计算机科学
COMPUTER SCIENCE

平衡可迁移与不可察觉的对抗攻击

康凯，王家宝，徐堃

引用本文

康凯，王家宝，徐堃. 平衡可迁移与不可察觉的对抗攻击[J]. 计算机科学，2025，52(6)：381-389.

KANG Kai, WANG Jiabao, XU Kun. Balancing Transferability and Imperceptibility for Adversarial Attacks [J]. Computer Science, 2025, 52(6): 381-389.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300083

平衡可迁移与不可察觉的对抗攻击

康 凯 王家宝 徐 堃
陆军工程大学指挥控制工程学院 南京 210007
(13913835075@139.com)

摘 要 基于数据驱动的深度学习模型由于无法覆盖所有可能样本数据,导致面临着精心设计的对抗样本的攻击问题。现有主流的基于RGB像素值的L_p范数扰动攻击方法虽然达到了很好的攻击成功率和迁移性,但是所生成的对抗样本存在极易被人眼感知的高频噪声,而基于扩散模型的攻击方法兼顾了迁移性和不可察觉性,但是其优化策略主要从对抗模型的角度展开,缺乏从代理模型的角度对可迁移性和不可察觉性的深入探讨和分析。为了进一步探索分析可迁移性和不可察觉性的控制来源,以基于代理模型的攻击方法为框架,提出了一种新的基于潜在扩散模型的对抗样本生成方法。该方法中,在基本的对抗损失约束条件下,设计了可迁移注意力约束损失和不可察觉一致性约束损失,实现了对可迁移性与不可察觉性的平衡。在ImageNet-Compatible,CUB-200-2011和Stanford Cars这3个公开数据集上,与已有方法相比,所提方法生成的对抗样本具有很强的跨模型迁移攻击能力和人眼不易觉察扰动的效果。

关键词:对抗攻击;扩散模型;可迁移性;不可察觉性;注意力机制
中图分类号 TP391

Balancing Transferability and Imperceptibility for Adversarial Attacks

KANG Kai,WANG Jiabao and XU Kun
College of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract Data-driven deep learning models face the problem of well-designed adversarial attacks due to their inability to cover all possible sample data. The existing main L_p-norm perturbation attack methods based on RGB pixel space have achieved great attack success rates and transferability,but the generated adversarial samples have high-frequency noise that is easily perceived by the human eye. The attack methods based on diffusion models balance transferability and imperceptibility,but their optimization strategies mainly focus on the perspective of adversarial models. Those researches lack deep exploration and analysis of transferability and imperceptibility from the perspective of surrogate model. In order to further explore and analyze the control sources of transferability and imperceptibility,a new adversarial sample generation method based on latent diffusion model is proposed within the framework of an attack method based on surrogate model. In this method,under the constraint of basic adversarial loss,transferable attention constraint loss and imperceptible consistency constraint loss are designed to achieve a balance between transferability and imperceptibility. On three publicly available datasets,ImageNet Compatible,CUB-200-2011,and Stanford Cars,compared with existing methods,the proposed method generates adversarial samples with strong cross-model transferable attack ability and the effect of imperceptible disturbance to the human eye.

Keywords Adversarial attacks,Diffusion model,Transferability,Imperceptibility,Attention mechanism

1 引言

当前,深度学习技术已经在诸多领域展现出惊人的性能,包括语言翻译[1]、自动驾驶[2]、医药图像分割[3]、遥感影像分析[4]等。但研究者也发现深度学习模型存在脆弱性,其对精心设计的对抗样本(也称攻击样本)会出现大幅度的性能下降[5]。对抗样本通过对原始样本增加人眼无法感知的微小扰动或视觉自然的小幅扰动,来实现误导深度学习模型预测结果的攻击能力,这种能力还可以泛化扩展到不同的深度模型架构,形成对深度学习模型系统的泛在威胁[6]。为了应对这些威胁,探索设计具有强泛化能力的对抗样本生成方法,以尽可能多地发现深度学习模型中存在的被攻击"盲点"和"漏洞"(即对抗样本),为增强深度学习模型的鲁棒性和可靠性提供借鉴和指导。

在商用和军用等情况下,智能模型的结构和参数通常都是无法预先获知的,可视为一种黑盒模型,针对黑盒模型的

到稿日期:2024-03-12 返修日期:2024-07-08
基金项目:江苏省自然科学基金(BK20200581)
This work was supported by the Natural Science Foundation of Jiangsu Province,China(BK20200581).
通信作者:王家宝(jiabao_1108@163.com)

攻击是困难的[7]。现有黑盒攻击方法从优化的角度可分为基于代理模型算法、基于元启发式算法、基于直接搜索算法和基于零阶优化算法[8]。其中，基于代理模型算法因其可直接针对代理模型设计优化目标，具有更好的训练学习效率而被广泛研究[5-6]。为了达成对抗样本在跨模型上的黑盒攻击迁移能力，早期方法基本都会假设训练代理模型和训练目标模型的数据具有一致的数据分布[9-10]。但是，这一强假设条件在很多时候是无法满足的，为此研究者进一步弱化了条件限制，在预训练模型和训练数据分布均无法获知的条件下来设计和实现更具迁移能力的对抗样本生成方法。目前，已知的此类方法包括 CDA[11]，BIA[5]等，但是这些方法都是在 L_p 范数扰动约束条件下生成的对抗样本，其迁移攻击能力有限，且存在易被觉察的高频噪声。

近来，研究者借助扩散模型[12-13]，在保证不可察觉的条件下生成具有较大扰动的对抗样本，实现了更强的跨模型迁移攻击能力[5]。与以往 L_p 范数扰动约束方法不同，基于扩散模型的方法通过扩散去噪过程来达成对图像内容的修改，在视觉上继承了去噪净化能力，具有更好的不可察觉特性，且攻击的迁移能力也很强。目前，较为高效的方法为 DiffAttack[6]，其基于扩散模型构建了对抗学习框架，并针对对抗模型(即扩散模型)设计了自注意力损失和结构一致性损失，以更好地保持生成对抗样本的可迁移能力和不可察觉效果。这一研究虽然得到了很好的性能结果，但是其设计的损失主要针对的是对抗模型，缺乏从代理模型的角度对可迁移性和不可察觉性的深入探讨和分析。

为了进一步探索分析可迁移能力和不可察觉效果的控制来源，本文借鉴现有针对代理模型的优化学习方法，设计并提出了一种新的基于潜在扩散模型的对抗样本生成方法。本文的主要贡献如下：

1)提出了一种新的基于潜在扩散模型的对抗样本生成方法。该方法生成的对抗样本具有很强的跨模型迁移攻击能力和人眼不易觉察扰动的效果。

2)设计了平衡可迁移与不可察觉的约束。设计了可迁移注意力约束损失和不可察觉一致性约束损失，从代理模型的角度探索了控制可迁移性和不可察觉性的可能性。

3)在 ImageNet-Compatible，CUB-200-2011 和 Stanford Cars 这 3 个公开数据集上与已有方法进行了对比实验，结果表明，所提方法生成的对抗样本具有更好的平衡可迁移性与不可察觉性。

