**引用格式**：张兆阳，孙芳慧，张明旭，等. 黑盒条件下生成式对抗攻击方法研究[J]. 信息对抗技术，2025，4(5)：1-21.[ZHANG Zhaoyang，SUN Fanghui，ZHANG Mingxu，et al. Research on generative adversarial attacks under black-box conditions[J]. Information Countermeasure Technology，2025，4(5)：1-21.(in Chinese)]

# 黑盒条件下生成式对抗攻击方法研究

张兆阳[1]，孙芳慧[1]，张明旭[2]，宋　伟[3]，王振邦[4]，王英琦[1]，张可卿[1]，王　莘[1*]

（1. 哈尔滨工业大学网络空间安全学院，黑龙江哈尔滨 150001；2. 中国电子学会，北京 100036；
3. 中移物联网有限公司，重庆 401336；4. 国网黑龙江省电力有限公司，黑龙江哈尔滨 150090）

**摘　要**　在进行图像对抗攻击时，针对目标模型进行的白盒攻击往往效果最佳，但实际中通常难以获取目标模型结构，这使得提高对抗样本的迁移性尤为关键。针对这一问题，提出一种基于生成对抗网络(generative adversarial network，GAN)的训练方法，用以生成具备强迁移性的对抗样本。研究发现，图像本身具有与模型无关的脆弱性，生成式方法正是通过挖掘这一特性进行攻击的。与传统方法在原图邻域内微调不同，该方法从其他类别分布中生成具有最大似然的图像，在视觉上接近真实图像，但能有效误导分类器。训练过程中，生成器生成对抗样本，判别器判断其标签的正确性，二者协同优化，不断提升样本的攻击性与真实度。实验表明，生成式对抗样本在多个模型上的攻击成功率显著高于传统方法，平均提升约 25%，展现出更强的跨模型泛化能力。该结果表明生成式对抗攻击不仅提升了黑盒攻击的实用性，也揭示了深度模型普遍存在的脆弱性，为后续防御机制设计提供了方向。

**关键词**　生成式对抗攻击；模型迁移性；黑盒攻击

**中图分类号**　TP 391　　**文章编号**　2097-163X(2025)05-0001-21

**文献标志码**　A　　**DOI**　10.12399/j.issn.2097-163x.2025.05.001

# Research on generative adversarial attacks under black-box conditions

ZHANG Zhaoyang[1]，SUN Fanghui[1]，ZHANG Mingxu[2]，SONG Wei[3]，WANG Zhenbang[4]，WANG Yingqi[1]，ZHANG Keqing[1]，WANG Shen[1*]

（1. School of Cybersecurity，Harbin Institute of Technology，Harbin 150001，China；2. China Electronics Society，Beijing 100036，China；3. China Mobile IoT Co.，Ltd.，Chongqing 401336，China；
4. State Grid Heilongjiang Electric Power Co.，Ltd.，Harbin 150090，China）

**Abstract**　In the context of image adversarial attacks，white-box attacks targeting the target model often yield the best performance. However，in practice，it is usually difficult to obtain the architecture of the target model，which makes improving the transferability of adversarial examples particularly crucial. To address this issue，a training method based on generative adversarial network (GAN) was proposed to generate adversarial examples with strong transferability. The study finds that images themselves possess model-agnostic vulnerabilities，and generative methods implement attacks precisely by exploiting this characteristic. Unlike traditional methods that perform fine-tuning within the neighborhood of the original image，

**收稿日期**：2025-07-07　　**修回日期**：2025-08-08

**通信作者**：王莘，E-mail：shen.wang@hit.edu.cn

**基金项目**：国防基础科研项目(JCKY2023603C043)；黑龙江省重点研发计划项目(2022ZX01C01)；黑龙江省自然科学基金资助项目(LH2024F023)

this method generates images with maximum likelihood from the distribution of other categories. These images are visually close to real images but can effectively mislead classifiers. During the training process, the generator produces adversarial examples, while the discriminator judges the correctness of their labels. The two components optimize collaboratively, continuously enhancing the adversarial potency and authenticity of the examples. Experiments show that the attack success rate of generative adversarial examples on multiple models is significantly higher than that of traditional methods, with an average improvement of approximately 25%, demonstrating stronger cross-model generalization ability. This result indicates that generative adversarial attacks not only enhance the practicality of black-box attacks but also reveal the widespread vulnerabilities of deep models, providing directions for the design of subsequent defense mechanisms.

**Keywords** generative adversarial attack; model transferability; black-box attack

## 0 引言

随着科技的发展，人工智能由于其易于操作、智能、省时等特点而越来越受到人们的欢迎。而作为人工智能的核心技术，深度学习也开始逐渐进入人们的视野。深度学习的核心思想是通过模拟人脑的神经网络结构，构建能够从海量数据中自动学习特征和模式的计算模型，从而实现复杂的认知任务。这项技术的突破性应用已经渗透到人们日常生活的方方面面。在计算机视觉领域，深度神经网络的表现已经超越了人类，能够实现高精度的图像识别、分类和目标检测[1]。在人们身边，深度学习也正在发挥不可或缺的作用：大语言模型（LLM）如 ChatGPT、DeepSeek[2]等的出现，帮助人们答疑解惑；随着智能家居出现，电子电器设备可以依靠声音完成控制[3]；AI 决策指挥车辆，让无人驾驶车从此面世[4-5]；在卫星遥感方面，AI 可以迅速识别图片内容[6]；在医疗诊断方面，智能医学图像分析可挖掘深层次的致病机理，提高医生读片效率，同时缓解医疗资源不足、分配不均的现实问题[7]等。

然而，随着深度学习模型在关键领域的深度集成，其安全性与鲁棒性面临严峻挑战。其中，对抗样本（adversarial examples）可以通过在原始样本上添加人类难以察觉的微小扰动来显著误导模型的决策。这揭示了深度神经网络存在的脆弱性，严重制约了 AI 系统在现实场景的可信部署。为提高模型的安全性，需要对于对抗攻击有更深入的研究。

为提升攻击的实用性，研究重心从白盒攻击（已知模型参数的攻击场景）转向更贴近实际的黑盒攻击（未知模型参数的攻击场景）。其核心在于提升对抗样本的迁移性，即攻击方法生成的样本对不同结构和参数的目标模型（实际被攻击的未知模型，对抗样本需在其上实现迁移攻击）的普遍有效性。当前面临的关键挑战在于，基于特定代理模型（攻击者可访问的白盒替代模型，用于生成对抗样本并尝试迁移到目标模型）的传统攻击方法生成的对抗样本，其攻击成功率严重依赖于该代理模型本身。尤其值得注意的是，当目标模型与代理模型在架构上存在差异时，攻击成功率常急剧下降，极大地限制了黑盒攻击的实际威胁范围和效果。

针对迁移效率低及代理模型依赖等核心问题，本文提出一种基于生成对抗网络（generative adversarial network，GAN）的强迁移性对抗攻击方法，通过生成器与判别器之间的相互监督来不断改进生成器。与传统梯度扰动方法在图片样本邻域微扰的范式不同，本文通过条件生成器直接学习目标错误类别的数据分布特征，生成具有错误类别特征但同时数据分布在原图片附近的样本。该方法具备高隐蔽性，生成的对抗样本严格遵循目标类别数据规律，但视觉上与正常样本高度一致；同时，还具备强迁移性。

## 1 对抗攻击背景

### 1.1 对抗攻击发展

随着人们越来越依赖人工智能与深度学习，一个安全隐患却逐渐凸显：深度学习神经网络对输入数据的微小扰动异常敏感。这些微小扰动

会让深度神经网络输出错误结果[8]，且由于扰动很微小，人类几乎不会发现这些扰动的存在。由此，SZEGEDY 等[9]首次提出对抗样本的概念，即对输入样本添加微小的扰动，使模型以高置信度输出错误结果。在计算机视觉中这类扰动常表现为噪声形式，但最新的研究也提出了存在如形状变换、颜色改变等非噪声形式的扰动。像这样生成对抗样本，并用对抗样本输入目标模型，误导目标模型输出错误结果的过程被称为对抗攻击。

对抗攻击原理的示意图如图 1 所示，图 1(左)是原始图像，图 1(右)是在原始图像上添加了扰动的对抗样本。由图 1 可观察到，右图相比左图有更多的噪声，看起来更加杂乱；当用被攻击的模型，即目标模型去识别这 2 个图像时，左边能顺利识别出图片的真实类别，但是对于右边的对抗样本，目标模型的识别就会出错，例如图 1(右)错误地将人的图片分类为车。对于这种目标模型识别对抗样本类别错误的情况，被视为此时对抗攻击成功。

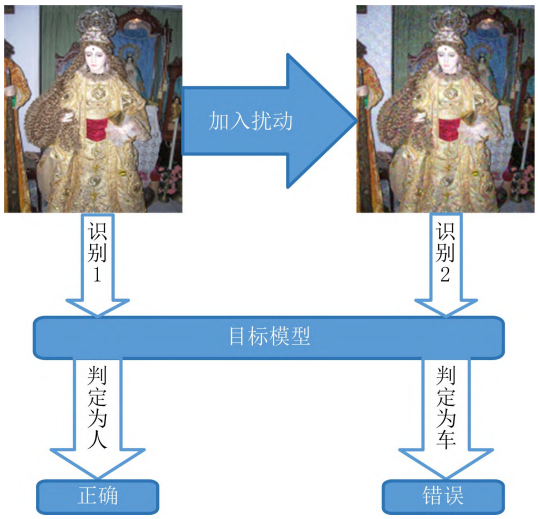


**图 1　对抗攻击原理示意图**

**Fig. 1　Illustration of the adversarial attack mechanism**

对抗攻击按照有无攻击目标分为 2 种：一种叫有目标攻击，这种攻击可以让目标模型识别对抗样本时输出攻击者事先设定好的目标类别，比如对于一个类别为猫的原始样本进行有目标对抗攻击，设定目标类别为狗，那么攻击成功时得到的对抗样本就会被识别为狗；另一种叫无目标攻击，这种攻击可以让目标模型输出错误的结果，攻击者未给出想要的类别，也就是说，如果原始样本类别为猫，只要对抗样本被目标模型识别出的类别不是猫而是其他类别，则攻击成功。按照攻击者获知的信息分类，对抗攻击可以被分为黑盒攻击和白盒攻击 2 大类。其中，白盒攻击[10-11]可以访问目标模型的架构、模型参数及训练数据集；黑盒攻击[12-13]无法获取目标模型的任何相关信息，因此更接近真实世界的场景。目前，研究的主要方向是黑盒攻击。

由于黑盒攻击无法获知目标模型的任何信息，而某个单一模型训练出的对抗样本在其他模型上的攻击成功率又很低，因此，如何提高对抗模型的迁移性就尤为重要。为了提高对抗样本的迁移性，本文提出了对生成式对抗样本的研究。该研究不仅可以帮助推进对抗攻击技术的发展，还揭示了模型的共性安全缺陷，有助于提高模型的鲁棒性，增强模型的防御能力。

### 1.2　对抗攻击分类

在对抗样本生成研究中，攻击者所掌握的模型信息在很大程度上决定了攻击方法的可行性与效果。根据攻击者对目标模型内部参数和结构的认知程度，通常将对抗攻击划分为白盒攻击与黑盒攻击。

#### 1.2.1　白盒攻击

白盒攻击假设攻击者拥有模型完整访问权限，可直接利用梯度设计扰动。SZEGEDY 等[9]最早提出基于优化的 L-BFGS(limited-memory Broyden-Fletcher-Goldfarb-Shanno)方法，通过最小化扰动实现误分类，但效率低且成功率有限；GOODFELLOW 等[14]的 FGSM(fast gradient sign method)则通过单次梯度计算快速生成对抗样本；后续 KURAKIN 等[10]的 BIM(I-FGSM)扩展为迭代攻击提升稳定性；MADRY 等[15]的 PGD(projected gradient descent)进一步引入随机初始化与范数投影，增强跨模型泛化性；Carlini-Wagner(C&W)方法[16]优化损失函数实现高成功率攻击，可突破防御性蒸馏[17]等防护机制。然而白盒攻击计算代价高、迁移性弱，当目标模型与代理模型不一致时效果显著下降。

