E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391.41; TP18　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)01-0062-20

论文引用格式:Zhang J W, Wu H, Yuan L, Wang H, Cheng X, Luo X Y, Ma B and Wang J W. 2026. Adversarial example generation, defense, and application. Journal of Image and Graphics, 31(1):0062-0081(张家伟, 吴昊, 袁霖, 王昊, 程鑫, 罗向阳, 马宾, 王金伟. 2026. 对抗样本生成、防御与应用研究综述. 中国图象图形学报, 31(1):0062-0081)[DOI:10.11834/jig.240732]

对抗样本生成、防御与应用研究综述

张家伟[1],吴昊[2],袁霖[1],王昊[3*],程鑫[4],罗向阳[2],马宾[5],王金伟[6]

1. 重庆邮电大学网络空间安全与信息法学院,重庆　400065; 2. 数学工程与先进计算国家重点实验室,郑州　450001;
3. 淮阴工学院计算机与软件工程学院,淮安　223003; 4. 南京信息工程大学计算机学院、网络空间安全学院,南京　210044;
5. 齐鲁工业大学网络空间安全学院,济南　250353; 6. 南开大学密码与网络空间安全学院,天津　300350

摘　要: 得益于愈发优异的性能,深度神经网络越来越多地应用在日常生活的各个方面,成为各领域中主导决策或辅助决策的重要角色。然而,对抗样本的发现对深度神经网络的可靠性和安全性构成重大威胁,极大限制了其在高度关注安全性的领域的应用。因此,研究者针对对抗样本展开了诸多研究。本文首先介绍对抗样本的研究背景及术语定义,并将现有的对抗攻击技术分成白盒攻击和黑盒攻击,其中白盒攻击包括基于优化、梯度和生成3种攻击,黑盒攻击包括基于迁移、查询两种攻击;其次,详细介绍现有的3种对抗防御策略,分别为对抗训练、对抗检测和对抗去噪;接着,列举对抗样本应用的两种典型技术,包括对抗隐写和对抗水印;最后,指出现有对抗样本在攻击、防御研究中存在的不足之处和值得进一步探索的领域。

关键词: 对抗样本(AE);对抗攻击;对抗防御;对抗隐写;对抗水印

Adversarial example generation, defense, and application

Zhang Jiawei[1], Wu Hao[2], Yuan Lin[1], Wang Hao[3*], Cheng Xin[4], Luo Xiangyang[2], Ma Bin[5], Wang Jinwei[6]

1. School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing　400065, China;
2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou　450001, China;
3. Department of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian　223003, China;
4. School of Computer Science, School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing　210044, China; 5. Department of Cyber Security, Qilu University of Technology, Jinan　250353, China;
6. College of Cryptology and Cyber Science, Nankai University, Tianjin　300350, China

Abstract: Due to their excellent performance, deep neural networks are increasingly utilized in various aspects of daily life, playing a crucial role in decision-making and providing assistance across various fields. However, the discovery of adversarial example poses a considerable threat to the reliability and security of deep neural networks, markedly limiting their application in safety-critical areas. Therefore, in recent years, researchers have conducted extensive studies on adver-

收稿日期:2024-12-16;**修回日期**:2025-05-20;**预印本日期**:2025-05-27

*通信作者:王昊　sa875923372@163.com

基金项目:国家重点研发计划资助(2021QY0700);国家自然科学基金项目(U24B20179,62472229,62072250,U23A20305,U23B2022,62371145,62072480,62172435,62302249,62272255,62302248,U20B2065);省部共建藏语智能信息处理及应用国家重点实验室(藏文信息处理教育部重点实验室)开放课题项目(N2024-Z-003);中原科技创新领军人才项目(214200510019);江苏省自然科学基金项目(BK20200750);江苏省研究生研究与实际创新项目(KYCX24_1513)

Supported by: National Key R&D Program of China (2021QY0700); National Natural Science Foundation of China (U24B20179, 62472229, 62072250, U23A20305, U23B2022, 62371145, 62072480, 62172435, 62302249, 62272255, 62302248, U20B2065)

sarial examples. This paper first introduces the research background and definitions related to adversarial examples, and categorizes existing adversarial attack techniques into white-box and black-box attacks. White-box attacks assume full transparency of the target model during the generation of adversarial examples, where the attack algorithm has access to the model's architecture, parameters, loss function, and gradients to craft adversarial examples. Black-box attacks refer to scenarios where the target model is treated as opaque during the generation of adversarial examples. The attack algorithm has no access to internal information, such as the model's architecture or parameters, and can only interact with the model through input-output queries. In contrast, white-box attacks can be categorized into three types based on their approach: optimization-based, gradient-based, and generation-based attacks. Specifically, optimization-based attacks formulate the generation of adversarial examples as a constrained optimization problem, which is approximately solved using optimization algorithms. Their main advantage lies in realizing high attack success rates with relatively small perturbations; however, they often incur substantial computational and time costs. Gradient-based attacks generate adversarial examples by computing perturbations in the direction of the loss function's gradient, scaled by a predefined step size. These methods notably improve the efficiency of adversarial example generation but require continuous access to the model's gradients. Meanwhile, generation-based attacks aim to produce adversarial examples by designing and training a generative model to attack the target model. Once trained, the generative model can efficiently generate adversarial examples without requiring access to the gradients or internal parameters of the target model, making it highly suitable for large-scale adversarial example generation. Black-box attacks can be categorized into two types: transfer-based and query-based attacks. Specifically, transfer-based attacks exploit the transferability of adversarial examples across different models by generating adversarial inputs against a surrogate model, which are then used to attack the target black-box model. The main advantage of this approach is that it does not require access to the target model or its internal information, resulting in low computational cost. However, differences between the surrogate and target models can lead to instability in attack performance. Query-based attacks work by gradually refining adversarial examples through repeated interactions with the target model. By submitting inputs and analyzing the corresponding outputs, these attacks aim to mislead the model without requiring any knowledge of its internal architecture. While they often achieve high success rates, their efficiency is hindered by the substantial number of queries required. With the continuous development of adversarial example generation algorithms, an increasing number of defense methods have emerged. The most widely adopted defense techniques can generally be categorized into three groups: adversarial training, adversarial detection, and adversarial denoising. Specifically, adversarial training is considered one of the most effective defense strategies against adversarial attacks. This training involves incorporating adversarial examples into the training process of the target model to enhance its robustness against adversarial perturbations. This approach is generally applicable and does not rely on specific attack methods; however, it suffers from high computational cost. In contrast to adversarial training, adversarial detection does not adjust the parameters of the target model. Instead, this detection introduces an auxiliary classifier to determine whether an input is an adversarial example. This approach offers good integrability with existing models; however, its generalization capability is often limited. Unlike adversarial training and adversarial detection, adversarial denoising focuses on eliminating adversarial perturbations from input data to prevent them from misleading the target model. The main advantage of adversarial denoising lies in the potential to restore adversarial examples to their original form; however, this process can occasionally degrade the quality of clean, unaltered inputs. Adversarial example techniques not only reveal the vulnerabilities of deep learning models but also inspire innovative applications in security-sensitive tasks. For instance, adversarial perturbations have been utilized to enhance system resilience in scenarios such as adversarial steganography, adversarial watermarking, model protection, and multimedia authentication. These cross-task explorations introduce new directions for adversarial example research and contribute to improving the practical deployment of artificial intelligence (AI) systems in complex environments. Therefore, this paper reviews various applications of adversarial examples within the fields of steganography and digital watermarking. Furthermore, a diverse set of adversarial example generation methods is selected and systematically evaluated in terms of attack success rate, robustness, and transferability across different target network architectures and benchmark datasets, including Caltech-256 and ImageNet. Finally, the study identifies current limitations in existing adversarial example research and highlights promising directions for future exploration, such as cross-domain adversarial attacks,

robustness evaluation, real-world attack scenarios, and the development of systematic defense frameworks.

Key words: adversarial example (AE); adversarial attack; adversarial defense; adversarial steganography; adversarial watermark

0 引 言

随着多媒体内容的爆炸式增长及计算机软、硬件性能的不断提高,人工智能等相关技术突飞猛进。其中,以深度神经网络(deep neural network, DNN)为代表的人工智能算法,在自动驾驶汽车、行人识别和刷脸支付等诸多与生活息息相关的领域中广泛投入使用,取得了令人瞩目的成果。虽然深度神经网络可以通过在大量数据上迭代训练以获取优秀的精度,但其认识、理解事物的方式仍然与人类存在着本质差异。由于越来越复杂的网络结构和愈加庞大的网络参数,深度神经网络表现得越发像一个"黑匣子"。

尽管其能够通过迭代训练学习到数据中存在的高度抽象特征,但模型的内部结构和决策过程对于人类而言是不透明且难以理解的。而且,由于当下深度神经网络仍然缺乏强有力的可解释性,这极大地限制了其在医疗、金融和司法等高度关注安全性的领域中的应用。同时,进一步的研究表明,原始样本在添加一些微弱的扰动后会形成对抗样本(adversarial example, AE),对抗样本能够以极高的成功率误导深度神经网络,使其做出错误的判断。如图1所示,当向原始样本中添加人眼几乎不可察觉的扰动后,深度神经网络会错误地将浣熊识别为骆驼。研究者将这种特殊的微弱扰动称为对抗扰动。

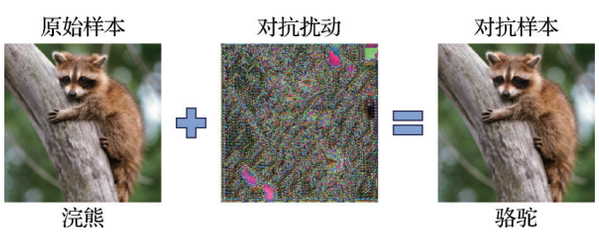

图1 对抗样本生成示意图

Fig. 1 Diagram of adversarial example generation

自对抗样本发现以来,越来越多的对抗样本生成算法不断提出,甚至部分算法已经能够生成应用于现实世界的对抗样本。如图2所示,研究人员通过在"停止"的交通指示牌上添加具有特定分布的对抗扰动,使得在多角度和距离条件下,汽车的道路识别神经网络将其识别为"限速40"的交通指示牌。如图3所示,研究人员将对抗样本生成技术与3D打印技术相结合,制作了一只"乌龟"的3D模型,而这个"乌龟"在多角度下均被深度神经网络识别为"步枪"。以上示例表明,对抗样本不仅只存在于实验室环境和数字世界中,它们已经能够直接应用在现实世界中并导致一系列严重的安全问题。

