

基于 Diff-AdvGAN 的图像对抗样本生成方法

赵 宏, 马宇轩, 宋馥荣

(兰州理工大学 计算机与通信学院, 兰州 730050)

摘 要: 针对基于生成对抗网络的对抗样本生成方法(AdvGAN)所存在的扰动偏离关键区域且可控性不强, 导致对抗样本攻击效果欠佳、真实性较低的问题, 提出了 Diff-AdvGAN 对抗样本生成方法。首先, 利用自适应空间特征融合模块(ASFF)融合图像的不同尺度特征图。其次, 将融合后的特征图输入生成器生成扰动, 并使用随机微分引导模块(SDGM)增强扰动的可控性, 生成对抗样本。最后, 将对抗样本输入判别器和目标模型, 迭代计算损失值并反馈至生成器, 以生成攻击性能较强的扰动。实验结果表明, Diff-AdvGAN 方法在 MNIST 数据集上对 LeNet C、VGG11 和 C&W model 模型的攻击成功率均大于 99%, 在 CIFAR-10 数据集上对 ResNet18 和 ResNet32 模型的攻击成功率分别为 96.17% 和 95.82%; 同时生成的扰动处于图像关键区域, 稀疏性高、幅度小, 均优于对比方法。

关键词: 对抗样本; 生成对抗网络; 扩散模型; 随机微分引导模块; 自适应空间特征融合

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2025)12-4052-11

DOI: 10.13229/j.cnki.jdxbgxb.20240555

Image adversarial examples generation based on Diff-AdvGAN

ZHAO Hong, MA Yu-xuan, SONG Fu-rong

(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: To address the problems of poor attack performance and low authenticity of adversarial examples caused by the perturbations generated by the adversarial example generation method based on generative adversarial networks (AdvGAN) deviating from key image regions and lacking controllability, a Diff-AdvGAN adversarial example generation method was proposed. Firstly, an Adaptively Spatial Feature Fusion (ASFF) module was employed to fuse feature maps of the images at different scales. Then, the fused feature maps were input into a generator to produce perturbations, and a Stochastic Differential Guide Module (SDGM) was used to enhance the controllability of the perturbations and generate adversarial examples. Finally, the adversarial examples are fed into a discriminator and a target model, the loss values were iteratively calculated and fed back to the generator to generate stronger perturbations with improved attack performance. Experimental results show that the Diff-AdvGAN method achieves attack success rates of over 99% on the MNIST dataset for the LeNet C, VGG11, and C&W models, and attack success rates of 96.17% and 95.82% for the ResNet18 and ResNet32 models on the CIFAR-10

收稿日期: 2024-05-20.

基金项目: 国家自然科学基金项目(62166025).

作者简介: 赵宏(1971-), 男, 教授, 博士. 研究方向: 并行与分布式处理、嵌入式系统、系统建模与仿真、深度学习、自然语言处理. E-mail: zhaoh@lut.edu.cn

dataset. Moreover, the perturbations generated by this method can accurately locate in the critical regions of the images, exhibiting high sparsity and small magnitudes, demonstrating significant advantages compared to comparison method.

Key words: adversarial examples; generative adversarial networks; diffusion models; stochastic differential guide module; adaptive spatial feature fusion

0 引 言

近年来,深度神经网络(Deep neural network, DNN)在计算机视觉领域取得了巨大成功,被广泛应用于图像分类^[1]、人脸识别^[2]、自动驾驶^[3]等场景。然而研究发现^[4],如果在图像中添加一定的扰动噪声生成对抗样本,DNN模型则会输出错误的结果,给实际应用DNN的系统带来极大安全隐患。例如在自动驾驶场景中,若攻击者对交通“停止”标志添加微小扰动^[5],自动驾驶系统可能会将其错误识别为“允许行驶”或其他标志,从而导致交通事故。因此,研究对抗样本的生成方法和过程,对检测和评价DNN模型的稳定性和安全性具有重大意义。

按照生成方式的不同,可将对抗样本生成方法分为基于梯度的生成方法、基于优化的生成方法和基于生成模型的生成方法。Goodfellow等^[6]提出了基于梯度的快速梯度符号法(Fast gradient sign method, FGSM),通过在梯度相反方向增加扰动,增大损失函数值,从而误导模型分类。Kurakin等^[7]在FGSM基础上提出了基础迭代方法(Basic iterative method, BIM),BIM能沿梯度增加的方向进行多次迭代,生成较为精准的扰动。基于梯度方法生成对抗样本的过程直观且速度较快,但有限的迭代次数不易构造有效扰动,导致攻击成功率较低。Carlini等^[8]提出基于优化的C&W方法,通过优化预定义的最佳目标函数生成微小的扰动。与C&W方法优化单一目标函数不同,Baluja等^[9]提出的对抗性变换网络(Adversarial transformation networks, ATNs)方法通过优化一组联合目标函数生成真实的对抗样本。相较于基于梯度的方法,基于优化方法生成的对抗样本真实性较高,但寻找目标函数的最佳超参数时间开销大,导致攻击效率较低。