#### 1.2.2　黑盒攻击

黑盒攻击假定攻击者无法获取模型参数和结构，只能通过模型的输入输出行为进行攻击，主要包含 3 类方法：基于梯度估计的攻击[18]，需构建辅助模型模拟梯度，代价高昂；基于局部搜索的攻击[19]，通过频繁查询寻找扰动，易被检测；

基于迁移性的攻击[13]，通过白盒替代模型生成扰动再迁移至黑盒目标，实用性更强。迁移攻击进一步分为 2 种方法：一种是实例相关方法（如动量机制[20]、图像变换[21]或梯度近似[22]），需逐样本生成扰动但易过拟合替代模型；另一种是实例无关方法（如 MOOSAVI-DEZFOOLI 等[23]提出的通用扰动，以及后续研究[24-25]优化的跨模型/任务迁移方案），仅需单次前向传播即可生成样本，兼具高效性与隐蔽性。因此，对抗攻击迁移性研究不仅可以推动攻击技术进步，更为提升模型鲁棒性提供关键方向。

## 2 图像分类背景

### 2.1 图像分类模型原理

#### 2.1.1 卷积神经网络结构

在图像领域，卷积神经网络（convolutional neural network，CNN）几乎是所有经典模型的核心架构。CNN 借助卷积操作，从图像中提取空间局部特征。

传统的全连接神经网络不能有效利用图像的空间信息，而卷积操作则可以通过滑动窗口（即卷积核）捕捉图像的局部模式。

一般的卷积操作可表示为：

$$\mathrm{FeatureMap}(i,j)=\sum_{m,n}\mathrm{Input}(i+m,j+n)\cdot\mathrm{Kernel}(m,n) \tag{1}$$

式中，$\mathrm{Input}(i+m,j+n)$是输入特征图上的像素值（或特征值），表示在第$(i+m,j+n)$位置的输入；$\mathrm{Kernel}(m,n)$是卷积核的权重矩阵，表示在位置$(m,n)$处的权重值；$\mathrm{FeatureMap}(i,j)$是卷积运算后输出特征图在$(i,j)$位置的值，即该位置局部感受野的加权和，表示在输入图像的局部区域，通过滑动小窗口（卷积核）逐像素相乘求和，得到局部特征表示，用于捕捉如边缘、角点等局部信息。

此外，卷积层之后通常会接一个非线性激活函数（如 ReLU）来增强网络的表达能力，即：

$$\mathrm{ReLU}(x)=\max(0,x) \tag{2}$$

式中，$x$ 是输入值，$\mathrm{ReLU}(x)$是激活后的输出值。池化层（如最大池化）用于下采样，降低特征图尺寸，提高计算效率，同时保留最重要的信息。

#### 2.1.2 网络训练机制与损失函数

所有模型都必须经过训练才能实现高精度预测，而训练依赖于损失函数和梯度优化。在分类任务中，最常用的损失函数是交叉熵损失（cross entropy loss），表示为：

$$\lambda_{\mathrm{CE}}=-\sum_i y_i\ln(\hat{y}_i) \tag{3}$$

式中，$y_i$ 是真实标签的 one-hot 编码，$\hat{y}_i$ 是模型输出的 softmax 概率值。softmax 的计算为：

$$\hat{y}_i=\frac{\mathrm{e}^{z_i}}{\sum_j \mathrm{e}^{z_j}} \tag{4}$$

式中，$\mathrm{e}^{z_i}$ 是对每个得分取的指数；$\sum_j \mathrm{e}^{z_j}$ 的作用是对所有类别得分取指数后求和，用作归一化，是最终得到的类别 $i$ 的预测概率。训练过程通过反向传播算法计算损失函数对每个参数的梯度，并通过优化器（如 SGD 或 Adam）进行更新，使得损失逐步减小。

同时，为了避免过拟合，很多网络会使用 Dropout、$\ell_2$ 正则化、Batch Normalization 等机制，使网络更具泛化能力。

#### 2.1.3 深层网络中的训练与结构改进

随着神经网络层数越来越深，训练变得更加困难，容易出现梯度消失或爆炸的问题。为此，一些模型引入了结构改进，具体改进方式有 2 种：

1）残差连接（ResNet）。引入了跳跃连接（skip connection），使得网络在每层之间学习残差，表示为：

$$y=F(x)+x \tag{5}$$

式中，$F$ 是当前层的变换，$x$ 是输入。这样即使某些层学不到有效表示，网络也能保持性能。

2）密集连接（DenseNet）。将每一层的输出都连接到后面所有层，避免信息遗失和梯度阻塞，表示为：

$$x_l=H([x_0,x_1,\cdots,x_{l-1}]) \tag{6}$$

式中，$x_l$ 是第 $l$ 层的输入，$x_0,x_1,\cdots,x_{l-1}$ 是前面所有层的输出，$H(\cdot)$是变换函数。这样可以保留更多的底层信息，缓解梯度消失。

#### 2.1.4 ViT 基础

ViT（vision transformer）是唯一一种非 CNN 结构的模型。它借鉴了 NLP（natural language processing）中的 Transformer 架构，使用自注意力机制（self-attention）处理图像。在 ViT 中，图像会被划分成多个 patch（小块），每个 patch 类似于一个词，然后通过 Transformer 层建模这些 patch 之间的关系。自注意力机制的核心计算为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (7)$$

式中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 分别为查询矩阵、键矩阵和值矩阵，$d_k$ 是维度缩放因子。ViT 虽然参数多，但具备更强的全局建模能力，对大规模训练数据表现更好。不过在小数据集上，通常不如 CNN 效果稳定。

### 2.2　GAN 原理

GAN 是 GOODFELLOW 等在 2014 年提出的深度生成模型架构。其核心思想是通过 2 个神经网络对抗博弈，训练出能“以假乱真”的生成器。简单来说，GAN 类似“造假者”与“鉴定者”的博弈，生成器伪造图像，判别器识别真假。该系统主要由生成器和判别器 2 部分构成。生成器以随机向量（如高斯或均匀分布噪声）为输入，输出“逼真”图像；判别器输入为图像，负责判断其是真实样本还是生成器合成的伪造图像。二者目标对立，生成器想“骗过”判别器，判别器则努力不被欺骗。

这种对抗的训练方式可以通过一个最小最大值的优化目标函数来描述，即：

$$\min_G\max_D V(D,G)=E_{x\sim p_{\text{data}}(x)}[\ln D(x)]+E_{z\sim p_z(z)}[\ln D(G(z))] \quad (8)$$

式中，$x$ 表示真实图像；$z$ 表示生成器输入的随机噪声；$G$ 表示生成器；$V$ 表示判别器；$G(z)$ 表示生成器输出的伪造图像，$D(x)$ 表示判别器对输入 $x$ 的输出；$V$ 是价值函数，定义了生成器和判别器博弈的规则，判别器欲使其最大，生成器欲使其最小。该目标函数的第 1 项表示判别器正确识别真实图像的概率，第 2 项表示它正确识别生成图像为假的概率。判别器想要最大化这个目标函数，而生成器则想让判别器“出错”，从而最小化它。训练过程中，首先固定生成器，训练判别器识别真假图片；接着固定判别器，训练生成器让它生成的图像更“真实”，逐步更新生成器和判别器。这种轮流优化的过程会导致生成器学会捕捉数据分布的特征，从而生成越来越像真实样本的图像。

GAN 的学习过程并非一次完成，与传统监督学习不同，它是零和博弈问题。训练中可能出现模式崩溃、判别器影响生成器学习、梯度消失等问题。训练成功时，GAN 生成的样本视觉质量高，能以假乱真。实际应用中，生成器用反卷积结构从低维噪声恢复高维图像，判别器用卷积结构判断图像“真伪”。随着研究推进，GAN 出现 DCGAN、WGAN、CycleGAN 等变种，在稳定性、多样性、收敛性等方面有改进扩展。

总体而言，GAN 基本原理体现“对抗学习”思想，通过生成器和判别器博弈逼近真实数据分布。它结构简单但生成能力惊人，是深度学习有影响力的模型之一。虽训练有困难，但理论清晰、应用广泛，是生成模型研究中的重要一环。

### 2.3　对抗攻击图像处理原理

对抗攻击是近年来深度学习研究重要方向，在图像分类任务中，研究者发现对输入图像施加微小、人眼不可感知的扰动，就能显著干扰模型判断。这揭示了深度神经网络在图像空间的脆弱性与非稳健性。所以，理解图像处理相关数学机制和变换基础对掌握对抗攻击至关重要。

在图像处理中，图像通常被表示为一个形如 $x\in\mathbf{R}^{H\times W\times C}$ 的多维张量，其中 $H$ 和 $W$ 分别表示图像的高度和宽度，$C$ 为颜色通道数。对于常见的 RGB 彩色图像，通道数为 3。对抗攻击的基本思路是通过人为构造扰动 $\delta$，使得添加扰动后的图像 $x_{\text{adv}}=x+\delta$ 能在保持视觉一致性的前提下诱导模型作出错误预测。为了限制扰动的不可感知性，常规定义扰动满足如式(9)的范数约束：

$$\|\delta\|_p\leqslant\epsilon \quad (9)$$

式中，$\|\delta\|_p$ 表示 $\ell_p$ 范数，$\epsilon$ 表示扰动强度。对抗攻击中常使用的范数包括 $\ell_\infty$ 范数和 $\ell_2$ 范数，它们分别对应最大像素值扰动和整体欧几里得距离的约束。

扰动的生成一般基于损失函数对图像输入的梯度信息。最基础的 FGSM 即利用损失函数 $\mathcal{L}(f(x),y)$ 对输入图像求梯度，并在该方向上施加扰动，生成的对抗样本为：

$$x_{\text{adv}}=x+\epsilon\cdot\text{sign}(\nabla_x L(f(x),y))) \quad (10)$$

式中，$f(x)$ 表示分类器的输出，$y$ 表示原始标签。式(10)说明，攻击者寻找能够使模型损失最大的扰动方向，并在该方向上进行线性扩展。

在视觉层面，为保持对抗样本与原图的感知一致性，常引入结构保留项，如 $\ell_2$ 重构损失，即：

$$\mathcal{L}_{\text{recon}}=\|x-x_{\text{adv}}\|_2^2 \quad (11)$$

式中，$x$ 是原始输入样本，$x_{\text{adv}}$ 是对抗样本。该损失函数约束扰动不要过度偏离原始图像，在实现攻击的同时，保持人眼无法分辨其差异，从而达到无感知攻击的目标。

综上所述，图像处理理论基础在对抗攻击中起支撑作用，决定扰动表达方式，影响攻击成功率与迁移效果。对抗样本可在保证可控性与视觉一致性的同时，有效干扰多种深度学习模型，展现出图像处理与对抗攻击在安全领域融合的广阔前景。

## 3 黑盒条件下生成式对抗攻击方法

传统的对抗攻击方法（如 FGSM、PGD 等），主要在原始样本的邻域内寻找微小扰动以误导模型输出，这类方法虽然有效，但在目标模型未知的黑盒场景下，其攻击样本的迁移能力较差，难以实现高效攻击。为了克服这一局限，本文提出一种基于 GAN 的攻击方法，通过构造生成器和判别器联合训练，在目标类别的统计分布下生成最具迷惑性的攻击样本，从根本上提升攻击样本在跨模型、跨结构间的迁移性。