为了解决对抗样本带来的安全威胁并提升深度神经网络的可靠性,研究者针对对抗样本形成的原因进行诸多探索。图4通过决策边界的差异展示了对抗样本存在的本质原因。图中金色的实线代表真实决策边界或人类的决策,在该决策下A类样本和B类样本能够被完全正确地区分。类似地,黑色的虚线代表通过大量数据训练后的深度神经网络的决策边界。通常情况下,该决策边界也能够较好地区分A类样本和B类样本。然而,当A类和B类样本添加对抗扰动后,它们会越过训练决策边界,从而导致深度神经网络误判。

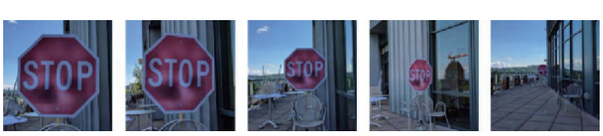

图2 多角度多光照条件下的道路指示牌对抗样本(Eykholt等,2018)

Fig. 2 Adversarial examples of road signs under multiple angles and multiple lighting conditions (Eykholt et al., 2018)

图3 3D打印技术构建的实体对抗样本(Athalye等,2018)

Fig. 3 Solid adversarial examples constructed by 3D printing technology (Athalye et al., 2018)

基于这一发现,研究者提出对抗训练,将对抗样本加入到深度神经网络的训练数据中,以此促进训练决策边界向真实决策边界逼近,从而提升深度神经网络的准确性。除此之外,对抗样本也被发现能与传统隐写术、数字水印结合,以提升隐写算法的抗隐写分析能力和数字水印的抗去除能力。因此,针

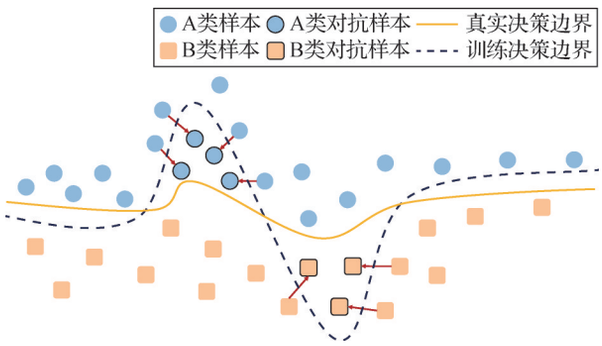

图4　对抗样本存在于真实决策边界与训练决策边界之间的差异空间

Fig. 4　Adversarial examples exist in the difference space between the real decision boundary and the training decision boundary

对对抗样本生成的相关研究不仅可以促进安全可靠深度神经网络的构建和实际应用，同时也可以作为一种全新的保护手段，在人工智能算法盛行的当下，进一步提升现有多媒体信息传递的安全性。综上，针对对抗样本的研究对于推动人工智能安全和多媒体信息安全领域的发展有着重大的意义和价值。

图5展示了对抗样本的研究领域。现有对抗样本研究主要分为对抗样本生成、对抗样本防御和对抗样本应用。对抗样本生成旨在通过不同的算法构建对抗样本，以此迷惑深度神经网络；对抗样本防御旨在利用不同的防御策略消除对抗样本给深度神经网络带来的负面影响；对抗样本应用旨在将对抗样本生成技术与诸如隐写或水印等技术结合，以提升算法的安全性、鲁棒性等。

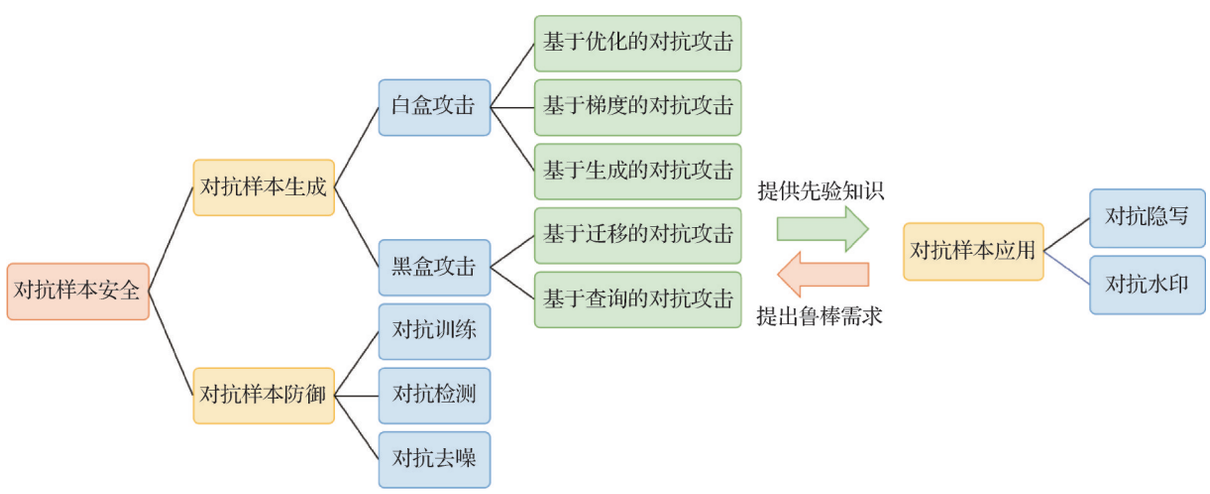

图5　对抗样本研究领域

Fig. 5　The field of adversarial example research

1　术语定义

与对抗样本相关的术语定义如下：

1）原始样本。泛指数据集中的干净图像、文本或其他类型数据，其未受到对抗扰动的干扰。通过训练的深度学习模型能够在该类数据上取得较高的精度表现。本文主要针对以图像为载体的对抗样本生成，因此本文中的原始样本及对抗样本等一系列术语都指代数字图像数据。

2）对抗扰动或对抗噪声。泛指对抗攻击方法所生成的具备特定分布的噪声。将该噪声添加到原始样本上会形成对抗样本。

3）通用对抗扰动。要求生成的某一特定分布的对抗扰动添加到任意的原始样本上都能误导深度学习模型。

4）扰动强度或扰动幅度。指对抗扰动在L_p范数度量下的大小。常用的范数有0范数、2范数和无穷范数。

5）不可察觉性、视觉质量。除了扰动强度、幅度外，现有研究也尝试通过定量的指标来衡量对抗扰动的不可察觉性或视觉质量。

6）对抗样本。泛指添加了对抗扰动的原始样本。将对抗样本输入通过原始样本训练的深度学习模型时，模型精度会发生大幅度退化。

7）攻击成功率。以错误分类的对抗样本占总样

本数量的百分比衡量对抗样本的攻击能力。

8)目标模型或威胁模型。通常指代对抗样本所误导的深度学习模型。

9)对抗攻击。泛指生成对抗样本的算法及过程。根据攻击是否使深度学习模型将对抗样本判断为指定结果,将其进一步分为有目标攻击和无目标攻击。根据攻击过程中是否能获取目标模型的参数,将其进一步分为白盒攻击和黑盒攻击。

10)有目标攻击。要求生成的对抗样本输入目标模型后,模型将输出某一指定的结果。

11)无目标攻击。要求生成的对抗样本输入目标模型后,模型将输出不同于输入原始样本的结果。通常,无目标攻击的难度低于有目标攻击。

12)白盒攻击。指目标模型在对抗样本的生成过程中是透明的,对抗攻击算法可以获取模型的结构、参数、损失和梯度等信息以生成对抗样本。主流的白盒攻击方法主要分为基于优化的攻击、基于梯度的攻击和基于生成的攻击。

13)黑盒攻击。指目标模型在对抗样本的生成过程中是不透明的,对抗攻击算法无法获取除模型输入数据和输出数据以外的任何信息。通常,黑盒攻击的难度高于白盒攻击,主流的黑盒攻击方法主要分为基于迁移性的攻击和基于查询的攻击。

14)查询次数。衡量了基于查询的攻击所需要获取模型对某一输入数据的输出结果的次数。当查询次数越少时,该黑盒攻击算法的效率越高。

15)代理模型。泛指结构和参数不同于目标模型的深度学习模型。该模型通常用于在目标模型未知或不可获取的条件下辅助对抗样本的生成,以完成对于目标模型的黑盒攻击。通常,代理模型可以是一个深度学习模型或多个深度学习模型的集成。

16)迁移性。衡量了针对某一个模型生成的对抗样本对于其他模型的泛化能力。对抗样本的迁移性越高则对其他模型的攻击成功率越高。生成具备高迁移性的对抗样本是达成黑盒攻击的一种重要方法和手段。

17)鲁棒性。衡量了对抗样本在经历诸如JPEG(Joint Photographic Experts Group)压缩等图像处理操作或旋转、裁剪等几何变换后的攻击能力。

18)对抗防御。泛指防止对抗样本对深度学习模型造成负面影响的算法。目前主流的对抗防御方法可分为对抗训练、对抗检测和对抗去噪。

2 数据集与评价指标

2.1 数据集

由于算力限制,早期的对抗样本研究主要集中在小尺寸的分类数据集上,诸如手写数字数据集(modified national institute of standards and technology, MNIST)、加拿大高级研究院10分类数据集(Canadian Institute for Advanced Research-10 classes, CIFAR-10)等。其中,MNIST包含70 000幅尺寸为28 × 28像素的灰度手写数字图像,这些图像由0～9共计10种类别构成,60 000幅为训练集,10 000幅为测试集。CIFAR-10包含60 000幅尺寸为32 × 32像素的彩色图像,由飞机、汽车、鸟、猫等共计10种类别构成,50 000幅为训练集,10 000幅为测试集。

随着研究的深入,研究者将对抗样本扩展到尺寸更大的数据集上,诸如ImageNet、美国加州理工学院256类数据集(California Institute of Technology 256 object categories, Caltech-256)等。ImageNet随着不断的更新迭代,目前已经由超过1 000万幅尺寸约300 × 300像素的图像构成,共分为1 000个类别。Caltech-256则由30 607幅尺寸约370 × 320像素的图像构成,共计257个类别。此外,研究者在不同任务上进行对抗样本研究,使用了名人人脸数据集(celebfaces attributes dataset, CelebA)、文本情感数据集(internet movie database, IMDB)和语音数据集Librispeech等。

2.2 评价指标

误分率(classification error rate, CER)通常用来衡量生成对抗样本的攻击能力,可定义为错误分类样本数量与总样本数量的比值。其值越高,说明对抗样本生成算法使得目标网络发生错误分类的样本数占总样本数的比例越高,也即该算法的攻击能力越强。

此外,峰值信噪比(peak signal-to-noise ratio, PSNR)、结构相似性(structural similarity, SSIM)和人眼感知度量(learned perceptual image patch similarity, LPIPS)(Zhang等,2018a)等指标可用来度量对抗样本的视觉质量。PSNR的取值范围为0到正无

穷，值越大表示对抗样本与原始样本越接近，对抗扰动越小，对抗样本视觉质量越好。SSIM 的取值范围为 0 到 1，值越接近 1 说明对抗样本与原始样本越接近，对抗样本视觉效果越好。