随着对抗样本相关研究的持续深入,预训练的生成模型可快速、有效地生成扰动。因此,GAN被广泛应用于对抗样本的生成任务。Zhao

等^[10]基于WGAN^[11]提出了一种两阶段的对抗样本生成方法,首先在输入图像的低维隐特征空间中寻找扰动隐变量,其次在隐特征空间中搜索对抗样本,由隐变量生成较自然的对抗样本。Zhang^[12]提出了一种图像到图像转换结构的对抗样本生成方法,无需输出中间扰动向量,直接由生成器将原始图像映射为对抗样本,由于扰动与原始图像的潜在边缘和形状更加匹配,因此生成的对抗样本更为自然。Xiao等^[13]提出一种基于GAN的对抗样本生成方法AdvGAN,通过生成器将原始图像中的噪声转化为扰动,再将扰动与原始图像叠加生成对抗样本,利用判别器区分对抗样本和原始图像,以保证生成的对抗样本与原始图像更接近。同时通过目标模型对对抗样本分类,计算损失值并反馈至生成器,使生成器生成更具攻击性的扰动^[14,15]。之后,Jandial等^[16]在AdvGAN基础上提出了AdvGAN++,该方法将输入由原始图像更换为更易受扰动影响的隐层特征向量,提升了攻击成功率,同时改变了生成器的编码-解码结构,减少了训练时间。扩散模型因其强大的图像生成能力^[17,18]也被运用到对抗样本的生成中,其主要思路是不改变图像的像素空间,对扩散模型压缩后的潜在映射进行扰动。Dai等^[19]提出了AdvDiff方法,在扩散模型逆向生成过程中进行稳定采样,以生成不受限制的对抗样本。Chen等^[20]通过对预测图像进行扰动,将其潜在代码引导到分类器的对抗样本空间,以此生成不受限制的对抗样本。Liu等^[21]提出了Adv-Diffusion框架,在潜在空间而非原始像素空间中生成难以察觉的扰动,从而成功攻击人脸识别模型。

基于GAN的方法在模型训练完成后,无需获取目标模型参数,通过生成器可快速批量地生成对抗样本,攻击效率高、实用性较强。但该方法仍然存在部分问题,如随着网络层数的加深,当AdvGAN获得图像的深层全局特征时,会丢失部分浅层细节特征,导致生成的扰动偏离图像关键区域,影响攻击效果。同时,AdvGAN生成的扰

动从图像噪声中采样得到,可控性不强,导致生成的对抗样本分布偏离原始图像,降低了对抗样本的真实性。

针对以上问题,本文基于特征融合和噪声扩散原理,提出了一种改进的图像对抗样本生成方法 Diff-AdvGAN (Diffusion-AdvGAN),以克服 AdvGAN 的缺陷。引入自适应空间特征融合模块 (Adaptively spatial feature fusion, ASFF) 融合深层和浅层特征图,生成的特征图包含全局特征和细节信息,使生成器在图像关键区域生成扰动,提升对抗样本的攻击效果。同时,利用随机微分引导模块 (Stochastic differential guide module, SDGM) 优化扰动,增强扰动可控性,生成真实性高的对抗样本。

1 Diff-AdvGAN 模型设计

1.1 AdvGAN 模型结构

AdvGAN^[13]模型结构如图 1 所示,由生成器、判别器、目标模型等模块构成。对抗样本生成过程为:将原始图像输入生成器生成扰动;然后,将扰动与原始图像叠加形成对抗样本;接着,将对抗样本输入判别器和目标模型,通过二者的损失函数计算损失值 L_{GAN} 和误判损失值 L_{adv} ,并将其反馈至生成器,使生成器生成更高质量的扰动。反复迭代以上过程,即可生成质量较高的对抗样本。

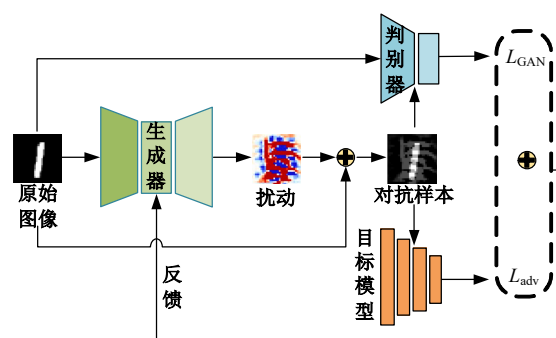


图 1 AdvGAN 模型结构

Fig. 1 AdvGAN model structure

1.2 Diff-AdvGAN 模型结构

本文提出的改进对抗样本生成方法 Diff-AdvGAN 模型结构如图 2 所示。Diff-AdvGAN 主要包含 ASFF、生成器 G 、SDGM、判别器 D 和目标模型 f 。ASFF 负责对不同尺度特征图进行自适应融合;生成器 G 负责生成扰动;SDGM 负责优化扰动并生成对抗样本;判别器 D 区分对抗样本和原始图像并计算对抗损失;目标模型 f 对对抗样本分类并计算误判损失,将对抗损失和误判损失反馈至生成器,使生成器生成与原始图像难以区分且具攻击性的对抗样本。Diff-AdvGAN 模型训练过程如算法 1 所示。

