基于双曲正切和矩的免疫防御

吴 昊[1] 王金伟[1),2)] 罗向阳[3)] 马 宾[4)]

[1)](南京信息工程大学计算机学院 南京 210044)

[2)](南京信息工程大学数字取证教育部工程研究中心 南京 210044)

[3)](中国人民解放军战略支援部队信息工程大学网络空间安全学院 郑州 450001)

[4)](齐鲁工业大学网络空间安全学院 济南 250353)

摘 要 对抗样本的发现与研究证实了深度神经网络的脆弱性.如果不对对抗样本的生成加以约束,那么触手可及的图像将不再安全并随时可能对不鲁棒的深度神经网络构成威胁.然而,现有的对抗防御主要旨在防止对抗样本成功攻击深度神经网络,而不是防止对抗样本的生成.因此,本文提出了一种新颖的对抗防御机制,该机制被称为免疫防御.免疫防御通过主动地在原始图像上添加难以察觉的扰动使得攻击者无法针对该图像制作出有效的对抗样本,从而同时保护了图像和深度神经网络.这种良性的扰动被称为免疫扰动,添加了免疫扰动的图像被称为免疫样本.在白盒免疫防御中,本文提出了双曲正切免疫防御(Hyperbolic Tangent Immune Defense,HTID)以制作高分类准确率、高防御性能和高视觉质量的白盒免疫样本;在黑盒免疫防御中,提出了基于矩的免疫防御(Moment-based Immune Defense,MID)以提升免疫样本的可迁移性,从而确保免疫样本对未知对抗攻击的防御性能.此外,本文还提出了免疫率以更加准确地衡量免疫样本的防御性能.在CIFAR-10、MNIST、STL-10和Caltech-256数据集上的大量实验表明,HTID和MID制作的免疫样本具有高分类准确率,在Inception-v3、ResNet-50、LeNet-5和Model C上的准确率均达到了100.0%,比原始准确率平均高出10.5%.制作的免疫样本同时具有高视觉质量,其SSIM最低为0.822,最高为0.900.实验也表明MID有着比HTID更高的可迁移性,MID在四个数据集上针对AdvGAN制作的免疫样本防御其他11种对抗攻击的平均免疫率分别为62.1%、52.1%、56.8%和48.7%,这比HTID高出15.0%、10.8%、17.5%和15.7%.

关键词 深度神经网络;对抗样本;对抗防御;免疫防御;可迁移性

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.01786

Immune Defense Based on Hyperbolic Tangent and Moments

WU Hao[1] WANG Jin-Wei[1),2)] LUO Xiang-Yang[3)] MA Bin[4)]

[1)](*School of Computer Science, Nanjing University of Information Science and Technology, Nanjing* 210044)

[2)](*Engineering Research Center of Digital Forensics Ministry of Education, Nanjing University of Information Science and Technology, Nanjing* 210044)

[3)](*School of Cyber Security, PLA Strategic Support Force Information Engineering University, Zhengzhou* 450001)

[4)](*School of Cyber Security, Qilu University of Technology, Jinan* 250353)

Abstract The vulnerability of deep neural networks to adversarial examples has been confirmed. If the generation of adversarial examples is unregulated, images within reach are no longer secure and pose a threat to non-robust DNNs. However, existing adversarial defenses primarily aim at preventing adversarial examples from attacking deep neural networks successfully, rather than

收稿日期:2023-05-30;在线发布日期:2024-05-11.本课题得到国家自然科学基金(No. 62072250, 62172435, U1804263, U20B2065, 61872203, 71802110, 61802212)、中国中原科技创新领军人才项目基金(No. 214200510019)、河南省网络空间态势感知重点实验室开放课题基金(No. HNTS2022002)资助.**吴 昊**,硕士,主要研究领域为人工智能安全、深度学习、图像处理. E-mail: howwooo@163.com. **王金伟**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能安全、多媒体取证、多媒体信息隐藏. E-mail: wjwei_2004@163.com. **罗向阳**(通信作者),博士,教授,主要研究领域为图像隐写、隐写分析. E-mail: luoxy_ieu@sina.com. **马 宾**,博士,教授,主要研究领域为可逆信息隐藏、多媒体安全、图像处理.

preventing their generation. Therefore, we propose a novel adversarial defense mechanism, which is referred to as immune defense. This mechanism applies carefully designed quasi-imperceptible perturbations to the raw images to prevent the generation of adversarial examples for the raw images thereby protecting both images and deep neural networks. Such perturbations are referred to as immune perturbations, and these perturbed images are referred to as immune examples. In the white-box immune defense, we propose Hyperbolic Tangent Immune Defense (HTID) to craft white-box immune examples with high classification accuracy, defensive performance, and visual quality. In the black-box immune defense, we propose Moment-based Immune Defense (MID) to enhance the transferability of immune examples, so as to ensure the defensive performance against unknown adversarial attacks. In addition, we propose immune rate to more accurately measure the defensive performance of immune examples. Extensive experiments on CIFAR-10, MNIST, STL-10, and Caltech-256 show that the immune examples crafted by HTID and MID have high classification accuracy, which reaches 100.0% and is 10.5% higher than the original accuracy on average. The immune examples also have high visual quality with SSIM between 0.822 and 0.900. The experiments also show that MID has higher transferability than HTID. The average immune rates of the immune examples crafted by MID against AdvGAN to defend against other 11 adversarial attacks on the two datasets are 62.1%, 52.1%, 56.8% and 48.7%, which are 15.0%, 10.8%, 17.5% and 15.7% higher than HTID, respectively.

Keywords deep neural network; adversarial example; adversarial defense; immune defense; transferability

1 引 言

对抗样本(Adversarial Examples)[1-7]的发现与研究证明了深度神经网络(Deep Neural Networks, DNNs)[8-12]是脆弱的. 攻击者可以通过向图像中添加精心设计的微小扰动轻而易举地制作出对抗样本. 当对抗样本被输入给DNNs时,它们会使得DNNs产生错误的输出,从而导致严重的安全问题.

对抗攻击(Adversarial Attacks)可以被分为以下三大类:基于梯度的攻击[2,13-17]、基于优化的攻击[1,18]和基于生成的攻击[19-23]. 具体来说,基于梯度的攻击直接利用DNNs反馈的梯度信息以增加分类损失,其具有攻击速度快和可迁移性较高的优点,但其攻击能力有限. 相比之下,基于优化的攻击通过优化一个多目标损失函数(包含分类的误差和对抗扰动的限制等)以制作出具有高视觉质量且强大的对抗样本,但其攻击速度慢且可迁移性较差. 此外,基于生成的攻击通过深度生成模型(Deep Generative Models)[24-26]直接生成多样化的对抗样本,其同时具备了生成速度快和攻击能力强的优点.

目前流行的对抗防御(Adversarial Defenses)包括对抗训练(Adversarial Training)[2,16,27-29]、防御蒸馏(Defensive Distillation)[30]、输入预处理[31-34]、基于检测的方法[35-36]和梯度混淆(Gradient Obfuscation)[31,33,37-38]. 这些防御方法通过提升模型鲁棒性、检测对抗样本或传递错误的梯度给攻击者以抵御对抗攻击. 除了梯度混淆,所有的对抗防御均是一种事后措施,即它们尝试防止对抗样本成功攻击DNNs但并不能避免对抗样本的产生. 如果对对抗样本的产生不加以控制,那么我们触手可及的图像将不再是安全的,它们可能随时被中间人(Man-in-the-Middle, MitM)攻击[53]制作成对抗样本并对不鲁棒的模型构成威胁. MitM攻击是一种能够拦截并操作通信数据的网络攻击,其可能发生在无线网络、网页浏览和即时通讯中. 例如,如图1(a)所示,Wang等人[54]指出在一些自动驾驶汽车或支持机器学习服务的移动应用中,相机将捕获到的图像发送到在线的DNN分类器进行分类,而MitM攻击可以在DNN获取数据前拦截并恶意修改数据以欺骗DNN模型. 此外,如图1(b)所示,Brown等人[55]表示对抗补丁可以被打印并张贴任何场景,该场景被拍照后能够使得分类器出错. 尽管预训练的

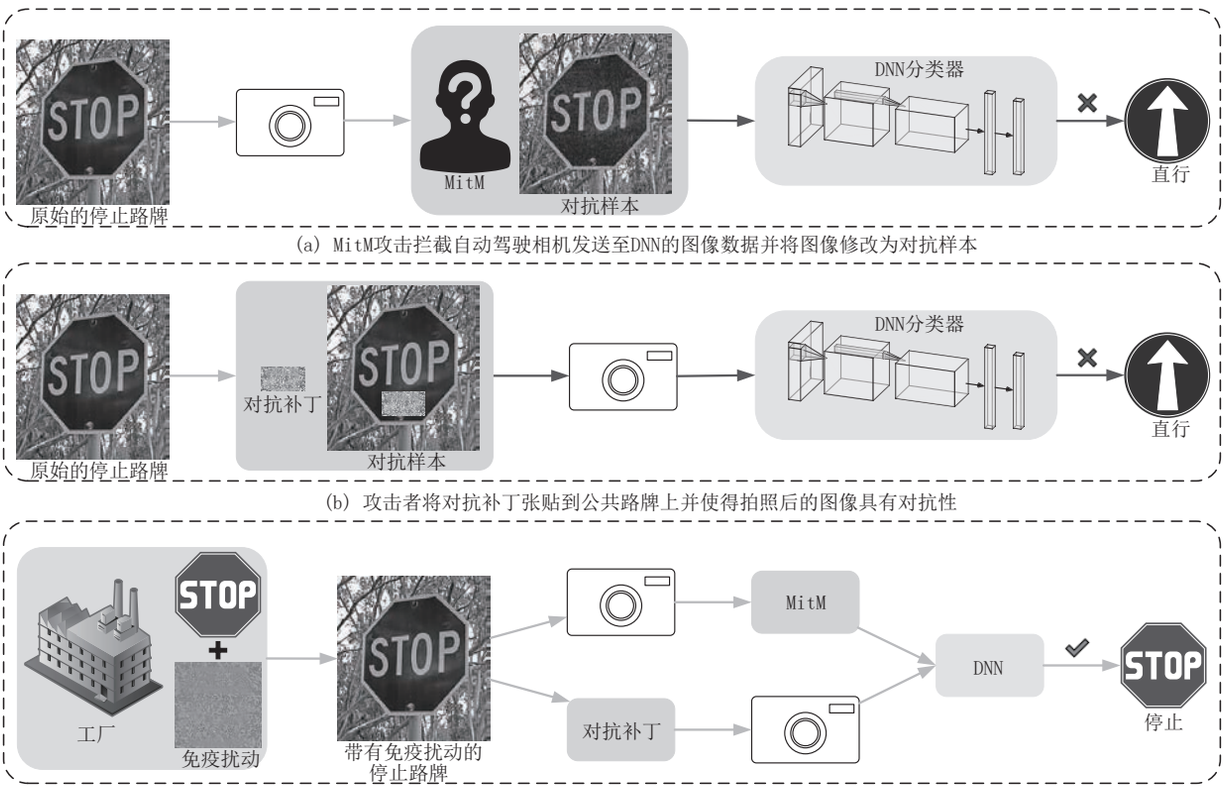

图1 对抗攻击和免疫防御的示意图

去噪器(Denoiser)[43,56-57]可以被放置在DNN之前以尝试去除对抗扰动,但Wang等人[54]通过改变去噪器的参数使得经过去噪器的图像变成对抗样本,从而使得去噪器失去去噪功能. 总的来说,现有的对抗防御无法有效地限制对抗样本的产生,因此无法保证图像的安全性和鲁棒性.

为了解决这一问题,我们提出了一种新颖的对抗防御机制,即免疫防御(Immune Defense). 如图1(c)所示,免疫防御通过向原始图像中添加精心制作的微小扰动以防止图像被制作成对抗样本,我们称添加的扰动为免疫扰动,并称添加了免疫扰动的图像为免疫样本. 该机制可以增强图像的安全性和鲁棒性,从而减轻了潜在对抗攻击的威胁. 在任务上,免疫防御不仅需要使得对抗扰动(Adversarial Perturbations)失效还需要维持图像的分类性能,这样的双重需求使得免疫防御的任务变得更加具有挑战性. 据我们所知,之前的工作没有以类似的方式研究过免疫防御,免疫防御是新颖且具有挑战性的.

在白盒(White-box)免疫防御中,防御者已知目标对抗攻击的全部信息(如扰动范数、最大扰动强度和攻击算法),从而可以相对轻松地针对目标对抗攻击制作出强大的白盒免疫样本. 在章节3中,我们提出了HTID以制作白盒免疫样本. HTID通过双曲正切函数将参数转换至低灵敏度空间,使得优化出的免疫扰动更加细微. 附录C中的图10展示了由HTID制作的免疫样本.

此外,由于免疫样本在面对未知的对抗攻击时仍需保证有效,所以更加复杂的黑盒(Black-box)免疫防御也应当被纳入考虑. 具体来说,我们考虑以下潜在的情况:防御者针对某个特定的对抗攻击制作出的白盒免疫样本受到了其他未知对抗攻击的破坏进而失效,这意味着未知的对抗攻击依旧能够针对白盒免疫样本制作出有效的对抗样本. 针对这一威胁,我们提出了MID以提升免疫样本的可迁移性(Transferability),从而使得免疫样本能够在多种对抗攻击之间保持有效性. MID借助切比雪夫不等式(Chebyshev's Inequality)[39]最小化了未知对抗攻击使得免疫样本失效的概率,从而提升了免疫样本的可迁移性,我们将在章节3中具体介绍该方法.

综上所述,本文的贡献如下:

(1)提出了一种新颖的对抗防御机制,即免疫防御. 免疫防御可以防止对抗样本的生成,从而保护图像和DNNs以免受对抗攻击的威胁.

(2)提出了HTID以制作白盒免疫样本.此外,我们还提出了MID以提升免疫样本的可迁移性,从而保证免疫样本在面对未知对抗攻击时依旧有效.

(3)实验表明HTID能够针对已知对抗攻击制作出较高性能的白盒免疫样本.与基线相比,MID在保证白盒性能的情况下能够有效地提升免疫样本防御未知攻击的迁移性.

2　相关工作

我们分别在章节2.1和章节2.2中介绍了几种流行的对抗攻击和对抗防御方法.

2.1　对抗攻击

对抗攻击指通过对图像进行人眼难以察觉的扰动以生成对抗样本并使得模型产生错误的输出.接下来,我们将介绍几种流行的对抗攻击.

快速梯度符号法(Fast Gradient Sign Method,FGSM)[2]是一种单步(One-step)的基于梯度的攻击,它沿着原始图像的梯度符号方向生成对抗扰动并添加到原始图像中.FGSM的攻击速度快,但攻击能力较弱.基本迭代法(Basic Iterative Method,BIM)[14]是一种迭代的基于梯度的攻击,它多次沿着当前样本的梯度符号方向添加对抗扰动,并在每次迭代后通过裁剪像素值使对抗扰动满足盒约束(Box Constraints).当迭代次数为1时,BIM退化为FGSM.与FGSM相比,BIM能够获得攻击能力更强和视觉质量更好的对抗样本,但计算成本也更高.投影梯度下降(Projected Gradient Descent,PGD)[16]是BIM的改进版本.迭代开始前,PGD在扰动范围内对原始图像进行随机扰动旨在获得多样化的对抗样本,这使得它成为一种通用的一阶对手(Universal First-order Adversary)[16].DNNs使用PGD生成的对抗样本进行对抗训练能够有效地提升DNNs对所有一阶攻击的鲁棒性.

由Carlini和Wagner名字命名的CW攻击[18]是一种基于优化的攻击,它不仅要求生成的对抗样本以κ的置信度被预测错误,而且还解决了对抗扰动的盒约束问题.CW攻击生成的对抗样本有着强大的攻击能力和较高的视觉质量,但其计算成本非常高,需要执行多次攻击以查询出最优的参数c.

对抗性生成对抗网络(Adversarial Generative Adversarial Networks,AdvGAN)[21],它将生成对抗网络(Generative Adversarial Networks,GAN)[24]和分类器结合并训练出一个生成对抗样本的生成器.与基于梯度和基于优化的攻击相比,AdvGAN能够快速地生成强大且视觉质量较高的对抗样本.可恢复生成对抗网络(Recoverable Generative Adversarial Networks,RGAN)[23]在AdvGAN的基础上增加了恢复器(Recover)和降维器(Dimension Reducer),恢复器可以将对抗样本恢复至原始图像,降维器旨在降低对抗扰动的复杂性,从而提升恢复器的恢复能力.通过生成自恢复对抗样本(Self-recoverable Adversarial Examples,SRAEs),RGAN可以成为社交网络中隐私安全的保护机制.