3 对抗样本生成

Szegedy 等人（2014）首次发现可以通过向图像添加人眼难以察觉的扰动使神经网络做出错误判断，并将添加了对抗扰动的样本称为对抗样本，同时这种生成对抗样本的方法称为对抗攻击。如今，新的对抗攻击方法不断提出，根据不同的攻击目标，对抗攻击可以分为有目标攻击和无目标攻击。有目标攻击下，生成的对抗样本需要使得目标神经网络模型做出预先指定类别的误判。而无目标攻击下，生成的对抗样本只需要使得目标神经网络模型做出错误的判断即可。除此之外，根据攻击时是否拥有访问目标神经网络模型的权限，对抗攻击可分为白盒攻击和黑盒攻击。在白盒攻击中，攻击算法拥有访问目标模型的权限，可以获取模型的结构、参数和梯度等信息。在黑盒攻击中，模型被视为一个“黑匣子”，攻击算法只拥有访问模型输入和输出的权限。因此，黑盒攻击的攻击难度通常要高于白盒攻击，而白盒攻击也能够利用对抗样本在不同模型之间的迁移性转化为黑盒算法。

3.1 白盒攻击

白盒攻击根据不同的攻击方法可进一步划分为基于优化的攻击、基于梯度的攻击和基于生成的攻击。

3.1.1 基于优化的攻击

基于优化的攻击将对抗样本生成问题转化为带约束的优化问题，并利用优化算法对其进行近似求解以生成对抗样本。其优势在于能够以较小的对抗噪声取得较高的攻击成功率，但计算开销和时间代价较大。Szegedy 等人（2014）首次提出基于优化的攻击，并进一步利用优化算法 L-BFGS（limited-memory Broyden-Fletcher-Goldfarb-Shanno）对其进行近似求解。然而 L-BFGS 算法求解效率较低，导致生成对抗样本需要消耗大量的时间。基于 L-BFGS，Carlini 和 Wagner（2017）提出一种新的基于优化的对抗样本生成方法 C&W（Carlini and Wagner），C&W 利用 tanh 和 arctanh 函数实现对样本空间的映射，以生成极难察觉的微弱扰动。同时，该扰动具有较强的攻击能力，能够误导经过防御性蒸馏的网络模型，但计算扰动所需要的时间仍然较长。为提升 C&W 方法生成对抗样本的视觉质量和鲁棒性，Zhao 等人（2020）将感知色彩距离引入到对抗样本的生成过程中，并提出两种对抗样本生成方法：感知颜色距离惩罚（perceptual color distance penalty，PerC-C&W）和感知色距交替损失（perceptual color distance alternating loss，PerC-AL）。相较于 RGB 色彩空间的对抗样本，PerC-C&W 和 PerC-AL 能够允许样本容纳更大的扰动量且不损失样本的视觉质量，然而该两种方法也同样面临优化时间复杂度过高这一问题。Zhong 等人（2022）利用阳光照射所产生的阴影构建对抗样本，相较于“贴纸式”、“涂鸦式”以及“补丁式”的方法（叶乙轩 等，2024），该攻击方案具备更好的隐蔽性。

3.1.2 基于梯度的攻击

基于梯度的攻击以损失函数的梯度为方向，按照指定步长计算对抗扰动并生成对抗样本。其优势在于能够极大地提高对抗样本生成的效率，但需要持续访问模型梯度。Goodfellow 等人（2015）首次提出快速梯度符号法（fast gradient sign method，FGSM），FGSM 根据损失函数对当前输入图像的梯度方向更新图像以生成对抗样本。虽然该方法大大提升了对抗样本的生成速度，但其扰动计算方式粗糙，导致攻击成功率较低，同时会引入过大的扰动使得原始样本发生较大的视觉质量退化。为进一步精细化扰动计算，Kurakin 等人（2018）对 FGSM 进行了优化，提出基础迭代法（basic iterative method，BIM）。BIM 将扰动计算划分为多个迭代过程，每次迭代采用较小的更新步长，使得生成的扰动在更加不易被察觉的同时也具备更高的攻击成功率。在 BIM 的基础上，越来越多针对攻击成功率、迁移性和鲁棒性的改进方法不断提出。例如，Papernot 等人（2016b）提出基于雅可比矩阵的显著图攻击（Jacobian-based saliency map attack，JSMA），该方法基于显著图选择对抗扰动添加的位置，有效地提升了攻击成功率，但时间复杂度较高。Moosavi-Dezfooli 等人（2016）提出 DeepFool，该方法利用泰勒展开式和投影公式求解生成对抗样本的最小扰动。随后，Moosavi-Dezfooli 等人（2017）进一步提出通用对抗扰动（universal adversarial perturbation，UAP），UAP 可以利用一种扰动实现对所有图像的对抗攻

击。Madry等人(2018)提出投影梯度下降法(projected gradient descent, PGD),通过引入随机初始化和扰动投影以解决对抗样本生成过程中的陷入局部最优的问题,同时PGD生成的对抗样本在加入对抗训练后能够更好地提升目标模型的鲁棒性。清华大学朱军教授的团队分别将动量(Dong等,2018)、输入多样性(Xie等,2019)以及转移不变性(Dong等,2019)引入对抗样本的生成过程中,并提出基于动量的迭代快速梯度符号法(momentum-based iterative fast gradient sign method, MI-FGSM)、多元输入迭代快速梯度符号法(diverse inputs iterative fast gradient sign method, DI^2-FGSM)以及平移不变基本迭代法(translation-invariant basic iterative method, TI-BIM),以进一步提升对抗样本在攻击能力、鲁棒性和迁移性上的表现。同样为了提升攻击能力,Rony等人(2019)提出解耦方向与范数法(decoupling direction and norm, DDN),该方法在计算扰动时将方向与范数解耦,从而在更小的扰动强度下达成更高的攻击成功率。王扬等人(2022)提出一种低感知性对抗样本生成算法,构造的对抗样本在保证较高攻击成功率的情况下具有更低的视觉感知性。Zhang等人(2024)基于人类视觉特性构造了一个噪声可见函数,以动态调整不同图像区域的对抗扰动强度,从而生成具备较好视觉质量和较强鲁棒性的对抗样本。

3.1.3 基于生成的攻击

基于生成的攻击通过构造并训练一个生成式模型来攻击目标模型以生成对抗样本。其优势在于生成式模型训练完成后不需要目标模型梯度等信息就能以极高的效率执行攻击,适合大规模生成对抗样本。Poursaeed等人(2018)基于Unet结构和残差结构提出两种生成对抗扰动方法:GAP-U(generative adversarial perturbations-Unet)和GAP-R(generative adversarial perturbations-resnet),以构造通用对抗样本和非通用对抗样本。Xiao等人(2018)提出基于生成对抗网络(generative adversarial network, GAN)(Goodfellow等,2020)的对抗攻击(adversarial GAN, AdvGAN)。架构AdvGAN通过构建生成器、判别器和目标分类器学习原始样本和对抗样本分布之间的映射,以此生成效果更逼真、质量更高的对抗样本。Jandial等人(2019)将目标模型的特征提取器获取的隐层向量作为GAN的输入,进一步提出AdvGAN的改进算法AdvGAN++。AdvGAN++中的先验特征赋予了其在多数据集多网络结构下更强的对抗攻击能力。Liu和Hisieh(2019b)提出基于谱归一化生成对抗网络(spectral normalization GAN, SNGAN)框架(Miyato等,2018)的鲁棒生成对抗网络(robust GAN, RobGAN)。RobGAN通过两个阶段训练出更鲁棒的判别器,以此提升生成器的鲁棒性,从而进一步提升生成对抗样本的质量。Lu等人(2021)提出基于对称凸度的自动编码器方法(symmetric saliency-based auto-encoder, SSAE),去除了现有方法中的判别器设置,直接通过对生成器进行监督学习来提升生成对抗样本的视觉质量。Hu等人(2022)提出一种基于环形裁剪的可扩展生成式攻击生成对抗纹理,以完成对现实世界中行人检测器的多角度攻击。Duan等人(2022)提出一种使用多个子网络模型生成多样本的方法。Zhang等人(2023)提出基于生成对抗自编码器的可恢复对抗样本生成框架,该方法能通过对抗样本的生成与恢复来保护社交网络用户的隐私安全。万鹏等人(2024)提出一种多域特征混合增强对抗样本迁移性方法MDFM(multiple domain feature mixup),以提高对抗样本在黑盒场景下的攻击成功率。吴昊等人(2024)提出的基于矩的免疫防御借助一阶原点矩和二阶中心矩将干净样本精心制作成免疫样本,以降低样本被制作成对抗样本的可能性。Chen等人(2025)提出利用扩散模型的生成能力和判别能力来生成对抗样本。

3.2 黑盒攻击

黑盒攻击根据不同的攻击策略可进一步划分为基于迁移性的攻击和基于查询的攻击。

3.2.1 基于迁移的攻击

基于迁移的攻击利用对抗样本在不同模型间的可迁移性,通过针对代理模型生成对抗样本直接攻击目标黑盒模型。其优势在于无需获取目标模型和其内部信息,计算成本低,但模型的差异性会导致该类攻击不稳定。Papernot等人(2016a)首先基于迁移性实现了对目标模型的黑盒攻击,他们通过构建一个目标模型的代理模型,并对该代理模型进行白盒攻击以生成对抗样本,再利用对抗样本在不同网络模型上的迁移性实现对于目标模型的攻击。随后,Liu等人(2017)针对对抗样本在大型模型和大规模数据集上的可迁移性进行了广泛研究。他们通过对集成模型进行白盒攻击的方式进一步提升对抗样

本的迁移性，以达到更高的黑盒攻击成功率。除此之外，Papernot 等人（2017）认为当代理模型的决策边界与目标模型越相近时，对抗样本的迁移性就会越好。他们提出利用目标模型的输入、输出对代理模型进行有监督训练，从而构建与目标模型相似的代理模型，以此提升黑盒攻击成功率。Naseer 等人（2019）通过提取域不变对抗来提升对抗样本的可迁移性。Nakka 和 Salzmann（2021）通过攻击模型的中间层特征来提升对抗样本的可迁移性。Zhang 等人（2022b）提出超越 ImageNet 攻击（beyond ImageNet attack，BIA），旨在突破现有对抗攻击在跨模型、跨数据集和跨任务的可迁移性。Wu 等人（2023）提出经验精确的 Nesterov 动量方法，该方法分配给动量一个能够预防早期过拟合的初始值，并且细化了预更新阶段，从而提升了对抗样本的可迁移性。Wang 等人（2024d）提出一种通过从高斯分布中采用模拟平滑损失函数的攻击方法，并计算了平滑损失函数的梯度增强动量以提升可迁移性。近年来，通过迁移性实现有效的黑盒攻击这一思路促使研究者不断探索诸如特征重要性（Wang 等，2021）、方差调整（Wang 和 He，2021）、全局动量初始化（Wang 等，2024a）和神经元归因（Zhang 等，2022a）等方法，进一步提升了对抗样本的迁移性。