1.3 自适应空间特征融合模块 ASFF

特征融合旨在利用不同层级的特征图信息,提高模型对输入图像的处理能力。当 AdvGAN 模型通过编码器获得输入图像的深层特征图时,

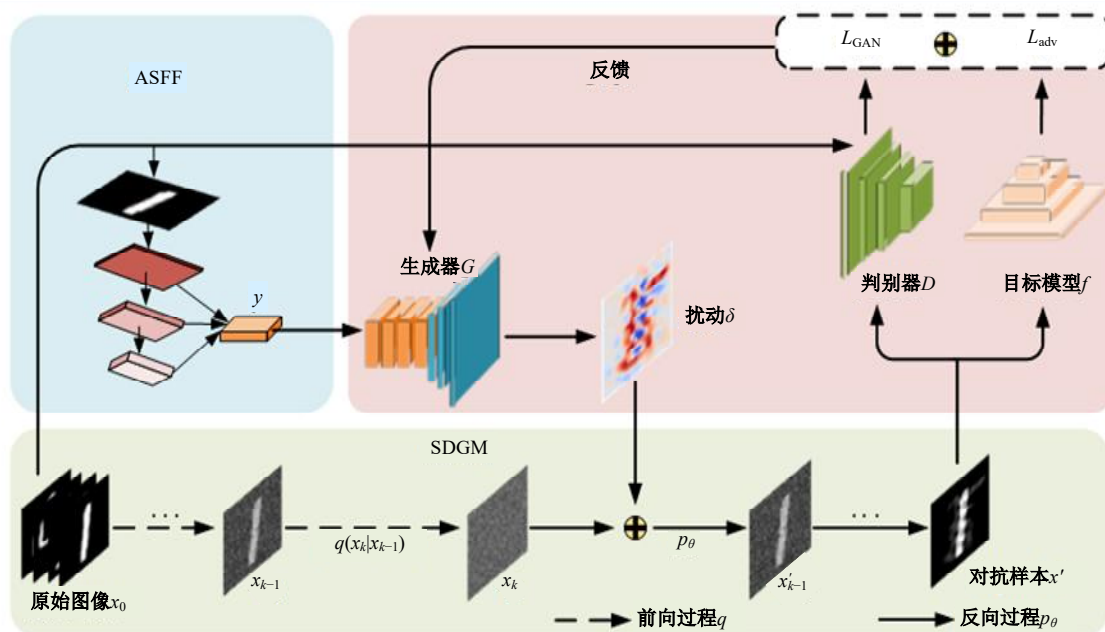


图 2 Diff-AdvGAN 模型结构

Fig. 2 Diff-AdvGAN model structure

算法 1 Diff-AdvGAN 模型训练过程

输入:原始图像 x_0

输出:对抗样本 x'

1. 将原始图像 x_0 输入 ASFF 中,生成包含全局特征和细节的融合特征图 y
2. 将 y 输入生成器 G 中生成扰动 δ
3. 将 δ 输入 SDGM 模块中进行优化,并生成对抗样本 x'
4. 将 x_0 和 x' 输入判别器 D 中判别,计算对抗损失 L_{GAN}
5. 将 x' 输入目标模型 f 中分类,计算误判损失 L_{adv}
6. 损失函数反向传播,更新生成器参数
7. 重复以上步骤,直至模型收敛

会丢失浅层特征图包含的细节信息。因此,在获取包含全局信息的深层特征过程中,需通过特征融合获取浅层特征的细节信息。ASFF^[22]解决了不同特征尺度不一致的问题,可自适应地学习每个尺度上特征映射的融合空间权重,且推理开销非常小。这使得 ASFF 可针对不同图像按照合适的权重大小灵活融合,而非简单地相加或级联特征图。

Diff-AdvGAN 模型在原始图像输入生成器前,引入自适应空间特征融合模块对不同尺度的特征图进行自适应融合,获取图像的全局特征和细节信息,使生成器更好地捕获图像的多尺度信息,从而在图像关键区域生成扰动,提升对抗样本的攻击效果。特征融合过程如图 3 所示。

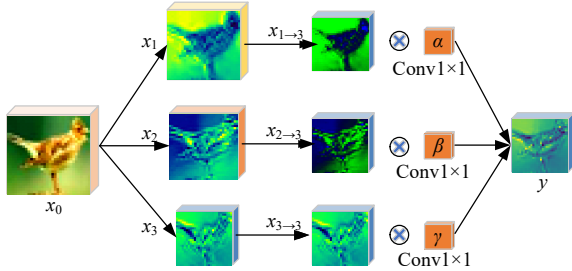


图 3 特征融合过程

Fig. 3 Feature fusion process

首先,原始图像 x_0 经卷积层的下采样得到 3 个待融合的特征图 x_1, x_2, x_3 。其次,以 x_3 分辨率特征为标准进行融合,改变 x_1 和 x_2 的通道数和分辨率。对 x_1 先执行步长为 2 的最大池化操作,再进行 3×3 卷积操作;对 x_2 使用一个 3×3 的卷积层(步长为 2)进行下采样,分别生成与 x_3 相同通道数和分辨率的特征图,记作 $x_{1 \rightarrow 3}, x_{2 \rightarrow 3}$ 。最后,对 $x_{1 \rightarrow 3}, x_{2 \rightarrow 3}, x_{3 \rightarrow 3}$ 特征图分别进行 1×1 卷积得到对应的权重参数 α, β, γ ,将缩放后的特征图 $x_{1 \rightarrow 3}, x_{2 \rightarrow 3}, x_{3 \rightarrow 3}$ 与各自权重相乘后再相加,得到