2 相关工作

2.1 可迁移的对抗攻击

针对黑盒攻击任务，很多方法采用基于代理模型的算法[8]，通过构造与目标模型类似的代理模型，使得由攻击代理模型生成的对抗样本能迁移至对目标模型的攻击。当黑盒任务中代理模型与目标模型的网络结构不同时，通常会出现明显的性能差距，不同的攻击技术也会导致对抗样本的迁移性差异，因此，提高对抗样本的迁移性是黑盒攻击研究的重点。Yuan 等[14]提出了一种元梯度对抗攻击方法，通过集成多个模型的梯度信息，提升了样本的迁移能力；Xiong 等[15]通过集成多个模型来寻找最佳优化方向，实现多模型的迁移攻击能力；Zhu 等[16]将注意力机制引入对抗攻击，以破坏图像中与目标语义相关的注意力显著特征；Huang 等[17]基于注意力热图技术，提出了针对敏感区域感知的对抗攻击方法，通过寻找关键像素并进行扰乱生成样本。这些通过集成思想或注意力机制的方法在提升对抗样本的迁移能力上具有增强作用，但集成模型通常涉及多个模型，具有较大的计算代价。

在无法获知目标模型训练数据的条件下，Huan 等[18]利用预训练模型的概率输出，最大化干净样本与对抗样本间的散度。Duan 等[19]通过多任务生成模型学习原始数据分布，再利用分布信号生成相应目标样本，之后用其训练的代理模型指导生成对抗样本。这些方法实现了在不同数据集下模型的有效训练，为相关研究提供了不同的思路。

2.2 不可察觉的攻击扰动

在通过扰动生成对抗样本的优化目标中，通常以 RGB 图像空间中的 L_p 范数对扰动进行约束，但是其在度量感知距离上并不是非常好，容易出现高频噪声。最近研究者开始关注不受约束条件下的生成，提出了不受限制但视觉不可察觉的攻击，例如 Qiu 等[20]和 Jia 等[21]通过修改图像(人脸)属性达成了不可察觉性；Yuan 等[22]通过构建颜色分布库，用其找到了一个成功的分布用于对抗攻击。这些方法虽然达成了不可察觉，但在迁移能力上无法与先前的基于像素的方法竞争。

近来，扩散模型[12-13]展现出了高质量的图像生成效果，其通过前向过程迭代加入噪声，通过反向过程逐步预测噪声并进行去噪，可生成高质量、多样性的图像。特别地，很多模型还可以由文本提示词进行指导生成，被广泛用于图像超分[23]、编辑等[24]任务中。当然，扩散模型优秀的去噪能力[25]也被用于提升攻击样本的不可察觉能力，例如 DiffAttack[6]利用扩散模型来生成对抗样本，同时达成了更好的可迁移性和不可察觉性；Liu 等[26]提出的 Adv-Diffusion，利用潜在扩散模型也实现了不可察觉的人脸攻击。以上方法实现了较 L_p 范数约束更好的对抗样本不可察觉效果，但是对不可察觉性的控制因素和效果影响等，还缺乏深入的探索和分析。

3 基于潜在扩散模型的不可察觉对抗性攻击方法

图 1 给出了所提出的基于潜在扩散模型的不可察觉对抗性攻击方法的框架。该框架主要包括一个用于生成对抗样本的对抗模型和一个用于代替不可预知的识别模型的代理模型。其中，前者基于干净图像生成对抗样本，后者对前者生成的样本进行识别评测，从而指导对抗模型优化生成的对抗样本。为了达成攻击方法不仅对代理模型有效，而且对其他模型也同等有效的泛化能力，即提升方法跨模型的可迁移攻击能力，基于识别模型中间层特征的跨通道注意力通常反映了

目标及其部件的可辨识区域信息[5]的现象,针对代理模型设计提出了可迁移注意力约束损失,以实现攻击代理模型等价于攻击其他模型的可迁移能力。同时,为了让对抗样本的篡改内容不易被人感知,针对代理模型设计了保持对抗样本特征与干净样本特征的不可察觉一致性约束损失,实现了对抗样本在视觉上的不可察觉效果。

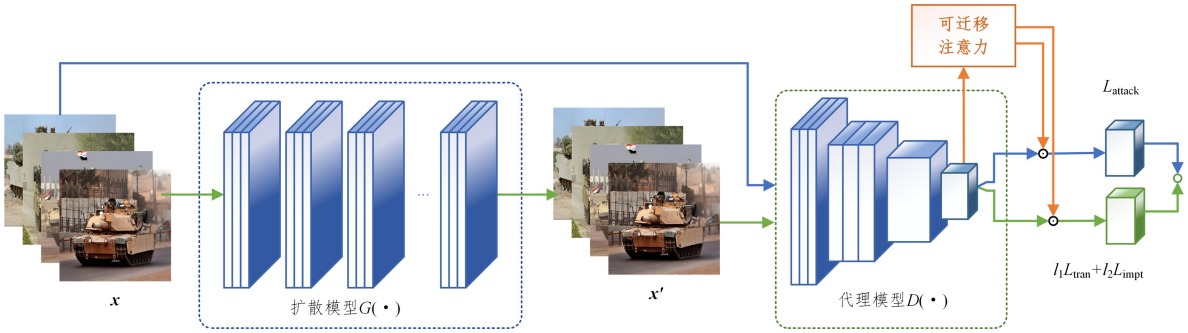


图 1　基于潜在扩散模型的对抗样本生成方法的框架

Fig. 1　Architecture of adversarial sample generation method based on latent diffusion model

形式化描述:给定在特定数据分布$\mathcal{X}$上训练的识别模型$T(\cdot)$,任务的目的是为来自服从$\mathcal{X}$的图像$\boldsymbol{x}$精心制作一个视觉自然的扰动,实现对模型$T(\cdot)$的欺骗。如果将制作扰动的过程定义为一个攻击方法$Attack(\cdot)$,则由$\boldsymbol{x}$制作对抗样本$\boldsymbol{x}'=Attack(\boldsymbol{x})$能成功欺骗模型$T(\cdot)$的目标可描述为:

$$\begin{aligned} &T(\boldsymbol{x}')\neq T(\boldsymbol{x})\\ &\text{s.t. } d(\boldsymbol{x}',\boldsymbol{x})\leqslant\tau \end{aligned} \tag{1}$$

其中,$d(\cdot,\cdot)$是用于度量样本$\boldsymbol{x}$与其生成的对抗样本$\boldsymbol{x}'$之间的距离,τ表示施加的最大扰动。

理论上,虽然无法获知目标模型$T(\cdot)$,以及训练$T(\cdot)$所使用数据的分布$\mathcal{X}$,但是现有图像领域的识别模型基本都是在 ImageNet 甚至更大规模的公开数据集上执行预训练,因此可以假设模型$T(\cdot)$具有 ImageNet 先验知识。基于以上假设,则可以通过在数据分布$\mathcal{Y}$和其上训练的替代预训练模型$D(\cdot)$来学习对抗模型,将目标域的学习任务转化为源域上的学习任务,即:

$$\begin{aligned} &D(\boldsymbol{x}')\neq D(\boldsymbol{x})\\ &\text{s.t. } d(\boldsymbol{x}',\boldsymbol{x})\leqslant\tau \end{aligned} \tag{2}$$