3.2.2　基于查询的攻击

基于查询的攻击通过反复向目标模型输入样本并观察其输出，迭代优化对抗样本以误导模型。其优势在于不依赖模型结构，攻击成功率较高，但大量的查询导致效率较低。Chen 等人（2017）首先引入了基于查询的方式以进行黑盒攻击。他们提出零阶优化方法（zeroth order optimization，ZOO）并利用有限差分（finite difference，FD）估计目标模型的梯度，接着根据估计的梯度进行白盒攻击以完成对目标模型的攻击。然而，该方法在估计梯度时需要反复多次查询目标模型的输入和输出，从而会消耗大量的时间和算力。为了进一步提升梯度估计的效率，Du 等人（2018）和 Bhagoji 等人（2018）分别通过圈定攻击区域和维度组合的方式降低算法查询目标模型的次数。除此之外，Ilyas 等人（2018）提出一种基于自然进化策略（natural evolutionary strategies，NES）的梯度估计方法，显著地降低了查询目标模型的次数。Tu 等人（2019）则设计了一个基于自编码器的零阶优化方法（autoencoder-based zeroth order optimization method，AutoZOOM），进一步缩小 ZOO 算法的搜索空间，有效地提升了生成对抗样本的效率。近年来，Rahmati 等人（2020）利用决策边界几何学进行黑盒攻击，他们以更小的扰动量和更低的查询次数达到了更高的黑盒攻击成功率。Li 等人（2020a）受到压缩感知理论的启发提出投影与概率驱动的黑盒攻击（projection & probability-driven black-box attack，PPBA），他们利用低频约束传感矩阵限制问题的解空间，进一步提升了查询的效率，该方法在进行黑盒攻击时只需要查询目标模型 300～1 200 次。Shi 等人（2023）提出定制迭代和采样的黑盒攻击，进一步提升了查询的效率。不同于其他方法，Park 等人（2024）采用了硬标签来代替现有查询中常使用的软标签，并使用预训练的代理模型来指导优化与查询的过程。Vo 等人（2024）提出基于分数的黑盒稀疏对抗攻击方法，他们利用贝叶斯算法提升了查询的速度和效率，从而能够更快地评估目标模型的鲁棒性。

4　对抗样本防御

随着对抗样本生成算法的不断提出，越来越多基于不同策略的对抗防御方法（周大为 等，2024）也不断涌现。现有的主流对抗防御方法可分为对抗训练、对抗检测和对抗去噪。

4.1　对抗训练

对抗训练被认为是最强的对抗防御策略之一，该策略将对抗样本加入到目标模型训练过程中，以提升模型抵抗对抗扰动的能力。其优势在于不依赖特定攻击方法，通用性较强，但计算开销极大。Szegedy 等人（2014）和 Goodfellow 等人（2015）在提出 L-BFGS 和 FGSM 攻击时首次提出对抗训练的概念。后来，Madry 等人（2018）在提出 PGD 攻击的同时，从深度模型的鲁棒性优化的角度对对抗训练进行了进一步的理论研究和证明。在此之后，对抗训练策略引起研究人员的广泛关注。例如，Wang 等人（2020）提出误分类感知对抗训练（misclassification aware adversarial training，MART）以区分正确分类的样本和错误分类的样本，并针对错误分类的样本设计正则项以提高对抗鲁棒性。Gowal 等人（2020）改变了原始样本在 StyleGAN（Karras 等，2021）下的分解表征，从而丰富训练样本以获得更好的对抗训练效果。

除此之外，Vivek和Babu（2020）解释了利用单步攻击生成的样本进行对抗训练时产生梯度掩蔽效应的原因，并提出一种基于dropout调度机制的单步对抗训练方法，有效地节省了对抗训练需要的计算资源，同时也确保了对抗鲁棒性。Xiao和Zheng（2020）提出一种增强训练样本的对抗训练方法，该方法利用额外的预训练模型生成对抗样本，并将该样本加入目标模型的对抗训练中以提升鲁棒性。Naseer等人（2020）提出一种自监督的对抗训练方法，他们证明了特征失真与对抗攻击成功率的关系，并基于此设计了特征空间损失，从而对目标模型的中间层进行攻击以生成对抗训练所需要的样本。Jia等人（2022）提出一种可学习攻击策略的对抗训练方式，通过自动产生攻击策略以提高模型的鲁棒性。Yue等人（2024）和Wang等人（2024e）进一步探索了在长尾数据分布和大规模数据场景下如何实现稳定且鲁棒的对抗训练。

4.2 对抗检测

不同于对抗训练，对抗检测不改变目标模型的参数，其通过添加额外分类器的方式判断输入的样本是否为对抗样本。其优势在于具有较好的可集成性，但泛化性存在一定限制。Qin等人（2020）重建输入样本的类别条件以检测对抗样本，据此他们提出一种新的重建式攻击以抵抗这种检测方式，同时他们发现胶囊网络（Sabour等，2017）使用的特征更符合人类的感知，在对抗攻击中的表现也总是优于传统的卷积神经网络。Deng等人（2021）基于贝叶斯神经网络（Bayesian neural network，BNN）提出轻量级贝叶斯精化（lightweight Bayesian refinement，LiBRe）以检测对抗样本，该方法将目标模型的最后几层替换为贝叶斯模块，并针对检测进行微调以保证目标模型能够在保持原始性能的同时实现对抗样本检测。受到人类辨别物体时会利用环境和上下文信息的启发，Li等人（2020b）基于图像中对抗扰动模式的区域不一致性设计了一组自编码器以检测对抗样本。除此之外，Tao等人（2018）提出一种识别与关键特征对应的神经元的方法。通过放大这些神经元的激活，他们构建了一个特征导向模型并根据目标模型和特征导向模型之间的不一致以检测对抗样本。Liu等人（2019a）从隐写分析的角度出发，提出空间富模型（spatial rich model，SRM）和增强空间富模型（enhanced spatial rich model，ESRM），通过估计图像修改概率实现对抗样本的检测。Qiu等人（2019）将目标模型解构为不同的功能模块，并进一步分析了模型在输入原始样本和对抗样本时不同的激活路径，从而提出利用激活路径相似性检测对抗样本的方法。Yin等人（2020）受到多分类问题的启发设计了多个二分类器，并根据输入数据的预测标签选择特定的二分类器以检测对抗样本。赵俊杰等人（2023）利用对抗噪声对于JPEG压缩的敏感性实现对不同对抗攻击方法的细粒度分类。

4.3 对抗去噪

不同于对抗检测和对抗训练，对抗去噪尝试去除对抗样本中的对抗扰动以避免其误导目标模型。其优势在于能够实现对抗样本的召回，但有可能会影响原始样本的质量。基于这一理念，Das等人（2017）、Guo等人（2018）以及Liu等人（2019c）先后提出诸多基于JPEG压缩的对抗去噪方法，他们发现压缩后的对抗样本极大可能会失去攻击能力。然而Raff等人（2019）发现使用JPEG压缩等操作处理对抗样本时，也会降低模型在原始样本上的分类精度。因此，为了保证目标模型的性能不受影响，诸多研究者尝试构建端到端的模型进行对抗去噪。Mustafa等人（2020）提出一种基于小波去噪和图像超分辨率重构的对抗扰动抹除方法，进一步提高修复图像的视觉质量。除此之外，Liao等人（2018）提出高阶表征引导去噪器（high-level representation guided denoiser，HGD），该方法以去噪后的图像与原始图像的误差为损失优化HGD，有效地实现了对抗扰动的去除。随后，Jin等人（2019）基于GAN网络结构设计出一种基于GAN的对抗扰动消除方法（adversarial perturbation elimination GAN，APE-GAN）以实现端到端的对抗扰动抹除，在多个网络结构和多个数据集下都取得了比HGD更好的对抗去噪效果。Zhou等人（2021）在此基础上进一步引入攻击不变特征，并据此设计出ARN（adversarial noise removing network）网络，实现了对于多种攻击方式所生成的对抗样本的有效修复。Nie等人（2022）和Wang等人（2022）提出基于去噪扩散概率模型（denoised diffusion probabilistic model，DDPM）结构（Ho等，2020）的对抗去噪方法，该类方法以一个较小的噪声进行前向扩散，再使用逆向生成将对抗样本净化为原始样本。Dai等人（2022）提出一种基于深度图像先验驱动的防御方法，通过提取图像丰富的统计特征以

实现对抗噪声的抹除。Wang等人(2025)在扩散模型的基础上设计了一种迭代窗口均值滤波器来实现对不同对抗攻击方法的通用稳定去噪。

5 对抗样本应用

除了对抗样本的生成方法和防御策略，研究者逐渐关注其在更广泛任务中的应用潜力，特别是在提升任务鲁棒性与系统稳健性方面的价值(王金伟等，2024)。对抗样本技术不仅揭示了模型的脆弱性，也启发了其在安全敏感任务中的创新应用，例如在对抗隐写、对抗水印、模型保护以及多媒体认证等场景中，利用对抗扰动增强系统对攻击的抵抗能力。这些跨任务的探索为对抗样本研究开辟了新的发展方向，也促进了人工智能系统在复杂环境下的实用性提升。本节主要介绍对抗样本在隐写术和数字水印技术两个典型安全任务上的应用发展。

5.1 对抗隐写

Volkhonskiy等人(2020)首先提出基于GAN(Goodfellow等，2020)的隐写方法(steganographic generative adversarial network，SGAN)，该方法在原始GAN网络架构上添加了隐写分析器作为判别网络，有效地提升了隐写的隐蔽性。为了提升SGAN生成隐写图像的视觉质量，Shi等人(2018)基于Wasserstein生成对抗网络(Wasserstein generative adversarial network，WGAN)结构(Arjovsky等，2017)提出基于GAN的安全隐写方法(secure steganography based on generative adversarial network，SSGAN)，通过替换生成网络结构使得生成的载体图像在隐写后更加逼真。Zhang等人(2018b)将对抗样本的概念引入到隐写中，通过构造强化的载体图像从而实现一种对抗隐写分析的隐写方案，该方法极大地提高了隐写的抗检测性。Tang等人(2019)提出一种基于卷积神经网络的对抗嵌入图像隐写方法，他们将隐写修改像素值的方向逼近于隐写分析器反向传播的梯度逆方向，以提高隐写的抗检测性和隐写信息的提取精度。Liu等人(2021)提出一种新的对抗嵌入策略，通过结合载体图像与隐写图像的多个梯度以决定损失的修改方向，并根据梯度的幅值确定最终嵌入的损失函数以生成对抗隐写图像。在此之后，Liu等人(2022a)进一步将这种对抗嵌入策略引入到有损的信道的隐写中，提升了隐写算法的安全性和对JPEG压缩的鲁棒性。Liu等人(2023)还通过隐写图像的批量生成和选择构建新的对抗嵌入策略，并通过不同的自适应高通滤波模组以获取隐写分析器的通用特征，即图像残差，从而应对隐写分析器被重新训练时的隐写安全性问题。Kishore等人(2022)则将用于隐写的深度神经网络的参数固定，并将隐写噪声以对抗噪声的方式添加到图像上，使得固定的深度神经网络可以从加噪的图像上解码出隐写信息。Luo等人(2023)在固定神经网络隐写的基础上引入了密钥机制，只有拥有正确密钥的接收者才能从隐写图像中解密出隐藏信息。Fan等人(2024)提出一种基于神经元归因的可迁移图像对抗隐写框架，通过寻找并干扰隐写分析器关注的关键特征，获取通用的抗隐写分析能力。