### 3.1 生成式对抗攻击方法实现的核心原理

在黑盒攻击的情况下，由于在某个具体模型上训练出的对抗样本一般只对自身的模型有较高的攻击成功率（attack success rate，ASR），所以为了让训练出的对抗样本能够对不同的目标模型都有更高的 ASR，需要提高攻击的迁移性，以达到更好的黑盒攻击成果。为此，实验选择用 GAN 生成式对抗样本来完成强迁移性攻击。生成式对抗样本不在原样本附近，但是和原始样本同分布，是在另一个类别的分布下似然最高的对抗样本，同时可以满足使模型识别结果出错的要求。传统的对抗攻击（如 FGSM、PGD 等）通常是在原始样本的邻域内找到一个微小扰动，使得模型误分类，即：

$$x_{\text{adv}} = x + \delta \tag{12}$$

式中，$x$ 为原始图像；$x_{\text{adv}}$ 为对抗样本；扰动 $\delta$ 很小，保证对人类来说难以察觉。

生成式对抗样本（如 GAN 生成的对抗样本）不再局限于原样本的邻域，而是直接生成一个新的样本，它与原样本不同，但仍然在数据分布中。生成式对抗样本是由 GAN 学习整个数据分布后生成的，因此它们仍然符合整体数据的统计特性。例如，如果数据集是猫和狗的图像，输入的原图片是猫，而目标标签是狗，那么 GAN 最终会生成一张在人类角度看来是猫，但是在目标模型上识别为狗的图片。这是因为生成的对抗样本在一轮轮的训练中被 $\ell_2$ 范数约束在原来的类别为猫的图片附近，因此看起来是猫，但是其特征分布却在训练中偏向于类别狗。GAN 生成的对抗样本不需要像 FGSM 那样保持与原样本极其接近，它可以生成完全不同但仍符合数据分布的图像，直接生成一个属于错误类别的样本，使分类器强烈误判。它比传统的微小扰动攻击更难检测，因为它看起来像是数据集中原生的样本，而不像是原始图片加了点奇怪的噪声。

GAN 的生成式对抗样本是在另一个类别的分布下似然最高的对抗样本，即 GAN 生成的对抗样本错误类别的概率最高，表示为：

$$P(x+\delta \mid y_{\text{false}}) > P(x+\delta \mid y_{\text{true}}) \tag{13}$$

式中，$y_{\text{false}}$ 和 $y_{\text{true}}$ 分别表示错误的分类类别和正确的分类类别。也就是说，原图像的分布服从于标签 $y_{\text{true}}$，而对抗样本的分布服从于标签 $y_{\text{false}}$。

强迁移性的对抗攻击不仅提升了黑盒攻击的实用性，还揭示了神经网络的通用脆弱性。研究迁移性有助于开发更高效的攻击方法，同时也有助于改进防御技术，使 AI 系统更安全。

### 3.2 基于 GAN 的生成式对抗攻击分析

为了得到用生成式模型来生成图像对抗样本，实验需要训练一个生成器和一个判别器，生成器生成对抗样本被用于欺骗判别器，而判别器的作用是分辨正确的图像和生成器伪造的图像。通过生成器和判别器之间的相互督促，生成器所生成的图片特征越来越像目标类别的图片，判别器的判别能力也会越来越强。最后，评估对抗样本的攻击效果，检验其是否能够欺骗目标模型。也就是说，判别器试图最大化真实样本的得分 $D(x,y)$ 并最小化生成样本的得分 $D(x_G,y)$。而 $D(x,y)$ 试图最大化判别器对 $x_G$ 的得分，使其更像真实样本。判别器试图让真实样本的分数尽量大（接近 1），同时让生成样本的分数尽量小（接近 −1）。表达式为：

$$\mathcal{L}_D = \frac{1}{2}E_{x\sim p_{\text{data}}}[\max(0,1-D(x,y))] + \frac{1}{2}E_{x_G\sim p_G}[\max(0,1+D(x_G,y))] \tag{14}$$

式中，$\mathcal{L}_D$ 为判别器损失函数，目标是尽量区分真实样本和生成样本，数值越小代表判别器判断越准确。$E_{x\sim p_{\text{data}}}$ 表示对服从真实数据集样本分布

的求数学期望，$E_{x_G\sim p_G}$ 对服从生成器生成分布的样本求数学期望，$y$ 是类别标签。对于真实样本 $x$，希望 $D(x,y)\geqslant 1$，如果 $D(x,y)<1$，就会有损失。对于生成样本 $x_G$，希望 $D(x_G,y)\leqslant -1$，如果 $D(x_G,y)>-1$，就会有损失。也就是说，判别器的目标是尽量识别真实图像为真，生成图像为假，尽量使自己的判断能力越来越强。

而生成器的优化目标不仅在于欺骗判别器，还需保证生成图像与原图在视觉上保持一致。因此，生成器的损失函数由 2 部分组成：一是对抗损失，通过二元交叉熵（binary cross-entropy，BCE）计算判别器对生成样本为真实图像的预测误差，从而引导生成器生成更具欺骗性的样本；二是重构损失，采用 $\ell_2$ 范数度量生成样本与原图之间的像素差异，确保生成结果在结构与语义上接近输入图像。最终的生成器总损失表达为：

$$\mathcal{L}_G=E_{z\sim p_z}[B_{\mathrm{BCE}}(D(G(z,y_t),y_t),1)]+\lambda\cdot\|G(z,y_t)-x\|_2 \tag{15}$$

式中，$G(z,y_t)$ 是生成器输出；$D(G(z,y_t),y_t)$ 是判别器输出，也就是对生成样本与标签的评分；$B_{\mathrm{BCE}}(\cdot)$ 是二元交叉熵损失，让判别器误认为是“真实图”；$\|G(z,y_t)-x\|_2$ 是 $\ell_2$ 重构损失；$x$ 是原图，用于保持图像语义结构一致；$\lambda$ 是 $\ell_2$ 损失系数，控制控制攻击性与图像质量之间的平衡。该损失设计在提升迁移攻击能力的同时，增强了生成样本的自然性和可解释性。

输入数据集以 CIFAR-10 为例，整体的训练过程原理图如图 2 所示。

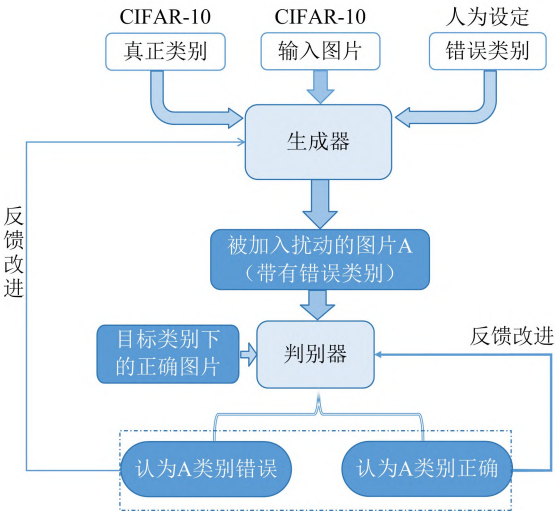


图 2　GAN 生成对抗样本算法原理图

Fig. 2　Diagram of the GAN-based adversarial example generation algorithm

由图 2 可以看到，首先生成器接收 CIFAR-10 的输入图片和对应的 CIFAR-10 的真正类别标签，同时，为了使生成的对抗样本具有对抗性，还要向生成器中输入人为设定的错误类别标签，生成器接收到这 3 个输入之后输出生成的图片 A，这些生成的学习了目标类别特征的图片带有目标标签。随后，将带有目标标签的生成图片 A 和带有目标类别的正确图片输入判别器，判别器就会根据同种图片推断这些生成图片的标签类别是否正确，输出的认为 A 正确和错误的权重都用于反馈改进判别器，其中被判别器发现类别错误的图片会将结果反馈回生成器，促进生成器反馈改进。

### 3.3　生成式对抗攻击算法

对于黑盒条件下生成式对抗攻击算法，接下来从生成器、判别器和它们之间的互相促进训练生成这 3 个方面进行介绍。

#### 3.3.1　生成器模块设计与任务

生成器模块是本方法中专门用于生成可攻击样本的核心组成部分。其设计基于 U-Net“编码器-解码器”结构，能够在保持输入图像主要特征的同时，通过多层下采样与上采样结构，在潜在空间中引导图像向攻击目标类别的方向发生特征偏移。生成器的输入包括原始图像和对应的标签与攻击者预设的目标标签，经过生成器后输出伪造样本。

与传统基于像素邻域的扰动方法不同，生成器模块采用端到端训练，由判别器提供对抗反馈，驱动其逐步学习生成具有误导性的图像分布。为了避免生成器输出出现无意义的扭曲或过强的扰动，训练中还引入了 $\ell_2$ 重构损失（即原图与生成样本的欧氏距离约束），确保输出样本在视觉上对人眼几乎不可察觉、具备较高自然性。

总体而言，生成器的主要任务是利用判别器反馈调整输出方向，不断提升欺骗判别器的能力；通过重构损失约束，维持样本的结构稳定性；产出既能攻击代理模型又具备跨模型迁移性、可用于黑盒测试的高质量对抗样本。

这种无梯度、基于分布学习的生成机制为跨架构攻击提供了有效支持。

#### 3.3.2　判别器模块与目标约束设计

判别器模块在本方法中承担着整体约束生

成器输出质量的核心任务。它的主要目标是通过区分输入图像是真实样本还是由生成器伪造出的样本，驱动生成器持续优化生成效果。具体来说，判别器接收的输入包括原始的真实样本与生成器输出的对抗样本，并通过输出置信值来表示其对样本真实性的判定。

在损失设计上，判别器采用基于 hinge 损失的对抗约束，即对于真实样本，惩罚判别器输出低于 1 的情况；对于伪造样本，惩罚判别器输出高于 −1 的情况。这种设计使得判别器能充分强化对真实分布的辨识能力，同时也为生成器设置了优化方向。换句话说，生成器的优化目标是在判别器面前“冒充”真实样本，而判别器则持续提升其甄别伪造样本的能力，两者形成典型的对抗博弈(adversarial game)关系。此外，判别器模块的存在不仅是为了对抗约束，还间接提供了输出样本的分布反馈，使生成器生成的样本在统计特性上更贴近真实类别分布。正是这种分布引导，使生成器在优化过程中能够避开过拟合单一模型梯度的局限，具备了更强的跨模型泛化与迁移能力。

### 3.3.3 生成器与判别器的协同对抗训练

本小节详细描述生成器与判别器如何通过交替优化，实现从“生成”到“攻击”的完整闭环。在整个生成式对抗攻击算法的优化过程中，生成器与判别器采用交替优化(alternating optimization)的方式进行循环训练。这一过程可以分为 2 个阶段：1) 固定生成器，重点优化判别器的判别能力；2) 固定判别器，重点优化生成器的生成与攻击能力。这种交替优化形成了典型的对抗博弈结构，二者在相互竞争中逐步逼近最优状态。

具体而言，在每一轮训练迭代中，首先保持生成器参数不变，将当前生成器生成的伪造样本与真实的目标类别样本一同输入判别器，由判别器进行区分，并计算 BCE 的损失。判别器的目标是最大化对真实样本与伪造样本的区分能力，也就是尽可能给真实样本打出高置信度，给生成样本打出低置信度。通过优化判别器参数，可以不断强化其鉴别能力，为生成器设置更高的对抗门槛。

接下来，固定判别器参数，仅优化生成器。此时，生成器以最小化生成器损失为目标，该损失函数由 2 部分组成：一是对抗损失，即生成器试图生成能够欺骗判别器、被误判为目标类别样本的伪造样本；二是重构或保持损失，用于约束生成样本与原始图像在像素或特征空间中的相似性，以保证生成的样本在感知质量上的自然性。

对于如何完成迁移性测试，本文的主要实验内容如图 3 所示。

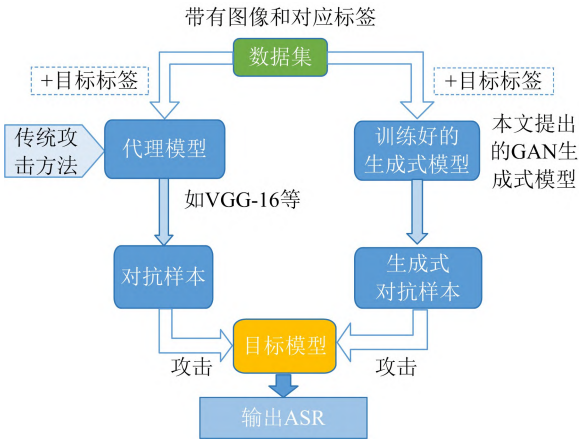


图 3 主要实验内容

Fig. 3 Overview of the main experiments

传统攻击方法，通过将数据集(含目标标签)和攻击算法输入代理模型(如 VGG-16)生成对抗样本，再攻击目标模型并输出不同目标模型对应的攻击成功率；而对于本文提出的 GAN 生成式方法，直接用训练好的生成模型生成对抗样本，用于攻击目标模型并输出 ASR。实验整体比较了基于代理模型的传统攻击和基于生成器的直接攻击 2 种思路。

