很好地提高 GAN 网络的性能，生成的对抗样本与原始图像的特征分布更加接近，视觉效果更好。

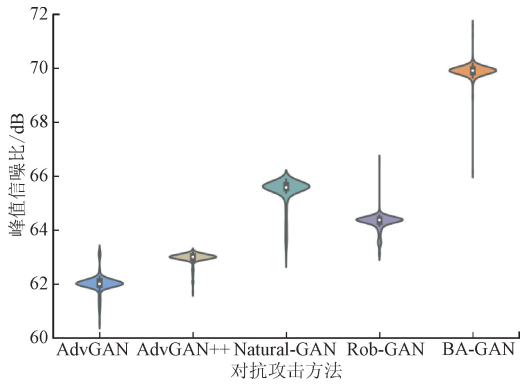

图 13 对抗攻击方法的峰值信噪指数对比

Fig. 13 PSNR comparison diagram of adversarial methods

3.3.2 消融实验及结果分析

为了验证本文 BA-GAN 的有效性，需要进一步分析对抗样本生成模型中关键模块对实验结果的影响。为此，设计相关的消融实验，主要探究在 BA-GAN 上移除扩张卷积的残差块(DR)、金字塔分割注意力机制(PSA)、AC-GAN 判别器(AC)和 PGD

攻击者(ATT)对对抗样本生成的影响，最后根据实验结果分析各模块的重要性。

为了确保验证的合理性，实验中用到的模型采用完全相同的训练参数，并且都设置在黑盒攻击场景下。针对目标攻击和无目标攻击两种方式，分别从对抗攻击成功率和对抗样本添加的平均扰动量来验证对抗样本生成模型中各个模块的必要性，实验结果如表 5 所示。

由表 5 可知，在移除扩张卷积的残差块和金字塔注意力机制后，攻击成功率明显下降，扰动也显著增加，说明在生成器中引入扩张卷积的残差块能在特征提取时增大感受野和提取多尺度上下文信息，比使用普通卷积的性能更好。引入的金字塔分割注意力机制能在多尺度通道注意力间建立一种长距离的特征依赖关系，从而弥补了仅使用扩张卷积获取的远距离信息没有相关性的弊端，因此缺失这两个模块后对于目标攻击任务和非目标攻击任务，对抗攻击成功率分别下降了 7%和 10%，平均扰动分别增加了 4.92 和 5.89。相比于移除 DR 和 PSA，移除 AC 和 ATT 的攻击效果略微下降，说明使用 AC-GAN 判别器和添加攻击者对对抗样本的攻击能力和生成质量有一定的提升，但效果有限。一方面是因为在 GAN 的判别器中添加分类网络以实现在分类特征空间中引起攻击的难度较大，另一方面是因为采用 PGD 攻击者虽然能控制扰动在一定的范围内，但需要对梯度进行多步迭代才能获取最优值。消融实验结果证明了本文提出的对抗样本生成模型中各模块的优越性，说明使用 BA-GAN 方法生成的对抗样本在对抗攻击时获得的高黑盒攻击成功率更加可靠，图像扰动更小。

表 5　消融实验结果

Table 5　Results of ablation experiment

移除模块	目标攻击		无目标攻击	
	对抗攻击成功率/%	平均扰动	对抗攻击成功率/%	平均扰动
BA-GAN	**68.95**	**13.75**	**79.34**	**10.75**
DR	65.91	15.57	74.17	12.48
PSA	64.43	17.12	72.34	14.37
AC	67.20	14.59	76.59	11.39
ATT	67.73	14.86	77.83	12.02
DR+PSA	61.19	18.67	69.96	16.64
DR+PSA+AC	59.65	19.45	68.58	17.38

注：表中加粗数据为最优值。

4　结　论

本文提出了一种基于 GAN 的肺部疾病诊断模型黑盒可迁移性对抗攻击方法 BA-GAN，在无目标对抗攻击和有针对性对抗攻击两种方式下，该方法能在查询次数有限、更严格的黑盒场景中实现可迁移性对抗攻击，并取得较高的攻击成功率。改进后的对抗样本生成方法相比于其他传统基于 GAN 的攻击方法更能捕捉肺部图像的语义特征信息和纹理细节，生成的对抗样本更加真实，并且可迁移性更高。

自从发现 DNN 存在对抗扰动的问题后，基于 DNN 的临床诊断模型更容易被攻击者欺骗，导致医疗模型产生致命性错误，进而误导医生做出错误的决策[27]，因此本文提出的对抗攻击方法是为了更好地帮助深度神经网络抵御未知的恶意攻击，进而训练出更稳健、更具解释性的模型，并且为加固深度学习模型在医疗领域的安全性和鲁棒性提供了参考方案。在进一步的研究中，需要测试更多的医疗诊断模型，优化训练方法，提高黑盒对抗成功率。但是，在开发用于医学成像的 DNN 模型及其实际应用时，需要更加谨慎仔细地考虑。

参考文献：

[1] 匡艳. 无监督肺结节良恶性辅助诊断研究 [D]. 成都：电子科技大学，2021：1-8.

[2] 刘雲. 全流程人工智能计算机辅助诊断在肺癌及食管癌中的应用 [D]. 上海：华东师范大学，2022：1-6.

[3] WANG Zizhou, SHU Xin, WANG Yan, et al. A feature space-restricted attention attack on medical deep learning systems [J]. IEEE Transactions on Cybernetics 2022,52(4): 2168-2275.