5.2 对抗水印

Jia等人(2020)首次提出一种对抗水印生成方法，认为对抗扰动是一种无特殊意义的噪声，直接添加在图像上会引起人们怀疑，对此，提出利用盆地跳跃演化优化算法(basin hopping evolution，BHE)在图像上添加可见水印形成对抗样本。Khachaturov等人(2021)发现图像修复技术可用于去除图像中的可见水印，因此针对图像修复深度神经网络进行对抗攻击，通过固定掩码和随机掩码两种方式添加扰动，从而有效地保护可见水印，防止其被图像修复技术去除。Liu等人(2022b)利用对抗扰动技术生成水印疫苗，以预防图像可见水印抹除。根据不同的攻击效果，该水印疫苗又可以进一步分为破坏水印疫苗(disrupting watermark vaccine，DWV)和不可擦除水印疫苗(inerasable watermark vaccine，IWV)。当使用水印去除网络抹除嵌入了DWV的图像中的水印时，图像的内容会被破坏，不再具备可读性和可使用性。而当使用水印去除网络抹除嵌入了IWV的图像中的水印时，图像中的水印能够有效抵御抹除。Chen等人(2023)提出一种通用水印疫苗，他们通过向全图添加噪声和只向可见水印位置添加噪声两种方式，取得了一种对抗扰动保护多幅水印图像的水印不被去除的效果。Lyu等人(2023)提出一种基于对抗攻击的鲁棒水印保护方法，以抵抗基于图像修复技术和盲水印去除技术的可见水印去除方法，他们设计了一个水印定位网络，利用图像超像素技术寻找领域内语义关联最小的区域，并利用差分进化算法以确定最优的水印掩码位置。Wang等人

(2024b)提出基于兴趣区域的对抗可见水印生成算法,有效地解决了可见水印容易被水印去除网络抹除的问题。Wang等人(2024c)还提出具有高度可迁移性的对抗不可见水印生成算法,通过基于注意力的Unet模型和可微噪声模拟赋予不可见水印抵抗去除、识别、分类、分割等一系列未授权操作的能力,进一步提升水印在深度学习时代下作为版权保护手段的安全性和可靠性。

6 实验对比

6.1 攻击能力对比

在实验中,本文在Caltech-256数据集上对比GAP(Poursaeed等,2018)、AdvGAN(adversarial generative adversarial network)(Xiao等,2018)、AdvGAN++(Jandial等,2019)、SSAE(symmetric saliency-based auto-encoder)(Lu等,2021)、RGAN(recoverable generative adversarial network)(Zhang等,2023)以及BIA(beyond imagenet attack)(Zhang等,2022b)的攻击能力、扰动强度和视觉质量。不同算法在4个Transformers和4个卷积神经网络(convolutional neural network, CNN)上的CER、PSNR、SSIM和LPIPS指标如表1所示。

从表1可以看出,基于生成的对抗攻击算法在针对卷积神经网络进行攻击时都能取得令人满意的攻击效果,攻击后的误分率达到99%左右,同时能够保证对抗样本的PSNR在30 dB以上。然而,在针对Transformer进行攻击时,现有方法的攻击性能出现较大幅度退化。尽管诸如AdvGAN和AdvGAN++等算法依然能够取得较高的误分率,但其需要极大扰动强度来构造对抗样本,这使得对抗样本的PSNR大幅度退化至20～25 dB。值得注意的是,诸如GAP-U、GAP-R和SSAE在内的对抗样本生成算法由于使用了无穷范数限制扰动强度,因此这些方法以Transformers为目标模型时攻击会崩溃。虽然其他的一些竞争方法可以针对Transformers在CER指标上达到97%～98%的性能,但往往需要极大的扰动强度,从而会使PSNR指标下降至18～24 dB。因此,在后续的研究中如何针对不同结构的目标模型实现有效且通用的攻击是值得进一步探索的问题。

6.2 鲁棒性对比

实验中,本文选择在Caltech-256数据集训练的DenseNet-121作为目标模型,对比GAP(Poursaeed等,2018)、AdvGAN(Xiao等,2018)、AdvGAN++(Jandial等,2019)、SSAE(Lu等,2021)、RGAN(Zhang等,2023)和BIA(Zhang等,2022b)对于JPEG压缩的鲁棒性。表2展示了不同算法在原始条件下及集成了JPEG掩码(Zhu等,2018)和小批量真实和模拟JPEG压缩(mini-batch of real and simulated jpeg compression, MBRS)(Jia等,2021)两种JPEG压缩模拟器下的鲁棒性。

如表2所示,JPEG压缩可以有效破坏现有的基于生成的对抗攻击方法所生成的对抗样本,并将CER从96%～99%降低到20%～30%。尽管使用JPEG模拟器JPEG-Mask和MBRS对这些方法进行微调,能够一定程度上提升其对于JPEG压缩的鲁棒性,但在质量因子为10和30时,CER仍会显著下降。值得注意的是,由于使用上述JPEG压缩模拟器进行训练会导致扰动强度不均匀,所以本文比较了PSNR在26～27 dB之间时的CER,该条件所有组合都能达到并收敛。

6.3 迁移能力对比

本文对GAP-R(Poursaeed等,2018)、CDA(cross domain attack)(Naseer等人,2019)、LTAP(learning transferable adversarial perturbations)(Nakka和Salzmann,2021)和BIA(Zhang等,2022b)跨模型、跨数据集的迁移性进行实验,结果如表3和表4所示。

在跨模型的实验中,如表3所示,本文选用在Caltech-256数据集上训练的Inception-v3网络作为代理模型,并测试不同方法生成的对抗样本对于IncRes-v2、ResNet-50和VGG-16的迁移能力。

在跨数据集的实验中,如表4所示,依然选用在Caltech-256数据集上训练的Inception-v3网络作为代理模型,并测试不同方法生成的对抗样本对于在ImageNet数据集上训练的Inception-v3、IncRes-v2、ResNet-50和VGG-16的迁移能力。

从表3和表4可以看出,无论是跨模型还是跨数据集,LTAP和BIA这类通过破坏中间层特征来构造对抗样本的方法都能够比早期的方法展现出更好的迁移性,尤其是在跨数据集的迁移性提升上,这类基于中间层特征破坏的方法提升的迁移性最高可达27.3%。

表 1　不同基于生成的攻击算法在 Caltech-256 数据集上的攻击表现

Table 1　Attack performance of different generation based attack algorithms on the Caltech-256 dataset

攻击算法	Transformers					CNN				
	目标模型	CER ↑ /%	PSNR ↑ /dB	SSIM ↑	LPIPS ↓	目标模型	CER ↑ /%	PSNR ↑ /dB	SSIM ↑	LPIPS ↓
GAP-U	ViT-B CER:6.50%	8.28	24.50	0.576 9	0.325 4	VGG-16 CER:27.35%	85.10	24.60	0.562 4	0.470 3
GAP-R		8.94	24.47	0.559 9	0.369 2		86.15	24.47	0.569 5	0.468 0
AdvGAN		97.78	20.00	0.487 8	0.567 0		98.37	27.77	0.785 8	0.480 2
AdvGAN++		98.29	24.25	0.696 3	0.371 6		99.07	29.48	0.826 2	0.383 1
SSAE		6.50	60.95	0.999 4	0.000 0		99.26	27.59	0.700 7	0.504 4
RGAN		97.54	22.29	0.645 2	0.444 9		98.48	28.49	0.806 3	0.473 1
BIA		10.78	24.39	0.728 4	0.345 5		99.34	28.61	0.719 5	0.503 0
GAP-U	DeiT-B CER:13.31%	18.72	24.52	0.576 8	0.418 7	DenseNet-121 CER:18.01%	96.92	36.33	0.921 6	0.051 5
GAP-R		38.05	24.48	0.628 0	0.334 2		96.84	36.33	0.920 8	0.047 9
AdvGAN		97.74	23.83	0.672 8	0.466 1		99.30	39.92	0.978 0	0.009 0
AdvGAN++		98.25	31.49	0.911 9	0.181 7		99.57	37.39	0.961 6	0.030 8
SSAE		12.92	56.70	0.999 2	0.000 0		98.29	36.43	0.920 0	0.078 3
RGAN		96.46	24.84	0.702 5	0.440 6		99.14	38.00	0.969 5	0.009 6
BIA		83.66	24.61	0.588 4	0.441 5		97.20	36.33	0.919 1	0.077 4
GAP-U	BEiT-B CER:4.55%	6.27	24.41	0.557 2	0.417 3	ResNet-50 CER:15.33%	99.33	36.36	0.916 1	0.055 2
GAP-R		5.99	24.46	0.631 7	0.281 8		99.41	36.37	0.916 5	0.068 8
AdvGAN		97.00	25.51	0.724 8	0.454 0		99.53	42.70	0.988 8	0.002 3
AdvGAN++		99.30	24.35	0.683 7	0.416 7		99.57	40.06	0.993 5	0.006 7
SSAE		4.55	65.61	0.999 8	0.000 0		99.53	34.65	0.896 4	0.083 3
RGAN		97.59	25.52	0.740 8	0.440 5		99.33	37.79	0.963 0	0.016 0
BIA		86.15	24.86	0.604 4	0.449 9		99.30	36.29	0.920 2	0.055 6
GAP-U	SwinT-B CER:6.85%	8.75	24.54	0.580 3	0.366 5	ConvNeXt-B CER:11.75%	19.30	24.42	0.594 7	0.340 5
GAP-R		9.22	24.45	0.651 9	0.318 0		28.17	24.42	0.606 8	0.320 2
AdvGAN		97.16	18.39	0.422 3	0.552 1		97.59	28.29	0.785 6	0.349 0
AdvGAN++		98.37	23.39	0.597 7	0.458 0		98.56	24.91	0.615 4	0.443 2
SSAE		6.03	60.23	0.999 5	0.000 0		96.19	26.75	0.670 3	0.395 0
RGAN		96.49	21.62	0.587 3	0.484 7		97.59	31.28	0.849 5	0.269 6
BIA		18.72	24.29	0.806 6	0.265 2		90.82	24.84	0.603 2	0.433 0
GAP-U	平均性能	10.50	24.49	0.572 8	0.382 0	平均性能	75.16	30.43	0.748 7	0.229 4
GAP-R		15.55	24.47	0.617 9	0.325 8		77.64	30.40	0.753 4	0.226 2
AdvGAN		97.42	21.93	0.576 9	0.509 8		98.70	**34.67**	0.884 6	0.210 1
AdvGAN++		**98.55**	**25.87**	**0.722 4**	**0.357 0**		**99.19**	32.96	0.849 2	0.216 0
SSAE		7.50	60.87	0.999 5	0.000 0		98.32	31.36	0.796 9	0.265 3
RGAN		97.02	23.57	0.669 0	0.452 7		98.64	33.89	**0.897 1**	**0.192 1**
BIA		49.83	24.54	0.682 0	0.375 5		96.66	31.52	0.790 5	0.267 3