2.2　对抗防御

对抗训练是对抗防御中常见的方法之一,其基本思想是在训练模型的过程中向训练集加入对抗样本,使得模型能够对对抗样本具有一定的鲁棒性.Goodfellow等人[2]首次引入对抗训练的概念,并发现使用基于FGSM的对抗性目标函数(Adversarial Objective Function)进行对抗训练能够有效地提升模型对对抗攻击的鲁棒性.该工作将对抗训练解释为一个最小-最大化问题,即当数据受到对抗攻击时,最小化最坏情况下的训练误差.Huang等人[27]提出了三种对抗攻击方法,即Adv_Alpha、Adv_Loss和Adv_Loss_Sign,并使用这三种对抗攻击对模型进行对抗训练.该方法被称为和对手一起学习(Learning with an Adversary,LWA).实验结果表明,与Goodfellow等人[2]所提出的方法相比,模型使用LWA方法训练获得了更高的鲁棒性.Kurakin等人[28]为了让对抗训练能够有效地应用于大尺寸的数据集,将对抗训练的每一个小批次(Mini-batch)以一定比例地划分为了干净图像和对抗样本.尽管上述工作对单步攻击表现出来的很强的鲁棒性,但Tramèr等人[29]表示经过对抗训练的模型仍然会受到简单的黑盒攻击.为了解决这一问题,他们提出了集成对抗训练(Ensemble Adversarial Training),集成对抗训练在训练集中加入了针对其他模型制作的对抗样本以提升模型对基于迁移的攻击(Transfer-based Attacks)的鲁棒性.此外,Madry等人[16]提出PGD并将其应用于对抗训练.PGD是一种通用的一阶对手,在扰动范围内能够生成多样化的对抗样本,从而使得基于PGD进行对抗训练的模型能够同时对多种攻击鲁棒.

常见的对抗防御还包括防御蒸馏[30]和输入预处理[31-34,43-44].与对抗训练相比,防御蒸馏需要训练两个模型,所以需要更长的训练时间和更多的计算资

源，而输入预处理会影响干净图像的分类精度并且防御效果有限. 因此，对抗训练仍是目前最常见和最有效的对抗防御方法.

然而，输入预处理与其他对抗性防御方法有着很好的兼容性. Guo等人[31]研究了位深度减少、JPEG压缩、总方差最小化和图像拼接防御对抗样本的有效性，他们发现总方差最小化和图像拼接是非常有效的防御方法. Liu等人[32]研究了一种精心设计的量化表以增强JPEG压缩对对抗样本的防御性能. Xie等人[33]提出了一种将输入图像进行随机调整大小和随机填充的方法以减轻对抗扰动的影响，该方法对于单步攻击和迭代攻击均有效，并且无需额外的训练或微调. Gupta等人[43]所提出的防御机制结合了类相关的图像修复方法和基于小波变换的图像去噪方法，其仅恢复对指定类别有用的小部分区域就能够抑制对抗样本. Mustafa等人[44]提出了一种计算高效的图像超分辨率恢复作为对抗防御方法，并且表明该方法能够将对抗性样本恢复到自然图像流形上，从而恢复到正确的类别.

3 免疫样本

章节3.1给出了问题定义，包括威胁模型、免疫样本的定义、白盒和黑盒免疫防御的概念等. 章节3.2提出了HTID. 章节3.3提出了MID. 章节3.4提出了免疫率以衡量免疫样本防御对抗攻击的性能.

3.1 问题定义

威胁模型. 对抗样本可以使得深度神经网络产生错误的预测(即无目标攻击)，甚至是指定错误的类别(即有目标攻击)，这使得添加了对抗扰动的图像不再安全并随时可能威胁到深度神经网络. 例如，MitM攻击有能力拦截自动驾驶汽车发送至DNN分类器的路牌图像数据，并将其修改为对抗样本，这种对抗样本很难被人眼感知却能够使得DNN分类错误[54]，从而对自动驾驶汽车的安全性构成了巨大的威胁，如图1(a)所示. 在这一过程中，攻击者可以根据对DNN的了解程度，采用白盒或黑盒攻击，而现有对抗防御无法阻止攻击者通过MitM攻击将路牌图像篡改为对抗样本. 因此，如图1(c)所示，我们可以通过在原始图像(如原始的路牌图像)上添加免疫扰动将图像制作成免疫样本，其上的免疫扰动能够使得对抗攻击无法针对图像生成有效的对抗样本，从而同时保障了图像的安全性和鲁棒性，并减轻潜在攻击对DNNs的威胁. 更多场景(如对抗补丁和数据集发布)下的威胁模型我们将在章节5中讨论.

首先，我们如下定义免疫样本.

定义1. 免疫样本. 设原始图像为$\boldsymbol{x}\in[0,1]^n$，其真实标签为$y\in\{1,\cdots,l\}$，深度神经网络$f:[0,1]^n\to\{1,\cdots,l\}$接收图像并返回图像的预测标签，对抗攻击算法$g:\boldsymbol{x}\mapsto g(\boldsymbol{x})$接收图像并返回对应的对抗样本$g(\boldsymbol{x})$，$g(\boldsymbol{x})$满足$f\big(g(\boldsymbol{x})\big)\neq y$或$f\big(g(\boldsymbol{x})\big)=y^*$，其中，$y^*$表示目标标签. 那么，满足

$$\begin{cases} f(\boldsymbol{x}^{IE})=y, \\ f\big(g(\boldsymbol{x}^{IE})\big)=y, \\ \text{s.t.}\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau \end{cases} \tag{1}$$

的样本$\boldsymbol{x}^{IE}$被称为$\boldsymbol{x}$的免疫样本. 其中，$\|\cdot\|_p$表示L_p范数($p=0,1,\cdots,\infty$)，$\boldsymbol{x}^{IE}-\boldsymbol{x}$被称为免疫扰动，$\tau$表示最大免疫扰动强度，其用于控制免疫扰动的不可见性.

基于上述免疫样本的定义，免疫样本在保证一定视觉质量的前提下，不仅可以被正确分类而且可以抵抗对抗攻击. 对防御者而言，对抗攻击算法g可能是已知的也可能是未知的，针对已知(或未知)对抗攻击制作免疫样本的过程被称为白盒(或黑盒)免疫防御，相应的免疫样本被称为白盒(或黑盒)免疫样本.

实际上，当防御者针对某一特定对抗攻击制作出的白盒免疫样本被攻击者获取后，攻击者发现无法使用该攻击制作出有效的对抗样本，其仍可能尝试使用其他对抗攻击制作对抗样本. 在这种情况下，攻击者尝试的其他对抗攻击对防御者而言是完全未知的. 为了确保免疫样本的有效性，免疫样本需要具有一定的可迁移性，从而有能力应对未知的对抗攻击. 因此，在章节3.2中，我们提出了HTID以制作白盒免疫样本；在章节3.3中，我们提出了MID以提升免疫样本在多种对抗攻击之间的可迁移性.

3.2 双曲正切免疫防御

设$P(y|\boldsymbol{x})$表示$\boldsymbol{x}$被f预测为类别y的概率，则当$P(y|\boldsymbol{x})>\max\limits_{t\neq y}P(t|\boldsymbol{x})$时，$\boldsymbol{x}$被正确分类，否则被错误分类，其中，$t\in\{1,\cdots,l\}\setminus\{y\}$，$\setminus$表示差集符号. 因此，我们定义损失函数$\mathcal{J}(\boldsymbol{x},y)$，即[18]

$$\mathcal{J}(\boldsymbol{x},y)=\max_{t\neq y}P(t|\boldsymbol{x})-P(y|\boldsymbol{x}) \tag{2}$$

则公式(1)可以被改写为

$$\begin{cases}\mathcal{J}(\boldsymbol{x}^{IE},y)<0,\\ \mathcal{J}(g(\boldsymbol{x}^{IE}),y)<0,\end{cases}\text{s.t.}\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau \tag{3}$$

在白盒免疫防御中，对抗攻击g的全部信息(如扰动范数、最大扰动强度和攻击算法)对防御者而言是已知的. 此外，我们假设防御者也具有目标模型的全部信息(如结构、参数和训练方式). 该假设基于制作免疫样本的防御者和模型发布者有着共同对手的事实. 在这一假设下，防御者可以将多个模型(如路牌识别模型)集成并为该集成模型将图像(如路牌图像)制作成免疫样本，从而不仅能够防止图像被制作成对抗样本，还让集成的每一个子模型的所处环境更加安全(在实验中，我们仅考虑单个模型而非集成模型). 根据这一假设、定义1和公式(3)，我们可以通过同时最小化$\mathcal{J}(\boldsymbol{x}^{IE},y)$、$\mathcal{J}(g(\boldsymbol{x}^{IE}),y)$和$\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p$以获取具有微小扰动的免疫样本，具体如下：

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\lambda_{ie}\cdot\mathcal{J}(\boldsymbol{x}^{IE},y)+\lambda_{adv}\cdot\mathcal{J}(g(\boldsymbol{x}^{IE}),y)+\\&\lambda_{pert}\cdot\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\end{aligned} \tag{4}$$

其中，λ_{ie}表示$\mathcal{J}(\boldsymbol{x}^{IE},y)$的权重，$\lambda_{adv}$表示$\mathcal{J}(g(\boldsymbol{x}^{IE}),y)$的权重，$\lambda_{pert}$表示$\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p$的权重.

然而，免疫样本依然存在盒约束(Box Constraints)，即$\boldsymbol{x}^{IE}\in[0,1]^n$. 尽管截断操作能够轻松地让免疫样本满足盒约束，但它会使得免疫样本陷入平坦区域，从而导致被截断维度的偏导数为0[18]. 针对这一问题，我们受CW攻击[18]的启发，将免疫样本表示为

$$\boldsymbol{x}^{IE}=\frac{1}{2}\left(\tanh(\boldsymbol{w})+1\right)\in(0,1)^n \tag{5}$$

其中，$\boldsymbol{w}\in\mathbb{R}^n$. 由于$\boldsymbol{w}$不再具有盒约束，并且能够保证$\boldsymbol{x}^{IE}\in[0,1]^n$，因此我们通过优化$\boldsymbol{w}$间接优化出$\boldsymbol{x}^{IE}$，从而自然地解决盒约束问题[1,18]. 具体如下：

$$\begin{aligned}\mathcal{L}(\boldsymbol{w},y)=&\lambda_{ie}\cdot\mathcal{J}\left(\frac{1}{2}\left(\tanh(\boldsymbol{w})+1\right),y\right)\\&+\lambda_{adv}\cdot\mathcal{J}\left(g\left(\frac{1}{2}\left(\tanh(\boldsymbol{w})+1\right)\right),y\right)\\&+\lambda_{pert}\cdot\left\|\frac{1}{2}\left(\tanh(\boldsymbol{w})+1\right)-\boldsymbol{x}\right\|_p\end{aligned} \tag{6}$$

$$\arg\min_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w},y) \tag{7}$$

我们将该方法称为双曲正切免疫防御(Hyperbolic Tangent Immune Defense，HTID)，其对应的算法流程如算法1所示.

算法1.　双曲正切免疫防御(HTID).

输入：真实标签为y的原始图像$\boldsymbol{x}$，深度神经网络f，目标对抗攻击算法g，分类损失$\mathcal{J}$(即公式(2))，优化器$\mathcal{G}$，迭代次数T

输出：免疫样本$\boldsymbol{x}^{IE}$

初始化$t\leftarrow1$，$\boldsymbol{w}\leftarrow\text{arctanh}(2\boldsymbol{x}-1)$；

WHILE $t\leqslant T$

　　通过公式(6)计算$\mathcal{L}(\boldsymbol{w},y)$；

　　根据公式(7)通过优化器$\mathcal{G}$更新$\boldsymbol{w}$；

　　$t\leftarrow t+1$；

WEND

RETURN $\boldsymbol{x}^{IE}\leftarrow\frac{1}{2}\left(\tanh(\boldsymbol{w})+1\right)$

3.3　基于矩的免疫防御

在黑盒免疫防御中，防御者事先并不知道攻击者会使用何种对抗攻击. 在这种情况下，所有被制作出的免疫样本需具备一定的可迁移性，从而有能力防御未知的对抗攻击. 因此，我们提出了矩优化以提升免疫样本的可迁移性.

免疫样本的可迁移性可以被描述为针对已知对抗攻击g制作的白盒免疫样本$\boldsymbol{x}^{IE}$仍可以使其他未知的对抗攻击g'失效，其中对抗样本$g(\boldsymbol{x}^{IE})$和$g'(\boldsymbol{x}^{IE})$满足(证明过程见附录A的定理1)

$$\left\|g(\boldsymbol{x}^{IE})-g'(\boldsymbol{x}^{IE})\right\|_{p'}\leqslant\rho \tag{8}$$

因此，未知的对抗攻击g'针对免疫样本$\boldsymbol{x}^{IE}$制作的对抗样本$g'(\boldsymbol{x}^{IE})$可以被表示为

$$g'(\boldsymbol{x}^{IE})=g(\boldsymbol{x}^{IE})+\boldsymbol{\delta} \tag{9}$$

其中，$\left\|\boldsymbol{\delta}\right\|_{p'}\leqslant\rho$. 免疫样本$\boldsymbol{x}^{IE}$要使任意对抗攻击$g'$失效，就要使得对于任意满足$\left\|\boldsymbol{\delta}\right\|_{p'}\leqslant\rho$的$\boldsymbol{\delta}$都有$\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)<0$. 因此，免疫样本$\boldsymbol{x}^{IE}$从已知对抗攻击$g$迁移至任意未知攻击$g'$的充要条件是

条件1.　$P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant0)=0$，其中，$P$表示事件发生的概率，$\mathcal{D}$表示$\boldsymbol{\delta}$所服从分布的概率密度函数：

$$\mathcal{D}(\boldsymbol{\delta})=\begin{cases}\dfrac{1}{V}, & \text{如果}\left\|\boldsymbol{\delta}\right\|_{p'}\leqslant\rho\\ 0, & \text{否则}\end{cases} \tag{10}$$

其中，V表示$|\boldsymbol{\delta}|_{p'}\leqslant\rho$的体积.

直接寻找满足条件1的免疫样本$\boldsymbol{x}^{IE}$相对困难，为了简化寻找难度，我们在此引入条件1的必要不充分条件(证明过程见附录A的定理2)：

条件2.　$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})<0$，其中，$\mathcal{D}$如公式(10)所示，$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE}):=\mathbb{E}_{\delta\sim\mathcal{D}}[\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)]$，$\mathbb{E}$表示

数学期望.

换言之,若免疫样本$\boldsymbol{x}^{IE}$满足条件1,就必须满足条件2. 因此,我们将直接寻找满足条件1的免疫样本$\boldsymbol{x}^{IE}$的问题转化为一个两阶段的问题,即先寻找满足条件2的免疫样本$\boldsymbol{x}^{IE}$(第一阶段),然后再寻找满足条件1的免疫样本$\boldsymbol{x}^{IE}$(第二阶段). 为此,我们首先通过最小化$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$直至寻找到满足$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})<0$的免疫样本$\boldsymbol{x}^{IE}$,即

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE}),\\&\text{s.t.}\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau.\end{aligned}\tag{11}$$

当通过优化公式(11)寻找到满足条件2的免疫样本$\boldsymbol{x}^{IE}$时,我们再去解决以下问题以寻找满足条件1的免疫样本$\boldsymbol{x}^{IE}$:

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\mathbb{P}_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0),\\&\text{s.t.}\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau.\end{aligned}\tag{12}$$

此时,根据切比雪夫不等式[39],有

$$\begin{aligned}&\mathbb{P}_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)\leqslant\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}\right]^2,\\&\text{s.t.}\,\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})<0,\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau.\end{aligned}\tag{13}$$

其中,$\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})=\sqrt{\mathbb{V}_{\delta\sim\mathcal{D}}[\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)]}$,$\mathbb{V}$表示方差,具体证明过程详见附录A的推论1. 因此,公式(12)可以被改写为(具体证明过程详见附录A的定理3)

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}\right]^2,\\&\text{s.t.}\,\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})<0,\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau.\end{aligned}\tag{14}$$

为了将公式(14)中分数结构的目标函数转化为简单的线性结构,我们同时最小化$\sigma^2_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$和$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$. 因此,公式(14)可以进一步改写为

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\lambda_\sigma\cdot\sigma^2_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})+\lambda_\mu\cdot\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE}),\\&\text{s.t.}\,\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})<0,\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau.\end{aligned}\tag{15}$$

其中,λ_σ为$\sigma^2_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$的权重,$\lambda_\mu$为$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$的权重.

由于公式(11)和公式(15)均包含最小化$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$,我们将两阶段的优化问题统一为如下公式:

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\lambda_\sigma\cdot\sigma^2_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})+\lambda_\mu\cdot\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE}),\\&\text{s.t.}\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_p\leqslant\tau.\end{aligned}\tag{16}$$

当$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})\geqslant 0$时,$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$作为惩罚项以约束自身优化至小于0的值. 当$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})<0$时,$\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})$帮助最小化$\mathbb{P}_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)$.