从代码实现角度，整个算法流程可划分为以下 7 个关键步骤：

1) 读取配置文件。加载配置文件(如 cifar10.yaml 等)，读取训练所需的超参数，包括批量大小、学习率、训练轮数等，确保训练过程按需调整。

2) 数据加载。调用 load_cifar10()等函数，加载 CIFAR-10 等数据集，并使用 DataLoader 构建批处理迭代器，保证数据高效流入训练器。

3) 初始化 GAN 训练器。包括加载生成器和判别器模型，配置优化器(如 Adam)、损失函数，并设定随机种子以确保结果的可重复性。

4) 加载预训练模型。若检测到已有的预训练模型权重存在(如之前保存的模型文件)，则优先加载这些权重，以便在已有训练基础上继续优化，而非从零开始。

5）训练过程。核心部分包括 2 步：一是训练判别器，从 DataLoader 中取出真实图片，让判别器同时接收真实与生成样本，并基于损失函数优化判别能力；二是训练生成器，生成伪造对抗样本，并优化其欺骗判别器与误导分类器的能力。

6）模型保存。每个训练轮次结束后，保存当前的生成器与判别器权重，以便后续测试、复现或继续训练。

7）模型评估。在训练过程中，定期利用生成器生成对抗样本，并调用 Tester 模块对这些样本进行评估，具体包括计算对抗攻击的 ASR 及迁移性能。

伪代码实现过程如算法 1 所示。

**算法 1　生成器与判别器的协同对抗训练机制**

**输入**：数据集 $S=\{(x^{(i)},y^{(i)})\}$，生成器，判别器，学习率 $\eta$，重构损失权重 $\lambda$，训练轮数 $T$

**输出**：最终训练完成的生成器

1. 初始化生成器参数 $\theta_G$，判别器参数 $\theta_D$
2. for $t=1$ to $T$ do
3. 　for 每个 mini-batch $\{(x,y)\}\subset D$ do
4. 　采样目标标签 $y_t\neq y$
5. 　生成对抗样本 $x_{\text{adv}}=G(x,y_t)$
6. 　判别器输出 $D(x,y)$，$D(x_{\text{adv}},y_t)$
7. 　计算判别器损失 $\mathcal{L}_D=0.5\cdot\text{ReLU}(1-D(x,y))+0.5\cdot\text{ReLU}(1+D(x_{\text{adv}},y_t))$
8. 　更新判别器参数 $\theta_D\leftarrow\theta_D-\eta\cdot\nabla_{\theta_D}\mathcal{L}_D$
9. 　计算生成器的判别器相关损失 $\mathcal{L}_{G,\text{disc}}=-\text{mean}(D(x_{\text{adv}},y_t))$
10. 　计算生成器重构损失 $\mathcal{L}_{G,\text{mse}}=\|x_{\text{adv}}-x\|_2^2$
11. 　生成器总损失 $\mathcal{L}_G=\mathcal{L}_{G,\text{disc}}+\lambda\cdot\mathcal{L}_{G,\text{mse}}$
12. 　更新生成器参数 $\theta_G\leftarrow\theta_G-\eta\cdot\nabla_{\theta_G}\mathcal{L}_G$
13. end for
14. end for

通过这种生成器和判别器之间的交替优化，模型在多轮迭代中逐步形成了一种动态平衡：判别器不断提升其区分能力，而生成器则不断提升其生成能力，最终达到生成器输出的样本既能够顺利通过判别器的检验，又能在目标分类器上成功实现攻击。

此外，这一训练机制的另一大优势在于可迁移性，由于生成器在训练过程中学习到的是目标类别的整体特征分布，而非单个模型的特定参数，因此生成的对抗样本在跨模型、跨架构环境下依然具备较高的 ASR，体现出生成式方法在黑盒攻击中的独特优势。

## 4　生成式对抗样本迁移性实验设计

为验证所提出的生成式对抗攻击方法在实际黑盒环境中的适应性与迁移能力，本文在 CIFAR-10 和 SVHN（street view house numbers）2 大标准图像分类数据集上设计并实施了一系列对抗攻击实验。通过与多个主流攻击方法的对比、不同扰动强度下的参数敏感性测试及多目标模型评估，全面考察生成式对抗样本的攻击成功率、迁移稳定性及视觉一致性。本节将详细展示实验设置、结果分析与性能对比。

### 4.1　实验环境

本文所需要的实验环境如表 1 所列。

表 1　实验环境

Tab. 1　Experimental environment

| 项目 | 配置 |
|---|---|
| 处理器 | AMD EPYC 7542 32-Core Processor |
| 硬件 | NVIDIA GeForce RTX 4090 |
| 操作系统 | Linux |
| Pytorch 版本 | 2.5.1 |
| Python 版本 | 3.11.10 |
| Anaconda 版本 | 23.7.4 |
| torch 版本 | 2.5.1 |
| torchvision 版本 | 0.20.1 |
| CUDA 版本 | 12.1 |
| pandas 版本 | 1.5.3 |
| 神经网络模型 | VGG-16，VGG-19，ResNet-18，ResNet-34，DenseNet-121，DenseNet-201，SENet |
| 相关依赖库 | torchdiffeq，geotorch，timm，gdown，autoattack，robustbench 等 |

### 4.2　实验所用数据集

#### 4.2.1　CIFAR-10 数据集

CIFAR-10 是多伦多大学提出的一个包含 10 类（飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车）32×32 像素 RGB 图像的基准数据集，每类提供 5 000 张训练样本和 1 000 张测试样本，适用于评估图像分类模型性能。该数据集通过 PyTorch 的 torchvision.datasets.CIFAR10 加载，并标准化处理（均值[0.491 4，0.482 2，0.446 5]，标准差[0.202 3，0.199 4，0.201 0]）。CIFAR-10 数

据集图片示例如图 4 所示。

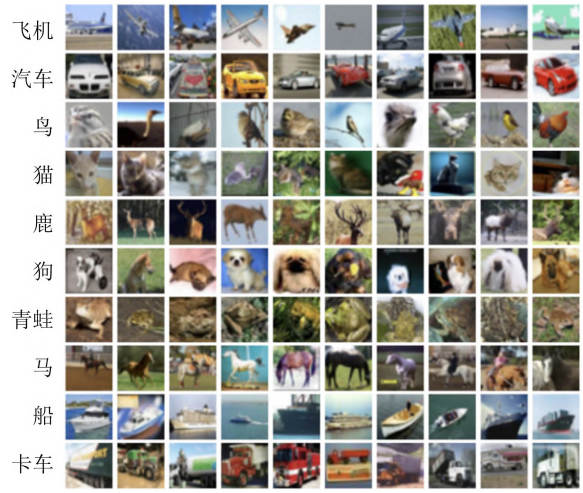


图 4 CIFAR-10 数据集图片示例

Fig. 4 Example images from the CIFAR-10 dataset

#### 4.2.2 SVHN 数据集

SVHN 数据集专用于街景门牌数字识别，提供 32×32 像素的 RGB 图像，其核心任务为识别街景图像中的中心数字。该数据集源自 Google Street View，其显著特征包含复杂背景干扰与图像模糊。训练集包含约 73 257 张样本，测试集约 26 032 张。实验通过 scipy.io.loadmat 加载.mat 格式数据，预处理包括维度调整及标准化，SVHN 数据集图片示例如图 5 所示。

图 5 SVHN 数据集图片示例

Fig. 5 Example images from the SVHN dataset

#### 4.2.3 实验数据集使用

数据集 CIFAR-10 和 SVHN 分别代表了图像分类中的自然图像场景与数字识别场景，CIFAR-10 中图像类别更加多样、语义丰富，而 SVHN 更偏向数字识别应用。它们对抗攻击实验的测试对象完全一致，使用的模型结构也统一为 VGG-16、VGG-19、ResNet-18、ResNet-34 和 DenseNet-121，以保证在多模型条件下评估对抗样本迁移攻击能力的公平性。

在训练过程中，CIFAR-10 和 SVHN 数据集都通过 DataLoader 加载，配置一致的 batch_size、shuffle 和 drop_last 参数，保证训练和测试过程中的样本分布均衡。生成式模型训练的图像输入来源主要为原始图像 $x$，原始图像对应标签及其目标标签 $y_t$，生成器接收这对输入生成条件对抗样本。在 Tester 模块中，针对 2 个数据集均可进行对多个模型的攻击效果进行评估，最终记录准确率、ASR 以及扰动强度。

CIFAR-10 与 SVHN 的组合，能够在实验中有效覆盖图像分类与数字识别 2 个主流任务场景，为验证所提出生成式模型的适用性与攻击迁移能力提供了多样化的数据支持。

### 4.3 实验所用对比方法

为了全面评估所提出的生成式对抗攻击方法的有效性和优势，实验中还引入了多种现有主流的对抗攻击算法作为对比方法，涵盖基于一阶梯度的方法（如 MI-FGSM[20]、VMI-FGSM[26]、EMI-FGSM[27]）、基于方向优化的策略（如 PGN[28]、AI-FGTM[29]）以及最新提出的 GRA[30] 方法，这些方法介绍如下：

1）MI-FGSM（momentum iterative FGSM）在 FGSM 基础上引入多步迭代和动量机制，通过历史梯度加权平均稳定扰动方向，显著提升跨模型迁移能力，成为迁移攻击的核心基线方法；

2）VMI-FGSM（variance-informed MI-FGSM）通过噪声采样估计梯度方差，动态调整扰动方向，增强黑盒场景下的稳定性和模型无关性；

3）PGN（prior-guided noise）利用类激活映射等模型响应生成先验引导扰动，无需梯度信息即可实现精准攻击，具备强无白盒依赖特性；

4）EMI-FGSM（ensemble MI-FGSM）对多个模型的梯度进行平均融合，生成通用对抗扰动以提升异构模型间的迁移效率；

5）AI-FGTM（attention-informed FGSM with targeted manipulation）结合注意力机制聚焦

关键区域扰动，以更低扰动幅度实现高目标 ASR，兼顾视觉隐蔽性；

6）GRA(gradient relevance attack)通过梯度空间重排与融合优化扰动方向，显著提升跨架构泛化迁移能力。

这些方法从梯度稳定性（MI-FGSM/VMI-FGSM）、无梯度引导（PGN）、多模型集成（EMI-FGSM）、注意力聚焦（AI-FGTM）和空间变换（GRA）等维度创新，为生成式对抗攻击的定量评估提供了攻击成功率、迁移能力与扰动范数等多维对比基准。

### 4.4　实验所用度量指标

在对抗样本实验中，为全面评估攻击方法的有效性，通常采用 2 类度量指标：一是攻击是否成功（分类准确率/ASR），二是攻击代价大小（扰动强度）。本实验中主要使用 ASR 和 $\ell_2$ 范数扰动强度作为评估标准。

#### 4.4.1　ASR

ASR 用于衡量对抗样本在目标模型上使预测结果发生预期偏移的能力。根据攻击类型的不同，计算方式略有差异：在无目标攻击（untargeted attack）中，若模型预测结果与原始标签不一致，则视为攻击成功；在有目标攻击（targeted attack）中，若模型预测结果等于攻击者指定的目标标签，则视为攻击成功。令总测试样本数为 $N$，目标标签为 $y_t$，模型对对抗样本 $x_{\text{adv}}^{(i)}$ 的预测表示为 $\hat{y}_{\text{adv}}^{(i)}$，则 ASR 计算公式为：

$$A_{\text{ASR}}=\frac{1}{N}\sum_{i=1}^{N}I(\hat{y}_{\text{adv}}^{(i)}=y_t) \tag{16}$$