包含全局特征和细节信息的融合特征图 y 。计算公式如下所示:

$$y = \alpha \cdot x_{1 \rightarrow 3} + \beta \cdot x_{2 \rightarrow 3} + \gamma \cdot x_{3 \rightarrow 3} \quad (1)$$

式中: α, β, γ 分别为不同尺度特征图的空间重要性权重, $\alpha, \beta, \gamma \in [0, 1]$ 且 $\alpha + \beta + \gamma = 1$ 。

1.4 生成器模块

生成器输入融合特征图后生成扰动,如图 4 所示,其包含 4 个残差块和 3 个反卷积层。残差块用于加强深层网络训练的稳定性,防止模型性能退化,增强特征表达能力。反卷积层将残差块输出的特征图转换为扰动,前两个反卷积层执行反卷积操作后先使用实例归一化加速模型收敛,再使用 ReLU 激活函数缓解梯度消失问题,最后一层反卷积层使用 Tanh 激活函数。

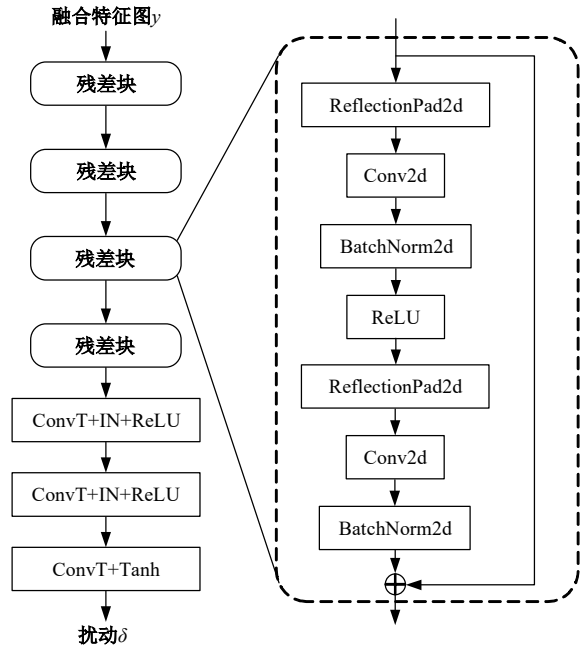


图 4 生成器结构

Fig. 4 Generator structure

1.5 随机微分引导模块 SDGM

引入扩散模型(Diffusion models, DMs)^[23]构造 SDGM, SDGM 引导扰动优化,增强扰动可控性,使生成的对抗样本与原始图像相似, SDGM 分为前向过程 q 和反向过程 p_θ 。定义原始图像数据分布服从 $x_0 \sim q(x_0)$, 其中 0 表示初始状态。将前向过程分为离散的时间步 T , 在前向过程中逐步将高斯噪声加入 x_0 , T 步生成噪声样本 $[x_1, x_2, \dots, x_t, \dots, x_T]$, 其遵循定义为 $q(x_t|x_{t-1}) = N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$ 的马尔可夫链。通过累

加单步前向过程,得到前向过程 q 的定义为:

$$q(x_t|x_0) = N\left(x_t; \sqrt{\bar{\alpha}_t} x_{t-1}, (1 - \bar{\alpha}_t)I\right) \quad (2)$$

式中: N 为高斯分布; β_t 为位于区间 $(0,1)$ 的固定值,其随着 t 的增加而增加; $\alpha_t = 1 - \beta_t$ 且 $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ 。噪声方差由 β_t 确定,均值由 β_t 和当前带噪的数据分布确定。

反向过程旨在通过去噪函数从高斯噪声数据分布 $x_T \sim N(0, I)$ 中恢复原始图像数据。利用重采样策略,可将训练修改的U-Net网络作为去噪器,将去噪过程视为参数 θ 的优化过程。反向过程 p_θ 定义为:

$$p(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) \quad (3)$$

$$p_\theta(x_{t-1}|x_t) \sim N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (4)$$

式中: θ 为可学习的参数; $\mu_\theta(x_t, t)$ 和 $\sigma_t^2 I$ 分别为反向过程的均值和方差。

为后续方便使用,定义噪声 $x_t \sim x_{t-1}$ 的去噪过程为 $x_{t-1} = P_\theta(x_t, t)$,其中 P_θ 为去噪函数。

SDGM模块优化扰动,生成对抗样本的过程为:前向过程运行 K 步得到原始图像 x_0 扩散后的分布 $x_K \sim q(x_K|x_0)$ 。接着,将生成器生成的扰动 δ 输入SDGM,并与前向过程产生的 x_K 相结合,即 $x_K + \delta$ 。最后,将 $x_K + \delta$ 作为输入,运行反向去噪过程 $x' = P_\theta(\dots(P_\theta(P_\theta(x_K + \delta, K), K-1)\dots, 0))$,优化扰动 δ 并生成对抗样本 x' 。由于推理时间和内存的限制,本文采用加速采样策略的去噪扩散隐式模型(Denoising diffusion implicit models, DDIM)^[24]。先将时间步 $T=1000$ 子采样得到 $T_d=40$,取 $K=0.2T=200$,子序列 $K_d=8$ 。接着,使用缩放后的 $K_d \ll K$ 运行SDGM模块优化扰动,生成真实性高的对抗样本。