在提升免疫样本可迁移性的同时,还需保证免疫样本的分类性能和扰动的不可见性. 因此,我们将免疫样本的分类损失$\mathcal{J}(\boldsymbol{x}^{IE},y)$和免疫扰动范数加入公式(16)中,并使用公式(5)转换参数空间得到最终的优化问题:

$$\begin{aligned}\mathcal{L}'(\boldsymbol{w},y)=&\lambda_{ie}\cdot\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)\\&+\lambda_{pert}\cdot\left\|\frac{1}{2}(\tanh(\boldsymbol{w})+1)-\boldsymbol{x}\right\|_p\\&+\lambda_\sigma\cdot\sigma^2_{\delta\sim\mathcal{D}}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)\\&+\lambda_\mu\cdot\mu_{\delta\sim\mathcal{D}}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right),\end{aligned}\tag{17}$$

$$\arg\min_{\boldsymbol{w}}\mathcal{L}'(\boldsymbol{w},y).\tag{18}$$

综上所述,通过解决公式(18)以求得可转移的免疫样本的方法被称为基于矩的免疫防御(Moment-based Immune Defense,MID). 对于MID能够提升免疫样本可迁移性的另一种解释是,如图2所示,MID最小化在$g(\boldsymbol{x}^{IE})$的$\boldsymbol{\delta}$邻域内的损失

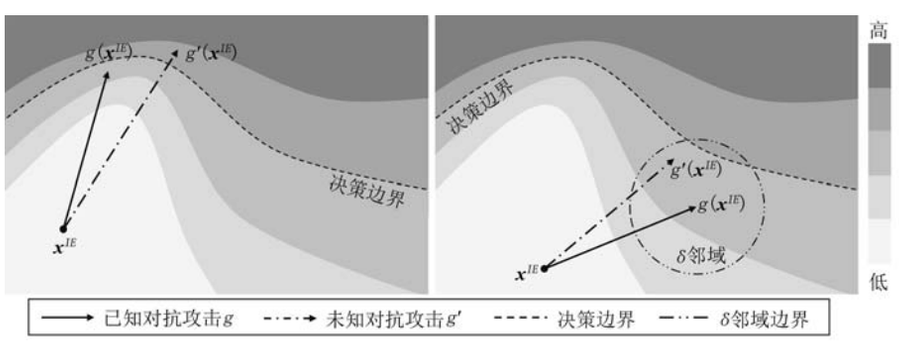

图2 HTID和MID制作的免疫样本防御对抗攻击的示意图(左图为HTID的示意图,右图为MID的示意图.$\boldsymbol{x}^{IE}$表示免疫样本,$g(\boldsymbol{x}^{IE})$表示已知的对抗样本,$g'(\boldsymbol{x}^{IE})$表示潜在的对抗样本. 图中颜色越深表示分类损失越大)

的期望和方差，使得生成的对抗样本$g(\boldsymbol{x}^{IE})$被优化至一个平坦的区域．与HTID相比，MID能够降低潜在对抗样本$g'(\boldsymbol{x}^{IE})$的损失，从而提升免疫样本的可迁移性．

由于$\mathcal{D}$连续，在实践中公式(17)包含的期望无法直接求解，所以我们使用蒙特卡罗方法(Monte Carlo Method)[40]求解期望的近似值．具体方案如下：首先，我们在$\|\boldsymbol{\delta}\|_{p'}\leqslant\rho$内随机采样出$N$个样本$\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_N$；其次，计算期望的近似，即

$$\mathbb{E}_{\delta\sim\mathcal{D}}\left[\mathcal{J}\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)+\boldsymbol{\delta},y\right)\right]\approx\frac{1}{N}\sum_{i=1}^{N}\mathcal{J}\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)+\boldsymbol{\delta}_i,y\right):=\hat{\mathbb{E}}\tag{19}$$

$$\mathbb{E}_{\delta\sim\mathcal{D}}\left[\mathcal{J}^2\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)+\boldsymbol{\delta},y\right)\right]\approx\frac{1}{N}\sum_{i=1}^{N}\mathcal{J}^2\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)+\boldsymbol{\delta}_i,y\right):=\hat{\mathbb{E}}^2\tag{20}$$

最后，我们使用公式(19)和(20)计算出$\sigma^2_{\delta\sim\mathcal{D}}$和$\mu_{\delta\sim\mathcal{D}}$的近似值，即

$$\sigma^2_{\delta\sim\mathcal{D}}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)\approx\hat{\mathbb{E}}^2-(\hat{\mathbb{E}})^2\tag{21}$$

$$\mu_{\delta\sim\mathcal{D}}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)\approx\hat{\mathbb{E}}\tag{22}$$

MID的算法流程被总结在算法2中．

算法2.　基于矩的免疫防御(MID)．

输入：真实标签为y的原始图像$\boldsymbol{x}$，深度神经网络f，目标对抗攻击算法g，分类损失$\mathcal{J}$(即公式(2))，优化器$\mathcal{G}$，迭代次数T

输出：免疫样本$\boldsymbol{x}^{IE}$

初始化$t\leftarrow1$，$\boldsymbol{w}\leftarrow\operatorname{arctanh}(2\boldsymbol{x}-1)$；

WHILE $t\leqslant T$

　　通过公式(21)和(22)计算出$\sigma^2_{\delta\sim\mathcal{D}}$和$\mu_{\delta\sim\mathcal{D}}$；

　　将$\sigma^2_{\delta\sim\mathcal{D}}$和$\mu_{\delta\sim\mathcal{D}}$带入公式(17)得到$\mathcal{L}'(\boldsymbol{w},y)$；

　　根据公式(18)通过优化器$\mathcal{G}$更新$\boldsymbol{w}$；

　　$t\leftarrow t+1$；

END WHILE

$\boldsymbol{x}^{IE}\leftarrow\frac{1}{2}(\tanh(\boldsymbol{w})+1)$；

RETURN $\boldsymbol{x}^{IE}$

3.4　免疫率

仅使用对抗样本的攻击成功率(Attack Success Rate，ASR)[13]在免疫防御前后的变化量衡量免疫样本的防御性能是不充分的，因为攻击成功率的变化量受免疫防御前的攻击成功率的限制．例如，当免疫防御前的攻击成功率本身就不高时，攻击成功率的变化量再多也不会超过该值．因此，我们提出免疫率准确衡量免疫样本的防御性能．

令所有带真实标签的原始图像组成的集合为$S=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，其中，元素$(\boldsymbol{x}_m,y_m)$表示真实标签为$y_m$的图像$\boldsymbol{x}_m$．设$f:\boldsymbol{x}\mapsto y$为分类器，其输入为图像，输出为预测标签．假设对抗攻击算法$g:\boldsymbol{x}\mapsto\boldsymbol{x}^{adv}$在$S$上制作了对抗样本并攻击$f$，$g$的输入为图像，输出为对应的对抗样本．防御者为保护图像和模型，在S上使用免疫防御算法$h:\boldsymbol{x}\mapsto\boldsymbol{x}^{IE}$为$f$制作免疫样本以防御对抗攻击$g$，其中，$h$的输入为图像，输出为对应的免疫样本．那么，免疫率可以被表示如下：

$$\text{免疫率}=\frac{ASR_1-ASR_2}{ASR_1}\times100\%=\left(1-\frac{\sum_{i=1}^{m}\mathbb{I}\left[f(g(\boldsymbol{x}_i))\neq y_i\right]}{\sum_{i=1}^{m}\mathbb{I}\left[f(g(h(\boldsymbol{x}_i)))\neq y_i\right]}\right)\times100\%\tag{23}$$

其中，$ASR_1=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left[f(g(\boldsymbol{x}_i))\neq y_i\right]$表示对抗攻击算法$g$针对原始图像制作对抗样本并攻击分类器$f$的攻击成功率，$ASR_2=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left[f(g(h(\boldsymbol{x}_i)))\neq y_i\right]$表示对抗攻击算法$g$针对免疫样本制作对抗样本并攻击分类器$f$的攻击成功率，$\mathbb{I}[\cdot]$表示指示函数，其具体定义为

$$\mathbb{I}[c]=\begin{cases}1, & \text{如果}c\text{为真}\\ 0, & \text{否则}\end{cases}\tag{24}$$

免疫率反映了“在免疫防御后，攻击成功率下降的数值”占“在免疫防御前，攻击成功率的原始数值”的比例．当免疫率为0时，说明免疫防御前后，攻击成功率没有变化；当免疫率为100%时，说明免疫防御后，攻击成功率降为0．免疫率的数值越大，免疫样本防御对抗攻击的性能越好．值得注意的是，免疫率的范围是$(-\infty,100\%]$，免疫率为负值说明免疫防御后，攻击成功率不降反升，此时的免疫样本是无效的．

根据公式(23)，我们可以通过免疫率实现由ASR_1到ASR_2换算：

$$ASR_2=(100\%-\text{免疫率})\times ASR_1\tag{25}$$

在章节4中，我们仅用免疫率展示免疫样本的

防御性能，而相应的ASR_2也可以根据公式(25)直接求得.

4 实验与分析

4.1 实验设置

(1)对抗攻击. 我们选择了12种对抗攻击算法作为免疫样本的防御目标，即AdvGAN[21]、RGAN[23]、FGSM[2]、BIM[14]、PGD[16]、CW[18]、SparseFool(SF)[45]、PGD_0[46]、TSAA[47]、Square[48]、SimBA[49]和SimBA-DCT[49]. 它们包含了基于生成的、基于梯度的、基于优化的、稀疏的白盒攻击和基于查询的黑盒攻击. 此外，我们假设攻击者针对目标模型实施无目标的白盒攻击. 因为，在这种情况下，攻击者可以轻易地制作强大的对抗样本，免疫样本必须在这种最坏的情况下仍然保证有效.

(2)数据集. 我们选择CIFAR-10[41]、MNIST[8]、STL-10[50]和Caltech-256[51]作为我们实验所用的数据集.

(3)分类器. 针对CIFAR-10和MNIST数据集，我们分别训练了两个分类器，即Inception-v3[12]和LeNet-5[8]，它们在对应数据集上的分类准确率分别为92.4%和99.2%. 针对STL-10和Caltech-256数据集，我们分别训练了两个ResNet-50[9]，它们对应的分类准确率为82.2%和84.2%.

(4)实验方案. 在白盒免疫防御中，我们使用HTID分别针对不同的对抗攻击为分类器制作白盒免疫样本，并测试它们的性能. 在黑盒免疫防御中，我们使用所提出的MID针对先进且生成速度快的AdvGAN制作免疫样本并将免疫样本迁移至其他对抗攻击. 由于之前的工作没有针对免疫样本及其可迁移性的研究，我们将HTID作为基线以验证MID在提升免疫样本可迁移性方面的有效性. 此外，我们还分别对HTID和MID的参数进行了消融实验以获取最佳的参数值.

(5)指标. 我们分别使用分类准确率和免疫率衡量免疫样本的分类性能和防御性能，并使用峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)[52]和结构相似性指数度量(Structural Similarity Index Measure，SSIM)[42]衡量免疫样本的视觉质量.

(6)参数设置. 所选对抗攻击在四种数据集上的参数设置和针对所选分类器的攻击成功率如附录B.1中的表13所示. 所提出的HTID和MID在CIFAR-10、MNIST、STL-10和Caltech-256上的参数设置如表1所示.

表1 HTID和MID在CIFAR-10和MNIST上的参数设置. 符号“×”表示不适用

数据集	防御	优化器	迭代次数	学习率	λ_{pert}	λ_{ie}	λ_{adv}	λ_{σ}	λ_{μ}	p	p'	ρ	N
CIFAR-10 STL-10	HTID	Adam	500	10^{-2}	0.0	10.0	1.0	×	×	2	2	×	×
Caltech-256	MID	Adam	500	10^{-2}	0.0	10^4	×	1.0	1.0	2	2	2.0	5
MNIST	HTID	Adam	500	10^{-2}	0.0	10.0	1.0	×	×	2	2	×	×
	MID	Adam	500	10^{-2}	0.0	10^4	×	1.0	1.0	2	2	2.0	5

4.2 双曲正切免疫防御的性能测试

我们使用在章节3.2中提出的HTID为分类器制作白盒免疫样本，并评估防御对抗攻击的性能. 实验设置与章节4.1中的保持一致. 此外，为了验证免疫样本不会影响其他分类器的性能，我们还分别为CIFAR-10和MNIST训练了分类准确率为95.6%的ResNet-50[9]和99.1%的Model C[21]以测试了免疫样本的分类准确率. 实验结果如表2所示.

从表2中可以看出，HTID针对不同对抗攻击制作的免疫样本在不同分类器上的准确率均达到了100%，与原始准确率相比，平均提升了3.4%. 此外，免疫样本在CIFAR-10上的免疫率也均达到了100.0%，平均PSNR和SSIM分别为32.7和0.863，最低PSNR和SSIM为针对RGAN时的26.2和0.806，最高SSIM为针对CW时的38.9和0.925. 在MNIST上的平均免疫率为69.5%，最低免疫率为针对BIM时的41.5%，最高免疫率为针对RGAN时的99.5%，平均PSNR和SSIM分别为22.4和0.924，最低PSNR和SSIM为针对AdvGAN时的20.5和0.900，最高PSNR和SSIM为针对CW时的24.7和0.956. HTID制作的免疫样本视觉效果图如附录C中的图10所示.

4.3 基于矩的免疫防御的可迁移性测试

我们对比了MID和不同参数设置下的HTID制作的免疫样本的可迁移性. 具体地，我们将

表2　HTID在CIFAR-10和MNIST上的分类准确率(%)、免疫率(%)、PSNR(dB)和SSIM. Avg. 表示一行数据的平均值

数据集	指标	AdvGAN	RGAN	FGSM	BIM	PGD	CW	Avg.
CIFAR-10	准确率(Inception-v3)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	准确率(ResNet-50)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	免疫率	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	PSNR	28.4	26.2	34.0	34.8	34.1	38.9	32.7
	SSIM	0.822	0.806	0.874	0.879	0.874	0.925	0.863
MNIST	准确率(LeNet-5)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	准确率(Model C)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	免疫率	84.4	99.5	75.8	41.5	55.7	60.1	69.5
	PSNR	20.5	21.0	23.6	23.2	21.5	24.7	22.4
	SSIM	0.900	0.908	0.936	0.930	0.912	0.956	0.924

HTID的参数设置得和MID的参数一致并将该参数设置下的HTID描述为$\mathrm{HTID}_{\mathrm{SAM}}$，而将原始参数设置下的HTID描述为$\mathrm{HTID}_{\mathrm{ORI}}$. 我们分别使用$\mathrm{HTID}_{\mathrm{ORI}}$、$\mathrm{HTID}_{\mathrm{SAM}}$和MID针对AdvGAN制作免疫样本，并测试免疫样本防御其他11种对抗攻击的免疫率，实验结果如表3所示.

表3　$\mathrm{HTID}_{\mathrm{ORI}}$、$\mathrm{HTID}_{\mathrm{SAM}}$和MID针对AdvGAN制作的免疫样本的分类准确率(%)、免疫率(%)、PSNR(dB)和SSIM，以及将免疫样本迁移至其他11种对抗攻击的免疫率(%)

数据集	防御	准确率	PSNR	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	SF	PGD_0	TSAA	Square	SimBA	SimBA-DCT	Avg.
CIFAR-10	$\mathrm{HTID}_{\mathrm{ORI}}$	**100.0**	28.4	0.822	**100.0**	50.4	77.5	43.1	57.7	23.3	75.5	77.4	22.4	10.3	12.5	15.1	47.1
	$\mathrm{HTID}_{\mathrm{SAM}}$	**100.0**	27.2	0.811	**100.0**	38.3	80.3	57.9	59.4	51.4	87.6	85.6	47.0	24.1	27.3	29.2	57.3
	MID	**100.0**	**29.9**	**0.836**	95.2	**57.2**	**84.0**	**60.7**	**62.6**	**58.5**	**88.9**	**87.1**	**52.3**	**30.0**	**31.6**	**36.5**	**62.1**
MNIST	$\mathrm{HTID}_{\mathrm{ORI}}$	**100.0**	**20.5**	**0.900**	84.4	71.3	52.5	29.5	22.3	30.7	35.7	53.4	47.8	19.6	25.7	22.3	41.3
	$\mathrm{HTID}_{\mathrm{SAM}}$	**100.0**	20.2	0.897	78.4	67.2	56.9	44.1	26.1	61.0	48.8	63.7	52.7	22.3	29.2	25.3	47.9
	MID	**100.0**	**20.5**	**0.900**	**85.9**	**74.4**	**58.7**	**45.6**	**27.1**	**65.8**	**50.0**	**69.4**	**56.7**	**28.4**	**32.0**	**30.7**	**52.1**
STL-10	$\mathrm{HTID}_{\mathrm{ORI}}$	**100.0**	30.6	0.847	**100.0**	31.0	48.6	29.8	31.5	19.7	16.5	55.3	54.1	23.7	28.8	33.0	39.3
	$\mathrm{HTID}_{\mathrm{SAM}}$	**100.0**	29.5	0.833	98.6	35.1	65.2	32.8	36.0	26.1	17.2	69.2	60.6	30.9	38.1	48.7	46.5
	MID	**100.0**	**32.4**	**0.873**	99.2	**35.4**	**87.5**	**35.1**	**42.1**	**32.4**	**23.6**	**87.5**	**80.0**	**51.1**	**44.3**	**63.7**	**56.8**
Caltech-256	$\mathrm{HTID}_{\mathrm{ORI}}$	**100.0**	34.0	0.863	**62.7**	21.7	35.1	22.7	23.1	32.1	24.0	53.3	46.1	17.1	21.2	36.7	33.0
	$\mathrm{HTID}_{\mathrm{SAM}}$	**100.0**	33.8	0.861	48.5	22.6	52.5	26.8	25.8	42.9	28.4	71.8	64.4	19.3	26.8	39.0	39.1
	MID	**100.0**	**34.4**	**0.867**	55.7	**24.5**	**77.0**	**31.3**	**32.5**	**54.3**	**33.5**	**90.5**	**74.9**	**32.4**	**35.4**	**42.5**	**48.7**

注：Avg. 表示免疫率的平均值. 加粗字体表示最佳数

从表3中可以看出，3种免疫防御制作出的免疫样本的准确率均达到了100%，与原始准确率相比，平均提升了10.5%. 由MID制作的免疫样本获得了最高的PSNR和SSIM，分别为29.9和0.836、20.5和0.900、32.4和0.873、34.4和0.867，平均比基线高出1.275和0.015，在MNIST和CIFAR-10上使用MID制作的免疫样本视觉效果图如附录C中的图10所示，在STL-10和Caltech-256数据集上的免疫样本如附录C中的图11所示. 此外，MID能够更好地防御其他11种对抗攻击，针对12种对抗攻击的平均免疫率比基线高出4.2%～18.3%.