其中,$\boldsymbol{x}\in\mathcal{Y}$,$\boldsymbol{x}'=Attack(\boldsymbol{x})$,$Attack(\cdot)$方法所使用的对抗模型为一个扩散模型$G$,其参数在源域上学习得到。

3.1　扩散对抗模型

所提方法采用开源的 Stable Diffusion[27]扩散模型作为对抗模型G。扩散模型的核心是扩散过程,即给定一个样本$\boldsymbol{x}_0=\boldsymbol{x}$,扩散过程就是渐进地加噪,生成噪声样本$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T$,其中$\boldsymbol{x}_t=\sqrt{\alpha_t}\boldsymbol{x}_{t-1}+\sqrt{1-\alpha_t}\,\epsilon,\epsilon\sim\mathcal{N}(0,1)$。当时间步$T$足够大时,最终的$\boldsymbol{x}_T$将趋于一个各向同性的高斯分布,该加噪过程也称为前向过程。这一渐进过程可通过重参数化直接转化为一步处理:

$$\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon,\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i \tag{3}$$

给定一个高斯分布噪声数据,扩散模型通过反向过程执行去噪操作来生成新样本。具体地,在每一步中通过一个去噪模块$\epsilon_\theta(\cdot,t)$预测的噪声对x_t进行去噪,该去噪模块通常为一个 U-Net 网络,并通过优化以下目标进行训练:

$$\min E_{x_0,t\sim u(t),\epsilon\sim N(0,1)}\parallel\epsilon-\epsilon_\theta(\boldsymbol{x}_t,t)\parallel_2^2 \tag{4}$$

以上扩散过程在原始图像空间上的训练代价极大。为此,研究者提出了潜在扩散模型[27],在一个相对低维的隐空间执行上述前向和反向过程,以减小原始像素空间的训练和采样代价。具体地,采用一个编码器$\mathcal{E}(\cdot)$将高维图像映射到低维隐空间,采用一个解码器$\mathcal{D}(\cdot)$将低维隐变量映射回高维图像空间。为了简化表示,下文中使用的$\boldsymbol{x}_t$均为隐空间表示。具体实施中,对由$\boldsymbol{x}_0$到$\boldsymbol{x}_t$的多个时间步反向执行 DDIM(Denoising Diffusion Implicit Models)确定性去噪过程[2],即可由$\boldsymbol{x}_t$恢复一个高质量的$\boldsymbol{x}_0$。

给定干净样本$\boldsymbol{x}$,其对应标签为y,由对抗模型生成的对抗样本为$\boldsymbol{x}'$。为了让对抗模型生成具有攻击能力的样本,训练过程中对潜在扩散模型的隐变量$\boldsymbol{x}_t$进行优化,优化目标为:

$$L_{\text{attack}}=-J(D(x'),\mathcal{Y}) \tag{5}$$

其中,$J(\cdot)$表示交叉熵损失,D表示代理模型,$x'=x_0'=\mathcal{R}\circ\cdots\circ\mathcal{R}(x_t)$,$\mathcal{R}(\cdot)$表示扩散去噪过程。

与基于像素的攻击方法[5]生成样本存在的极易感知的高频噪声相比,基于扩散模型生成的对抗样本包含更多的高级语言线索,且更加自然,主要原因是扩散模型本身就是一个去噪过程,可以有效减少易被感知的高频噪声。已有研究表明,这些语义丰富的扰动不仅不易被感知,而且有利于增加样本攻击的可迁移性[6]。

3.2　平衡可迁移与不可察觉的约束

3.2.1　可迁移注意力约束损失

为了进一步缩小对抗样本在代理模型与目标模型之间的性能差异,与已有方法[6]在扩散模型(即对抗模型)的隐空间进行注意力设计以提升泛化能力不同,我们认为对代理模型进行注意力设计可以更加有效地对目标模型产生作用,因此基于已有研究[4-5]表明的集成策略可以避免模型陷入局部极值、改变泛化能力的结论,我们直接对代理模型的中间层输出特征进行跨通道集成,形成一个跨通道的可迁移注意力,将该注意力作用于对抗样本与干净样本,构建可迁移注意力约束目标。

跨通道的可迁移注意力基于这样一个现象假设,即不同

的深度学习识别模型，对同一图像提取的中间层特征在跨通道融合后，会反映出基本一致的目标及其关键部件区域[5]。这意味着利用跨通道注意力对所提取的特征进行加权可以获得更精确的目标及部件特征描述，这将使得不同模型对同一图像所提取的特征更趋于一致。基于此，对于由这种趋于一致的特征所进行的识别决策而言，对代理模型的成功攻击也意味着对其他模型有着可迁移的成功攻击。

具体地，假设代理模型 $D(\cdot)$ 在第 m 个中间层提取的特征 $F=f(x)$，则可以通过通道注意力机制对特征按通道进行集成，生成跨通道的可迁移注意力 A。

$$A=\frac{|\sum_{i=0}^{C}F_i|}{C} \tag{6}$$

其中，C 为特征图 F 的通道数，F_i 为特征图 F 的第 i 个通道特征。

经过以上操作获得的注意力在不同模型之间会呈现类似的效果，具有鲁棒的跨模型无关属性[5]。为了提升对抗样本的跨模型攻击能力，在前述学习过程中，我们将可迁移注意力 A 应用至中间层特征，并构建对抗样本和干净样本对目标的注意力加权后的差异损失 L_{tran}。

$$L_{\text{tran}}=J_{\cos}(g(A\odot f(x)),g(A\odot f(x'))) \tag{7}$$

其中，$J_{\cos}$ 表示余弦相似度，$g(\cdot)$ 表示将高维张量特征拉伸为一维向量，$\odot$ 表示哈达玛积。上式最小化损失 L_{tran} 等价于最小化对抗样本的注意力特征 $A\odot f(x')$ 与干净样本的注意力特征 $A\odot f(x)$ 的余弦相似度，意味着使这两个注意力特征尽可能不同，那么由这两个注意力特征预测的结果也会尽可能不同，即对抗攻击更可能成功。由于跨通道的可迁移注意力 A 在不同模型上具有类似的关注目标及部件区域的一致性，因此由该注意力加权生成干净样本注意力特征 $A\odot f(x)$ 或对抗样本特征注意力特征 $A\odot f(x')$，在不同模型之间都更具有一致性。当对抗样本成功攻击代理模型时，也意味着其更可能成功攻击其他模型，即对抗样本具有更好的跨模型迁移攻击能力。

3.2.2 不可察觉一致性约束损失

鉴于基于像素的对抗攻击方法生成的对抗样本存在人极易感知的高频噪声（见图 2），因此在考虑攻击成功率的同时，增加对抗样本的不可察觉性也非常重要。与 L_p 范数约束的攻击前后样本的像素级变化不同，我们从保持对抗样本与干净样本的特征一致性角度设计损失，使生成对抗样本具有不可察觉特性。借鉴超分辨率方法[28]中的感知一致性损失，利用代理模型的中间层输出特征，设计了关于约束对抗样本与干净样本的不可察觉一致性的约束损失。