1.6 判别器模块

判别器的功能区分输入为原始图像还是对抗样本,结构如图5所示,包含4个卷积层用于提取图像特征。其中,卷积核大小为 3×3 ,步长为1。每层执行卷积操作后,使用批归一化加速训练增强模型的泛化能力,使用LeakyReLU激活函数对输出特征进行非线性变换。最后一层将图像映射为单一标量值,通过Sigmoid函数输出对抗样本的置信度,Sigmoid函数输出值越接近1,表示输入越接近原始图像;输出值越接近0,表示输入越

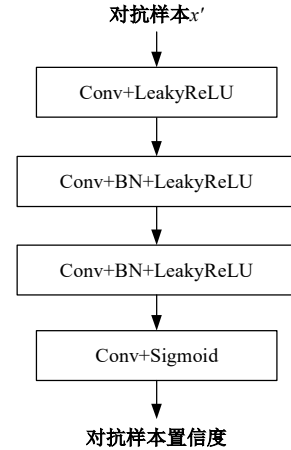


图5 判别器结构

Fig. 5 Discriminator structure

接近对抗样本。

1.7 损失函数

Diff-AdvGAN模型训练中的损失函数包括:对抗性损失 L_{GAN} 、目标模型 f 的误判损失 L_{adv} 和限制扰动大小的软铰链损失 L_{hinge} 。对抗损失 L_{GAN} 指导生成器生成的样本难以被判别器区分。 L_{GAN} 的计算公式如下所示:

$$L_{GAN} = E_{x_0} \log D(x_0) + E_{x_0} \log(1 - D(x')) \quad (5)$$

式中: $D(\cdot)$ 为判别器的输出; x_0 为原始图像; x' 为对抗样本。

根据目标模型 f 的输出计算误判损失 L_{adv} , L_{adv} 指导生成器生成攻击性强的对抗样本,鼓励对抗样本的预测概率分布与原始图像的分布存在显著差异,使目标模型产生错误预测。误判损失 L_{adv} 计算公式如下所示:

$$L_{adv} = E_{x_0} l_f(x') \quad (6)$$

式中: l_f 为训练目标模型 f 的损失函数。

为限制扰动的大小,保持对抗样本的真实性和自然性,在训练过程中使用基于 l_2 范数的软铰链损失稳定生成模型的训练。 L_{hinge} 计算如下所示:

$$L_{hinge} = E_{x_0} \max(0, \|\delta\|_2 - c) \quad (7)$$

式中:常数 c 为优化间距; δ 为扰动。

因此,Diff-AdvGAN模型在训练过程中的损失函数 L 如下所示:

$$L = L_{adv} + \alpha' L_{GAN} + \beta' L_{hinge} \quad (8)$$

式中: α' 、 β' 为超参数,用于调节对抗损失和软铰链损失的相对权重,以便更好地平衡生成器的目标。

2 实验设计和结果分析

2.1 数据集

为评估 Diff-AdvGAN 模型生成对抗样本的性能,选择 MNIST、CIFAR-10、ImageNet-1K 数据集进行实验。其中,MNIST 数据集共有 7 万张灰度手写数字图片,包含训练集 6 万张、测试集 1 万张,每张图片大小为 28×28 。CIFAR-10 数据集包含 10 个类别的 RGB 彩色图片,其中训练集 5 万张、测试集 1 万张,每张图片大小均为 32×32 。ImageNet-1K 数据集为 ImageNet 的子集,包含 120 多万个训练样本、5 万个验证样本,涵盖 1 000 个类别。

2.2 评价指标

实验采用分类准确率(Accuracy, ACC)、攻击成功率(Attack success rate, ASR)、 l_0 、 l_1 和 l_2 范数衡量模型性能。其中,ACC 用于衡量目标模型对原始图像的分类能力,分类准确率越高,说明目标模型的分类能力越强,能更好地评估攻击模型的性能。假设共有 N 个原始图像,其中目标模型成功分类 m 个,则分类准确率 ACC 如下所示:

$$ACC = \frac{m}{N} \times 100\% \quad (9)$$

ASR 用于衡量攻击模型对目标模型的误导能力,攻击成功率越高,说明攻击模型对目标模型的攻击能力越强。假设攻击模型共生成 N_{adv} 个对抗样本,其中 n 个对抗样本成功误导目标模型,则攻击成功率 ASR 为:

$$ASR = \frac{n}{N_{adv}} \times 100\% \quad (10)$$

l_0 范数指对抗样本相对原始图像修改的像素个数,其限制了可更改的像素数量,但不限制每个像素更改的程度。 l_0 范数的值越小,说明图像中改动的像素点数量越少,扰动的稀疏性越高。

l_1 范数指对抗样本与原始图像各个元素的绝对差值之和。 l_1 范数的值越大,说明对抗样本与原始图像的差异越明显。

l_2 范数指对抗样本相对原始图像像素变化量的平方和的平方根,其限制了累积的更改,在更改数量与程度之间达到一种平衡状态。 l_2 范数的值越小,说明添加在图像中的扰动幅度越小。