4.4　消融实验

我们将在本章节使用基于坐标下降的调参方法对MID的参数进行了消融实验以快速获得较佳的参数值，附录B.2对该调参方法进行了额外的说明，针对HTID的消融实验放置在附录B.3中. 此外，我们还探索了损失函数$\mathcal{J}$的影响.

4.4.1　MID的参数消融实验

我们预先设置MID的优化器为Adam，迭代次数为500，学习率为10^{-2}，$p=2$. 我们针对AdvGAN使用MID制作免疫样本并将免疫样本迁移至RGAN、FGSM、BIM、PGD和CW. 在本小节中，我们对MID的参数λ_{pert}、λ_{ie}、λ_{σ}、λ_{μ}、ρ和N进行消融.

(1) 参数λ_{pert}. 参数λ_{pert}对应了损失$\left\|\frac{1}{2}(\tanh(\boldsymbol{w})+1)-\boldsymbol{x}\right\|_p$在$\mathcal{L}'(\boldsymbol{w},y)$中的重要性，主要影响着免疫样本的视觉质量. 我们预先设置$\lambda_{ie}=$

$\lambda_\sigma=\lambda_\mu=1.0$、$\rho=2.0$、$N=5$，并分别测试$\lambda_{pert}$为0.0、1.0和10.0时，MID的性能．实验结果如表4所示．实验结果表明，随着λ_{pert}值的增加，免疫样本的准确率和免疫率呈现下降或不变的趋势，而SSIM呈现上升趋势．我们取$\lambda_{pert}=0.0$以获得最佳的免疫率．此时，在两个数据集上，免疫样本的准确率均为100.0%，SSIM分别为0.827和0.896，防御6种对抗攻击的平均免疫率分别为43.4%和48.9%．

表4 MID在不同的λ_{pert}值下针对AdvGAN制作的免疫样本的分类准确率(%)、免疫率(%)和SSIM，以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

数据集	λ_{pert}	准确率	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	Avg.
CIFAR-10	0.0	**100.0**	0.827	**100.0**	**46.5**	**58.8**	**12.3**	**13.6**	**29.1**	**43.4**
	1.0	**100.0**	0.846	85.2	27.3	47.4	2.6	2.8	5.7	28.5
	10.0	**100.0**	**0.936**	78.8	10.2	24.5	0.0	0.0	0.0	18.9
MNIST	0.0	**100.0**	0.896	**79.3**	**69.6**	**47.0**	**28.3**	**24.3**	**44.8**	**48.9**
	1.0	**100.0**	0.948	57.1	41.1	41.8	20.9	9.1	40.2	35.0
	10.0	99.2	**0.989**	−0.9	−0.2	−7.9	−3.7	−1.7	29.0	2.4

注：Avg. 表示免疫率的平均值．加粗字体表示最佳数据

(2) 参数λ_{ie}．参数λ_{ie}对应了损失$\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)$在$\mathcal{L}'(\boldsymbol{w},y)$中的重要性，主要影响着免疫样本的分类准确率．我们预先设置$\lambda_{pert}=0.0$、$\lambda_\sigma=\lambda_\mu=1.0$、$\rho=2.0$、$N=5$，并分别测试$\lambda_{ie}$为0.0、1.0、10.0、$10^2$、$10^3$、$10^4$和$10^5$时，MID的性能．实验结果如表5所示．实验结果表明，随着λ_{ie}值的增加，免疫样本的分类准确率和SSIM均呈现上升趋势，而免疫样本迁移至其他5种对抗攻击的免疫率呈现先上升后下降的趋势．我们取$\lambda_{ie}=10^4$以获取最佳的平均免疫率．此时，在两个数据集上，免疫样本的分类准确率均为100.0%，SSIM分别为0.836和0.900，防御6种对抗攻击的平均免疫率分别为69.7%和59.6%．

表5 MID在不同的λ_{ie}值下针对AdvGAN制作的免疫样本的分类准确率(%)、免疫率(%)和SSIM，以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

数据集	λ_{ie}	准确率	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	Avg.
CIFAR-10	0.0	54.0	0.818	**100.0**	16.6	−14.4	0.6	1.8	11.5	19.4
	1.0	**100.0**	0.827	**100.0**	46.5	58.8	12.3	13.6	29.1	43.4
	10.0	**100.0**	0.831	**100.0**	46.5	72.5	24.8	24.8	37.8	51.1
	10^2	**100.0**	0.834	**100.0**	50.8	79.4	48.1	49.6	46.9	62.5
	10^3	**100.0**	0.835	98.8	52.9	79.4	60.3	60.9	51.3	67.3
	10^4	**100.0**	0.836	95.2	**57.2**	**84.0**	**60.7**	**62.6**	**58.5**	**69.7**
	10^5	**100.0**	**0.837**	88.8	50.8	74.8	60.3	62.2	49.1	64.3
MNIST	0.0	99.1	0.896	69.2	64.7	43.5	26.2	22.7	42.9	44.9
	1.0	**100.0**	0.896	79.3	69.6	47.0	28.3	24.3	44.8	48.9
	10.0	**100.0**	0.896	82.7	72.5	54.2	29.9	25.5	45.3	51.7
	10^2	**100.0**	0.897	84.4	72.8	56.9	39.2	26.4	47.9	54.6
	10^3	**100.0**	0.898	85.7	73.9	56.7	43.0	26.6	53.7	56.6
	10^4	**100.0**	0.900	**85.9**	**74.4**	**58.7**	**45.6**	**27.1**	**65.8**	**59.6**
	10^5	**100.0**	**0.906**	**85.9**	73.9	53.6	45.5	**27.1**	62.6	58.1

注：Avg. 表示免疫率的平均值．加粗字体表示最佳数据

(3) 参数λ_σ．参数λ_σ对应了损失$\sigma^2_{\delta\sim\mathcal{D}}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)$在$\mathcal{L}'(\boldsymbol{w},y)$中的重要性，主要影响着分类损失在对抗样本$g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)$的$\delta$邻域内的平坦度．我们预先设置$\lambda_{pert}=0.0$、$\lambda_{ie}=10^4$、$\lambda_\mu=1.0$、$\rho=2.0$、$N=5$，并分别测试$\lambda_\sigma$为0.0、1.0和10.0时，MID的性能．实验结果如表6所示．实验结果表明，随着λ_σ值的增加，免疫样本的分类准确率均为100.0%，SSIM均呈现上升趋势，而免疫样本迁移至其他5种对抗攻击的免疫率呈现先上升后下降的趋势．我们取$\lambda_\sigma=1.0$以获取最

表6　MID在不同的λ_σ值下针对AdvGAN制作的免疫样本的分类准确率(%)、免疫率(%)和SSIM，以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

数据集	λ_σ	准确率	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	Avg.
CIFAR-10	0.0	**100.0**	0.819	**100.0**	53.6	80.7	59.3	60.9	56.3	68.1
	1.0	**100.0**	0.836	95.2	**57.2**	**84.0**	**60.7**	**62.6**	**58.5**	**69.7**
	10.0	**100.0**	**0.847**	87.2	42.4	76.2	55.3	50.4	53.6	54.0
MNIST	0.0	**100.0**	0.899	79.2	69.8	57.0	43.7	25.5	64.0	56.5
	1.0	**100.0**	0.900	**85.9**	**74.4**	**58.7**	**45.6**	**27.1**	**65.8**	**59.6**
	10.0	**100.0**	**0.917**	72.3	59.0	53.8	40.7	18.3	57.6	50.3

注：Avg. 表示免疫率的平均值. 加粗字体表示最佳数据

佳的平均免疫率. 此时，在两个数据集上，免疫样本的分类准确率均为100.0%，SSIM分别为0.836和0.900，防御6种对抗攻击的平均免疫率分别为69.7%和59.6%.

(4) 参数λ_μ. 参数λ_μ对应了损失$\mu_{\delta\sim\mathcal{D}}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)$在$\mathcal{L}'(\boldsymbol{w},y)$中的重要性，主要影响着对抗样本$g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right)$的$\delta$邻域内的平均分类损失. 我们预先设置$\lambda_{pert}=0.0$、$\lambda_{ie}=10^4$、$\lambda_\sigma=1.0$、$\rho=2.0$、$N=5$，并分别测试$\lambda_\mu$为0.0、1.0和10.0时，MID的性能. 实验结果如表7所示. 实验结果表明，随着λ_μ值的增加，免疫样本的分类准确率均为100.0%，SSIM均呈现下降趋势，而免疫样本迁移至其他5种对抗攻击的免疫率呈现先上升后下降的趋势. 我们取$\lambda_\mu=1.0$以获取最佳的平均免疫率. 此时，在两个数据集上，免疫样本的分类准确率均为100.0%，SSIM分别为0.836和0.900，防御6种对抗攻击的平均免疫率分别为69.7%和59.6%.

表7　MID在不同的λ_μ值下针对AdvGAN制作的免疫样本的分类准确率(%)、免疫率(%)和SSIM，以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

数据集	λ_μ	准确率	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	Avg.
CIFAR-10	0.0	**100.0**	**0.920**	5.9	3.3	6.2	0.0	0.0	20.2	5.9
	1.0	**100.0**	0.836	**95.2**	**57.2**	**84.0**	**60.7**	**62.6**	**58.5**	**69.7**
	10.0	**100.0**	0.805	89.5	51.2	75.4	52.4	52.6	48.4	61.6
MNIST	0.0	**100.0**	**0.962**	−0.2	3.6	5.1	−4.0	−0.4	59.6	10.6
	1.0	**100.0**	0.900	**85.9**	**74.4**	**58.7**	**45.6**	**27.1**	**65.8**	**59.6**
	10.0	**100.0**	0.896	84.7	73.3	57.3	44.0	26.9	53.7	56.7

注：Avg. 表示免疫率的平均值. 加粗字体表示最佳数据

(5)参数ρ. 参数ρ是已知对抗样本和潜在对抗样本之间$L_{p'}$距离($p'=2$)的估计，其直接影响了对未知对抗攻击的估计，从而影响着免疫样本迁移至未知对抗攻击的性能. 我们预先设置$\lambda_{pert}=0.0$、$\lambda_{ie}=10^4$、$\lambda_\sigma=\lambda_\mu=1.0$、$N=5$，并分别测试$\rho$为1.0、2.0、3.0、4.0和5.0时，MID的性能. 实验结果如表8所示. 实验结果表明，随着ρ值的增加，免疫样本的分类准确率均为100.0%，SSIM均呈现上升趋势，而免疫样本针对6种不同对抗攻击的免疫率呈现先上升后下降的趋势. 我们取$\rho=2.0$以获取最佳的免疫率. 此时，在两个数据集上，免疫样本的分类准确率均为100.0%，SSIM分别为0.836和0.900，防御6种对抗攻击的平均免疫率分别为69.7%和59.6%.

(6)参数N. 参数N是蒙特卡罗方法采样的次数，其影响期望估计的准确性. 一般而言，根据大数定律[58]可知，N越大，期望估计得越准确. 我们预先设置$\lambda_{pert}=0.0$、$\lambda_{ie}=10^4$、$\lambda_\sigma=\lambda_\mu=1.0$、$\rho=2.0$，并分别测试$N$为1.0、2.0、…、10.0时，MID的性能. 实验结果如图3、图4和图5所示. 实验结果表明，随着N值的增加，免疫样本的分类准确率均为100.0%，SSIM呈现下降趋势，免疫样本针对6种不同对抗攻击的免疫率呈现先快速上升后缓慢上升的趋势. 由于增加N的同时也增加了计算成本，所以我们取$N=5$以平衡免疫率和计算成本. 此时，在两个数据集上，免疫样本的分类准确率均为100.0%，SSIM分别为0.836和0.900，防御6种对

表8 MID在不同的ρ值下针对AdvGAN制作的免疫样本的分类准确率(%)、免疫率(%)和SSIM,以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

数据集	ρ	准确率	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	Avg.
CIFAR-10	1.0	**100.0**	0.834	94.0	56.6	83.2	59.2	60.8	57.9	68.6
	2.0	**100.0**	0.836	**95.2**	**57.2**	**84.0**	**60.7**	**62.6**	**58.5**	**69.7**
	3.0	**100.0**	0.839	93.5	56.2	83.2	60.2	61.4	57.8	68.7
	4.0	**100.0**	0.842	92.6	55.5	82.8	59.1	59.6	56.5	67.7
	5.0	**100.0**	**0.849**	90.9	55.3	81.2	57.4	58.1	55.5	66.4
MNIST	1.0	**100.0**	0.900	83.1	72.6	57.9	**45.6**	26.3	64.1	58.3
	2.0	**100.0**	0.900	**85.9**	**74.4**	**58.7**	**45.6**	**27.1**	**65.8**	**59.6**
	3.0	**100.0**	0.901	82.3	72.2	**58.7**	45.2	26.9	63.8	58.2
	4.0	**100.0**	0.901	80.9	72.1	58.4	44.7	26.7	62.8	57.6
	5.0	**100.0**	**0.902**	79.5	72.0	57.2	44.5	26.1	62.6	57.0

注:Avg. 表示免疫率的平均值. 加粗字体表示最佳数据

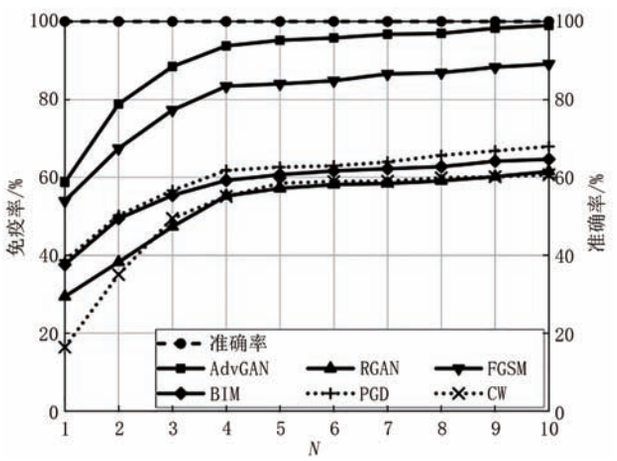

图3 在CIFAR-10上,MID在不同的N值下针对AdvGAN制作的免疫样本的分类准确率(%)和免疫率(%),以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

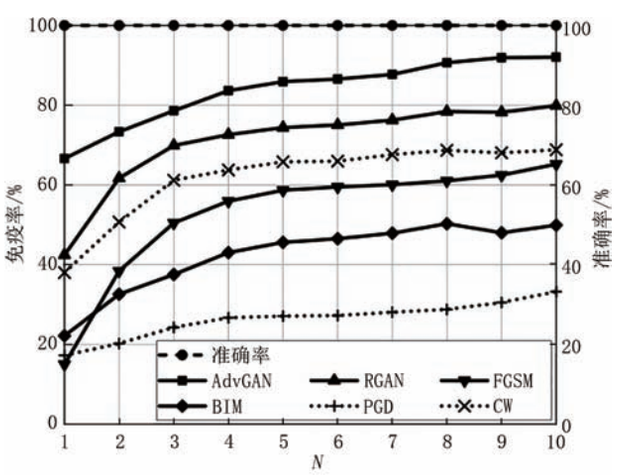

图4 在MNIST上,MID在不同的N值下针对AdvGAN制作的免疫样本的分类准确率(%)和免疫率(%),以及将免疫样本迁移至其他5种对抗攻击的免疫率(%)

抗攻击的平均免疫率分别为69.7%和59.6%.