注：加粗字体表示在攻击成功（CER>70%）时各平均性能的最优结果。“↑”表示值越高越好，“↓”表示值越低越好。

表2　不同基于生成的攻击算法对于JPEG压缩的误分率对比

Table 2　Comparison of classification error rate for JPEG compression using different generation based attack algorithms

/%

攻击方法	JPEG压缩模拟器	质量因子					无压缩	平均性能
		10	30	50	70	90		
GAP-U	无模拟器	34.75	21.79	19.84	19.42	20.86	96.92	35.60
GAP-R		34.09	21.98	19.88	19.49	20.82	96.84	35.52
AdvGAN		34.16	21.67	19.88	22.88	86.30	99.30	47.37
AdvGAN++		35.02	22.02	24.28	38.95	75.76	99.57	**49.27**
SSAE		34.28	21.21	19.96	19.11	22.88	98.29	35.95
RGAN		34.44	21.25	20.12	28.44	91.56	99.14	49.16
BIA		34.36	21.40	20.12	19.65	25.02	97.20	36.29
GAP-U	JPEG掩码	35.64	42.22	75.88	68.21	89.42	91.60	67.16
GAP-R		36.89	31.63	35.60	43.04	90.74	90.66	54.76
AdvGAN		54.51	94.59	97.82	98.56	98.88	99.03	**90.57**
AdvGAN++		40.86	95.91	97.86	98.79	99.07	99.22	88.62
SSAE		54.01	79.42	94.71	98.48	98.91	98.60	87.35
RGAN		43.31	92.30	98.17	98.37	98.95	98.95	88.34
BIA		52.26	74.98	90.47	95.25	96.85	97.20	84.50
GAP-U	MBRS	36.38	43.50	56.54	48.48	99.53	99.96	64.07
GAP-R		36.03	24.13	98.29	99.57	99.96	99.96	76.32
AdvGAN		75.95	88.60	94.05	98.21	99.30	99.34	**92.57**
AdvGAN++		67.24	91.21	97.78	99.18	99.53	99.57	92.42
SSAE		43.77	75.14	94.05	96.85	98.76	99.57	84.69
RGAN		52.84	93.58	97.12	97.78	99.22	99.53	90.01
BIA		59.81	71.56	75.64	77.74	97.39	99.49	80.27

注：加粗字体表示各组平均性能的最优结果。

表3　不同攻击方法在不同模型上的误分率

Table 3　The classification error rate of different attack methods on different models

/%

方法	Caltech-256数据集			
	Inception-v3	IncRes-v2	ResNet-50	VGG-16
GAP-R	98.6	85.1	94.0	96.2
CDA	96.2	81.2	90.1	95.7
LTAP	99.5	**95.0**	**97.7**	99.3
BIA	**99.6**	93.2	96.4	**99.4**

注：加粗字体表示各列最优结果。

表4　不同攻击方法针对Caltech-256数据集上训练的Inception-v3生成的对抗样本在ImageNet数据集上的误分率

Table 4　The classification error rate of adversarial examples generated by Inception-v3 trained on the Caltech-256 dataset using different attack methods on the ImageNet dataset

/%

方法	ImageNet数据集			
	Inception-v3	IncRes-v2	ResNet-50	VGG-16
GAP-R	95.5	71.2	92.9	97.5
CDA	86.7	59.5	87.4	94.8
LTAP	99.9	**86.8**	**98.1**	**99.8**
BIA	**100.0**	86.0	96.7	99.5

注：加粗字体表示各列最优结果。

7 结 语

本文首先介绍了与对抗样本相关的术语，随后介绍了白盒攻击和黑盒攻击以及对抗训练、对抗检测和对抗去噪 3 种对抗防御策略，最后介绍了对抗隐写和对抗水印两种对抗样本应用领域。尽管这些研究取得了一定进展和突破，但在该领域仍有许多研究值得进一步深入探索，例如跨领域的对抗攻击、鲁棒性评价、现实场景下的对抗攻击和系统化防御框架。

1）跨领域的对抗攻击。目前，对抗样本的相关研究正逐步从计算机视觉领域扩展到诸如自然语言处理（natural language processing，NLP）、语音识别、物联网（internet of things，IoT）等更多更复杂的领域。在自然语言处理领域，对抗样本的生成与攻击主要针对文本分类、情感分析、机器翻译等任务。例如，攻击者通过对输入文本进行微小修改（如同义词替换、字符扰动等）使模型产生错误预测，但这种扰动往往难以被人类察觉。在语音识别中，通过添加微弱的噪声，使得语音识别系统误解输入音频，从而产生错误的输出。在物联网领域，由于设备资源有限且安全防护薄弱，基于对抗样本的物理层攻击、数据投毒攻击等也给其带来了严峻的挑战。未来，跨领域的对抗样本研究将重点关注如何在不同模态下生成高效的对抗样本，同时探索跨模态对抗攻击。例如，从图像到文本或语音的迁移攻击。

2）鲁棒性评价。目前，针对对抗样本的鲁棒性评估尚缺乏统一的标准，现有方法通常关注单一模型、单类攻击方法和单次攻击的防御效果。然而，随着对抗攻击与防御技术的不断发展，构建完善的鲁棒性评估体系成为亟待解决的问题。未来研究方向将着重于提出全面的评估指标体系，涵盖不同攻击类型（如白盒攻击、黑盒攻击、物理攻击）和不同攻击次数（单次攻击、多次攻击、组合攻击）的防御效果。此外，还需考虑评估模型在面对不同模态、跨任务对抗攻击时的鲁棒性表现。例如，设计通用的测试基准，衡量模型在多种扰动强度下的表现，从而系统地评估深度模型的安全性。

3）现实场景下的对抗攻击。尽管现有对抗攻击算法在理论上取得了较大的突破，但其在现实场景中的部署研究仍然较为匮乏。已有的一些现实场景下的对抗攻击也集中在激光、阴影和雨滴等比较简单的场景中。然而，现实场景需要考虑环境光照、视角变化以及噪声干扰等更多更复杂的因素的影响。未来研究也将不限于以现实世界中的相机为对象进行对抗攻击，更应该关注如何对于传感器、雷达和 X 光等其他对象进行有效的攻击。以上对于确保人工智能（artificial intelligence，AI）系统能够在真实场景中始终保持高安全性和稳定性。并应用于自动驾驶、智能医疗以及物联网等高安全需求领域尤为关键。

4）系统化防御框架。随着对抗样本防御研究的深入，未来的发展将逐步从对抗样本检测到对抗样本细粒度分类，最终实现对于不同对抗攻击方法针对性修复的系统化防御框架。在对抗样本检测阶段，研究将重点关注高效、鲁棒的检测方法，利用多尺度特征提取、频域分析以及模型行为异常性等手段，对输入数据中隐藏的对抗扰动进行精准识别。然而，单纯的检测无法提供有效的防御策略，因此细粒度分类将成为下一步关键研究方向。细粒度分类旨在区分不同类型的对抗样本，例如区分该对抗样本由什么攻击方法生成、对抗扰动强度等，从而为后续防御策略提供有针对性的依据。基于细粒度分类的结果，对于不同攻击方法的针对性修复将得到进一步发展。针对不同类型的对抗扰动，设计适配性强的修复策略，如输入重构（通过去噪网络或压缩机制重建正常输入）、模型自适应修正（通过激活层调整或动态梯度抑制）以及多模态融合等方法，从根本上消除对抗样本的影响。这一框架不仅能够提高防御的效果和效率，还能够实现对抗攻击的可解释性分析和系统化修复，为实际应用场景中复杂、多样的对抗威胁提供全面的解决方案。

参考文献（Reference）

Arjovsky M，Chintala S and Bottou L. 2017. Wasserstein generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning. Sydney，Australia：PMLR：214-223

Athalye A，Engstrom L，Ilyas A and Kwok K. 2018. Synthesizing robust adversarial examples//Proceedings of the 35th International Conference on Machine Learning. Stockholm，Sweden：PMLR：284-293

Bhagoji A N，He W，Li B and Song D. 2018. Practical black-box attacks on deep neural networks using efficient query mechanisms//Proceedings of the 15th European Conference on Computer Vision.