具体地，假设代理模型 $D(\cdot)$ 在第 m 个中间层输出的特征 $F=f(x)$，则不可察觉一致性的约束损失 L_{impt} 为：

$$L_{\text{impt}}=-J_{\cos}(g(F),g(F')) \tag{8}$$

其中，$J_{\cos}(\cdot,\cdot)$ 表示余弦相似度，$g(\cdot)$ 表示将高维张量特征拉伸为一维向量，$F'=f(x')$ 表示对抗样本经代理模型提取的特征。该约束利用代理模型约束对抗样本与干净样本在特征感知层的一致性，能更直接达成对抗样本的不可察觉性。

综上，模型迁移攻击能力的强弱与对抗模型对干净图像扰动幅度有关，扰动幅度大则攻击成功率高，但是扰动变化极易被人类感知发现，因此在不可察觉条件下增强模型的迁移攻击能力更为重要。综合学习损失为：

$$L_{\text{total}}=L_{\text{attack}}+\lambda_1 L_{\text{tran}}+\lambda_2 L_{\text{impt}} \tag{9}$$

其中，参数 λ_1 和 λ_2 为对应损失项的加权因子。

4 实验

4.1 数据、模型及设置

基于已有方法[6,22]的实验设置，我们主要采用 ImageNet-Compatible 数据集进行评测，该数据集包括 1 000 幅 299×299 分辨率的彩色图像。由于原始 Stable Diffusion 模型无法处理原始的图像大小，因此实验统一将图像缩放至 224×224 像素大小。同时，还评测了细粒度图像分类数据集 CUB-200-2011[29]和 Stanford Cars[30]。

评测具体采用了两类共 9 种不同的网络结构，包括卷积架构中的 ResNet-50（Res-50）[31]，VGG-19[32]，Inception-v3（Inc-v3）[33]，MobileNet-v2（Mob-v2）[34]和 ConvNeXt[35]，以及 Transformer 架构中的 ViT-B/16（ViT-B）[36]，Swin-B[37]，DeiT-B[38]和 DeiT-S[38]模型。此外，还评测了由对抗样本训练的模型，包括 Adv-Inc-v3[39]，Inc-v3ens3[40]，Inc-v3ens4[40]和 IncRes-v2ens[40]。

实验采用 Ubuntu 18.04 操作系统，Intel Xeon E5-2673 v4 CPU 16 块，内存 128 GB，英伟达 RTX 3090 显卡，显存 24 GB，CUDA 版本 11.3，PyTorch 版本 1.12.0。采用 DDIM[41]作为 Stable Diffusion 模型的采样器。扩散步骤设置为 20，并应用 5 次 DDIM 逆步骤处理干净图像。在逆向过程中，指导尺度设置为 0，在去噪过程中设置为 2.5，采用 AdamW 优化器，初始学习率为 0.01，迭代次数设置为 30。无特殊说明，参数 $\lambda_1=0.2$，$\lambda_2=0.2$。

实验采用 Top-1 精度来评测攻击方法的性能，精度越低表示攻击成功率越高，同时采用 FID（Frechet Inception Distance）[42]指标通过计算真实数据与生成样本之间的距离，来评价对抗样本的不可察觉效果，数值越低则表明图像质量越高，多样性越好。

4.2 主体实验

为了评测所提方法的基本性能，我们对比了正常训练的模型，包括 5 种基于像素的攻击方法（MI-FGSM[43]，DI-FGSM[44]，TI-FGSM[45]，PI-FGSM[46]和 S2I-FGSM[47]），以及两种无限制攻击方法（PerC-AL[48]和 NCF[22]），这些方法遵从其原始最优参数设置，并统一将输入图像分辨率调整至 224×224 大小。生成对抗样本的对抗模型包括 Res-50，VGG-19，Mob-v2，Inc-v3，ConvNeXt 和 Swin-B。实验结果如表 1 所列。其中，“S.”表示代理（Surrogate）模型，“T.”表示目标（Target）模型。当代理模型和目标模型一致时，对应数据白盒攻击（背景底色为灰色），否则为黑盒攻击。其中“AVG(w/o self)”表示去除与代理模型一致的模型结果后计算的平均精度，即对所有黑盒模型攻击后的平均精度；“FID”由 1 000 幅图像与 ImageNet 验证集计算得到，表中最佳结果用粗体表示，次优结果用下划线表示。