综上所述,在黑盒免疫防御中,我们设置MID针对AdvGAN制作可迁移的免疫样本,参

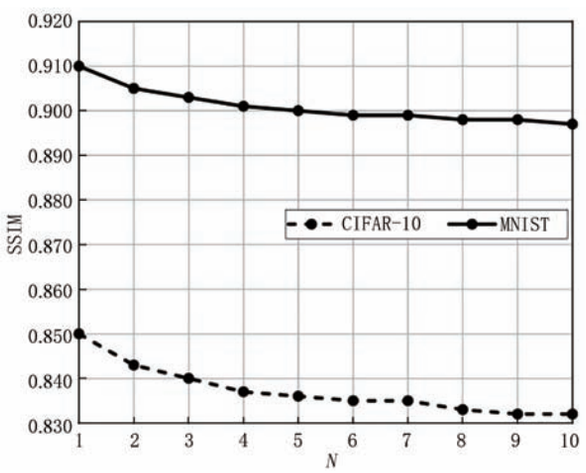

图5 在CIFAR-10和MNIST上,MID在不同的N值下针对AdvGAN制作的免疫样本的SSIM

数$\lambda_{pert}=0.0$、$\lambda_{ie}=10^4$、$\lambda_{\sigma}=\lambda_{\mu}=1.0$、$\rho=2.0$和$N=5$. 此时,在两个数据集上,免疫样本的平均准确率均达到100%,SSIM分别为0.836和0.900,防御6种对抗攻击的平均免疫率分别为69.7%和59.6%.

4.4.2 损失函数$\mathcal{J}$的影响

在本章节我们对比了损失函数$\mathcal{J}$为公式(2)和无目标的CW损失时免疫样本的性能. 无目标的CW损失可以被描述为

$$\mathcal{J}(\boldsymbol{x},y)=\max_{t\neq y} Z(t|\boldsymbol{x})-Z(y|\boldsymbol{x}), \tag{26}$$

其中,Z表示logits输出. 我们在CIFAR-10数据集上针对AdvGAN分别使用具有公式(2)和公式(26)的MID制作免疫样本,并将免疫样本迁移至RGAN、FGSM、BIM、PGD、CW、SF、PGD_0和TSAA,然后记录免疫率,实验结果如表9所示. 实验结果表明,公式(2)和公式(26)对应的平均免疫率分别为71.8%和72.0%,它们之间的绝对误差仅有0.2%,这表明了本文所使用的损失函数和CW

表9　在CIFAR-10上，使用MID在不同损失下针对AdvGAN制作的免疫样本的免疫率(%)以及将免疫样本迁移至其他8种对抗攻击的免疫率(%)

损失函数	AdvGAN*	RGAN	FGSM	BIM	PGD	CW	SF	PGD_0	TSAA	Avg.
公式(2)	95.2	57.2	84.0	60.7	62.6	58.5	88.9	87.1	52.3	71.8
公式(26)	94.3	52.4	86.6	62.6	66.4	60.0	86.0	87.6	52.0	72.0

注：Avg. 表示免疫率的平均值. 符号“*”表示代理攻击

损失函数具有相当的效果.

4.5　攻击强度对免疫防御性能的影响

4.5.1　对抗扰动强度对免疫防御性能的影响

我们选择容易调控扰动强度的L_∞攻击(FGSM、BIM和PGD)和L_2攻击(AdvGAN、RGAN和CW)作为对手以探索不同扰动强度的攻击对免疫率的影响. 不同攻击具有不同范数和量级的扰动强度，为了使得实验的变量一致，我们引入缩放因子S作为统一的变量. 通过缩放因子乘以扰动或扰动强度就可以实现对不同攻击的扰动强度的统一调控. 实验的具体方案如下，我们在四种数据集上使用MID针对AdvGAN制作可迁移的免疫样本，然后将免疫样本迁移至多种具有不同强度的攻击，最后根据不同的S值和免疫率绘制出对抗扰动强度与

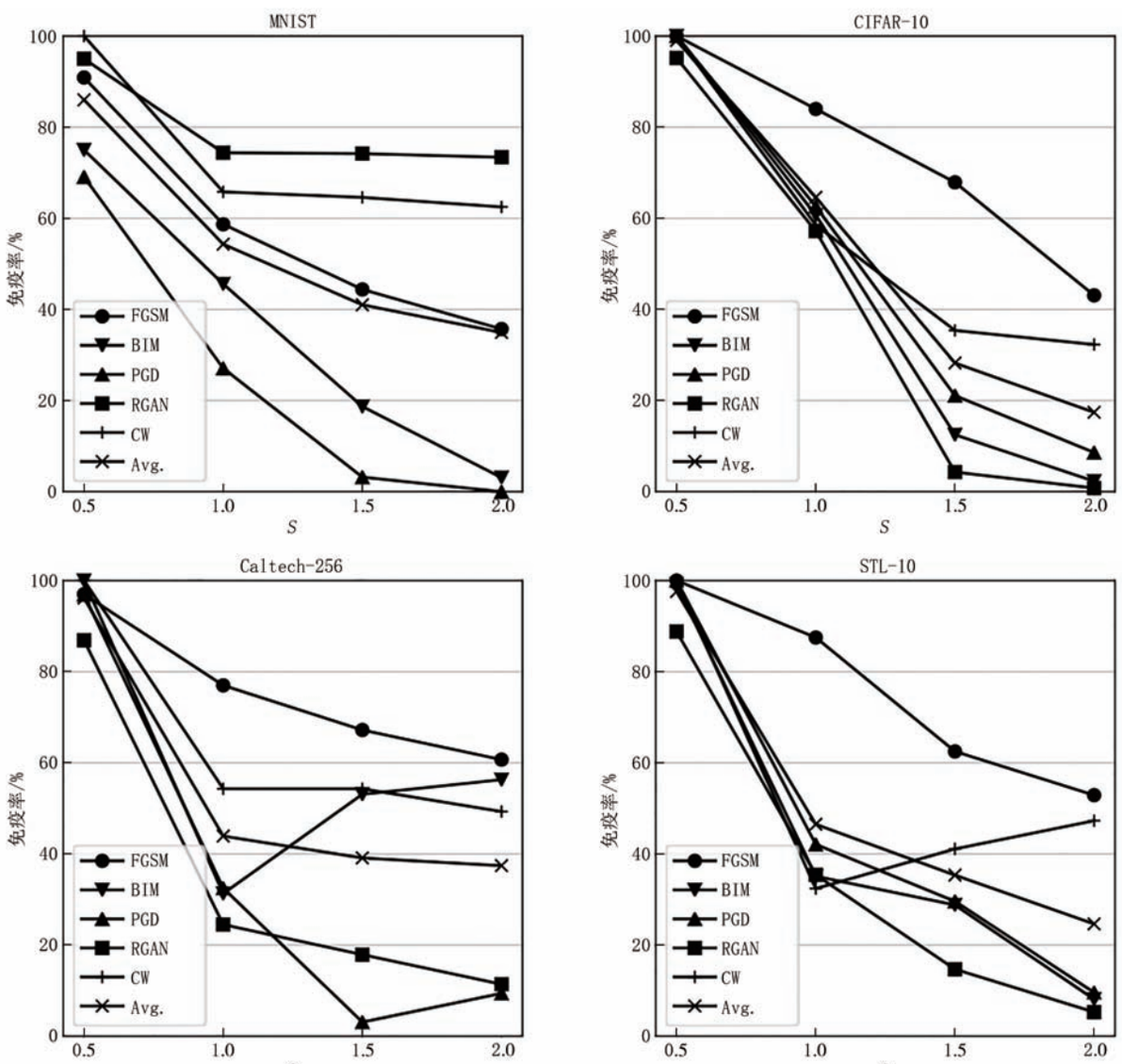

图6　在MNIST、CIFAR-10、STL-10和Caltech-256上，使用MID针对AdvGAN制作的免疫样本并将免疫样本迁移至其他多种不同扰动强度的对抗攻击的免疫率(%)(Avg. 表示平均免疫率值)

免疫率的关系图，如图6所示．实验结果表明，随着对抗扰动强度的增大，免疫样本迁移至其它攻击的平均免疫率呈现下降趋势．当$S=0.5$时，在四个数据集上，MID将免疫样本迁移至其它攻击的平均免疫率分别86.0%、99.0%、97.6%和96.2%，这比原有性能（当$S=1.0$时）分别上升了31.7%、34.4%、51.1%和52.2%．当$S=2.0$时，在四个数据集上，MID将免疫样本迁移至其它攻击的平均免疫率分别34.9%、17.4%、24.6%和37.4%，这比原有性能（当$S=1.0$时）分别下降了19.4%、47.2%、21.9%和6.5%．

4.5.2 查询次数对免疫防御性能的影响

我们还探索了基于查询的黑盒攻击的查询次数对免疫样本性能的影响．我们在四种数据集上使用MID针对AdvGAN制作可迁移的免疫样本，然后将免疫样本迁移至三种基于查询的攻击（即Square、SimBA和SimBA-DCT）．我们通过不断调整黑盒攻击的查询次数并记录免疫率的变化，绘制出查询次数与免疫样本性能的趋势图，结果如图7所示．实验结果表明，随着查询次数从1000到7000，免疫率整体上呈现下降趋势，并在查询次数达到4000后出现平缓的趋势，这是由于黑盒攻击在经过4000次查询后的攻击能力逐渐饱和，所以对免疫样本的影响也在逐渐减弱．

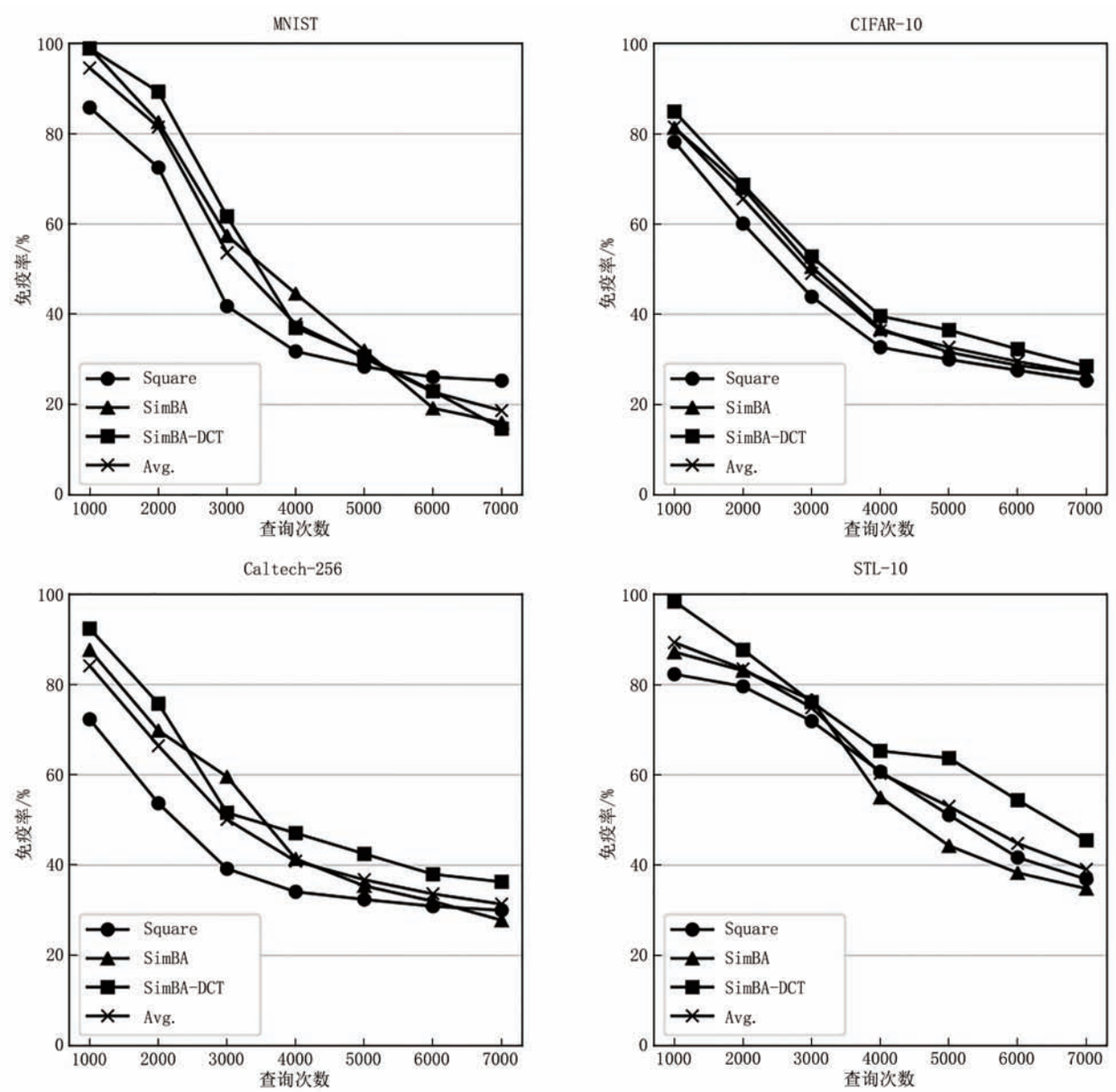

图7 在MNIST、CIFAR-10、STL-10和Caltech-256上，使用MID针对AdvGAN制作的免疫样本并将免疫样本迁移至其他多种不同查询次数的黑盒攻击的免疫率（%）（Avg. 表示平均免疫率值）

4.6　免疫样本的鲁棒性测试

在免疫样本传输或保存时，JPEG压缩的量化过程会使得图像的高频信息产生损失，这可能会影响到免疫样本的性能．为了评估免疫样本对JPEG压缩的鲁棒性，我们设计了如下实验方案：首先，我们在MNIST、CIFAR-10、STL-10和Caltech-256数据集上使用MID针对AdvGAN制作出免疫样本；然后，我们将免疫样本分别经过质量因子为90、80和70的JPEG压缩和解压缩并得到压缩后的免疫样本；最后，我们将经过JPEG压缩的免疫样本迁移至其他攻击并记录免疫率，实验结果如图8所示．实验结果表明，随着质量因子的增加，平均免疫率不断降低．当质量因子为70时，免疫样本在MNIST、CIFAR-10、STL-10和Caltech-256数据集上的平均免疫率分别为27.6%、26.5%、18.7%和27.5%，这比经过JPEG压缩的免疫样本的性能下降了28.3%、42.4%、34.3%和24.8%．

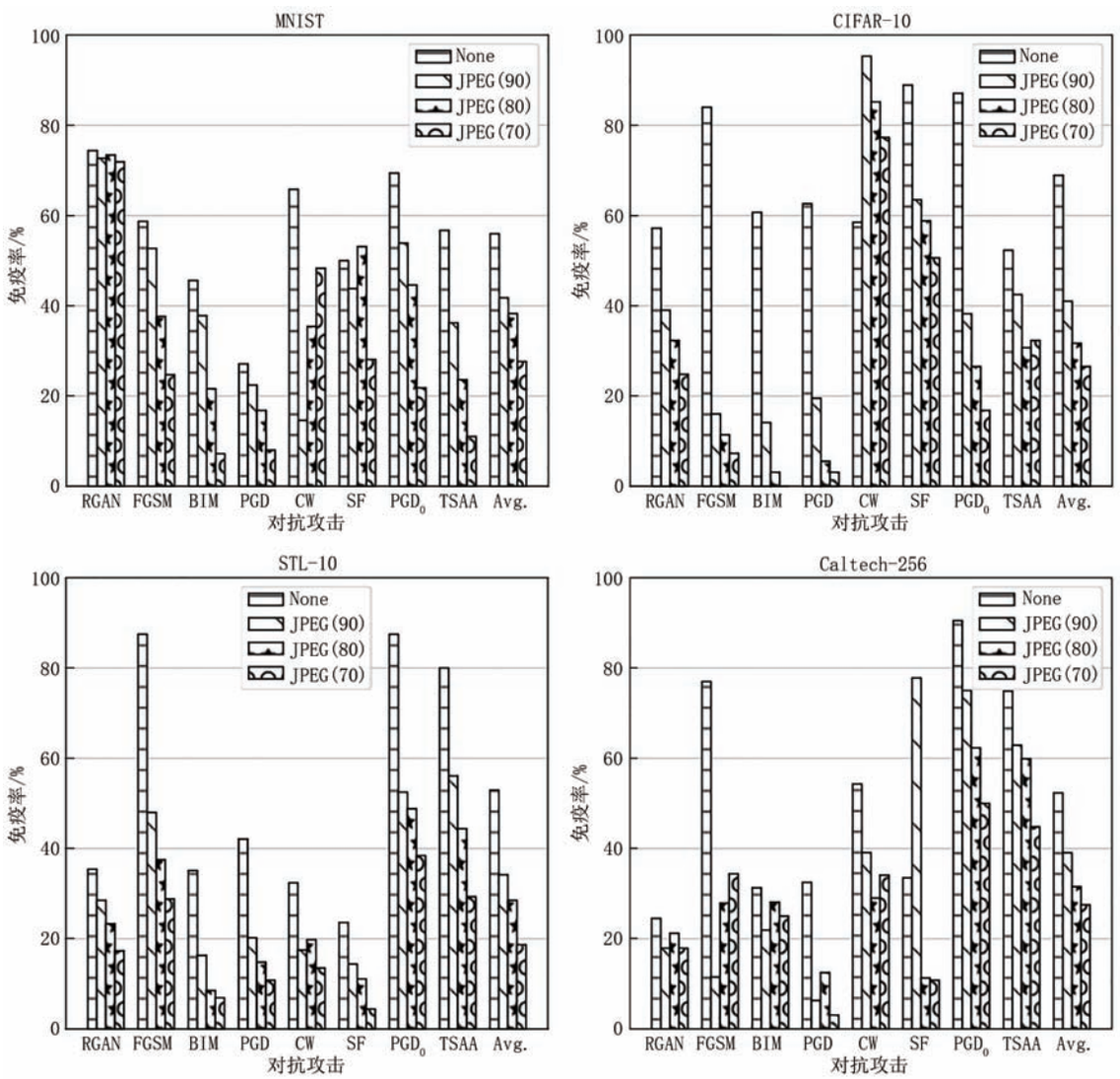

图8　在MNIST、CIFAR-10、STL-10和Caltech-256上，使用MID针对AdvGAN制作的免疫样本并经过不同质量因子的JPEG压缩后的免疫率(%)(Avg. 表示平均免疫率值)

此外，我们还测试了使用PGD这种不会产生浮点数输出的优化方法得到的免疫样本的鲁棒性．具体来说，为了让MID的优化问题适配PGD的优化方法，我们将原优化问题微调为

$$\begin{aligned}\mathcal{L}'(\boldsymbol{x}^{IE},y)=&\lambda_{ie}\cdot\mathcal{J}(\boldsymbol{x}^{IE},y)\\&+\lambda_{\sigma}\cdot\sigma^2_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE},y)\\&+\lambda_{\mu}\cdot\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE},y)\end{aligned}\tag{27}$$

$$\arg\min_{\boldsymbol{x}^{IE}}\mathcal{L}'(\boldsymbol{x}^{IE},y),\ \text{s.t.}\left\|\boldsymbol{x}^{IE}-\boldsymbol{x}\right\|_{\infty}\leqslant\tau\tag{28}$$

为了让优化的扰动不包括浮点数，我们取消了PGD的随机重启(Random Restarts)并使用梯度符号下降法求解上式，即

$$\boldsymbol{x}_0^{IE}=\boldsymbol{x} \tag{29}$$

$$\boldsymbol{x}_{t+1}^{IE}=\mathrm{Clip}_{(\boldsymbol{x},\tau)}\left\{\boldsymbol{x}_t^{IE}+\alpha\cdot\mathrm{sign}\left(\nabla\mathcal{L}'(\boldsymbol{x}_t^{IE},y)\right)\right\} \tag{30}$$

其中，α表示迭代步长，sign(·)表示符号函数，$\mathrm{Clip}_{(\boldsymbol{x},\tau)}\{\cdot\}$表示将输入截断到$\boldsymbol{x}$的$\tau$球内.