式中，$I(\cdot)$为指示函数，当条件为真时取 1，反之为 0。该指标反映对抗样本误导模型到指定类别的能力，值越高说明攻击越成功。

#### 4.4.2　范数扰动强度

为了衡量对抗样本在输入空间中的修改幅度，$\ell_2$ 范数被广泛采用。对于每个样本，其扰动为 $\delta^{(i)}=x_{\text{adv}}^{(i)}-x^{(i)}$，则该样本的扰动强度为：

$$\|\delta^{(i)}\|_2=\Big(\sum_{j=1}^{HWC}(x_{\text{adv}}^{(i)}-x_j^{(i)})^2\Big)^{\frac{1}{2}} \tag{17}$$

式中，$H$，$W$，$C$ 分别为图像的高、宽和通道数。为了整体评估攻击方法的扰动代价，常对所有样本求平均，表示为：

$$A_{\text{avg},\ell_2}=\frac{1}{N}\sum_{i=1}^{N}\|x_{\text{adv}}^{(i)}-x^{(i)}\|_2 \tag{18}$$

$\ell_2$ 范数值越小表示攻击越隐蔽，越接近于原始图像，更难被人眼察觉。在实际评估中，ASR 反映攻击方法的有效性，而 $\ell_2$ 范数反映攻击的隐蔽性。两者结合能够全面刻画一个攻击方法在实际应用中的性能表现。在本实验中，算法以固定批次的样本数在不同模型上计算上述指标，记录并比较生成式对抗攻击方法与其他攻击方法的优劣。

### 4.5　实验流程设计

实验流程设计分为以下步骤：

1）目标（代理）模型训练。首先，在指定数据集（如 CIFAR-10、SVHN）上分别对多个主流深度神经网络进行分类任务训练，包括 VGG-16、VGG-19、ResNet-18、ResNet-34、SENet、DenseNet-121 和 DenseNet-201。这些模型被用作代理模型和目标模型，代理模型用于生成对抗样本，目标模型用于测试迁移性能。训练过程中采用标准交叉熵损失，结合常规的随机梯度下降优化算法，直至验证集上收敛。

2）生成式模型训练。在训练好的代理模型基础上，构建基于条件生成对抗网络（CGAN）的生成器。生成器以原始图像、真实标签与攻击目标标签为输入，生成对抗样本。判别器则负责判别输入样本是否为目标类别真实样本。通过联合优化生成器和判别器，多轮迭代训练使生成器逐渐学会生成既符合目标类别分布又具备误导分类器能力的高质量对抗样本。模型在训练过程中记录各个训练轮次的生成器参数，并保存供后续评估使用。

3）测试生成式模型。在生成式模型的测试阶段，每隔若干轮迭代，使用最新的生成器对代理模型生成的对抗样本进行攻击性能评估。具体做法是将生成的对抗样本输入代理模型，计算有目标攻击及无目标攻击的 ASR，同时记录不同扰动强度下的攻击表现，以便跟踪生成器训练过程中攻击效果的演化。

4）迁移性测试。在最终评估阶段，为了公平比较，分别在传统攻击方法（如 MI-FGSM、VMI-FGSM、EMI-FGSM、PGN、AI-FGTM、GRA）与生成式方法之间设计一致的迁移攻击流程。对于传统方法，流程是先将数据集样本输入代理模型，结合代理模型梯度和攻击算法生成对抗样本，再将这些对抗样本输入步骤 1）中训练好的所有的目标模型，计算其分类准确率，评估迁移攻

击效果。而对于生成式方法，直接使用步骤 2）中训练好的生成器独立生成对抗样本，无须借助代理模型或额外优化过程，直接输入目标模型进行分类测试和准确率计算。

5）数据分析。实验通过比较不同攻击方法在各目标模型上的 ASR、迁移性能和计算效率，评估生成式方法在黑盒攻击任务中的表现。在白盒攻击场景下，分析代理模型与目标模型一致时的攻击成功率，作为性能基准。在黑盒攻击场景下，重点评估对抗样本在代理模型（如 VGG-16）与目标模型（如 ResNet-18）不一致时的迁移性。

具体地，实验具体分为白盒攻击和黑盒攻击 2 种攻击场景。其中，白盒攻击场景为代理模型与目标模型一致（如均为 VGG-16），攻击算法直接利用目标模型的梯度信息生成对抗样本，测试攻击算法的理论效能；而黑盒攻击为代理模型与目标模型不同（如代理模型为 VGG-16，目标模型为 ResNet-18），对抗样本在代理模型上生成后迁移至目标模型，测试其跨模型攻击效果。黑盒攻击是本研究的重点，旨在评估对抗样本的迁移性在实际场景中的表现，模拟攻击者无法直接访问目标模型的现实情况。

## 5 实验结果及分析

### 5.1 迁移性测试

#### 5.1.1 CIFAR-10 数据集上的迁移性实验

经过380轮迭代训练后，生成器生成的基于 CIFAR-10 数据集的部分对抗样本如图 6 所示。可以看到输出的图片经肉眼观察无法发现明显噪声，图片内容比较清晰。

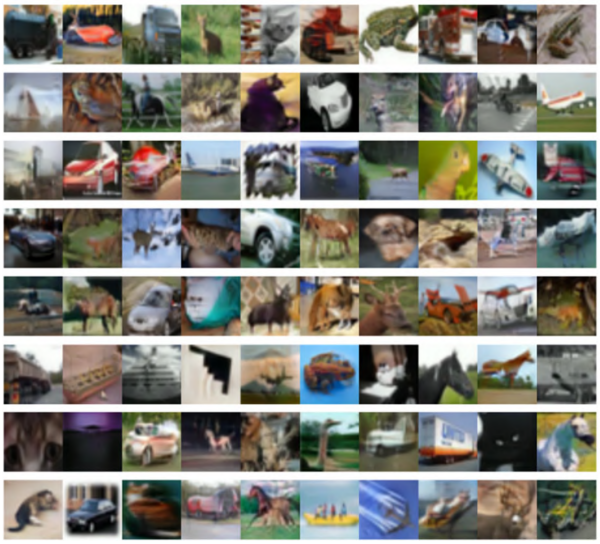

图 6 生成器在第 380 次迭代后生成的 CIFAR-10 部分对抗样本示例

Fig. 6 Examples of partial adversarial samples from CIFAR-10 generated after the 380th iteration

为了进一步验证生成式对抗样本的迁移能力，本文在 CIFAR-10 数据集上进行了迁移性测试实验。实验分别比较了不同攻击方法和不同代理模型下所生成的对抗样本，对多个目标模型进行有目标攻击的成功率表现。实验过程中控制扰动强度（即 $\ell_2$ 范数）为 15，ASR 的测试结果见表 2 所列。

表 2 CIFAR-10 数据集上对抗样本迁移性测试结果

Tab. 2 Transferability test results of adversarial examples on the CIFAR-10 dataset

/%

| 攻击方法 | 代理模型 | 目标模型 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | VGG-16 | VGG-19 | ResNet-18 | ResNet-34 | SENet | DenseNet-121 | DenseNet-201 |
| 生成式对抗攻击 | 无需代理模型 | **47.90** | **48.70** | **56.00** | **49.00** | **37.60** | **45.30** | **43.90** |
| MI-FGSM | VGG-16 | 99.60 | 33.60 | 20.20 | 8.10 | 9.10 | 9.20 | 8.70 |
| | VGG-19 | 32.30 | 97.40 | 19.50 | 8.00 | 8.60 | 9.10 | 7.60 |
| | ResNet-18 | 22.40 | 22.10 | 97.60 | 12.30 | 10.10 | 10.20 | 10.00 |
| | ResNet-34 | 25.90 | 24.80 | 30.50 | 84.80 | 10.50 | 13.20 | 12.60 |
| | SENet | 5.40 | 4.20 | 4.40 | 2.60 | 100 | 6.10 | 5.10 |
| | DenseNet-121 | 4.90 | 5.20 | 5.70 | 2.80 | 6.70 | 99.80 | 7.30 |
| | DenseNet-201 | 6.30 | 4.80 | 5.30 | 3.10 | 6.60 | 6.40 | 99.80 |

续表

| 攻击方法 | 代理模型 | 目标模型 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | VGG-16 | VGG-19 | ResNet-18 | ResNet-34 | SENet | DenseNet-121 | DenseNet-201 |
| VMI-FGSM | VGG-16 | 99.00 | 41.00 | 23.10 | 10.80 | 10.50 | 10.90 | 9.80 |
| | VGG-19 | 38.40 | 96.00 | 22.50 | 8.40 | 9.10 | 8.40 | 9.20 |
| | ResNet-18 | 24.40 | 25.20 | 97.30 | 12.70 | 10.10 | 10.30 | 10.90 |
| | ResNet-34 | 25.70 | 24.00 | 30.30 | 85.20 | 10.00 | 14.70 | 12.60 |
| | SENet | 6.20 | 5.60 | 5.90 | 3.30 | 100 | 6.70 | 6.40 |
| | DenseNet-121 | 8.00 | 5.80 | 6.80 | 4.00 | 8.50 | 99.60 | 8.90 |
| | DenseNet-201 | 7.40 | 5.30 | 6.90 | 3.30 | 7.10 | 7.80 | 99.70 |
| PGN | VGG-16 | 94.70 | 39.80 | 22.30 | 11.70 | 11.20 | 11.50 | 12.20 |
| | VGG-19 | 36.80 | 90.80 | 20.70 | 9.40 | 10.80 | 9.70 | 10.90 |
| | ResNet-18 | 23.30 | 22.30 | 83.40 | 10.70 | 11.20 | 10.40 | 12.40 |
| | ResNet-34 | 25.40 | 24.60 | 28.80 | 79.80 | 11.90 | 15.20 | 13.80 |
| | SENet | 7.70 | 6.60 | 6.90 | 3.10 | 97.70 | 8.50 | 6.80 |
| | DenseNet-121 | 8.30 | 7.20 | 8.70 | 4.50 | 10.20 | 96.60 | 12.00 |
| | DenseNet-201 | 8.60 | 6.50 | 7.80 | 4.30 | 8.90 | 8.80 | 94.90 |
| EMI-FGSM | VGG-16 | 99.60 | 34.20 | 18.70 | 9.30 | 9.70 | 9.40 | 10.40 |
| | VGG-19 | 31.30 | 98.80 | 17.70 | 8.10 | 9.70 | 8.70 | 7.70 |
| | ResNet-18 | 20.30 | 19.80 | 97.70 | 11.70 | 10.40 | 11.20 | 10.80 |
| | ResNet-34 | 20.80 | 20.00 | 24.40 | 88.60 | 10.80 | 11.50 | 11.10 |
| | SENet | 6.50 | 5.20 | 5.30 | 2.90 | 100 | 6.60 | 6.50 |
| | DenseNet-121 | 6.70 | 5.20 | 5.70 | 3.20 | 7.50 | 99.50 | 6.70 |
| | DenseNet-201 | 7.20 | 5.40 | 7.20 | 3.40 | 7.50 | 8.10 | 99.60 |
| AI-FGTM | VGG-16 | 98.80 | 26.30 | 16.00 | 8.90 | 8.20 | 8.10 | 7.80 |
| | VGG-19 | 25.50 | 96.00 | 15.30 | 6.60 | 7.30 | 7.60 | 7.80 |
| | ResNet-18 | 18.40 | 18.70 | 93.90 | 11.30 | 8.80 | 9.10 | 10.20 |
| | ResNet-34 | 21.80 | 20.00 | 26.50 | 82.40 | 9.10 | 11.70 | 11.20 |
| | SENet | 5.40 | 4.60 | 4.30 | 2.70 | 99.70 | 4.70 | 6.30 |
| | DenseNet-121 | 5.80 | 5.90 | 5.90 | 3.00 | 6.60 | 99.00 | 7.10 |
| | DenseNet-201 | 6.70 | 5.00 | 6.50 | 3.20 | 6.70 | 7.00 | 99.20 |
| GRA | VGG-16 | 96.80 | 43.70 | 24.80 | 12.40 | 11.30 | 11.60 | 10.90 |
| | VGG-19 | 40.10 | 91.20 | 24.60 | 9.60 | 10.60 | 10.00 | 9.70 |
| | ResNet-18 | 25.80 | 26.00 | 88.80 | 12.90 | 10.80 | 10.20 | 11.60 |
| | ResNet-34 | 26.90 | 26.50 | 31.40 | 84.20 | 11.80 | 15.50 | 14.50 |
| | SENet | 7.90 | 6.00 | 6.10 | 3.10 | 97.50 | 7.20 | 7.50 |
| | DenseNet-121 | 9.70 | 7.90 | 8.60 | 4.60 | 9.90 | 97.70 | 10.90 |
| | DenseNet-201 | 8.70 | 6.50 | 8.50 | 4.10 | 8.70 | 9.70 | 95.70 |