表 1　正常训练模型的评估结果

Table 1　Evaluation results of normally trained models

S. \ T.	Attacks	CNNs/%					Transformers/%				AVG (w/o self)/%	FID
		Res-50	VGG-19	Mob-v2	Inc-v3	ConvNeXt	ViT-B	Swin-B	DeiT-B	DeiT-S		
	Clean	92.7	88.7	86.9	80.5	97.0	93.7	95.9	94.5	94	91.5	57.8
Res-50	MI-FGSM	0.0	19.9	20.2	28.9	57.8	67.3	67.0	72.4	67.0	50.1	81.2
	DI-FGSM	0.0	21.2	20.5	34.5	71.6	82.0	75.3	80.5	76.0	57.7	85.3
	TI-FGSM	0.0	42.4	37.1	46.0	83.6	81.6	83.7	84.5	79.0	67.2	66.0
	PI-FGSM	0.0	14.1	15.0	24.0	72.5	65.3	77.5	76.7	65.0	51.3	97.9
	S^2I-FGSM	0.0	9.2	6.6	18.6	44.1	63.9	52.0	65.9	59.0	39.9	79.8
	PerC-AL	6.5	83.1	80.2	76.4	96.0	93.9	94.8	94.4	93.0	89.0	**58.2**
	NCF	11.3	30.5	30.3	52.6	78.3	65.7	76.8	75.1	67.0	59.5	70.9
	DiffAttack	3.7	24.4	22.9	31.0	41.0	48.8	43.8	49.5	45.0	38.3	62.6
	Ours	0.5	23.7	21.3	28.8	41.4	49.4	43.6	47.4	44.3	**37.5**	63.3
VGG-19	MI-FGSM	22.7	0.0	15.4	33.5	53.2	73.2	63.3	74.7	68.0	50.5	82.4
	DI-FGSM	32.2	0.0	23.9	46.5	67.2	84.7	71.9	84.8	80.0	61.4	70.9
	TI-FGSM	44.5	0.0	32.8	47.4	77.8	81.4	79.3	83.6	79.0	65.7	66.6
	PI-FGSM	22.7	0.0	16.4	29.8	68.3	68.0	75.7	79.5	68.0	53.6	96.4
	S^2I-FGSM	17.9	0.0	11.3	31.8	49.5	74.1	57.9	76.0	68.0	48.3	82.9
	PerC-AL	87.5	4.6	79.0	76.1	95.1	94.2	94.0	94.3	93.0	89.2	**57.9**
	NCF	38.3	6.8	31.5	52.4	80.5	67.5	77.6	77.4	71.0	62.0	70.4
	DiffAttack	21.1	2.7	19.4	29.7	43.1	52.9	41.6	51.3	45.0	**38.0**	63.9
	Ours	20.6	0.2	17.7	30.1	44.1	52.1	41.4	52.1	45.6	**38.0**	63.3
Mob-v2	MI-FGSM	26.4	18.7	0.0	31.0	62.0	69.5	65.2	71.6	63.0	50.9	76.4
	DI-FGSM	28.7	18.9	0.0	33.9	73.4	79.9	71.4	79.6	75.0	57.6	78.6
	TI-FGSM	47.2	37.9	0.0	45.2	83.0	79.9	80.9	81.8	76.0	66.5	65.6
	PI-FGSM	21.1	13.3	0.0	27.6	74.4	65.3	77.0	77.4	66.0	52.8	98.7
	S^2I-FGSM	21.0	13.4	0.0	27.2	64.3	74.1	62.6	75.2	68.0	50.7	79.4
	PerC-AL	88.2	84.2	5.9	76.8	96.2	93.9	94.2	94.3	94.0	90.2	**58.1**
	NCF	36.0	29.4	7.4	51.9	77.4	67.2	76.1	76.1	68.0	60.3	69.7
	DiffAttack	23.6	23.4	1.8	31.6	50.3	51.4	45.8	53.4	46.0	40.7	62.9
	Ours	23.2	23.4	0.3	30.5	48.2	52.2	44.4	51.4	45.1	**39.8**	62.8
Inc-v3	MI-FGSM	42.8	36.6	34.4	0.0	79.8	75.3	79.4	78.6	73.0	62.5	80.5
	DI-FGSM	61.7	57.4	51.9	0.2	89.9	84.6	86.8	86.7	82.0	75.1	67.1
	TI-FGSM	76.0	70.1	66.7	0.1	93.8	88.7	91.2	89.7	88.0	83.0	62.8
	PI-FGSM	37.9	22.4	28.4	0.0	81.0	74.3	83.0	81.9	72.0	**60.1**	92.5
	S^2I-FGSM	52.3	47.8	43.3	0.0	86.3	80.8	84.1	83.8	78.0	69.6	72.5
	PerC-AL	90.8	87.0	85.8	7.7	97.5	93.6	95.1	94.2	94.0	92.3	**58.4**
	NCF	52.6	45.8	46.2	17.4	85.7	75.9	83.4	82.7	76.0	68.5	66.7
	DiffAttack	59.5	55.6	55.4	13.9	76.9	75.2	72.8	74.0	71.0	67.6	62.3
	Ours	60.1	56.8	56.1	3.6	78.1	74.8	75.5	75.9	72.7	68.8	61.9
ConvNeXt	MI-FGSM	34.5	22.4	26.5	41.9	0.0	63.4	18.0	56.5	56.0	39.9	84.5
	DI-FGSM	33.6	24.3	29.8	46.6	0.0	71.0	18.8	62.2	64.0	43.8	79.6
	TI-FGSM	50.7	37.3	41.1	51.8	0.0	70.9	38.8	68.6	69.0	53.5	73.5
	PI-FGSM	23.6	14.2	17.1	22.4	0.0	43.0	37.2	48.7	43.0	31.2	101.8
	S^2I-FGSM	13.6	9.6	11.9	20.2	0.0	35.4	4.2	31.0	31.0	**19.6**	99.4
	PerC-AL	89.0	84.5	84.0	77.5	88.9	92.8	90.0	92.4	92.0	87.8	**57.7**
	NCF	47.1	41.4	39.2	54.7	41.4	61.6	63.9	64.8	62.0	54.3	67.0
	DiffAttack	20.9	24.8	21.8	25.8	1.9	26.7	11.4	21.6	24.0	22.1	73.3
	Ours	19.1	23.0	20.9	24.1	1.1	25.4	10.4	18.5	23.4	20.6	75.4
Swin-B	MI-FGSM	55.7	42.3	42.7	55.2	42.5	70.6	0.9	64.2	64.0	54.7	72.8
	DI-FGSM	52.7	43.0	44.5	56.4	33.9	66.6	2.7	57.2	58.0	51.5	65.7
	TI-FGSM	71.9	61.7	56.9	60.2	66.0	76.3	1.9	72.2	72.0	67.2	65.9
	PI-FGSM	38.3	21.6	25.8	35.7	54.8	48.4	0.6	52.4	47.0	40.5	89.7
	S^2I-FGSM	47.4	37.8	35.4	45.3	26.8	48.5	1.0	46.2	45.0	41.6	68.2
	PerC-AL	92.2	87.4	85.5	78.5	94.6	94.0	6.3	94.1	93.0	89.9	**57.9**
	NCF	49.5	44.9	44.9	60.5	70.1	63.7	36.9	66.0	63.0	57.8	65.5
	DiffAttack	43.5	42.1	40.7	41.4	34.0	39.0	9.9	35.0	37.0	**39.1**	65.5
	Ours	43.4	43.1	41.2	40.6	33.2	39.5	9.9	34.6	37.2	**39.1**	66.5

同时,由表 1 可以观测到,所提方法在不同的模型架构上都实现了最佳的迁移攻击效果。其中,与 DiffAttack 同作为基于扩散模型的攻击方法,较基于像素的攻击方法在 Res-50,VGG-19,Mob-v2,ConvNeXt 上的性能提升了近 10%。当然,Inc-v3 模型的攻击性能与个别基于像素的攻击方法相比略差,但其对应的 FID 性能要好很多。从对抗样本的不可察觉角度来看,PerC-AL 具有最佳的 FID 性能,但是其对黑盒模型的迁移能力最差,其对应的 AVG(w/o self)性能与干净(Clean)图像的 91.5 非常接近,几乎没有什么攻击能力。因此,该方法不具有太大的比较意义。

我们可视化了不同攻击方法生成的对抗样本,结果如图 2 所示。由图 2 可以发现,MI-FGSM,DI-FGSM,TI-

FRGSM, PI-FGSM, S2I-FGSM 方法的攻击样本均呈现出明显的高频噪声，很容易被人感知，相比较而言，所提方法视觉上更具不可察觉性，与 NCF 相比，颜色空间更加自然。虽然 PerC-AL 也呈现出非常好的不可察觉性，甚至较所提方法更好，但是其迁移攻击能力最差。可见，所提方法在视觉自然的条件下达到了更优的攻击性能。