2.3 实验设置

实验环境为 Linux 操作系统,系统版本为 Ubuntu 18.04.6 LTS 64 位,处理器为 Intel(R)

Xeon(R) Gold 5218 CPU @ 2.30 GHz 24-core processor 48,内存 176 GB,所用 GPU 为 NVIDIA A100。

在训练目标模型的过程中,使用 Adam 作为优化器,epoch 设为 120, batchsize 设为 128,初始学习率设为 0.001,并在第 50 个和第 80 个 epoch 后将学习率降为之前的 10%。

在训练攻击模型 Diff-AdvGAN 的过程中, batchsize 设为 64,使用初始学习率为 0.001 的 Adam 训练 100 个 epoch,并在第 50 个和第 80 个 epoch 后将学习率衰减 10 倍,扰动阈值设为 0.3。

2.4 实验结果分析

在无目标攻击的情况下,各方法在 MNIST 测试集上攻击 LeNet C 模型生成 500 张对抗样本的所需时间如表 1 所示,计算时忽略模型和数据集加载时间,仅记录对抗样本生成过程所需的时间。AdvGAN 方法生成速度最快,不足 0.01 s 即可生成对抗样本。扩散模型的推理速度使得 AdvDiff 的生成时间较长。C&W 方法最慢,生成对抗样本所需时间高达 3 h 左右。其中,Diff-AdvGAN 方法慢于 AdvGAN 方法,但远快于 AdvDiff 和 C&W 方法,这是因为 Diff-AdvGAN 方法在生成对抗样本时增加了特征融合和优化扰动操作,导致模型复杂度比 AdvGAN 方法高,所以生成时间明显增加。

表 1 各方法生成对抗样本所需时间

Table 1 Time required for each algorithm to generate adversarial examples

方法	FGSM	C&W	AdvGAN	AdvDiff	Diff-AdvGAN
时间	0.06 s	>3 h	<0.01 s	>1 min	0.5 s

2.4.1 无防御攻击评估

在未设置防御的情况下,评估各类生成算法在 MNIST 和 CIFAR-10 数据集上生成对抗样本的攻击成功率。在 MNIST 数据集上攻击 LeNet C、VGG11、C&Wmodel (C&W 方法中使用的 MNIST 分类模型)时,Diff-AdvGAN 方法攻击成功率均大于 99%。在 CIFAR-10 数据集上攻击 ResNet18 和 ResNet32 模型时,Diff-AdvGAN 方法攻击成功率均大于 95%。如表 2 所示,相较于其他攻击方法,Diff-AdvGAN 方法对各目标模型的攻击都取得了最优结果。其中,FGSM 对所有目标模型的攻击成功率最低,如攻击 VGG11 模型时,准确率仅为 56.61%,攻击成功率远低于其

表 2 各方法在 MNIST 和 CIFAR-10 上的攻击成功率

Table 2 Attack success rate of each method on MNIST and CIFAR-10

数据集	目标模型	分类	攻击成功率/%						
		准确率/%	FGSM	DeepFool	C&W	AdvGAN	AdvGAN++	AdvDiff	Diff-AdvGAN
MNIST	LeNet C	99.23	94.04	95.16	96.67	97.90	98.40	94.20	99.48
	VGG11	99.4	56.61	86.04	87.50	99.65	99.58	93.85	99.69
	C&Wmodel	99.47	82.30	84.78	85.72	98.30	98.97	92.70	99.43
CIFAR-10	ResNet18	89	90.69	93.22	93.75	94.52	96.01	91.14	96.17
	ResNet32	87.5	90.06	91.67	90.30	94.70	95.29	90.29	95.82

他方法。AdvGAN++将 AdvGAN 的输入更换为隐层特征向量,使攻击成功率有所提升。Diff-AdvGAN 方法相比 AdvGAN、AdvGAN++、AdvDiff 方法有更高的攻击成功率。具体分析,相比 AdvGAN 和 AdvDiff 方法,Diff-AdvGAN 在 MNIST 数据集上攻击 LeNet C 模型时,攻击成功率分别提升了 1.58% 和 5.28%,在 CIFAR-10 数据集上攻击 ResNet18 模型时,攻击成功率分别提升了 1.65% 和 5.03%。以上结果说明,Diff-AdvGAN 方法生成的对抗样本对目标模型的误导更有效。

同时,实验统计了各攻击方法生成对抗样本的范数值,以评估扰动的变化程度。分别以在 MNIST 数据集上攻击 C&Wmodel 模型生成对抗

样本和在 CIFAR-10 数据集上攻击 ResNet32 模型生成对抗样本为例。如表 3 所示,所有方法中,FGSM 表现最差。Diff-AdvGAN 方法在 l_0 、 l_1 、 l_2 范数上的表现几乎全优于 AdvGAN、AdvGAN++ 和 Zhang^[12] 方法,与 C&W 方法均获得两项最优的范数值,C&W 方法生成的对抗样本真实性高,范数值很小。其中,AdvDiff 方法在 l_2 范数上的表现优于 Diff-AdvGAN,由于 AdvDiff 可在扩散逆生成过程中进行稳定采样,因此扰动幅度较小。Diff-AdvGAN 方法在 MNIST 数据集上的 l_0 范数仅为 148.68,在 CIFAR-10 数据集上 l_1 范数仅为 11.6,远远低于其他方法,这表明 Diff-AdvGAN 方法生成的扰动稀疏性高,构造了较真实的对抗样本。