为了提升数据的可读性与对比分析的系统性，本文对原始数据进行了清洗与整合，整理得到图 7～10 共 4 张统计图。基于这些统计图，可以清晰地对比各类攻击方法与代理模型在迁移攻击任务中的表现差异。

图 7 为各类攻击方法在迁移攻击场景中的平均 ASR。从图中可以看到，生成式对抗攻击方法在所有方法中表现最为优越，平均 ASR 达到了 47%，显著高于传统方法如 MI-FGSM（23%）、VMI-FGSM（25%）、PGN（24%）等。这一结果表明，生成式对抗攻击方法所生成的对抗样本不仅具备更强的攻击能力，而且能够有效跨越不同模型结构，在多种目标模型上保持稳定且较高的干扰效果，展现出良好的迁移性。

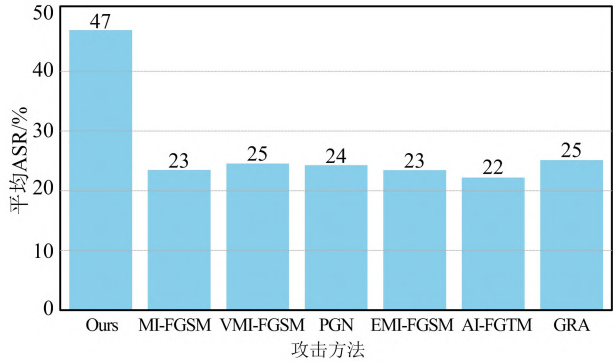


图 7 不同对抗攻击算法在 CIFAR-10 数据集上的平均 ASR

Fig. 7 Average ASR of different adversarial attack algorithms on the CIFAR-10 dataset

进一步分析图 8 中在目标模型与代理模型不一致，即黑盒攻击的情况下的平均 ASR 可知，生成式对抗样本在该场景下依然保持了 47% 的攻击率，而其他方法如 GRA、AI-FGTM 等则均下降至 15%以下。可以明显观察到，生成式对抗攻击在跨模型的迁移攻击中展现出显著优势，较之前柱状图中的差距进一步拉大。这一结果进一步验证了生成式对抗样本具备较强的迁移能力，能够脱离对特定结构或同类网络的依赖，提升在异构模型间的攻击效果。

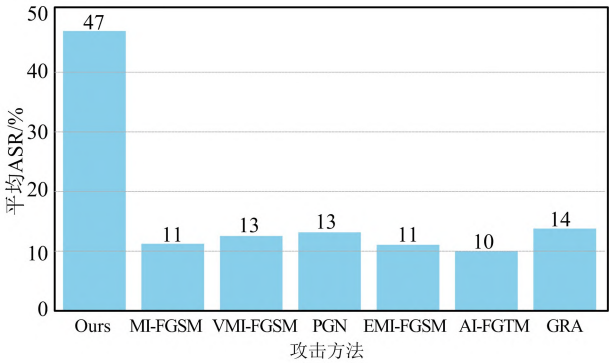


图 8 代理模型和目标模型不一致时，不同对抗攻击算法在 CIFAR-10 数据集上的平均 ASR

Fig. 8 Average ASR of different adversarial attack algorithms on the CIFAR-10 dataset under mismatched proxy and target models

在图 9 所示的不同代理模型下的平均 ASR 中，生成式对抗攻击的平均成功率显著高于传统代理模型（如 VGG-16 为 29%，ResNet-18 为 26%）。这表明生成式攻击能够跳脱局部扰动的限制，在更大的搜索空间中寻找有效的攻击方向，从而显著提升攻击的覆盖范围与效果。

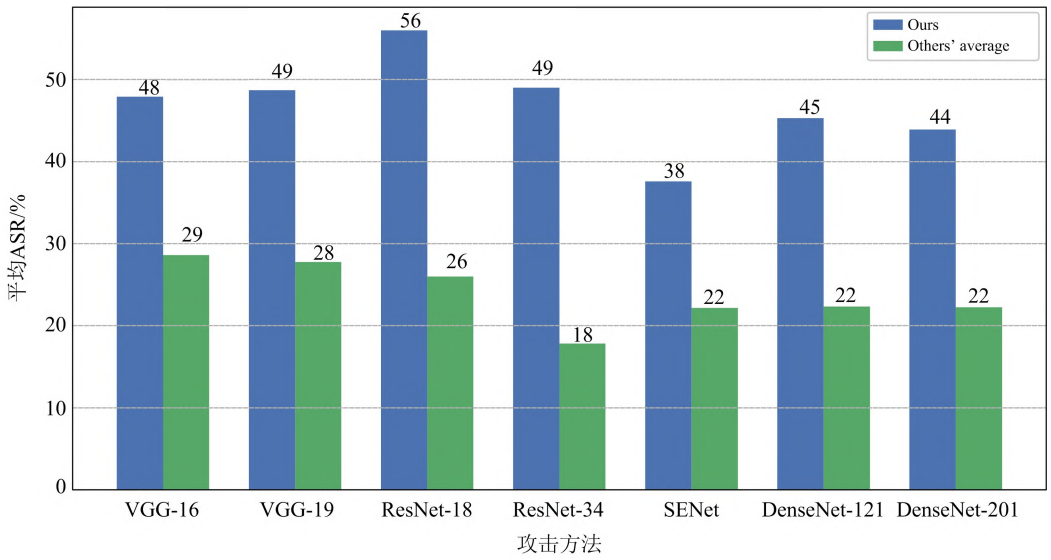


图 9 不同对抗攻击算法在 CIFAR-10 数据集上对各目标模型的平均 ASR

Fig. 9 Average ASR of different adversarial attack algorithms on various target models using the CIFAR-10 dataset

进一步地，图 10 展示了在黑盒攻击的情况下，生成式攻击方法依然保持了稳定而高效的攻击性能，凸显出其突出的跨模型泛化能力。而传统代理模型方法在该迁移场景下的 ASR 显著下降，仅在 8%～20%区间内波动。因此，可以更加明显地看出，生成式对抗攻击方法在异构模型环境下展现出更为突出的性能优势。

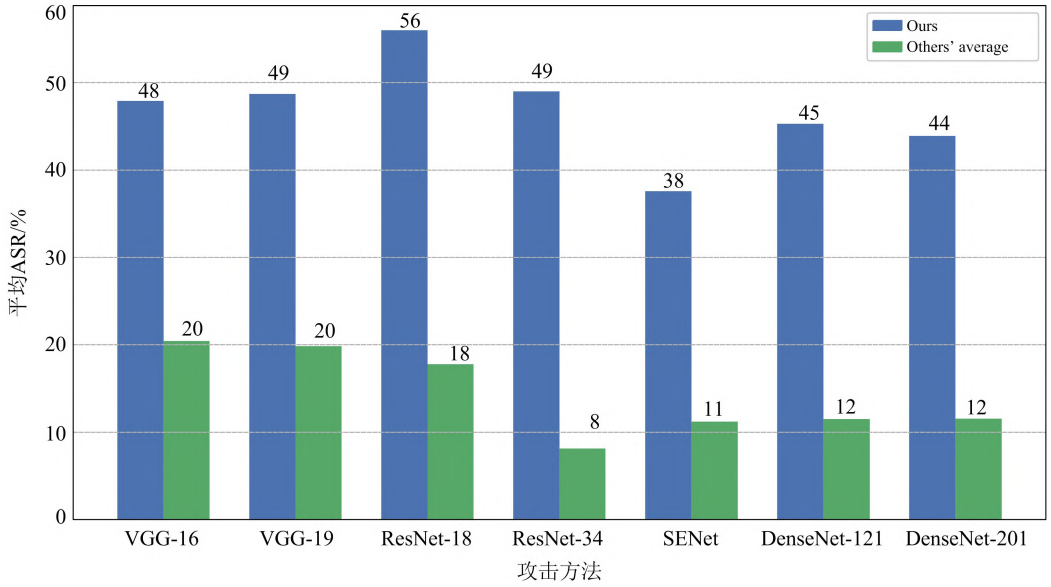


图 10　黑盒攻击时，不同对抗攻击算法在 CIFAR-10 数据集上对各目标模型的平均 ASR

Fig. 10　Average ASR of adversarial attack algorithms on target models in black-box scenarios using the CIFAR-10 dataset

### 5.1.2　SVHN 数据集上的迁移性实验

为了更全方面地验证生成式对抗攻击方法对于多种数据集均有较强的迁移性，本文也在数据集 SVHN 上进行了实验。由于 SVHN 图片来源于街拍的数字图像，因此相比于其他 32×32 像素图片的噪声会更多，使得噪声的存在并不明显。如图 11 所示，可以看到迭代 75 轮后的生成器生成的图片无明显噪声，内容比较清晰，与原图差别不大。

图 11　生成器在第 75 次迭代后生成的 SVHN 部分对抗样本示例

Fig. 11　Sample adversarial examples from SVHN generated after the 75th iteration

为了对生成式对抗攻击方法进行更深入的了解，将数据集更改为 SVHN 来探究生成样本在 SVHN 上的迁移性。实验保证 $\ell_2$ 范数值也为 15，有目标攻击的迁移性测试的攻击成功率结果见表 3 所列。

类似地，对表格 3 中的数据进行分析整理，结果如图 12～13 所示。图 12 为黑盒攻击场景下，生成式对抗攻击与其他传统攻击方法的平均 ASR 统计图。由图 12 可知，生成式对抗样本在黑盒攻击环境下的 ASR 达到 69%，而其他方法的迁移 ASR 均降至 40%以下，举例的这几种传统攻击方法的 ASR 均在 24%～40%之间。并且在此情况下，可以观察到，生成式对抗攻击的方法在跨模型之间的攻击效果领先其他攻击方法 30%以上，该结果进一步验证了生成式的对抗样本具备高度的模型间迁移性。

观察发现，传统攻击方法在 SVHN 上的攻击表现受限于代理模型和目标模型之间的结构一致性。当攻击模型与代理模型结构相似时，攻击效果较好；但当结构差异增大时，ASR 便显著下降。说明 MI-FGSM 的迁移能力有限，更适用于白盒或同类模型场景，而不适合实际中面对未知模型的黑盒攻击。