Munich, Germany: Springer: 158-174 [DOI: 10.1007/978-3-030-01258-8_10]

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

Chen J B, Liu X W, Liang S Y, Jia X J and Xun Y. 2023. Universal watermark vaccine: universal adversarial perturbations for watermark protection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada: IEEE: 2322-2329 [DOI: 10.1109/CVPRW59228.2023.00228]

Chen J Q, Chen H, Chen K Y, Zhang Y L, Zou Z X and Shi Z W. 2025. Diffusion models for imperceptible and transferable adversarial attack. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47 (2): 961-977 [DOI: 10.1109/TPAMI.2024.3480519]

Chen P Y, Zhang H, Sharma Y, Yi J F and Hsieh C J. 2017. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA: ACM: 15-26 [DOI: 10.1145/3128572.3140448]

Dai T, Feng Y, Chen B, Lu J and Xia S T. 2022. Deep image prior based defense against adversarial examples. Pattern Recognition, 122: #108249 [DOI: 10.1016/j.patcog.2021.108249]

Das N, Shanbhogue M, Chen S T, Hohman F, Chen L, Kounavis M E, et al. 2017. Keeping the bad guys out: protecting and vaccinating deep learning with JPEG compression [EB/OL]. [2024-12-16]. https://arxiv.org/pdf/1705.02900.pdf

Deng Z J, Yang X, Xu S Z, Su H and Zhu J. 2021. LiBRe: a practical Bayesian approach to adversarial detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 972-982 [DOI: 10.1109/CVPR46437.2021.00103]

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L, et al. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Dong Y P, Pang T Y, Su H and Zhu J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4307-4316 [DOI: 10.1109/CVPR.2019.00444]

Du Y L, Fang M, Yi J F, Cheng J and Tao D C. 2018. Towards query efficient black-box attacks: an input-free perspective//Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security. Toronto, Canada: ACM: 13-24 [DOI: 10.1145/3270101.3270106]

Duan M X, Li K L, Deng J Y, Xiao B and Tian Q. 2022. A novel multi-sample generation method for adversarial attacks. ACM Transactions on Multimedia Computing, Communications, and Applications, 18(4): #112 [DOI: 10.1145/3506852]

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C W, et al. 2018. Robust physical-world attacks on deep learning visual classification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1625-1634 [DOI: 10.1109/CVPR.2018.00175]

Fan Z X, Chen K J, Zeng K, Zhang J S, Zhang W M and Yu N H. 2024. Natias: neuron attribution-based transferable image adversarial steganography. IEEE Transactions on Information Forensics and Security, 19: 6636-6649 [DOI: 10.1109/TIFS.2024.3421893]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/3422622]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Gowal S, Qin C L, Huang P S, Cemgil T, Dvijotham K, Mann T, et al. 2020. Achieving robustness in the wild via adversarial mixing with disentangled representations//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1208-1217 [DOI: 10.1109/CVPR42600.2020.00129]

Guo C, Rana M, Cissé M and van der Maaten L. 2018. Countering adversarial images using input transformations//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #574

Hu Z H, Huang S Y, Zhu X P, Sun F C, Zhang B and Hu X L. 2022. Adversarial texture for fooling person detectors in the physical world//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13297-13306 [DOI: 10.1109/CVPR52688.2022.01295]

Ilyas A, Engstrom L, Athalye A and Lin J. 2018. Black-box adversarial attacks with limited queries and information//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 2142-2151

Jandial S, Mangla P, Varshney S and Balasubramanian V. 2019. Adv-GAN++: harnessing latent layers for adversary generation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 2045-2048 [DOI: 10.1109/ICCVW.2019.00257]

Jia X J, Wei X X, Cao X C and Han X G. 2020. Adv-watermark: a

novel watermark perturbation for adversarial examples//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1579-1587 [DOI: 10.1145/3394171.3413976]

Jia X J, Zhang Y, Wu B Y, Ma K, Wang J and Cao X C. 2022. LAS-AT: adversarial training with learnable attack strategy//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13388-13398 [DOI: 10.1109/CVPR52688.2022.01304]

Jia Z Y, Fang H and Zhang W M. 2021. MBRS: enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 41-49 [DOI: 10.1145/3474085.3475324]

Jin G Q, Shen S W, Zhang D M, Dai F and Zhang Y D. 2019. APE-GAN: adversarial perturbation elimination with GAN//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 3842-3846 [DOI: 10.1109/ICASSP.2019.8683044]

Karras T, Laine S and Aila T. 2021. A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43 (12): 4217-4228 [DOI: 10.1109/TPAMI.2020.2970919]

Khachaturov D, Shumailov I, Zhao Y R, Papernot N and Anderson R J. 2021. Markpainting: adversarial machine learning meets inpainting//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR: 5409-5419

Kishore V, Chen X Y, Wang Y, Li B Y and Weinberger K Q. 2022. Fixed neural network steganography: train the images, not the network//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Kurakin A, Goodfellow I J and Bengio S. 2018. Adversarial examples in the physical world//Yampolskiy R V, ed. Artificial Intelligence Safety and Security. New York, USA: Chapman and Hall/CRC: 99-112

Li J, Ji R R, Liu H, Liu J Z, Zhong B N, Deng C, et al. 2020a. Projection a probability-driven black-box attack//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 359-368 [DOI: 10.1109/CVPR42600.2020.00044]

Li S S, Zhu S T, Paul S, Roy-Chowdhury A, Song C Y, Krishnamurthy S, et al. 2020b. Connecting the dots: detecting adversarial perturbations using context inconsistency//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 396-413 [DOI: 10.1007/978-3-030-58592-1_24]

Liao F Z, Liang M, Dong Y P, Pang T Y, Hu X L and Zhu J. 2018. Defense against adversarial attacks using high-level representation guided denoiser//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1778-1787 [DOI: 10.1109/CVPR.2018.00191]

Liu J Y, Zhang W M, Zhang Y W, Hou D D, Liu Y J, Zha H Y, et al. 2019a. Detection based defense against adversarial examples from the steganalysis point of view//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4820-4829 [DOI: 10.1109/CVPR.2019.00496]

Liu M L, Fan H Y, Wei K K, Luo W Q and Lu W. 2022a. Adversarial robust image steganography against lossy JPEG compression. Signal Processing, 200: #108668 [DOI: 10.1016/j.sigpro.2022.108668]

Liu M L, Luo W Q, Zheng P J and Huang J W. 2021. A new adversarial embedding method for enhancing image steganography. IEEE Transactions on Information Forensics and Security, 16: 4621-4634 [DOI: 10.1109/TIFS.2021.3111748]

Liu M L, Song T T, Luo W Q, Zheng P J and Huang J W. 2023. Adversarial steganography embedding via stego generation and selection. IEEE Transactions on Dependable and Secure Computing, 20(3): 2375-2389 [DOI: 10.1109/TDSC.2022.3182041]

Liu X Q and Hsieh C J. 2019b. Rob-GAN: generator, discriminator, and adversarial attacker//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11226-11235 [DOI: 10.1109/CVPR.2019.01149]

Liu X W, Liu J, Bai Y, Gu J D, Chen T, Jia X J, et al. 2022b. Watermark vaccine: adversarial attacks to prevent watermark removal//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 1-17 [DOI: 10.1007/978-3-031-19781-9_1]

Liu Y P, Chen X Y, Liu C and Song D. 2017. Delving into transferable adversarial examples and black-box attacks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Liu Z H, Liu Q, Liu T, Xu N, Lin X, Wang Y Z, et al. 2019c. Feature distillation: DNN-oriented JPEG compression against adversarial examples//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 860-868 [DOI: 10.1109/CVPR.2019.00095]

Lu S H, Xian Y Q, Yan K, Hu Y, Sun X, Guo X W, Huang F Y, et al. 2021. Discriminator-free generative adversarial attack//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 1544-1552 [DOI: 10.1145/3474085.3475290]

Luo Z C, Li S, Li G B, Qian Z X and Zhang X P. 2023. Securing fixed neural network steganography//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 7943-7951 [DOI: 10.1145/3581783.3611920]

Lyu M Z, Huang Y and Kong A W K. 2023. Adversarial attack for robust watermark protection against inpainting-based and blind watermark removers//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 8396-8405

[DOI: 10.1145/3581783.3612034]

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

Miyato T, Kataoka T, Koyama M and Yoshida Y. 2018. Spectral normalization for generative adversarial networks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Moosavi-Dezfooli S M, Fawzi A, Fawzi O and Frossard P. 2017. Universal adversarial perturbations//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 86-94 [DOI: 10.1109/CVPR.2017.17]

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

Mustafa A, Khan S H, Hayat M, Shen J B and Shao L. 2020. Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing, 29: 1711-1724 [DOI: 10.1109/TIP.2019.2940533]

Nakka K K and Salzmann M. 2021. Learning transferable adversarial perturbations//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #1069

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

Naseer M, Khan S, Khan M H, Khan F S and Porikli F. 2019. Cross-domain transferability of adversarial perturbations//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1156

Nie W L, Guo B, Huang Y J, Xiao C W, Vahdat A and Anandkumar A. 2022. Diffusion models for adversarial purification//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 16805-16827

Papernot N, McDaniel P and Goodfellow I. 2016a. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples [EB/OL]. [2024-12-16]. https://arxiv.org/pdf/1605.07277.pdf

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik Z B and Swami A. 2017. Practical black-box attacks against machine learning//Proceedings of 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates: ACM: 506-519 [DOI: 10.1145/3052973.3053009]

Papernot N, McDaniel P, Jha S, Fredrikson M, Celik Z B and Swami A. 2016b. The limitations of deep learning in adversarial settings//Proceedings of 2016 IEEE European Symposium on Security and Privacy. Saarbruecken, Germany: IEEE: 372-387 [DOI: 10.1109/EuroSP.2016.36]

Park J, Miller P and McLaughlin N. 2024. Hard-label based small query black-box adversarial attack//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3974-3983 [DOI: 10.1109/WACV57701.2024.00394]

Poursaeed O, Katsman I, Gao B C and Belongie S. 2018. Generative adversarial perturbations//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4422-4431 [DOI: 10.1109/CVPR.2018.00465]

Qin Y, Frosst N, Sabour S, Raffel C, Cottrell G W and Hinton G E. 2020. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net

Qiu Y X, Leng J W, Guo C, Chen Q, Li C, Guo M Y and Zhu Y H. 2019. Adversarial defense through network profiling based path extraction//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4772-4781 [DOI: 10.1109/CVPR.2019.00491]

Raff E, Sylvester J, Forsyth S and Mclean M. 2019. Barrage of random transforms for adversarially robust defense//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6521-6530 [DOI: 10.1109/CVPR.2019.00669]

Rahmati A, Moosavi-Dezfooli S M, Frossard P and Dai H. 2020. GeoDA: a geometric framework for black-box adversarial attacks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8443-8452 [DOI: 10.1109/CVPR42600.2020.00847]

Rony J, Hafemann L G, Oliveira L S, Ben Ayed I, Sabourin R, et al. 2019. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4317-4325 [DOI: 10.1109/CVPR.2019.00445]

Sabour S, Frosst N and Hinton G E. 2017. Dynamic routing between capsules//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 3859-3869

Shi H C, Dong J, Wang W, Qian Y L and Zhang X Y. 2018. SSGAN: secure steganography based on generative adversarial networks//Proceedings of the 18th Pacific Rim Conference on Multimedia. Harbin, China: Springer: 534-544 [DOI: 10.1007/978-3-319-

77380-3_51］

Shi Y C，Han Y H，Hu Q H，Yang Y and Tian Q. 2023. Query-efficient black-box adversarial attack with customized iteration and sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence，45(2)：2226-2245［DOI：10.1109/TPAMI.2022.3169802］

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

Tang W X，Li B，Tan S Q，Barni M and Huang J W. 2019. CNN-based adversarial embedding for image steganography. IEEE Transactions on Information Forensics and Security，14(8)：2074-2087［DOI：10.1109/TIFS.2019.2891237］

Tao G H，Ma S Q，Liu Y Q and Zhang X Y. 2018. Attacks meet interpretability：attribute-steered detection of adversarial samples//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal，Canada：Curran Associates Inc.：7728-7739