在实验中，我们将使用Adam优化的MID称为MID-Adam，将使用上述PGD优化的MID称为MID-PGD. 此外，我们将MID-PGD的α设置为1/255，迭代次数设置为和MID-Adam相同的500. 实验在MNIST和CIFAR-10上进行，分别使用MID-PGD和MID-Adam针对AdvGAN制作免疫样本，并记录了免疫样本在经过不同质量因子的JPEG压缩后防御其他攻击的平均免疫率(%)，实验结果如图9所示.

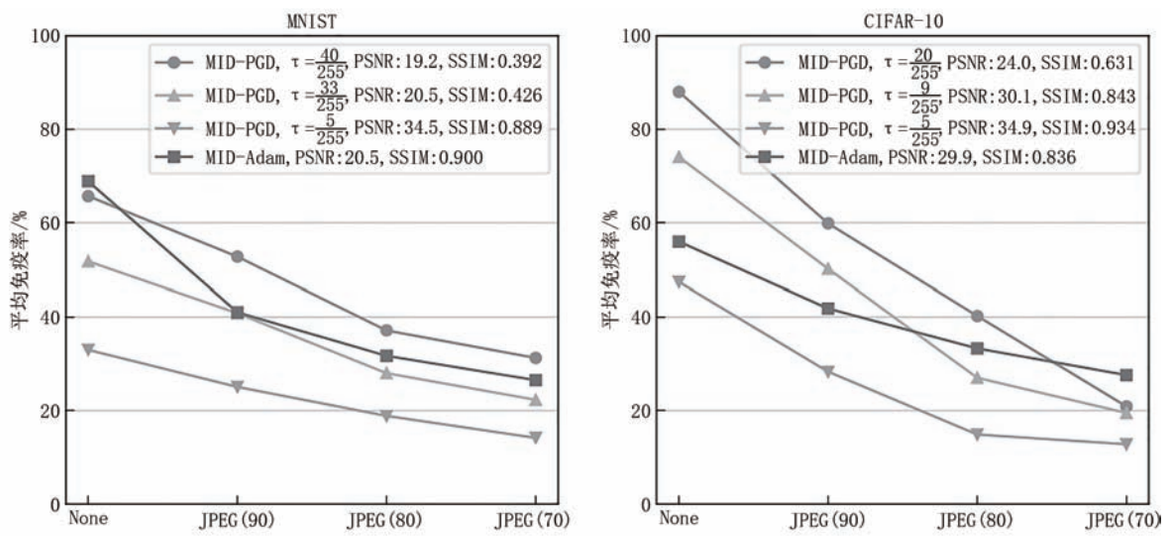

图9 MID-PGD和MID-Adam对不同质量因子的JPEG压缩的鲁棒性

实验结果表明，随着质量因子的下降，MID-PGD和MID-Adam的平均免疫率都呈现下降趋势，这表明JPEG压缩对免疫扰动具有破坏性. 此外，我们发现MID-PGD免疫样本对JPEG压缩的鲁棒性很大程度上取决于最大免疫扰动强度τ的值，即随着τ的增大，MID-PGD对JPEG压缩的鲁棒性逐渐变强，而带来的负面效果是MID-PGD免疫样本的PSNR和SSIM在下降. 例如，当MNIST数据集上的$\tau=40/255$或CIFAR-10数据集上的$\tau=20/255$时，MID-PGD对JPEG压缩的鲁棒性总体上好于MID-Adam，但它们的免疫样本的视觉质量大幅度下降.

因此，为了更加公平地对比，我们通过控制MID-PGD的τ值使得MID-PGD与MID-Adam免疫样本的视觉质量相似后再去对比MID-PGD和MID-Adam的鲁棒性. 图9中的结果表明，当MID-PGD和MID-Adam免疫样本的PSNR最相近时(此时，MNIST上的$\tau=33/255$，CIFAR-10上的$\tau=9/255$)，MID-PGD和MID-Adam免疫样本对不同程度的JPEG压缩的鲁棒性相当；然而，在MNIST数据集上，当MID-PGD和MID-Adam免疫样本的SSIM最相近时(此时，$\tau=5/255$)，MID-Adam表现出比MID-PGD更好的鲁棒性.

综上所述，MID-PGD免疫样本对JPGE压缩的鲁棒性很大程度上取决于最大免疫扰动强度τ. 当MID-PGD和MID-Adam免疫样本的PSNR相似时，它们对JPEG压缩的鲁棒性也相近，但MID-Adam可能获得更高的SSIM(如在MNIST数据集上).

4.7 免疫防御与输入预处理的对比实验

输入预处理是一种常见的对抗防御方法，与免疫防御相同，其也是一种应用于图像的防御机制. 输入预处理和免疫防御的区别之一是实施的阶段不同. 输入预处理处在对手攻击后的阶段，其旨在对对抗样本执行图像变换使其不再具备攻击能力，而免疫防御处在对手攻击前的阶段，其旨在对干净图像预先添加免疫扰动使得后续添加的对抗扰动失效. 因此，我们对比了免疫防御和输入预处理防御的防御性能. 我们选择特征蒸馏(Feature Distillation，FD)[32]、CIIDefence[43]和小波去噪(Wavelet Denoising，WD)[44]作为基线和免疫防御

MID进行对比，我们依旧使用免疫率作为指标从数值上对比免疫防御和输入预处理防御的性能．此外，我们还将免疫防御和输入预处理进行结合以验证两者之间的兼容性，实验结果如表10所示．实验结果表明，在免疫防御和输入预处理防御对比时，免疫防御在两个数据集上分别获得了最高的71.8%和59.3%的平均免疫率，比基线高出23.2%～30.9%和8.3%～40.2%．在输入预处理防御和“免疫防御＋输入预处理”的集成防御对比时，集成防御在两个数据集上的平均免疫率均比集成中的单个输入预处理高，分别高出9.7%～43.4%和27.4%～34.0%．此外，MID ＋ CIIDefence在所有防御中获得了最高的84.3%和78.8%的平均免疫率，这说明免疫防御和输入预处理防御有良好的兼容性．

表10　MID针对AdvGAN制作的免疫样本的免疫率(%)和三种输入预处理防御的对比

数据集	防御	AdvGAN	RGAN	FGSM	BIM	PGD	CW	SF	PGD_0	TSAA	Avg.
CIFAR-10	FD	10.9	50.7	63.4	*75.0*	*75.0*	75.8	28.2	18.8	36.2	48.2
	CIIDefence	22.2	56.6	28.5	14.1	25.0	71.9	57.6	41.2	51.2	40.9
	WD	25.5	51.5	47.3	60.2	68.8	*81.3*	47.1	26.5	29.1	48.6
	MID(所提方案)	***95.2****	*57.2*	*84.0*	60.7	62.6	58.5	***88.9***	*87.1*	*52.3*	*71.8*
	MID ＋ FD(所提方案)	46.2*	58.2	81.2	82.8	80.5	84.4	11.8	32.4	43.3	57.9
	MID ＋ CIIDefence(所提方案)	86.8*	**79.9**	**95.5**	87.5	91.4	**100.0**	45.9	**100.0**	**71.7**	**84.3**
	MID ＋ WD(所提方案)	90.1*	65.7	**95.5**	**96.1**	**96.1**	99.2	21.2	**100.0**	48.8	79.2
MNIST	FD	10.0	18.6	16.5	12.6	12.8	50.0	12.5	28.9	10.4	19.1
	CIIDefence	53.6	*83.6*	33.3	28.8	23.2	*97.9*	15.6	35.7	*87.4*	51.0
	WD	10.8	17.0	15.4	11.7	19.6	75.0	15.6	15.2	12.4	21.4
	MID(所提方案)	***85.9****	74.4	*58.7*	*45.6*	*27.1*	65.8	*50.0*	*69.4*	56.7	*59.3*
	MID ＋ FD(所提方案)	53.6*	75.8	58.1	64.9	50.4	91.7	31.3	33.1	18.7	53.1
	MID ＋ CIIDefence(所提方案)	81.6*	**93.8**	**73.1**	**67.6**	**57.6**	**98.3**	**68.8**	**77.8**	**90.6**	**78.8**
	MID ＋ WD(所提方案)	32.0*	28.9	41.9	53.2	40.0	95.8	43.8	64.6	38.6	48.8

注：Avg. 表示免疫率的平均值．符号“*”表示已知的代理攻击上的数据，斜体数据表示单个防御中的最佳数据，加粗数据表示单个防御和集成防御中的最佳数据

4.8　MID在不同代理攻击下的泛化性测试

为了验证MID在不同代理攻击下的泛化性，我们在MNIST、CIFAR-10、STL-10和Caltech-256数据集上分别针对RGAN、FGSM和PGD制作免疫样本，然后将免疫样本迁移至其他攻击并测试免疫率，实验结果如表11所示．实验结果表明，在四个数

表11　MID针对四种不同代理攻击制作的免疫样本迁移至其他攻击的PSNR(dB)、SSIM和免疫率(%)

数据集	PSNR	SSIM	AdvGAN	RGAN	FGSM	BIM	PGD	CW	SF	PGD_0	TSAA	Avg.
MNIST	20.5	0.900	85.9*	74.4	58.7	45.6	27.1	65.8	50.0	69.4	56.7	59.3
	23.9	**0.942**	86.0	93.8*	58.1	45.0	48.0	77.1	25.1	58.9	47.2	59.9
	22.6	0.930	24.0	47.7	69.9*	50.8	59.6	85.4	28.1	59.2	33.1	50.9
	21.3	0.910	24.8	76.6	67.7	61.3	53.6*	91.7	50.0	78.1	59.8	**62.6**
CIFAR-10	29.9	0.836	95.2*	57.2	84.0	60.7	62.6	58.5	88.9	87.1	52.3	71.8
	32.4	**0.902**	73.1	95.7*	90.2	63.3	65.6	71.9	27.1	83.3	48.8	68.8
	32.2	0.899	81.4	52.4	100.0*	91.4	88.3	72.0	12.9	69.1	52.8	68.9
	32.0	0.898	44.7	54.9	100.0	99.2	99.2*	76.5	58.8	78.8	52.8	**73.9**
STL-10	32.4	0.873	99.2*	35.4	87.5	35.1	42.1	32.4	23.6	87.5	80.0	58.1
	32.5	**0.875**	59.3	98.3*	81.7	32.0	51.5	30.0	19.8	63.6	59.6	55.1
	32.2	0.871	52.5	30.2	100.0*	64.8	70.3	41.1	22.0	64.5	60.6	56.2
	32.4	0.874	55.1	32.8	99.0	94.5	95.3*	53.6	24.2	72.7	69.7	**66.3**
Caltech-256	**34.4**	**0.867**	55.7*	24.5	77.0	31.3	32.5	54.3	33.5	90.5	74.9	52.7
	34.4	**0.867**	59.2	78.5*	77.8	71.9	62.5	54.3	55.7	64.4	74.8	66.6
	33.7	0.849	86.4	44.2	96.7*	87.5	31.3	89.9	88.9	79.3	90.4	77.2
	34.1	0.862	89.8	37.6	100.0	100.0	100.0*	94.9	88.3	89.5	95.2	**88.4**

注：Avg. 表示免疫率的平均值，符号“*”表示已知的代理攻击上的数据，加粗数据表示最佳数据

据集上，RGAN作为代理攻击时制作的免疫样本具备最佳的视觉质量，PSNR和SSIM分别为23.9和0.942、32.4和0.902、32.5和0.875、34.4和0.867. 此外，PGD作为代理攻击时制作的免疫样本具备最高的平均免疫率，比AdvGAN作为代理攻击时制作的免疫样本分别高出3.3%、2.1%、8.2%和35.7%. 针对PGD作为代理攻击时的免疫率最高的现象，我们的解释是PGD所应用的随机重启能够产生丰富多样的一阶对抗样本，从而使得针对这些多样化的对抗样本制作出的免疫样本更有可能迁移至其他攻击.

5 讨 论

在本章节我们讨论了免疫样本在建设鲁棒的公共路牌系统和公共数据集这两个方面的应用场景，具体如下.

(1)鲁棒的公共路牌系统. 攻击者不仅可以通过MitM攻击在数字世界制作对抗样本，还可以在物理世界中直接为路牌张贴上对抗补丁. 如图1(a)(b)所示，前者的扰动是微小或难以察觉的，而后者的补丁是明显的和区域性的，两者之间存在明显差异，但它们都可以被描述成定义1中的符号g. 因此，免疫样本的定义同时适用于二者. 在具体场景下，如图1(c)所示，工厂在生产路牌时，可以为路牌图像添加上免疫扰动以抵御对抗补丁或微小扰动. 因此，免疫防御在建设鲁棒的公共路牌系统方面具备很大的潜力.

接下来，我们通过实验验证这一应用场景. 具体来说，我们选择了德国交通标志检测基准(GTSDB)[59]作为实验所用的数据集，并训练了一个对Meta路牌图像分类准确率为100%的Inception-v3分类器. Meta路牌图像的大小均为100×100. 随后，我们使用MID(代理攻击为AdvGAN)将路牌图像制作成免疫样本. 如表12的第2行所示，这些路牌免疫样本依旧能被Inception-v3分类器100%地准确识别.

在现实世界中，对手很有可能通过在路牌上张贴对抗补丁的方式以达到欺骗自动驾驶汽车的目的. 因此，我们制作出了大小为路牌大小的5%的对抗补丁. 我们在表12的第3和4行中记录了在免疫防御前后，张贴了对抗补丁的路牌的分类准确率. 实验结果显示免疫防御能够帮助分类器提升62.8%的分类准确率.

此外，在路牌识别的场景下，摄像机变换(如视角、光照、距离等)带来的影响需要考虑. 因此，我们分别利用“随机视角变换(失真比例为0%～50%)”、“调整亮度(强度因子为0.3)”、“随机缩放(缩放比例为0%～50%)”和“ISO噪声(色调变化方差为0.01～0.05，颜色和亮度噪声的因子为0.1～0.5)”模拟出了视角变化、光照、距离变化和相机传感器噪声. 在实验中，我们分别测试了在有无免疫防御时，张贴了对抗补丁的路牌在经过摄像机变换后的分类准确率，结果如表12的第5和6行所示. 实验结果表明，在有免疫防御时，路牌的分类准确率均有所提高，平均提高8.1%.