**表 3 SVHN 数据集上对抗样本迁移性测试结果**

**Tab. 3 Transferability test results of adversarial examples on the SVHN dataset**

/%

| 攻击方法 | 代理模型 | 目标模型 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | VGG-16 | VGG-19 | ResNet-18 | ResNet-34 | SENet | DenseNet-121 | DenseNet-201 |
| **生成式对抗攻击** | **无需代理模型** | **73.10** | **73.80** | **73.40** | **72.20** | **65.00** | **64.40** | **62.90** |
| MI-FGSM | VGG-16 | 96.00 | 60.10 | 48.60 | 51.10 | 28.30 | 30.00 | 29.20 |
| | VGG-19 | 64.90 | 95.70 | 53.70 | 53.60 | 31.70 | 31.50 | 30.10 |
| | ResNet-18 | 53.90 | 56.40 | 99.40 | 73.10 | 30.40 | 32.70 | 31.20 |
| | ResNet-34 | 53.80 | 54.90 | 74.20 | 98.00 | 29.70 | 31.20 | 30.50 |
| | SENet | 11.00 | 12.90 | 8.10 | 9.10 | 99.60 | 16.60 | 14.80 |
| | DenseNet-121 | 9.70 | 11.10 | 7.50 | 7.90 | 16.60 | 99.60 | 15.00 |
| | DenseNet-201 | 11.90 | 12.30 | 7.90 | 9.20 | 16.80 | 16.40 | 99.70 |
| VMI-FGSM | VGG-16 | 94.50 | 62.70 | 51.70 | 53.40 | 35.70 | 38.40 | 36.90 |
| | VGG-19 | 68.20 | 94.20 | 59.40 | 60.30 | 36.30 | 37.00 | 34.50 |
| | ResNet-18 | 56.80 | 59.40 | 99.50 | 74.90 | 33.80 | 37.10 | 37.10 |
| | ResNet-34 | 57.30 | 57.50 | 75.20 | 97.40 | 32.80 | 37.10 | 35.40 |
| | SENet | 19.50 | 21.80 | 14.60 | 14.40 | 99.90 | 26.60 | 23.20 |
| | DenseNet-121 | 17.40 | 18.70 | 14.70 | 13.40 | 23.50 | 99.60 | 23.20 |
| | DenseNet-201 | 17.50 | 19.20 | 13.30 | 13.20 | 25.20 | 24.20 | 99.80 |
| PGN | VGG-16 | 90.00 | 61.90 | 49.70 | 50.80 | 33.90 | 37.30 | 35.30 |
| | VGG-19 | 66.70 | 92.80 | 58.50 | 59.80 | 35.80 | 39.20 | 35.10 |
| | ResNet-18 | 47.20 | 47.30 | 98.50 | 65.90 | 25.70 | 28.50 | 27.30 |
| | ResNet-34 | 51.20 | 53.30 | 71.50 | 95.20 | 31.40 | 33.60 | 32.70 |
| | SENet | 12.90 | 16.70 | 11.40 | 12.10 | 99.30 | 21.00 | 17.90 |
| | DenseNet-121 | 12.30 | 14.10 | 9.10 | 9.50 | 18.90 | 99.40 | 17.80 |
| | DenseNet-201 | 13.30 | 14.90 | 10.90 | 10.50 | 18.00 | 19.60 | 99.10 |
| EMI-FGSM | VGG-16 | 96.40 | 54.90 | 41.60 | 43.90 | 26.00 | 25.00 | 27.10 |
| | VGG-19 | 58.90 | 96.90 | 43.20 | 46.40 | 27.00 | 24.90 | 25.40 |
| | ResNet-18 | 37.20 | 38.40 | 99.90 | 53.00 | 22.70 | 22.80 | 23.40 |
| | ResNet-34 | 44.50 | 44.70 | 66.10 | 98.80 | 26.00 | 27.70 | 27.30 |
| | SENet | 11.90 | 13.10 | 8.80 | 9.00 | 99.80 | 17.40 | 15.30 |
| | DenseNet-121 | 10.70 | 10.80 | 7.90 | 9.30 | 17.20 | 99.70 | 15.10 |
| | DenseNet-201 | 10.90 | 12.90 | 8.20 | 8.90 | 14.30 | 16.60 | 99.90 |
| AI-FGTM | VGG-16 | 93.30 | 47.20 | 34.00 | 35.50 | 22.60 | 21.30 | 21.50 |
| | VGG-19 | 50.30 | 92.90 | 40.10 | 42.00 | 23.40 | 24.50 | 21.30 |
| | ResNet-18 | 33.30 | 36.70 | 98.80 | 46.70 | 21.60 | 22.40 | 23.10 |
| | ResNet-34 | 37.00 | 39.30 | 55.90 | 96.00 | 22.10 | 22.70 | 22.20 |
| | SENet | 10.40 | 12.90 | 9.20 | 9.30 | 99.70 | 17.20 | 14.00 |
| | DenseNet-121 | 10.70 | 11.50 | 8.60 | 9.80 | 17.90 | 98.60 | 15.50 |
| | DenseNet-201 | 12.50 | 13.90 | 10.90 | 10.60 | 17.50 | 17.00 | 99.20 |

续表

| 攻击方法 | 代理模型 | 目标模型 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | VGG-16 | VGG-19 | ResNet-18 | ResNet-34 | SENet | DenseNet-121 | DenseNet-201 |
| GRA | VGG-16 | 91.50 | 67.30 | 57.40 | 58.60 | 40.90 | 41.90 | 41.30 |
| | VGG-19 | 71.30 | 93.00 | 62.80 | 64.10 | 39.50 | 44.10 | 40.30 |
| | ResNet-18 | 61.30 | 63.20 | 99.20 | 77.20 | 37.10 | 43.10 | 40.80 |
| | ResNet-34 | 63.10 | 63.40 | 76.90 | 96.80 | 39.20 | 44.20 | 41.60 |
| | SENet | 17.50 | 21.60 | 14.40 | 15.70 | 99.80 | 25.50 | 22.90 |
| | DenseNet-121 | 15.10 | 18.30 | 12.20 | 12.50 | 24.80 | 99.70 | 24.80 |
| | DenseNet-201 | 18.70 | 19.80 | 14.30 | 14.90 | 23.50 | 24.30 | 99.70 |

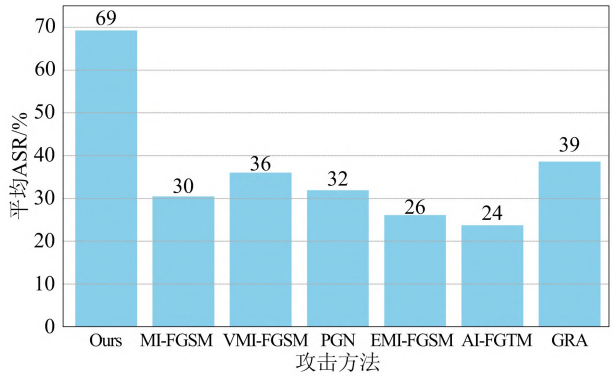


**图 12　黑盒攻击时，不同对抗攻击算法在 SVHN 数据集上的平均 ASR**

**Fig. 12　Average ASR of different adversarial attack algorithms on the SVHN dataset in black-box scenarios**

图 13 为黑盒攻击场景下，生成式对抗攻击与其他传统攻击方法在攻击各个模型时的平均 ASR。由图 13 可知，在代理模型与目标模型结构差异显著的情况下，生成式对抗攻击方法依旧保持高效攻击性能（例如在攻击 VGG-16 时 ASR 为 73%），而对于传统代理模型，比如 ResNet 系列和 VGG 系列攻击成功率则在 34%～35%之间，其余的较复杂的模型的成功率则降至 30%以下。整体而言，生成式对抗样本的迁移性则没有因为模型的复杂度上升而下降。因此，生成式对抗攻击方法的优势在这种情况下非常明显。

与 CIFAR-10 数据集的实验结果相比，生成式对抗样本在 SVHN 上整体 ASR 更高。这一现象可以归因于 SVHN 数据集中图像背景较为统一、样本结构规整，生成器更容易建模其类间边界，从而生成高质量的类间对抗样本。而 CIFAR-10 由于图像内容复杂、背景干扰大，对抗样本更难保持跨模型的一致性，迁移 ASR 相对偏低。

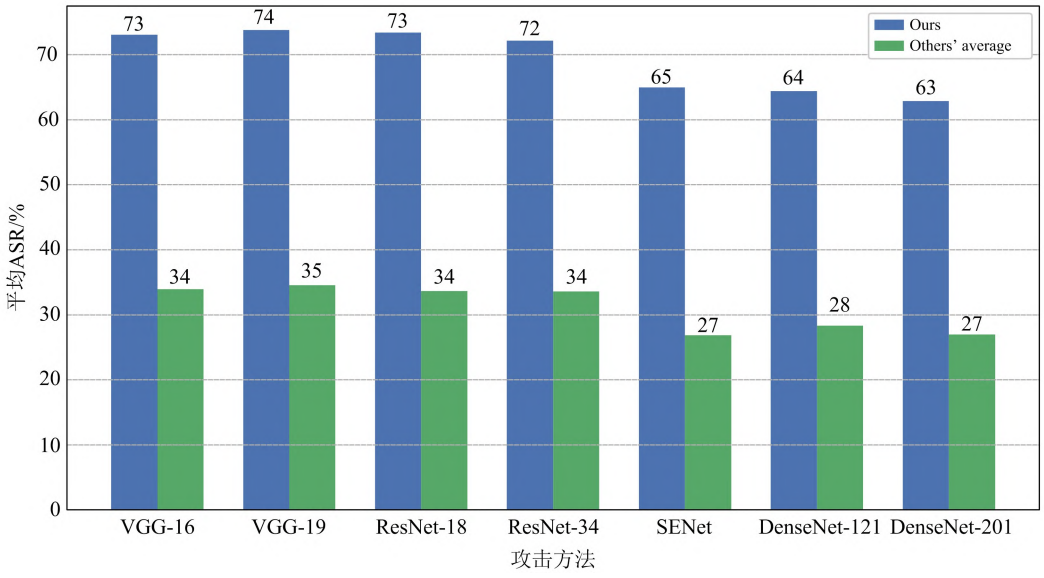


**图 13　黑盒攻击时，不同对抗攻击算法在 SVHN 数据集上对各目标模型的平均 ASR**

**Fig. 13　Average ASR of different adversarial attack algorithms on various target models using the SVHN dataset in black-box scenarios**

因此，生成式对抗攻击方法在 SVHN 上展现出优越的迁移攻击能力和结构不敏感性，其生成的样本不仅能有效误导单一模型，在异构模型间仍具有良好的攻击效果，验证了其在黑盒攻击场景中的实用价值。这进一步凸显了生成式对抗方法相较于传统扰动方法的优势所在。

综合以上分析可知，生成式对抗攻击方法在跨模型的场景中表现出了非常显著的迁移能力。生成式方法不仅在性能上超越了基于梯度的传统攻击技术，而且在实际应用中特别适合于黑盒攻击以及在未知模型环境下的攻击任务。这种优势使得生成式攻击方法成为当前研究和实际应用中非常有前景的方向。

### 5.2 参数敏感性实验

为验证生成式对抗攻击方法各模块对攻击性能的贡献，本研究设计了参数敏感性实验，控制目标变量，改变扰动强度和生成轮数，观察 ASR 与扰动强度的变化。该方法有助于明确模型设计中各部分的有效性，提升方法的可解释性，为对抗攻击策略的优化提供实证支持。

#### 5.2.1 攻击强度对迁移性的影响

本文把 $\ell_2$ 范数，也就是对图片的干扰强度，称为攻击强度。攻击强度越高，对图片的干扰就越强，越容易被发现，但是相应的攻击更容易成功。对于不同强度的攻击强度测试图如图 14 所示。

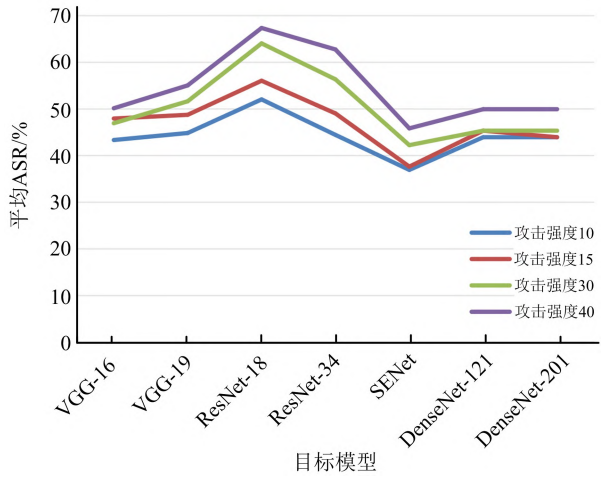


**图 14 不同攻击强度对各模型 ASR 的影响**

**Fig. 14 Impact of different attack strengths on ASR of various models**

从图 14 可以看出，随着攻击强度（即扰动幅度）的增大，各模型的 ASR 整体呈上升趋势，表明模型对更大幅度的对抗扰动更为敏感。尤其是 ResNet-18 和 ResNet-34，这 2 类残差结构模型对攻击强度变化反应最为明显：在攻击强度为 10 时，平均攻击成功率分别为 52.00%和 44.40%；而在攻击强度为 40 时则迅速上升至 67.30%和 62.70%，展现出较低的鲁棒性。