表 3 不同方法生成对抗样本与原始图像之间的范数平均值

Table 3 Norm average between different methods generate adversarial examples and original images

范数	数据集	FGSM	DeepFool	C&W	AdvGAN	AdvGAN++	AdvDiff	Zhang	Diff-AdvGAN
l_0	MNIST	781.31	462.39	783.79	456.44	389.10	150.70	169.34	148.68
	CIFAR-10	3 053.40	3 051.70	2 952.00	3 051.50	3 049.35	2 989.31	3 052.80	3 021.10
l_1	MNIST	372.36	32.87	39.47	82.96	56.71	42.17	20.92	22.58
	CIFAR-10	521.30	24.85	56.50	222.86	129.87	64.35	142.32	11.60
l_2	MNIST	13.61	2.27	1.89	4.03	3.31	2.08	2.72	2.85
	CIFAR-10	9.58	0.65	1.43	4.30	4.14	1.86	4.66	3.60

2.4.2 有防御攻击评估

防御情况下的攻击指对目标模型采取防御手段,若对抗样本仍能成功误导目标模型,则证明对抗样本在防御情况下的攻击依然有效。对抗训练是常用的防御策略,指将攻击模型生成的对抗样本和原始图像同时作为训练集,对目标模型进行对抗训练,以提高目标模型的鲁棒性。使用 3 种对抗训练方法进行防御:标准 FGSM 对抗训练(Adv)^[6]、集成对抗训练(Ens)^[25]、迭代训练(Iter-Adv)^[26],将原始图像与对抗样本按照 1:1 组成新的数据集对目标模型进行对抗训练,以验证 Diff-AdvGAN 方法在防御情况下的有效性。

图 6 为 FGSM、AdvGAN、AdvGAN++、

Diff-AdvGAN 方法在不同对抗训练策略下的攻击成功率。从图 6 中可以看出,当攻击执行防御策略的目标模型时,Diff-AdvGAN 方法取得的攻击成功率最高。如图 6(c) 所示,防御情况下在 CIFAR-10 数据集上攻击 Wide-ResNet34 模型效果最为显著,Diff-AdvGAN 相比 AdvGAN++ 方法,面对 Adv 防御时攻击成功率提升了 33.18%,面对 Ens 防御时成功率提升了 29.73%,面对 Iter-Adv 防御时成功率提升了 21.5%。其他情况下,Diff-AdvGAN 方法的攻击效果也均优于 FGSM、AdvGAN 方法,以上结果充分证明了 Diff-AdvGAN 方法在防御条件下的攻击有效性。

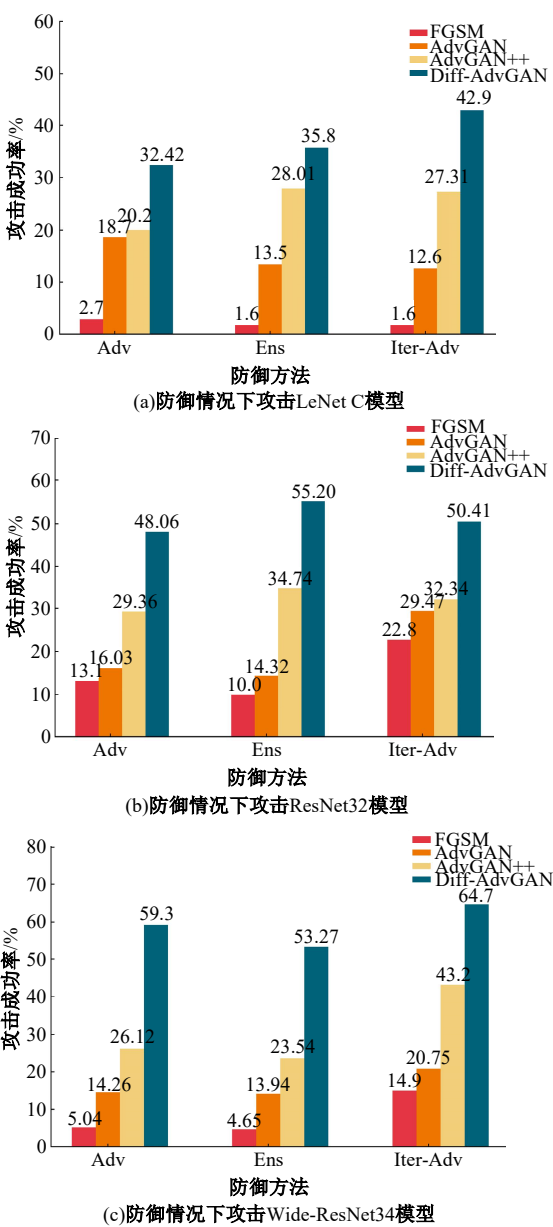


图 6 防御下各方法攻击成功率

Fig. 6 Attack success rate of each method under defense

2. 4. 3 消融实验

为了进一步验证 Diff-AdvGAN 方法中关键模块的作用,在 MNIST 数据集上针对 LeNet C 模型、在 CIFAR-10 数据集上针对 ResNet32 模型进行消融实验。在 AdvGAN 基准模型基础上,将只采用自适应空间特征融合模块的实验命名为 +ASFF,将只采用随机微分引导模块的实验命名为 +SDGM。