表12 不同设置下的路牌的分类准确率(%)

行号	原始路牌	免疫扰动	对抗补丁	变换	分类准确率
1	√	×	×	×	**100.0**
2	√	√	×	×	**100.0**
3	√	×	√	×	30.2
4	√	√	√	×	**93.0**
5	√	×	√	视角变化√	81.4
				光照√	60.5
				距离变化√	81.4
				相机传感器噪声√	46.5
				平均	67.5
6	√	√	√	视角变化√	**86.0**
				光照√	**72.1**
				距离变化√	**88.4**
				相机传感器噪声√	**55.8**
				平均	**75.6**
7	√	×	×	视角变化√	97.7
				光照√	95.3
				距离变化√	90.7
				相机传感器噪声√	58.1
				平均	85.5
8	√	√	×	视角变化√	**100.0**
				光照√	**97.7**
				距离变化√	**93.0**
				相机传感器噪声√	**67.4**
				平均	**89.5**

注：符号“√”表示有该设置，符号“×”表示无该设置

此外，我们也测试了在没有对抗补丁的情况下，摄像机变换对免疫样本的影响，结果呈现在表12的第7和8行. 实验结果表明，在摄像机变换后，原始路牌的分类准确率平均下降了10.5%. 而经过免疫防御后，该准确率上升了4.0%.

综上所述，免疫防御能够防止路牌图像被制作成对抗样本，而且在有摄像机变换的情况下，免疫防御依旧能够帮助分类器提高分类准确率，证明了免疫防御的有效性.

(2)鲁棒的公共数据集. 尽管公共数据集(如ImageNet、Caltech-256、CIFAR-10、MNIST等)得到了学术界广泛的认可和使用，但公共数据集的安全性和鲁棒性尚未得到关注. 攻击者可以借助MitM攻击在发布者上传数据集或用户下载数据的过程中毒化数据集，被毒化的数据集图像具有对抗性，但人眼难以察觉图像上的对抗扰动，从而导致公共数据环境被不知不觉地破坏. 发布者可以将数据集图像制作成免疫样本后再发布至网络上以防止数据集被毒化. 因此，免疫防御在建设鲁棒的公共数据集方面也具备很大的潜力.

6 总　结

在这项工作中，我们提出了一种新颖的对抗防御机制，即免疫防御，以防止对抗样本的生成. 首先，我们引入免疫样本和免疫防御的定义. 然后，我们提出了HTID以针对已知对抗攻击制作白盒免疫样本. 此外，我们还探索了黑盒免疫防御并提出了MID以提升免疫防御在多种对抗攻击之间的可转移性. 最后，实验结果表明，HTID制作的免疫样本具有高准确率、高视觉质量和高免疫率；MID在保证其他性能的情况下能够有效地提升免疫样本的可迁移性. 这项工作有望进一步保障图像和深度神经网络的安全性. 然而，由于免疫扰动可能被攻击者检测或破坏，我们将在未来的工作中重点研究免疫样本的鲁棒性.

致　谢　感谢同事的技术帮助与指导，感谢评审专家的宝贵意见.

参　考　文　献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013

[2] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

[3] Huang S, Papernot N, Goodfellow I, et al. Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284, 2017

[4] Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277, 2016

[5] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, UAE, 2017: 506-519

[6] Tramèr F, Papernot N, Goodfellow I, et al. The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453, 2017

[7] Zhang Si-Si, Zuo Xin, Liu Jian-Wei. The adversarial example problem in deep learning. Chinese Journal of Computers, 2019, 42(8): 1886-1904 (in Chinese)
(张思思，左信，刘建伟. 深度学习中的对抗样本问题. 计算机学报, 2019, 42(8): 1886-1904)

[8] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[11] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 1-7

[12] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2818-2826

[13] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[14] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world// Roman V. Yampolskiy. Artificial Intelligence Safety and Security. UK: Chapman and Hall/CRC, 2018: 99-112

[15] Lin J, Song C, He K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks//Proceedings of the International Conference on Learning Representations. Virtual, 2020: 1-12

[16] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[17] Wang X, He K. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1924-1933

[18] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2017: 39-57

[19] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 1-9

[20] Hayes J, Danezis G. Learning universal adversarial perturbations with generative models//Proceedings of the IEEE Security and Privacy Workshops (SPW). San Francisco, USA, 2018: 43-49

[21] Xiao C, Li B, Zhu J Y, et al. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 3905-3911

[22] Jandial S, Mangla P, Varshney S, et al. Advgan++ : Harnessing latent layers for adversary generation//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea, 2019: 1-4

[23] Zhang J, Wang J, Wang H, et al. Self-recoverable adversarial examples: A new effective protection mechanism in social networks. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(2): 562-574

[24] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144

[25] Masci J, Meier U, Cireşan D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction//Proceedings of the 21st International Conference on Artificial Neural Networks. Espoo, Finland, 2011: 52-59

[26] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015

[27] Huang R, Xu B, Schuurmans D, et al. Learning with a strong adversary. arXiv preprint arXiv:1511.03034, 2015

[28] Kurakin A, Goodfellow I J, Bengio S. Adversarial machine learning at scale//Proceedings of the International Conference on Learning Representations. Vancouver, Canada 2017:1-17

[29] Tramèr F, Boneh D, Kurakin A, et al. Ensemble adversarial training: Attacks and defenses//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-20

[30] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks// Proceedings of the IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2016: 582-597

[31] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-12

[32] Liu Z, Liu Q, Liu T, et al. Feature distillation: Dnn-oriented jpeg compression against adversarial examples//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 860-868

[33] Xie C, Wang J, Zhang Z, et al. Mitigating adversarial effects through randomization//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018:1-16

[34] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155, 2017

[35] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA, 2017: 3-14

[36] Feinman R, Curtin R R, Shintre S, et al. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410, 2017

[37] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples// Proceedings of the 37th International Conference on Machine Learning. Stockholm, Sweeden, 2018: 274-283

[38] Song Y, Kim T, Nowozin S, et al. PixelDefend: Leveraging generative models to understand and defend against adversarial examples// Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-20

[39] Tchebichef P. Des valeurs moyennes. Journal de Mathèmatiques Pures et Appliquèes, 1867, 12(2): 177-184

[40] Caflisch R E. Monte carlo and quasi-Monte carlo methods. Acta Numerica, 1998, 7: 1-49

[41] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009

[42] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[43] Gupta P, Rahtu E. Ciidefence: Defeating adversarial attacks by fusing class-specific image inpainting and image denoising// Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 6708-6717

[44] Mustafa A, Khan S H, Hayat M, et al. Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing, 2019, 29: 1711-1724

[45] Modas A, Moosavi-Dezfooli S M, Frossard P. Sparsefool: A few pixels make a big difference//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9087-9096

[46] Croce F, Hein M. Sparse and imperceivable adversarial attacks// Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 4724-4732

[47] He Z, Wang W, Dong J, et al. Transferable sparse adversarial attack//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 14963-14972

[48] Andriushchenko M, Croce F, Flammarion N, et al. Square attack: A query-efficient black-box adversarial attack via random search// Proceedings of the 16th European Conference on Computer Vision. Virtual, 2020: 484-501

[49] Guo C, Gardner J, You Y, et al. Simple black-box adversarial attacks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 2484-2493

[50] Coates A, Ng A, Lee H. An analysis of single-layer networks in unsupervised feature learning//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Lauderdale, USA, 2011: 215-223

[51] Griffin G, Holub A, Perona P. Caltech-256 object category dataset. 2007

[52] Almohammad A, Ghinea G. Stego image quality and the reliability of PSNR//Proceedings of the 2nd International Conference on Image Processing Theory. Paris, France, 2010: 215-220

[53] Conti M, Dragoni N, Lesyk V. A survey of man in the middle attacks. IEEE Communications Surveys & Tutorials, 2016, 18(3): 2027-2051

[54] Wang D, Li C, Wen S, et al. Man-in-the-middle attacks against machine learning classifiers via malicious generative models. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2074-2087

[55] Brown T B, Mané D, Roy A, et al. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017

[56] Zhou D, Wang N, Peng C, et al. Removing adversarial noise in class activation feature space//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 7878-7887

[57] Zhou D, Liu T, Han B, et al. Towards defending against adversarial examples via attack-invariant features//Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021: 12835-12845

[58] Hsu P L, Robbins H. Complete convergence and the law of large numbers. Proceedings of the National Academy of Sciences, 1947, 33(2): 25-31

[59] Houben S, Stallkamp J, Salmen J, et al. Detection of traffic signs in real-world images: The German traffic sign detection benchmark// Proceedings of the International Joint Conference on Neural Networks. Dallas, USA, 2013: 1-8

附录A.

定理1. 对于任意对抗攻击g和g'，输入为$\boldsymbol{x}$，都有

$$\left\| g(\boldsymbol{x})-g'(\boldsymbol{x}) \right\|_{p'} \leqslant \rho \tag{31}$$

证明.

$$\begin{aligned}\left\| g(x)-g'(x) \right\|_{p'} &= \left\| [g(x)-x]-[g'(x)-x] \right\|_{p'} \\ &\leqslant \left\| g(x)-x \right\|_{p'} + \left\| g'(x)-x \right\|_{p'} = \epsilon+\epsilon' := \rho\end{aligned} \tag{32}$$

其中，ϵ和ϵ'分别为g和g'在$L_{p'}$范数约束下的最大对抗扰动强度. 证毕.

定理2. 条件2是条件1的必要不充分条件.

证明.

(1)必要性证明：

由条件1的$\mathbb{P}_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)=0$可知，对于任意的$\boldsymbol{\delta}\sim\mathcal{D}$，都有$\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)<0$，所以$\mathbb{E}_{\delta\sim\mathcal{D}}[\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)]<0$. 因此，条件2是条件1的必要条件.

(2)不充分性证明：

对于不充分性证明，我们只需找出一个满足条件2但不满足条件1的特例. 我们将满足$\|\boldsymbol{\delta}\|_{p'}\leqslant\rho$的$\boldsymbol{\delta}$的全体集合表示为一个无限集$\{\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\boldsymbol{\delta}_3,\cdots\}$. 当$\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0$，$\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta}_2,y)=-\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta}_1,y)<0$ 并且$\sum_{i=3}^{\infty}\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta}_i,y)<0$时，满足$\mathbb{E}_{\delta\sim\mathcal{D}}[\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)]=\sum_{i=1}^{\infty}\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta}_i,y)<0$，即满足条件2. 但此时存在$\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta}_1,y)\geqslant 0$，即不满足条件1的$P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)=0$. 也就是说，存在满足条件2但不满足条件1的特例，所以条件2是条件1的不充分条件.

因此，条件2是条件1的必要不充分条件. 证毕.

引理1. 切比雪夫不等式(Chebyshev's Inequality)[39]. 对于任意$b>0$，随机变量X满足

$$P(|X-\mathbb{E}X|\geqslant b)\leqslant\frac{\mathbb{V}X}{b^2}, \tag{33}$$

其中，$\mathbb{P}$表示概率，$\mathbb{V}$表示方差，$\mathbb{E}$表示期望.

推论1. 若随机变量X满足$\mathbb{E}X<0$，则有

$$P(X\geqslant 0)\leqslant\frac{\mathbb{V}X}{[\mathbb{E}X]^2}, \tag{34}$$

其中，$\mathbb{P}$表示概率，$\mathbb{V}$表示方差，$\mathbb{E}$表示期望.

证明. 根据$\mathbb{E}X<0$和引理1，有

$$\begin{aligned}P(X\geqslant 0) &= \mathbb{P}(X-\mathbb{E}X\geqslant -\mathbb{E}X) \\ &\leqslant P(|X-\mathbb{E}X|\geqslant -\mathbb{E}X)\leqslant\frac{\mathbb{V}X}{[\mathbb{E}X]^2}\end{aligned} \tag{35}$$

证毕.

定理3. 若条件2成立，则公式(14)的解等于公式(12)的解.

证明. 令

$$\hat{\boldsymbol{x}}^{IE}=\arg\min_{\boldsymbol{x}^{IE}}\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}\right]^2 \tag{36}$$

则

$$\begin{aligned}0\leqslant P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\hat{\boldsymbol{x}}^{IE})+\boldsymbol{\delta},y)\geqslant 0) &\leqslant\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\hat{\boldsymbol{x}}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\hat{\boldsymbol{x}}^{IE})}\right]^2 \\ &=\inf_{\boldsymbol{x}^{IE}}\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}\right]^2=0\end{aligned} \tag{37}$$

可以推出

$$\begin{aligned}&P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\hat{\boldsymbol{x}}^{IE})+\boldsymbol{\delta},y)\geqslant 0)=0 \\ &=\inf_{\boldsymbol{x}^{IE}}P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)\end{aligned} \tag{38}$$

所有$\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}\right]^2$和$P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)$有着相同的最小值点，即

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{IE}}\left[\frac{\sigma_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}{\mu_{\delta\sim\mathcal{D}}(\boldsymbol{x}^{IE})}\right]^2 \\ &=\arg\min_{\boldsymbol{x}^{IE}}P_{\delta\sim\mathcal{D}}(\mathcal{J}(g(\boldsymbol{x}^{IE})+\boldsymbol{\delta},y)\geqslant 0)\end{aligned} \tag{39}$$

最后，我们为公式(14)加上条件2作为约束条件. 证毕.

附录B.

B.1 对抗攻击的参数设置

所选对抗攻击在四种数据集上的参数设置和针对所选分类器的攻击成功率如表13所示.

B.2 基于坐标下降的调参方法

基于坐标下降的调参方法是指每次迭代只优化一个参

表13 12种对抗攻击在CIFAR-10、MNIST、STL-10和Caltech256上的参数设置和攻击成功率

数据集	攻击	范数	最大扰动强度	优化器	步长/学习率	迭代次数	攻击成功率
CIFAR-10	AdvGAN	L_2	×	Adam	10^{-3}	150	94.4%
	RGAN	L_2	×	Adam	10^{-3}	150	93.5%
	FGSM	L_∞	8/255	signGD	8/255	1	87.4%
	BIM	L_∞	8/255	signGD	1/255	10	100.0%
	PGD	L_∞	8/255	signGD	2/255	7	100.0%
	CW	L_2	×	Adam	10^{-1}	100	100.0%
	SF	L_0	×	LinearSolver	×	×	98.6%
	PGD_0	L_0	50	GD	60 000/255	20	98.0%
	TSAA	L_0	×	Adam	10^{-3}	150	69.7%
	Square	L_∞	0.3	随机搜索	×	5000	100.0%
	SimBA	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	95.7%
	SimBA-DCT	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	95.7%
MNIST	AdvGAN	L_2	×	Adam	10^{-3}	150	90.3%
	RGAN	L_2	×	Adam	10^{-3}	150	96.9%
	FGSM	L_∞	32/255	signGD	32/255	1	64.9%
	BIM	L_∞	32/255	signGD	1/255	36	80.9%
	PGD	L_∞	32/255	signGD	1/255	40	90.2%
	CW	L_2	×	Adam	10^{-2}	200	97.7%
	SF	L_0	×	LinearSolver	×	×	97.5%
	PGD_0	L_0	50	GD	60 000/255	20	97.1%
	TSAA	L_0	×	Adam	10^{-3}	150	99.2%
	Square	L_∞	0.3	随机搜索	×	5000	99.2%
	SimBA	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	80.6%
	SimBA-DCT	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	73.3%
STL-10	AdvGAN	L_2	×	Adam	10^{-3}	150	92.2%
	RGAN	L_2	×	Adam	10^{-3}	150	90.7%
	FGSM	L_∞	8/255	signGD	8/255	1	81.2%
	BIM	L_∞	8/255	signGD	1/255	10	99.9%
	PGD	L_∞	8/255	signGD	2/255	5	99.9%
	CW	L_2	×	Adam	10^{-3}	200	99.4%
	SF	L_0	×	LinearSolver	×	×	91.1%
	PGD_0	L_0	50	GD	60 000/255	20	94.5%
	TSAA	L_0	×	Adam	10^{-3}	150	66.8%
	Square	L_∞	0.1	随机搜索	×	5000	99.8%
	SimBA	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	95.4%
	SimBA-DCT	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	94.6%
	AdvGAN	L_2	×	Adam	10^{-3}	150	92.2%
Caltech-256	AdvGAN	L_2	×	Adam	10^{-3}	150	91.8%
	RGAN	L_2	×	Adam	10^{-3}	150	95.2%
	FGSM	L_∞	16/255	signGD	16/255	1	95.3%
	BIM	L_∞	16/255	signGD	1/255	20	100.0%
	PGD	L_∞	16/255	signGD	2/255	10	100.0%
	CW	L_2	×	Adam	10^{-3}	200	100.0%
	SF	L_0	×	LinearSolver	×	×	84.6%
	PGD_0	L_0	50	GD	60 000/255	20	93.0%
	TSAA	L_0	×	Adam	10^{-3}	150	62.3%
	Square	L_∞	0.05	随机搜索	×	5000	99.3%
	SimBA	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	91.1%
	SimBA-DCT	L_2	$0.2\times\sqrt{5000}$	设计的查询算法	×	5000	82.7%

注：符号“×”表示不适用

数(对应一个坐标)，而固定其他参数. 通过逐个优化每个参数逐步逼近最优解，即

$$\begin{cases}\lambda_{pert}^{*}=\arg\min_{\lambda_{pert}}\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert},\lambda_{ie},\lambda_{\mu},\lambda_{\sigma},\rho,N),\\ \lambda_{ie}^{*}=\arg\min_{\lambda_{ie}}\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie},\lambda_{\mu},\lambda_{\sigma},\rho,N),\\ \lambda_{\mu}^{*}=\arg\min_{\lambda_{\mu}}\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu},\lambda_{\sigma},\rho,N),\\ \lambda_{\sigma}^{*}=\arg\min_{\lambda_{\sigma}}\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma},\rho,N),\\ \rho^{*}=\arg\min_{\rho}\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma}^{*},\rho,N),\\ N^{*}=\arg\min_{N}\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma}^{*},\rho^{*},N),\end{cases}\tag{40}$$

该过程满足

$$\begin{aligned}&\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert},\lambda_{ie},\lambda_{\mu},\lambda_{\sigma},\rho,N)\\ &\geqslant\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie},\lambda_{\mu},\lambda_{\sigma},\rho,N)\\ &\geqslant\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu},\lambda_{\sigma},\rho,N)\\ &\geqslant\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma},\rho,N)\\ &\geqslant\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma}^{*},\rho,N)\\ &\geqslant\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma}^{*},\rho^{*},N)\\ &\geqslant\mathcal{L}'(\boldsymbol{x}^{IE},y;\lambda_{pert}^{*},\lambda_{ie}^{*},\lambda_{\mu}^{*},\lambda_{\sigma}^{*},\rho^{*},N^{*}).\end{aligned}\tag{41}$$

该方法的核心思想是贪心算法，当参数量较大时，有着调参速度快的优点，能在参数空间内快速地寻找到较优的参数.