[4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C/OL]//2nd

International Conference on Learning Representations (ICLR). New York, USA: ICLR, 2014 [2022-03-01]. https://nyuscholars.nyu.edu/en/publications/intriguing-properties-of-neural-networks.

[5] HIRANO H, MINAGI A, TAKEMOTO K. Universal adversarial attacks on deep neural networks for medical image classification [J]. BMC Medical Imaging, 2021, 21(1): 9.

[6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [C]//3rd International Conference on Learning Representations (ICLR). New York, USA: ICLR, 2015: 20193407343614.

[7] XIE Cihang, WU Yuxin, MAATEN L V D, et al. Feature denoising for improving adversarial robustness [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 501-509.

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

[9] XIAO Chaowei, LI Bo, ZHU Junyan, et al. Generating adversarial examples with adversarial networks [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2018: 3905-3911.

[10] BAI Tao, ZHAO Jun, ZHU Jinlin, et al. AI-GAN: attack-inspired generation of adversarial examples [C]//2021 IEEE International Conference on Image Processing (ICIP). Piscataway, NJ, USA: IEEE, 2021: 2543-2547.

[11] ZHANG Jie, LI Bo, XU Jianghe, et al. Towards efficient data free blackbox adversarial attack [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2022: 15094-15104.

[12] TRUONG J B, MAINI P, WALLS R J, et al. Data-free model extraction [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 4769-4778.

[13] 左龙，张鹏，荆树旭，等. 用于图像超分辨率重建的双通道残差网络 [J]. 西安交通大学学报，2022，56(1)：158-164.
ZUO Long, ZHANG Peng, JING Shuxu, et al. Dual-channel residual network for image super-resolution reconstruction [J]. Journal of Xi'an Jiaotong University, 2022, 56(1): 158-164.

[14] TIAN Hongfeng, JI Bai, QUAN Wei, et al. MPA-net: multi-scale pyramid attention network for liver tumor segmentation [C]//2021 International Conference on Electronic Information Engineering and Computer Science (EIECS). Piscataway, NJ, USA: IEEE, 2021: 658-661.

[15] 王小银，吕硕，孙家泽，等. 基于生成对抗网络的医学诊断模型知识蒸馏对抗攻击方法 [J]. 西安交通大学学报，2022，56(7)：76-85.
WANG Xiaoyin, LÜ Shuo, SUN Jiaze, et al. Knowledge distillation adversarial attack method based on generative adversarial network for medical diagnosis model [J]. Journal of Xi'an Jiaotong University, 2022, 56(7): 76-85.

[16] WANG Wenxuan, YIN Bangjie, YAO Taiping, et al. Delving into data: effectively substitute training for black-box attack [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 4759-4768.

[17] WANG Zi. Learning fast converging, effective conditional generative adversarial networks with a mirrored auxiliary classifier [C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE, 2021: 2565-2574.

[18] LIU Yichen. A lane line detection method based on squeeze and excitation network [C]//2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE). Piscataway, NJ, USA: IEEE, 2022: 117-121.

[19] RAHMAN T, KHANDAKAR A, QIBLAWEY Y, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images [J]. Computers in Biology and Medicine, 2021, 132: 104319.

[20] WANG Xiaosong, PENG Yifan, LU Le, et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 3462-3471.

[21] YANG Huan, CHEN Lili, CHENG Zhiqiang, et al. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study [J]. BMC Medicine, 2021, 19(1): 80.

[22] DEVNATH L, LUO Suhuai, SUMMONS P, et al. Deep ensemble learning for the automatic detection of

(下转第 220 页)

Communications, 2021, 42(9): 218-230.

[28] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. The Journal of Machine Learning Research, 2010, 11: 3371-3408.

[29] 侯青松，郭英，王布宏，等. 共形阵列天线振动条件下稳健的 DOA 估计及位置误差校正 [J]. 信号处理，2010，26(11)：1756-1760.
HOU Qingsong, GUO Ying, WANG Buhong, et al. Robust direction finding and position errors calibration for conformal array antenna in the presence of vibration [J]. Signal Processing, 2010, 26(11): 1756-1760.

[30] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks [C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Sardinia, Italy: PMLR, 2011: 315-323.

（编辑　陶晴）

（上接第 212 页）

pneumoconiosis in coal worker's chest X-ray radiography [J]. Journal of Clinical Medicine, 2022, 11(18): 5342.

[23] 余艳杰，孙嘉琪，葛思擘，等. CycleGAN-SN：结合谱归一化和 CycleGAN 的图像风格化算法 [J]. 西安交通大学学报，2020，54(5)：133-141.
YU Yanjie, SUN Jiaqi, GE Sibo, et al. CycleGAN-SN: image stylization algorithm combining spectral normalization and CycleGAN [J]. Journal of Xi'an Jiaotong University, 2020, 54(5): 133-141.

[24] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning [C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery, 2017: 506-519.

[25] ZHOU Mingyi, WU Jing, LIU Yipeng, et al. DaST: data-free substitute training for adversarial attacks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 231-240.

[26] 许新征，常建英，丁世飞. 基于 StarGAN 和类别编码器的图像风格转换 [J]. 软件学报，2022，33(4)：1516-1526.
XU Xinzheng, CHANG Jianying, DING Shifei. Image style transfering based on StarGAN and class encoder [J]. Journal of Software, 2022, 33(4): 1516-1526.

[27] HIRANO H, TAKEMOTO K. Simple iterative method for generating targeted universal adversarial perturbations [J]. Algorithms, 2020, 13(11): 268.

（编辑　陶晴）