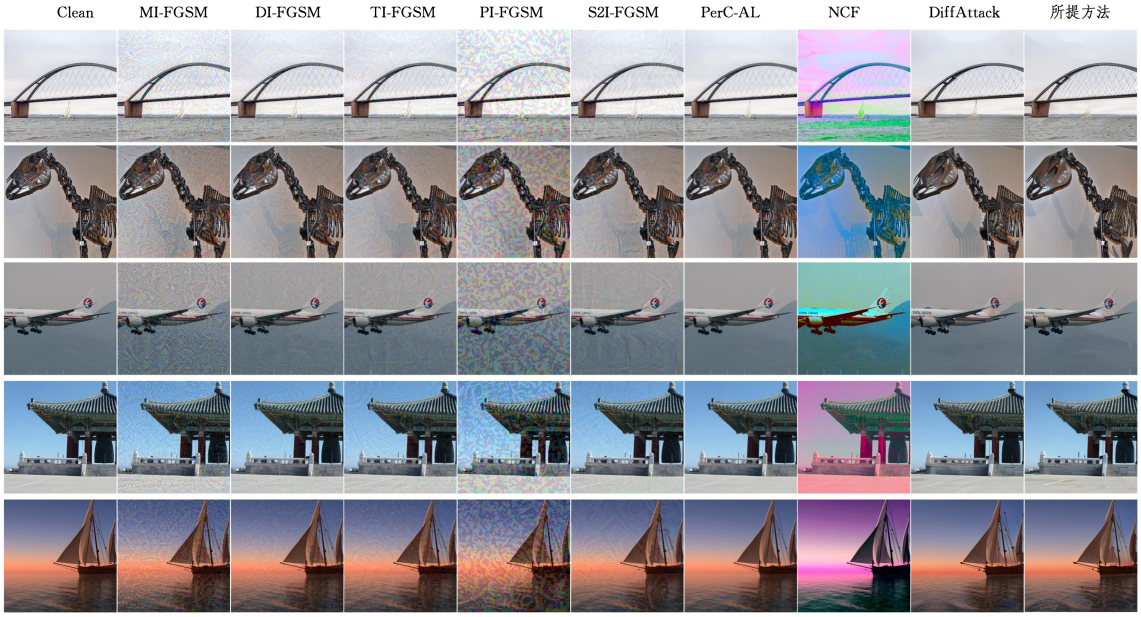


图 2 生成对抗样本效果对比图

Fig. 2 Comparison of generated adversarial sample effects

为了进一步验证模型的迁移攻击能力，我们对多个采用对抗样本训练的防御方法进行了评测。借鉴前人的规范，我们对比了 4 种防御方法，包括 Adv-Inc-v3, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ 和 IncRes-v2$_{ens}$。所提方法采用 Inc-v3 作为代理模型，以保证更加公平的对比，结果如表 2 所列。

表 2 防御模型的评测结果

Table 2 Evaluation results of defense models

(%)

A.	D.				
	Inc-v3$_{normal}$	Adv-Inc-v3	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
MI-FGSM	0.0	51.1	49.8	53.7	70.5
DI-FGSM	0.2	64.2	58.5	61.5	74.9
TI-FGSM	0.1	66.1	62.4	64.5	76.8
PI-FGSM	0.0	**42.3**	45.0	44.6	62.0
S2I-FGSM	0.0	51.8	47.0	52.2	67.7
PerC-AL	7.7	80.8	76.7	75.4	88.6
NCF	17.4	48.8	47.2	49.0	60.5
DiffAttack	13.9	46.0	43.8	43.1	**58.3**
Ours	3.6	43.7	**41.1**	**42.7**	**58.3**

表 2 中，“A.”表示攻击(Attack)模型，“D.”表示防御(Defense)模型，“Inc-v3$_{normal}$”是普通训练的模型对比基准，最佳结果用粗体表示，次优结果用下划线表示。由表 2 可以发现，所提方法对于无防御白盒模型(Inc-v3$_{normal}$)具有极好的攻击效果，攻击成功率较 DiffAttack 方法提升了 10.3%；对防御黑盒模型而言，攻击效果则差很多，但所提方法与其他方法相比，具有更好的攻击性能，且对不同防御模型而言攻击效果差不多，展现出了较好的鲁棒性。

此外，在 CUB-200-2011 和 Stanford Cars 两个更具挑战性的细粒度识别数据集上进行了评测。根据文献[6]的设置，选取 1000 幅图像生成对抗样本进行评测，模型采用 Zhang 等[5]预训练的 ResNet-50(Res-50)，SENet-154(SE-154)和 SE-ResNet101(SR-50)，对应结果如表 3 所列。其中，“AVG”表示去除与代理模型一致的模型结果后计算的平均精度；同时，根据表 1 所列结果，考虑到 PerC-AL 方法的迁移性太差，在此未将其包含在对比方法中。

表 3 细粒度分类攻击的评测结果

Table 3 Evaluation results of fine-gained classification attack

T. / S.	Attacks	CUB-200-2011					Stanford Cars				
		Res-50%	SE-154%	SR-101%	AVG/%	FID	Res-50%	SE-154%	SR-101%	AVG/%	FID
Res-50	Clean	75.7	80.5	76.6	77.6	11.1	73.9	76.4	74.4	74.9	11.6
	MI-FGSM	3.1	40.7	32.7	36.7	31.7	0.1	33.5	25.9	29.7	41.3
	DI-FGSM	0.3	42.7	33.8	38.3	20.9	0.1	33.3	29.3	31.3	28.7
	TI-FGSM	2.8	50.6	43.9	47.3	21.1	0.1	46.9	41.0	44.0	23.2
	PI-FGSM	9.1	35.2	26.2	30.7	34.8	1.5	31.5	23.2	27.4	53.2
	S^2I-FGSM	0.7	35.1	28.1	31.6	24.3	0.1	25.7	24.4	25.1	34.4
	NCF	0.2	22.7	13.9	18.3	35.2	6.6	46.0	38.4	42.2	24.1
	DiffAttack	3.3	19.3	16.7	18.0	**20.6**	0.1	15.1	13.1	14.1	**17.8**
	Ours	0.1	17.0	12.3	**14.7**	21.8	0.0	16.0	11.3	**13.7**	18.4

（续表）

S. \ T.	Attacks	CUB-200-2011 Res-50%	CUB-200-2011 SE-154%	CUB-200-2011 SR-101%	CUB-200-2011 AVG/%	CUB-200-2011 FID	Stanford Cars Res-50%	Stanford Cars SE-154%	Stanford Cars SR-101%	Stanford Cars AVG/%	Stanford Cars FID
	Clean	75.7	80.5	76.6	77.6	11.1	73.9	76.4	74.4	74.9	11.6
SE-154	MI-FGSM	41.7	0.2	42.3	42.0	37.9	32.4	0.0	33.8	33.1	41.3
	DI-FGSM	54.5	0.2	48.9	51.7	23.5	45.6	0.1	45.5	45.6	29.1
	TI-FGSM	60.1	0.3	56.2	58.1	20.8	54.1	0.1	53.2	53.7	23.0
	PI-FGSM	30.5	0.0	33.1	31.8	46.5	21.9	0.0	26.3	**24.1**	59.6
	S^2I-FGSM	43.2	0.0	34.0	38.6	25.4	27.7	0.0	25.7	26.7	33.6
	NCF	13.5	6.8	17.6	**15.5**	35.0	38.5	20.7	41.6	40.1	23.3
	DiffAttack	53.8	2.5	51.3	52.6	**17.9**	37.3	0.9	32.5	34.9	16.2
	Ours	54.1	2.4	51.0	52.6	19.0	37.7	0.8	34.7	35.2	**16.1**
SR-101	MI-FGSM	32.8	36.0	0.1	34.4	41.0	25.5	27.6	0.0	26.6	44.6
	DI-FGSM	39.4	38.0	0.2	38.7	23.5	28.1	29.3	0.2	28.7	28.5
	TI-FGSM	53.4	55.3	0.2	54.4	**21.8**	48.7	49.8	0.0	49.3	22.5
	PI-FGSM	21.7	29.8	0.0	25.8	45.5	18.5	29.3	0.0	23.9	59.9
	S^2I-FGSM	30.4	31.5	0.0	31.0	26.7	20.5	17.1	0.1	18.8	36.9
	NCF	9.4	20.2	3.1	**14.8**	33.3	33.4	46.8	12.1	40.1	24.0
	DiffAttack	27.0	23.5	3.9	25.3	22.4	17.5	16.0	0.3	16.8	**18.0**
	Ours	27.7	23.9	3.6	25.8	25.0	15.7	15.8	0.1	**15.8**	18.1