VGG-16 与 VGG-19 同样呈现随攻击强度增强而攻击效果提升的趋势，分别从 43.30%和 44.80%上升至 50.10%和 55.00%。相较之下，SENet 在 4 个攻击强度下表现相对稳定，ASR 变化较小，说明其结构在一定程度上具有更好的抗扰动能力。而 DenseNet-121 和 DenseNet-201 在所有强度下的成功率变化幅度较小，在 43.90%～49.90%之间浮动，体现出其对攻击强度的响应较为平稳。

综合而言，图 14 揭示了生成式攻击在扰动强度与 ASR 之间的正相关性，验证了对抗样本随着攻击强度增加能显著提高误导模型的能力。同时也反映出不同网络结构对攻击强度的敏感性差异，残差结构（ResNet 系列）最易被攻击，而 DenseNet 和 SENet 相对更稳健。该结论有助于后续在攻击强度控制与模型防御方面做更有针对性的策略设计。

#### 5.2.2 生成轮数对迁移性的影响

图 15 中展示的是生成器在前 40 个训练 epoch 中的表现，包含 3 个关键指标：对抗样本在原标签下的准确率（clean accuracy，蓝色线），越低说明攻击越有效；目标标签攻击的成功率（attack accuracy，橙色线），越高越好；对抗扰动的大小（$\ell_2$ 范数，黄色线），用于评估图像被修改的程度。

从整体趋势上看，随着训练轮次的推进，生成器逐渐学会如何在较小扰动的条件下生成具有欺骗性的对抗样本，攻击能力逐步增强。在初始阶段（epoch 1～10），attack accuracy 波动明显，平均值较低，仅在部分 epoch（如第 6、9 和 10 轮）突破了 14%。与此同时，$\ell_2$ 范数平均维持在 20 以上，说明此时生成器尚未收敛，生成的扰动幅度较大，攻击性能也不稳定。从 epoch 11 开始，ASR 逐渐提高，特别是在 epoch 12～20 区间，attack accuracy 多次突破 20%，甚至在 epoch 19 达到 39.06%。与此同时，$\ell_2$ 范数逐步下降，平均降低到 15 左右，表明生成器开始以较低扰动生成

更高质量的攻击样本，显示出对抗样本生成效果的提升。进入 epoch 21～40 后，ASR 整体表现更加稳定，部分 epoch 均维持在 30% 以上，与之相对应的 $\ell_2$ 范数在这一阶段大多稳定在 12～14 区间，说明生成器能够以中等扰动水平实现较强的攻击效果。尤其在 epoch 26，attack accuracy 达到 30%，而 $\ell_2$ 范数降至 11.98，为低扰动高攻击率的典型代表。此外，clean accuracy 始终维持在较低水平，多数轮次处于 10% 以下，这进一步说明对抗样本已经有效破坏了原有模型对正常类别的识别能力。这种 clean accuracy 的抑制现象与高 attack accuracy 的提升相辅相成，表明生成器在保持扰动压制的同时成功欺骗了目标模型。

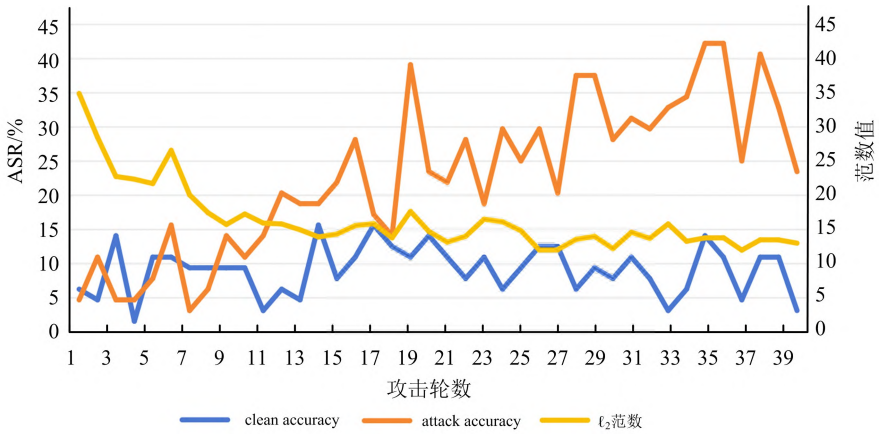


**图 15　对抗样本不同生成轮数对各模型 ASR 的影响**

**Fig. 15　Impact of different generation iterations of adversarial examples on the attack success rate of various models**

综上所述，从前 40 个 epoch 的训练结果可以观察到，随着训练轮数的增加，生成器的攻击能力逐步增强，扰动控制逐步趋稳。

## 6　结束语

本文围绕生成式对抗攻击在迁移性攻击中的应用展开研究，提出了一种基于 GAN 的方法，并在 CIFAR-10 与 SVHN 数据集上进行了系统的对比与参数敏感性实验。研究结果表明，该方法兼具高迁移性、隐蔽性与图像质量，摆脱了对模型结构与梯度的依赖，生成的对抗样本在迁移攻击成功率上显著优于传统方法，在图像结构上保持较好的自然性与可解释性。同时，提出的方法为黑盒攻击提供了新路径，对对抗样本的可解释性与防御机制设计提供了理论支持与应用前景。

## 参 考 文 献

[1] 沙嘉强. 针对图像分类的有目标生成式对抗攻击研究[D]. 杭州：浙江科技大学，2024.
SHA Jiaqiang. Research on targeted generative adversarial attacks for image classification[D]. Hangzhou: Zhejiang University of Science and Technology, 2024. (in Chinese)

[2] 赵正平. 人工智能大语言模型和 AI 芯片的新进展(续)[J]. 微纳电子技术，2025，62(4)：040101.
ZHAO Zhengping. New advances in AI large language models and AI chips(continued)[J]. Micronanoelectronic Technology, 2025, 62(4): 040101. (in Chinese)

[3] GUO X, SHEN Z J, ZHANG Y J, et al. Review on the application of artificial intelligence in smart homes[J]. Smart Cities, 2019, 2(3): 402-420.

[4] AKHTAR M, MORIDPOUR S. A review of traffic congestion prediction using artificial intelligence[J]. Journal of Advanced Transportation, 2021, 2021(1): 8878011.

[5] 鲁思迪，何元恺，施巍松. 车计算：自动驾驶时代的新型计算范式[J]. 计算机研究与发展，2025，62(1)：2-21.
LU Sidi, HE Yuankai, SHI Weisong. Vehicle computing: an emerging computing paradigm for the autonomous driving era[J]. Journal of Computer Research and Development, 2025, 62(1): 2-21. (in Chinese)

[6] ZHANG L F, ZHANG L P. Artificial intelligence for remote sensing data analysis: a review of challenges and opportunities[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 270-294.

[7] 樊琳，龚勋，郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. 电子学报，2024，52(7)：2341-2355.
FAN Lin, GONG Xun, ZHENG Cenyang. A multi-modal medical image analysis algorithm based on text guidance[J]. Acta Electronica Sinica, 2024, 52(7):

2341-2355. (in Chinese)

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

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2025-07-20]. https://arxiv.org/abs/1312.6199.

[10] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]// YAMPOLSKIY R V. Artificial intelligence safety and security. New York: Chapman & Hall/CRC, 2018: 99-112.

[11] 纪守领,杜天宇,邓水光,等. 深度学习模型鲁棒性研究综述[J]. 计算机学报,2022,45(1):190-206.
JI Shouling, DU Tianyu, DENG Shuiguang, et al. Robustness certification research on deep learning models: a survey[J]. Chinese Journal of Computers, 2022, 45(1): 190-206. (in Chinese)

[12] 陶卿,高乾坤,姜纪远,等. 稀疏学习优化问题的求解综述[J]. 软件学报,2013,24(11):2498-2507.
TAO Qing, GAO Qiankun, JIANG Jiyuan, et al. Survey of solving the optimization problems for sparse learning[J]. Journal of Software, 2013, 24(11): 2498-2507. (in Chinese)

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

[14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. [2025-07-20]. https://arxiv.org/abs/1412.6572.

[15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2025-07-20]. https://arxiv.org/abs/1706.06083.

[16] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. [S.l.]: IEEE, 2017: 39-57.

[17] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of 2016 IEEE Symposium on Security and Privacy. San Jose: IEEE, 2016: 582-597.

[18] TU C C, TING P S, CHEN P Y, et al. Autozoom: autoencoder-based zeroth order optimization method for attacking black-box neural networks [C]// Proceedings of 2019 AAAI Conference on Artificial Intelligence. [S.l.: s.n.], 2019: 742-749.

[19] NARODYTSKA N, KASIVISWANATHAN S P. Simple black-box adversarial perturbations for deep networks[EB/OL]. [2025-07-20]. https://arxiv.org/abs/1612.06299.

[20] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 9185-9193.

[21] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2730-2739.

[22] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4312-4321.

[23] MOOSAVI-DEZFOOLI S-M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 1765-1773.

[24] POURSAEED O, KATSMAN I, GAO B C, et al. Generative adversarial perturbations[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 4422-4431.

[25] NASEER M, KHAN S, KHAN M H, et al. Cross-domain transferability of adversarial perturbations [C]//Proceedings of the 32nd Aunnual Conference on Neural Information Processing Systems. [S.l.: s.n.], 2019: 12905-12915.

[26] WANG X S, HE K. Enhancing the transferability of adversarial attacks through variance tuning [C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 1924-1933.

[27] WANG X S, LIN J D, HU H, et al. Boosting adversarial transferability through enhanced momentum[EB/OL]. [2025-07-20]. https://arxiv.org/abs/2103.10609.

[28] GE Z J J, LIU H Y, WANG X S, et al. Boosting adversarial transferability by achieving flat local maxima[C]//Proceedings of the 36th Aunnual Conference

on Neural Information Processing Systems. [S. l. : s. n. ], 2023: 70141-70161.

[29] ZOU J H, DUAN Y X, LI B Y, et al. Making adversarial examples more transferable and indistinguishable [C]// Proceedings of 2022 AAAI Conference on Artificial Intelligence. [S. l. : s. n. ], 2022: 3662-3670.

[30] ZHU H G, REN Y C, SUI X Y, et al. Boosting adversarial transferability via gradient relevance attack [C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 4741-4750.

## 作 者 简 介

**张兆阳**

男，1996 年生，博士，副研究员，研究方向为人工智能安全、数字信号处理

E-mail: zhaoyang_zhang@stu. hit. edu. cn

**孙芳慧**

女，1989 年生，博士，助理研究员，研究方向为网络空间安全与逆向分析

E-mail: sunfanghui@hit. edu. cn

**张明旭**

女，1986 年生，工程师，研究方向为数字信号处理与通信系统

E-mail: zhangmingxu@cie. org. cn

**宋　伟**

女，1985 年生，工程师，研究方向为数字特征提取及智能模型测试

E-mail: songwei@cmiot. chinamobile. com

**王振邦**

男，1981 年生，博士，高级工程师，研究方向为电力监控网络安全

E-mail: zhenbangw@163. com

**王英琦**

男，1997 年生，博士研究生，研究方向为人工智能安全、多媒体信号处理

E-mail: wangyqcbw@163. com

**张可卿**

女，2002 年生，硕士研究生，研究方向为人工智能安全、信号处理

E-mail: 1344548187@qq. com

**王　莘**

男，1980 年生，博士，教授，研究方向为人工智能安全、数字水印技术

E-mail: shen. wang@hit. edu. cn

**责任编辑**　殷文卓