Tu C C，Ting P，Chen P Y，Liu S J，Zhang H，Yi J F，et al. 2019. AutoZOOM：autoencoder-based zeroth order optimization method for attacking black-box neural networks//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu，USA：AAAI：742-749［DOI：10.1609/aaai.v33i01.3301742］

Vivek B S and Babu R V. 2020. Single-step adversarial training with dropout scheduling//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：947-956［DOI：10.1109/CVPR42600.2020.00103］

Vo V Q，Abbasnejad E and Ranasinghe D C. 2024. Brusleattack：a query-efficient score-based black-box sparse adversarial attack//Proceedings of the 12th International Conference on Learning Representations. Vienna，Austria：OpenReview.net

Volkhonskiy D，Nazarov I and Burnaev E. 2020. Steganographic generative adversarial networks//Proceedings of the 12th International Conference on Machine Vision. Amsterdam，the Netherlands：SPIE：991-1005［DOI：10.1117/12.2559429］

Wan P，Hu C and Wu X J. 2024. Multi-domain feature mixup boosting adversarial examples transferability method. Journal of Image and Graphics，29(12)：3670-3683（万鹏，胡聪，吴小俊. 2024. 多域特征混合增强对抗样本迁移性方法. 中国图象图形学报，29(12)：3670-3683）［DOI：10.11834/jig.230895］

Wang H R，Sun R X，Chen C J，Xue M H，Soon L K and Wang S. 2025. Iterative window mean filter：thwarting diffusion-based adversarial purification. IEEE Transactions on Dependable and Secure Computing，22（2）：1827-1844［DOI：10.1109/TDSC.2024.3472569］

Wang J F，Chen Z Y，Jiang K X，Yang D K，Hong L Y，Guo P X，et al. 2024a. Boosting the transferability of adversarial attacks with global momentum initialization. Expert Systems with Applications，255：#124757［DOI：10.1016/j.eswa.2024.124757］

Wang J W，Huang W Y，Zhang J W，Luo X Y and Ma B. 2024b. Adversarial watermark：a robust and reliable watermark against removal. Journal of Information Security and Applications，82：#103750［DOI：10.1016/j.jisa.2024.103750］

Wang J W，Jiang X L，Tan G F and Luo X Y. 2024. Frontier research and prospect of watermarking for generated images. Journal of Cybersecurity，2(1)：50-62（王金伟，姜晓丽，谭贵峰，罗向阳. 2024. 生成图水印的前沿研究与展望. 网络空间安全科学学报，2(1)：50-62）［DOI：10.20172/j.issn.2097-3136.240104］

Wang J Y，Lyu Z Y，Lin D H，Dai B and Fu H F. 2022. Guided diffusion model for adversarial purification［EB/OL］.［2022-05-30］. https://arxiv.org/pdf/2205.14969v1.pdf

Wang J W，Wang H H，Zhang J W，Wu H，Luo X Y and Ma B. 2024c. Invisible adversarial watermarking：a novel security mechanism for enhancing copyright protection. ACM Transactions on Multimedia Computing，Communications and Applications，21（2）：#43［DOI：10.1145/3652608］

Wang J W，Wang M Y，Wu H，Ma B and Luo X Y. 2024d. Improving transferability of adversarial attacks with gaussian gradient enhance momentum//Proceedings of the 6th Chinese Conference on Pattern Recognition and Computer Vision. Xiamen，China：Springer：421-432［DOI：10.1007/978-981-99-8546-3_34］

Wang X S and He K. 2021. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：1924-1933［DOI：10.1109/CVPR46437.2021.00196］

Wang Y，Cao T Y，Yang J B，Zheng Y F，Fang Z and Deng X T. 2022. A perturbation constraint related weak perceptual adversarial example generation method. Journal of Image and Graphics，27(7)：2287-2299（王杨，曹铁勇，杨吉斌，郑云飞，方正，邓小桐. 2022. 结合扰动约束的低感知性对抗样本生成方法. 中国图象图形学报，27(7)：2287-2299）［DOI：10.11834/jig.200681］

Wang Y S，Zou D F，Yi J F，Bailey J，Ma X J and Gu Q Q. 2020. Improving adversarial robustness requires revisiting misclassified examples//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa，Ethiopia：OpenReview.net

Wang Z B，Guo H C，Zhang Z F，Liu W X，Qin Z and Ren K. 2021. Feature importance-aware transferable adversarial attacks//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：7619-7628［DOI：10.1109/ICCV48922.2021.00754］

Wang Z Y，Li X H，Zhu H R and Xie C H. 2024e. Revisiting adversarial training at scale//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：24675-24685［DOI：10.1109/CVPR52733.2024.02330］

Wu H, Wang J W, Luo X Y and Ma B. 2024. Immune defense based on hyperbolic tangent and moments. Chinese Journal of Computers, 47(8): 1786-1812 (吴昊, 王金伟, 罗向阳, 马宾. 2024. 基于双曲正切和矩的免疫防御. 计算机学报, 47(8): 1786-1812) [DOI: 10.11897/SP.J.1016.2024.01786]

Wu H, Wang J W, Zhang J W, Wu Y F, Ma B and Luo X Y. 2023. Improving the transferability of adversarial attacks through experienced precise Nesterov momentum//Proceedings of 2023 International Joint Conference on Neural Networks. Gold Coast, Australia: IEEE: 1-8 [DOI: 10.1109/IJCNN54540.2023.10191329]

Xiao C and Zheng C X. 2020. One man's trash is another man's treasure: resisting adversarial examples by adversarial examples//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 409-418 [DOI: 10.1109/CVPR42600.2020.00049]

Xiao C W, Li B, Zhu J Y, He W, Liu M Y and Song D. 2018. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial. Stockholm, Sweden: ijcai.org: 3905-3911

Xie C H, Zhang Z S, Zhou Y Y, Bai S, Wang J Y, Ren Z, et al. 2019. Improving transferability of adversarial examples with input diversity//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2725-2734 [DOI: 10.1109/CVPR.2019.00284]

Ye Y X, Du X, Chen S, Zhu S Z and Yan Y. 2024. Sparse adversarial patch attack based on QR code mask. Journal of Image and Graphics, 29(7): 1889-1901 (叶乙轩, 杜侠, 陈思, 朱顺痣, 严严. 2024. 二维码掩膜下的稀疏对抗补丁攻击. 中国图象图形学报, 29(7): 1889-1901) [DOI: 10.11834/jig.230453]

Yin X W, Kolouri S and Rohde G K. 2020. GAT: generative adversarial training for adversarial example detection and robust classification//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net

Yue X L, Mou N P, Wang Q and Zhao L C. 2024. Revisiting adversarial training under long-tailed distributions//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24492-24501 [DOI: 10.1109/CVPR52733.2024.02312]

Zhang J P, Wu W B, Huang J T, Huang Y Z, Wang W X, Su Y X and Lyu M R. 2022a. Improving adversarial transferability via neuron attribution-based attacks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14973-14982 [DOI: 10.1109/CVPR52688.2022.01457]

Zhang J W, Wang J W, Wang H and Luo X Y. 2023. Self-recoverable adversarial examples: a new effective protection mechanism in social networks. IEEE Transactions on Circuits and Systems for Video Technology, 33(2): 562-574 [DOI: 10.1109/TCSVT.2022.3207008]

Zhang J W, Wang J W, Wang H, Luo X Y and Ma B. 2024. Trustworthy adaptive adversarial perturbations in social networks. Journal of Information Security and Applications, 80: #103675 [DOI: 10.1016/j.jisa.2023.103675]

Zhang Q L, Li X D, Chen Y F, Song J K, Gao L L, He Y, et al. 2022b. Beyond ImageNet attack: towards crafting adversarial examples for black-box domains//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y W, Zhang W M, Chen K J, Liu J Y, Liu Y J and Yu N H. 2018b. Adversarial examples against deep neural network based steganalysis//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. Innsbruck, Austria: ACM: 67-72 [DOI: 10.1145/3206004.3206012]

Zhao J J, Wang J W and Wu J F. 2023. Adversarial attack method identification model based on multi-factor compression error. Journal of Image and Graphics, 28(3): 850-863 (赵俊杰, 王金伟, 吴俊凤. 2023. 基于多质量因子压缩误差的对抗样本攻击方法识别. 中国图象图形学报, 28(3): 850-863) [DOI: 10.11834/jig.220516]

Zhao Z Y, Liu Z R and Larson M. 2020. Towards large yet imperceptible adversarial image perturbations with perceptual color distance//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1036-1045 [DOI: 10.1109/CVPR42600.2020.00112]

Zhong Y Q, Liu X M, Zhai D M, Jiang J J and Ji X Y. 2022. Shadows can be dangerous: stealthy and effective physical-world adversarial attack by natural phenomenon//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15324-15333 [DOI: 10.1109/CVPR52688.2022.01491]

Zhou D W, Liu T L, Han B, Wang N N, Peng C L and Gao X B. 2021. Towards defending against adversarial examples via attack-invariant features//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 12835-12845

Zhou D W, Xu Y B, Wang N N, Liu D C, Peng C L and Gao X B. 2024. Generalized adversarial defense against unseen attacks: a survey. Journal of Image and Graphics, 29(7): 1787-1813 (周大为, 徐一搏, 王楠楠, 刘德成, 彭春蕾, 高新波. 2024. 针对未知攻击的泛化性对抗防御技术综述. 中国图象图形学报, 29(7): 1787-1813) [DOI: 10.11834/jig.230423]

Zhu J R, Kaplan R, Johnson J and Li F F. 2018. HiDDeN: hiding data with deep networks//Proceedings of the 15th European Conference

on Computer Vision. Munich, Germany: Springer: 682-697 [DOI: 10.1007/978-3-030-01267-0_40]

作者简介

张家伟，男，博士，讲师，主要研究方向为对抗样本取证和JPEG压缩取证。E-mail: zjwei_2020@163.com

王昊，通信作者，男，博士，讲师，主要研究方向为信息安全和JPEG压缩取证。E-mail: sa875923372@163.com

吴昊，男，博士研究生，主要研究方向为对抗样本取证和JPEG压缩取证。E-mail: howwooo@163.com

袁霖，男，副教授，主要研究方向为图像视频安全与隐私保护。E-mail: yuanlin@cqupt.edu.cn

程鑫，男，博士研究生，主要研究方向为信息安全和JPEG压缩取证。E-mail: chengxin0314@126.com

罗向阳，男，教授，主要研究方向为图像隐写和隐写分析技术。E-mail: luoxy_ieu@sina.com

马宾，男，教授，主要研究方向为可逆信息隐藏、多媒体取证、隐写与隐写分析。E-mail: sddxmb@126.com

王金伟，男，教授，主要研究方向为人工智能安全和多媒体取证。E-mail: wjwei_2004@163.com