由表 3 可知，所提方法在两个细粒度识别数据集上与基于像素的攻击方法相比，取得了相对较优的 AVG 和 FID 性能，且与 DiffAttack 的性能非常接近，这说明对于细粒度识别任务，所提方法从代理模型角度可以达成与从对抗模型角度进行优化同等的效果，具有较强的不可察觉性和迁移攻击能力。横向对比 3 种不同的预训练细粒度识别模型，可以发现 SE-154 在保持与 Res-50 和 SR-101 差不多 FID 的条件下，攻击成功率出现了较大的差距，反映出了该模型的跨模型迁移攻击能力相对较差，主要原因可能是和模型的深度有关，相对复杂的深度模型可能更容易拟合训练数据，存在迁移能力不足的缺点；同时，在 SE-154 模型下，各类攻击方法的迁移性和不可察觉性是矛盾的两面，当迁移攻击成功率高（AVG 数值低）时，对应不可察觉效果较好（FID 数值高）；反之亦然。

4.3　消融实验

为了评测所提出的迁移能力与不可察觉的平衡约束，将仅保留攻击损失的方法作为基准，分别添加可迁移注意力约束损失和不可察觉一致性约束损失，测试各模块的效果和作用，评测采用 Res-50 作为代理模型，结果如表 4 所列。其中，“AVG”为排除 Res-50 后剩余的 8 种正常训练的模型和 4 种防御模型的平均结果。

表 4　消融实验的评测结果

Table 4　Evaluation results of ablation experiments

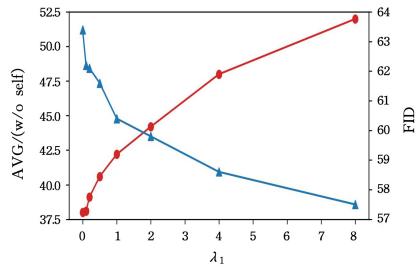
L_{attack}	L_{tran}	L_{impt}	AVG/%	FID
√	×	×	38.0	63.4
√	√	×	**37.4**	64.7
√	×	√	39.1	**62.1**
√	√	√	37.8	63.3

从表 4 可以发现，单独使用 L_{tran} 虽然可以提升迁移攻击效果，但是对应的 FID 分值增大，类似的单独使用 L_{impt} 虽然可以提升 FID，但是对应的迁移攻击性能下降；综合使用两者后 AVG 和 FID 均有提升。由此可见，攻击的迁移性和不可察觉性是一对矛盾，需要平衡两者进行具体实施。

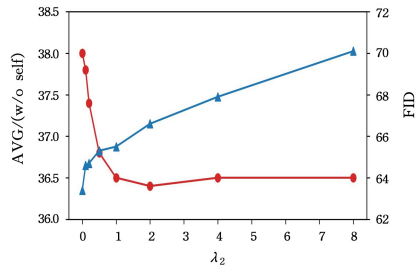
4.4　参数分析

本节对综合损失中的损失项权重参数 λ_1 和 λ_2 进行分析，主要按以下规则进行：先在一个大的范围进行评测，再在相对较小的范围搜索。

由于可迁移注意力约束损失和不可察觉一致性约束损失均采用余弦距离度量，同时考虑到 CE 损失的一般范围，采用 0.0，0.1，0.2，0.5，1，2，4，8 分别对权重 λ_1 和 λ_2 进行初步评测，评测采用 Res-50 作为代理模型，实验结果如图 3(a)和图 3(b)所列。可以发现，当 λ_1 增加时，AVG 不断增大，FID 不断降低，表明可迁移注意力约束损失权重越大，迁移能力越差，但不可察觉性越好；当 λ_2 增加时，AVG 总体在不断减小，FID 不断增大，表明不可察觉一致性约束损失权重越大，迁移能力越好，但不可察觉性越差。由此可以发现，这两个损失分别从两个不同的方面进行约束。



(a)



(b)

图 3　不同参数的性能评测结果

Fig. 3　Performance evaluation results of different parameters

综合考虑参数 λ_1 和 λ_2 的影响，为了平衡可迁移性和不可察觉性，并考虑图 2 展示的不可察觉效果，选择 $\lambda_1=0.2$ 和 $\lambda_2=0.2$ 作为一个较好的可行解。

结束语　针对对抗样本的跨模型迁移能力和视觉不可觉察效果难以平衡的问题，基于潜在扩散模型提出了一种从代

理模型角度设计损失进行优化学习的新方法。所提方法与从对抗模型角度设计损失进行优化学习的 DiffAttack 方法相比，在具有相近的不可察觉效果条件下，取得了更优的攻击迁移性能。研究结果表明，生成兼具可迁移和不可察觉的对抗样本除了从对抗模型（扩散模型）角度进行优化设计，还可以从代理模型角度进行优化设计，且优化效果相当甚至更好。

基于潜在扩散模型的对抗样本生成方法，虽然达成了很好的迁移性和不可察觉性，但是扩散模型本身的扩散过程极其耗时，导致单幅对抗样本的生成耗时需要 30 s 左右，且与基于 RGB 像素值的 L_p 范数扰动攻击方法相比，还存在非常大的优化空间。对此，未来将借鉴更加快速的扩散模型相关研究工作，以提升样本的生成速度。

参 考 文 献

[1] LI Y, LI J, JIANG J, et al. P-transformer: Towards better document-to-document neural machine translation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3859-3870.

[2] FENG S, SUN H, YAN X, et al. Dense reinforcement learning for safety validation of autonomous vehicles[J]. Nature, 2023, 615: 620-627.

[3] ZHANG Y, XIE F, SONG X, et al. Dermoscopic image retrieval based on rotation-invariance deep hashing[J]. Medical Image Analysis, 2022, 77: 102301.

[4] CHEN J, CHEN K, CHEN H, et al. Contrastive learning for fine-grained ship classification in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[5] ZHANG Q, LI X, CHEN Y, et al. Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains[C]// Proceedings of the International Conference on Learning Representations, 2022.

[6] CHEN J, CHEN H, CHEN K, et al. Diffusion Models for Imperceptible and Transferable Adversarial Attack[C]// Proceedings of the International Conference on Learning Representations. 2024.