B.3　HTID的参数消融实验

我们预先设置HTID的优化器为Adam，迭代次数为500，学习率为10^{-2}，并对参数λ_{pert}、λ_{ie}和λ_{adv}进行消融.

(1) 参数λ_{pert}. 参数λ_{pert}对应了损失$\left\|\frac{1}{2}(\tanh(\boldsymbol{w})+1)-\boldsymbol{x}\right\|_p$在$\mathcal{L}(\boldsymbol{w},y)$中的重要性，主要影响着免疫样本的视觉质量. 我们预先设置$\lambda_{ie}=\lambda_{adv}=1.0$并分别测试$\lambda_{pert}$为0.0、1.0和10.0时，HTID的性能. 实验结果如表14所示. 实验结果表明，随着λ_{pert}值的增加，免疫样本的准确率和免疫率呈现下降或不变的趋势，而SSIM呈现上升趋势. 这是因为当λ_{pert}增加时，控制视觉质量的损失$\left\|\frac{1}{2}(\tanh(\boldsymbol{w})+1)-\boldsymbol{x}\right\|_p$的比重上升，而其他损失的比重相对下降，导致了除SSIM在上升外，其他指标均下降或不变. 由于免疫样本的免疫率至关重要，所以我们选择免疫率最高时λ_{pert}的值，即$\lambda_{pert}=0.0$. 此时，在两个数据集上，免疫样本的平均准确率均达到100%，平均免疫率分别为100.0%和68.0%，平均SSIM分别为0.871和0.926. 可以发现，即使

表14　HTID在不同的λ_{pert}值下的分类准确率(ACC，%)、免疫率(IR，%)和SSIM

数据集		CIFAR-10			MNIST		
λ_{pert}		0.0	1.0	10.0	0.0	10	10.0
AdvGAN	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.2
	IR	**100.0**	97.9	13.1	**84.2**	57.8	−0.9
	SSIM	0.825	0.993	**0.998**	0.901	0.949	**0.986**
RGAN	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	100.0	98.9	4.8	**99.4**	79.4	−0.2
	SSIM	0.810	0.992	**0.998**	0.908	0.980	**0.997**
FGSM	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	100.0	88.6	54.2	**75.0**	64.4	−7.6
	SSIM	0.879	0.990	**0.998**	0.938	0.976	**0.996**
BIM	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	**100.0**	89.9	0.0	**41.0**	17.8	−3.5
	SSIM	0.880	0.972	**0.995**	0.932	0.973	**0.996**
PGD	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	**100.0**	79.6	0.0	**55.4**	3.8	−2.0
	SSIM	0.878	0.973	**0.991**	0.912	0.967	**0.992**
CW	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.8
	IR	**100.0**	65.2	0.0	**52.9**	27.3	−2.1
	SSIM	0.932	0.994	**0.999**	0.966	0.990	**0.999**

注：加粗字体表示最佳数据

$\lambda_{pert}=0.0$，免疫样本依旧有着良好的视觉质量．这是由于公式(5)拥有较小的梯度，即$\frac{\mathrm{d}\boldsymbol{x}^{IE}}{\mathrm{d}\boldsymbol{w}}=\frac{1}{2}(-\tanh^2(\boldsymbol{w})+1)$，这导致了$\boldsymbol{w}$的变化将引起$\boldsymbol{x}^{IE}$较小程度的变化．这种低灵敏度使得优化出的免疫扰动更加细微，从而使得免疫样本拥有较高的视觉质量．(2)参数λ_{ie}. 参数λ_{ie}对应了损失$\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)$在$\mathcal{L}(\boldsymbol{w},y)$中的重要性，主要影响着免疫样本的分类准确率．我们预先设置$\lambda_{pert}=0.0$和$\lambda_{adv}=1.0$并分别测试λ_{ie}为0.0、1.0、10.0和100.0时，HTID的性能．实验结果如表15所示．实验结果表明，随着λ_{ie}值的增加，免疫样本的分类准确率和SSIM分别呈现上升和下降趋势．值得注意的是，免疫样本在CIFAR-10上的免疫率均达到了100%从而无法观察出免疫率的变化趋势．然而，免疫样本在MNIST上的免疫率呈现先上升后下降的趋势，这是由于$\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)$和$\mathcal{J}\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right),y\right)$之间存在耦合．一般地，当$\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)$在减少时，$\mathcal{J}\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right),y\right)$也会相应地减少，所以我们在增加$\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)$的权重$\lambda_{ie}$至临界值前，也间接地增加了$\mathcal{J}\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right),y\right)$的权重，所以免疫率有所上升．然而，当$\lambda_{ie}$增加至临界值后，随着$\lambda_{ie}$的增加，优化过程将极端地关注$\mathcal{J}\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1),y\right)$损失，而几乎不关注其他损失，从而导致免疫率下降．同样地，我们更加关注免疫率这一指标，在两个数据集上，我们均设置$\lambda_{ie}=10.0$. 此时，在两个数据集上，免疫样本的平均准确率均达到100%，平均免疫率分别为100.0%和69.5%，平均SSIM分别为0.863和0.924. (3)参数λ_{adv}. 参数λ_{adv}对应了损失$\mathcal{J}\left(g\left(\frac{1}{2}(\tanh(\boldsymbol{w})+1)\right),y\right)$在$\mathcal{L}(\boldsymbol{w},y)$中的重要性，主要影响着免疫样本的免疫率．我们预先设置$\lambda_{pert}=0.0$和$\lambda_{ie}=10.0$并分别测试λ_{adv}为0.0、1.0和10.0时，HTID的性能．实验结果如表16所示．实验结果表明，随着λ_{adv}值的增加，免疫样本的分类准确率均保持在100.0%，SSIM逐渐降低．当$\lambda_{adv}=0.0$时，优化过程不关注对抗样本的损失从而导致免疫率较差．当$\lambda_{adv}=1.0$时，两个数据集上的免疫率达到最高．当$\lambda_{adv}=10.0$时，免疫率保持不变或降低．我们取$\lambda_{adv}=1.0$以获取最佳的免疫率．此时，在两个数据集上，免疫样本的平均准确率均达到100%，平均免疫率分别为100.0%和69.5%，平均SSIM分别为0.863和0.924. 综上所述，在白盒免疫防御中，我们设置HTID的参数$\lambda_{pert}=0.0$、$\lambda_{ie}=10.0$和$\lambda_{adv}=1.0$以获取最佳的免疫率．此时，在两个数据集上，免疫样本的平均准确率均达到100%，平均免疫率分别为100.0%和69.5%，平均SSIM分别为0.863和0.924.

表15 HTID在不同的λ_{ie}值下的分类准确率(ACC，%)、免疫率(IR，%)和SSIM

数据集		CIFAR-10				MNIST			
λ_{ie}		0.0	1.0	10.0	100.0	0.0	1.0	10.0	100.0
AdvGAN	ACC	78.6	**100.0**	**100.0**	**100.0**	99.1	**100.0**	**100.0**	**100.0**
	IR	**100.0**	**100.0**	**100.0**	**100.0**	83.2	84.2	**84.4**	82.6
	SSIM	**0.832**	0.825	0.822	0.819	**0.902**	0.901	0.900	0.900
RGAN	ACC	77.8	**100.0**	**100.0**	**100.0**	99.6	**100.0**	**100.0**	**100.0**
	IR	**100.0**	**100.0**	**100.0**	**100.0**	99.2	99.4	**99.5**	98.9
	SSIM	**0.815**	0.810	0.806	0.803	**0.909**	0.908	0.908	0.908
FGSM	ACC	83.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	100.0	100.0	100.0	100.0	74.0	75.0	**75.8**	72.9
	SSIM	**0.880**	0.879	0.874	0.871	**0.940**	0.938	0.936	0.934
BIM	ACC	75.2	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	**100.0**	**100.0**	**100.0**	**100.0**	40.0	41.0	**41.5**	39.2
	SSIM	**0.883**	0.880	0.879	0.872	**0.935**	0.932	0.930	0.928
PGD	ACC	71.3	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	**100.0**	**100.0**	**100.0**	**100.0**	53.4	55.4	**55.7**	53.2
	SSIM	**0.880**	0.878	0.874	0.866	**0.913**	0.912	0.912	0.911
CW	ACC	66.4	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	**100.0**	**100.0**	**100.0**	**100.0**	22.2	52.9	**60.1**	56.0
	SSIM	**0.942**	0.932	0.925	0.920	**0.974**	0.966	0.956	0.948

注：加粗字体表示最佳数据

表16　HTID在不同的λ_{adv}值下的分类准确率(ACC,%)、免疫率(IR,%)和SSIM

数据集		CIFAR-10			MNIST		
λ_{adv}		0.0	1.0	10.0	0.0	1.0	10.0
AdvGAN	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	7.8	**100.0**	**100.0**	−0.9	**84.4**	83.9
	SSIM	**0.910**	0.822	0.819	**0.957**	0.900	0.890
RGAN	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	3.7	**100.0**	**100.0**	12.2	**99.5**	99.4
	SSIM	**0.815**	0.806	0.801	**0.929**	0.908	0.903
FGSM	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	78.3	**100.0**	**100.0**	**18.2**	**75.8**	74.1
	SSIM	**0.891**	0.874	0.858	**0.947**	0.936	0.935
BIM	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	49.1	100.0	100.0	−5.1	**41.5**	40.0
	SSIM	**0.898**	0.879	0.874	**0.952**	0.930	0.929
PGD	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	67.5	**100.0**	**100.0**	1.2	**55.7**	55.4
	SSIM	**0.892**	0.874	0.857	**0.953**	0.912	0.912
CW	ACC	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	IR	83.2	**100.0**	**100.0**	57.0	**60.1**	53.6
	SSIM	**0.936**	0.925	0.922	**0.979**	0.956	0.953

注：加粗字体表示最佳数据

附录C.

HTID和MID在CIFAR-10和MNIST上制作的免疫样本如图10所示. 在STL-10和Caltech-256上使用MID针对AdvGAN制作的免疫样本如图11所示.

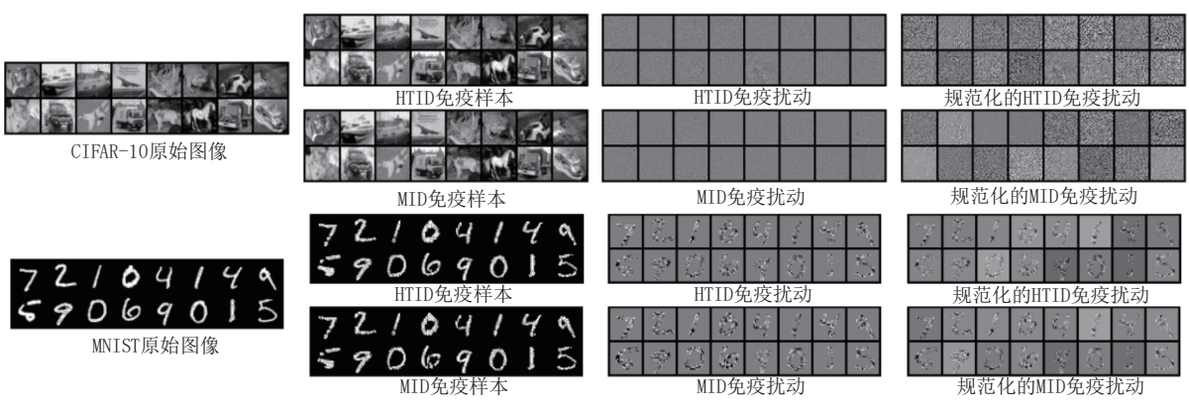

图10　HTID和MID在CIFAR-10和MNIST上制作的免疫样本

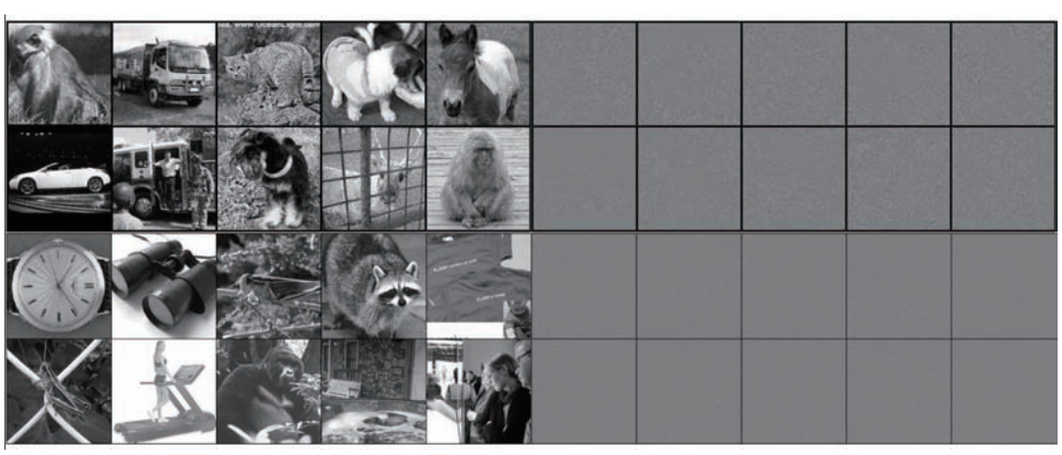

图11　在STL-10和Caltech-256数据集上使用MID针对AdvGAN制作的免疫样本(左上图：STL-10上的免疫样本，右上图：对应的免疫扰动. 左下图：Caltech-256上的免疫样本，右下图：对应的免疫扰动)

WU Hao, M. S. His current research interests focus on AI security, deep learning, and image processing.

WANG Jin-Wei, Ph. D., professor. His research interests include AI security, multimedia forensics, and multimedia information hiding.

LUO Xiang-Yang, Ph. D., professor. His research interests are image steganography and steganalysis.

MA Bin, Ph.D., professor. His research interests include reversible data hiding, multimedia security, and image processing.

Background

The problem addressed in this paper falls within the field of adversarial defense and aims to prevent the generation of adversarial examples. Currently, related works have made significant progress in adversarial defense, e. g., adversarial training, defensive distillation, input preprocessing, detection of adversarial examples, and gradient obfuscation. However, preventing the generation of adversarial examples is still a challenge that needs to be overcome.

This study aims to advance the current state of the field by crafting immune examples for images to prevent the generation of adversarial examples. We define the immune example. In brief, it is the image after adding immune perturbations to the original image, and immune perturbations can invalidate the adversarial perturbations, thereby protecting the image and the deep neural networks. Through HTID and MID, this research endeavors to achieve immune examples with excellent classification accuracy, high visual quality, strong defense performance, and outstanding transferability.

The research team involved in this study has a strong track record in the field of adversarial attack and defense. Our previous works have encompassed the detection of adversarial examples, the generation of adversarial examples, and the generation of enhanced samples, which have been published in IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, International Workshop on Digital-forensics and Watermarking-2020/2022, China Multimedia-2021, International Conference on Information Security Practice and Experience-2021, and Multimedia Systems. These accomplishments have established the team's expertise and credibility in the field and have paved the way for the current study.

It is worth noting that this research aligns with the objectives and initiatives of the national project, i. e., research on the key technology of hyper complex number deep forensics for color images. The participation of this study in the project not only strengthens its research foundation but also underscores its importance and relevance within the research landscape of adversarial defense. In addition, this study was supported by the National Natural Science Foundation of China (No. 62072250, 62172435, U1804263, U20B2065, 61872203, 71802110, 61802212), the Zhongyuan Science and Technology Innovation Leading Talent Project of China (No. 214200510019), Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (No. HNTS2022002).

Overall, this study aims to contribute to the efforts in preventing the generation of adversarial examples in the field of adversarial defense. By building upon the achievements of the research team and aligning with the national project, this research strives to make a significant impact on adversarial defense.