[7] BRENDEL W, RAUBER J, BETHGE M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models[C]// Proceedings of the International Conference on Learning Representations, 2018.

[8] WU Y, LIU J. A Survey on Black-box adversarial attack in image analysis[J]. Journal of Computer Science, 2024(5): 1138-1178.

[9] WANG X, HE X, WANG J, et al. Admix: Enhancing the Transferability of Adversarial Attacks Through Variance Tuning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16138-16147.

[10] ZHU Y, CHEN Y, LI X, et al. Toward understanding and boosting adversarial transferability from a distribution perspective[J]. IEEE Transactions on Image Processing, 2022, 31: 6487-6501.

[11] NASEER M M, KHAN S H, KHAN M H, et al. Cross-domain Transferability of Adversarial Perturbations[C]// Advances in Neural Information Processing Systems. 2019: 12885-12895.

[12] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics[C]// Proceedings of the International Conference on Machine Learning. 2015: 2256-2265.

[13] HO J, JAIN A, ABBEEL P. Denoising Diffusion Probabilistic Models[C]// Advances in Neural Information Processing Systems. 2020: 6840-6851.

[14] YUAN Z, ZHANG J, JIA Y, et al. Meta Gradient Adversarial Attack[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7728-7737.

[15] XIONG Y, LIN J, ZHANG M, et al. Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14963-14972.

[16] ZHU J, DAI F, YU L, et al. Attention-guided transformation-invariant attack for black-box adversarial examples[J]. International Journal of Intelligent Systems, 2022, 37(5): 3142-3165.

[17] HUANG L, WEI S, GAO C, et al. Cyclical adversarial attack pierces black-box deep neural networks[J]. Pattern Recognition, 2022, 131: 108831.

[18] HUAN Z, WANG Y, ZHANG X, et al. Data-free Adversarial Perturbations for Practical Black-box Attack[C]// Advances in Knowledge Discovery and Data Mining. 2020: 127-138.

[19] DUAN M, LI K, DENG J, et al. A novel multi-sample generation method for adversarial attacks[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2022, 18(4): 1-21.

[20] QIU H, XIAO C, YANG L, et al. Semanticadv: Generating Adversarial Examples via Attribute-Conditioned Image Editing[C]// Proceedings of the European Conference on Computer Vision. 2020: 19-37.

[21] JIA S, YIN B, YAO T, et al. Adv-attribute: Inconspicuous and Transferable Adversarial Attack on Face Recognition[C]// Proceedings of the 36rh Conference on Neural Information Processing Systems. 2022.

[22] YUAN S, ZHANG Q, GAO L, et al. Natural Color Fool: Towards Boosting Black-box Unrestricted Attacks[C]// NeurIPS 2022. 2022.

[23] SAHARIA C, HO J, CHAN W, et al. Image super-resolution via iterative refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4713-4726.

[24] PARMAR G, SINGH K K, ZHANG R, et al. Zero-shot Image-to-image Translation[C]// Proceedings of the ACM SIGGRAPH Conference. 2023: 1-11.

[25] NIE W, GUO B, HUANG Y, et al. Diffusion Models for Adversarial Purification[C]// Proceedings of the International Conference on Machine Learning. 2022: 16805-16827.

[26] LIU D, WANG X, PENG C, et al. Adv-Diffusion: Imperceptible Adversarial Face Identity Attack via Latent Diffusion Model[C]// Proceedings of the Conference on Artificial Intelligence. 2024: 3585-3593.

[27] ROMBACH R,BLATTMANN A,LORENZ D,et al. High-resolution Image Synthesis with Latent Diffusion Models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:10674-10685.

[28] JOHNSON J,ALAHI A,FEI-FEI L. Perceptual Losses for Real-time Style Transfer and Super-resolution[C]// Proceedings of the European Conference on Computer Vision. 2016:694-711.

[29] WAH C,BRANSON S,WELINDER P,et al. The caltech-ucsd birds-200-2011 dataset:Tech. Rep. CNS-TR-2011-001[R]. California Institute of Technology,2011.

[30] KRAUSE J,STARK M,DENG J,et al. 3d Object Representations for Fine-grained Categorization[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013:554-561.

[31] HE K,ZHANG X,REN S,et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[32] SIMONYAN K,ZISSERMAN A. Very Deep Convolutional Networks for Large-scale Image Recognition[C]// Proceedings of the International Conference on Learning Representations. 2015.

[33] SZEGEDY C,VANHOUCKE V,IOFFE S,et al. Rethinking the Inception Architecture for Computer Vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818-2826.

[34] SANDLER M,HOWARD A,ZHU M,et al. Mobilenetv2:Inverted Residuals and Linear Bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4510-4520.

[35] LIU Z,MAO H,WU C Y,et al. A Convnet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:11966-11976.

[36] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al. An Image is Worth 16×16 Words:Transformers for Image Recognition at Scale[C]// Proceedings of the International Conference on Learning Representations. 2020.

[37] LIU Z,LIN Y,CAO Y,et al. Swin Transformer:Hierarchical Vision Transformer using Shifted Windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:9992-10002.

[38] TOUVRON H,CORD M,DOUZE M,et al. Training Data-efficient Image Transformers & Distillation through Attention [C]// Proceedings of the International Conference on Machine Learning. 2021:10347-10357.

[39] KURAKIN A,GOODFELLOW I,BENGIO S,et al. Adversarial Attacks and Defences Competition[C]// Advances in Neural Information Processing Systems. 2018:195-231.

[40] TRAMÈR F,KURAKIN A,PAPERNOT N,et al. Ensemble Adversarial Training:Attacks and Defenses[C]// Proceedings of the International Conference on Learning Representations. 2018.

[41] SONG J,MENG C,ERMON S. Denoising Diffusion Implicit Models[C]// Proceedings of the International Conference on Learning Representations. 2021.

[42] HEUSEL M,RAMSAUER H,UNTERTHINER T,et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium[C]// Advances in Neural Information Processing Systems. 2017:6626-6637.

[43] DONG Y,LIAO F,PANG T,et al. Boosting Adversarial Attacks with Momentum[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:9185-9193.

[44] XIE C,ZHANG Z,ZHOU Y,et al. Improving Transferability of Adversarial Examples with Input Diversity[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:2730-2739.

[45] DONG Y,PANG T,SU H,et al. Evading Defenses to Transferable Adversarial Examples by Translation-invariant Attacks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4312-4321.

[46] GAO L,ZHANG Q,SONG J,et al. Patch-wise Attack for Fooling Deep Neural Network[C]// Proceedings of the European Conference on Computer Vision. 2020:307-322.

[47] LONG Y,ZHANG Q,ZENG B,et al. Frequency Domain Model Augmentation for Adversarial Attack[C]// Proceedings of the European Conference on Computer Vision. 2022:549-566.

[48] ZHAO Z,LIU Z,LARSON M. Towards Large Yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:1036-1045.

KANG Kai, born in 1986, Ph.D candidate. His main research interests include adversarial attack and so on.

WANG Jiabao, born in 1985, Ph.D, associate professor. His main research interests include computer vision and machine learning.

(责任编辑:喻藜)