如表 4 所示,Diff-AdvGAN 在各评价指标上均取得最优结果,说明 Diff-AdvGAN 方法的特征融合策略和优化扰动策略可提升对抗样本的攻击

成功率,提高扰动稀疏性,降低扰动幅度。具体分析,+ASFF、+SDGM 相比 AdvGAN 方法,在 MNIST 数据集上的攻击成功率分别提升了 1.2% 和 0.66%,在 CIFAR-10 数据集上分别提升了 0.74% 和 0.31%,说明结合全局特征和细节信息的特征融合策略能使生成器在图像关键区域生成扰动,生成的对抗样本更易误导目标模型。同时,+SDGM 相比 +ASFF, l_0 、 l_1 、 l_2 范数均下降,说明通过 SDGM 模块优化扰动,增强了扰动的可控性,生成的扰动稀疏性高、整体幅度小,提高了对抗样本的真实性。

表 4 在 MNIST 和 CIFAR-10 数据集消融实验

Table 4 Ablation experiments on MNIST and CIFAR-10 datasets

数据集	方法	ASR/%	l_0	l_1	l_2
MNIST	AdvGAN	97.9	456.440	82.96	4.03
	+ ASFF	99.1	212.3	37.5	3.83
	+ SDGM	98.56	188.13	28.74	3.08
	Diff-AdvGAN	99.48	148.68	22.58	2.85
CIFAR-10	AdvGAN	94.7	3 051.5	222.86	4.30
	+ ASFF	95.44	3 025.6	33.97	4.15
	+ SDGM	95.01	3 029.48	20.68	3.81
	Diff-AdvGAN	95.82	3 021.1	11.6	3.60

2. 4. 4 对抗样本可视化

图 7 和图 8 分别展示了 MNIST 和 CIFAR-10 数据集上生成的对抗样本示例。从图 7(b)和图 8(b)中可以看出,AdvGAN 方法生成的扰动偏离图像关键区域,条状噪声较多,扰动幅度过大。Diff-AdvGAN 方法融合了图像全局特征和细节信息,使生成器在图像关键区域生成扰动,且经 SDGM 模块优化,增强了扰动可控性。如图 7(c)和图 8(c)所示,扰动大多分布在图像关键区域,稀疏性高,幅度小且生成的对抗样本真实性高。

为验证 Diff-AdvGAN 方法对高分辨率图像的攻击性能,在 ImageNet 数据集上攻击了预训练的 VGG16_BN(top-1 错误率为 23.7%,top-5 错误率为 6.8%)模型,将 ImageNet 验证集作为测试集生成了 5 万张对抗样本。经过多次测试,Diff-AdvGAN 方法的攻击成功率均大于 99%。如图 9 所示,生成的扰动在决定图像类别的关键区域,扰动稀疏性高、幅度小,生成的对抗样本人眼不易区分。

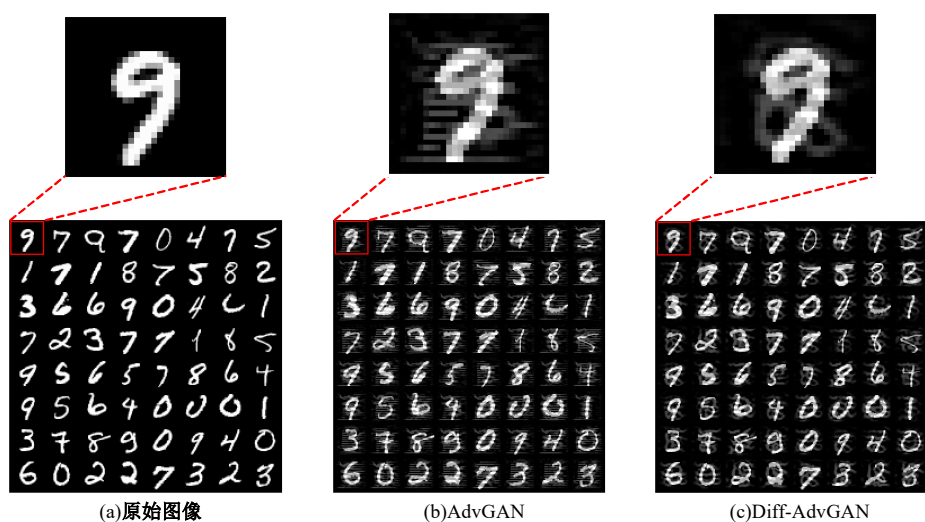


图 7 MNIST 数据集上攻击 C&Wmodel 生成对抗样本

Fig. 7 Attack C&Wmodel to generate adversarial examples on the MNIST dataset

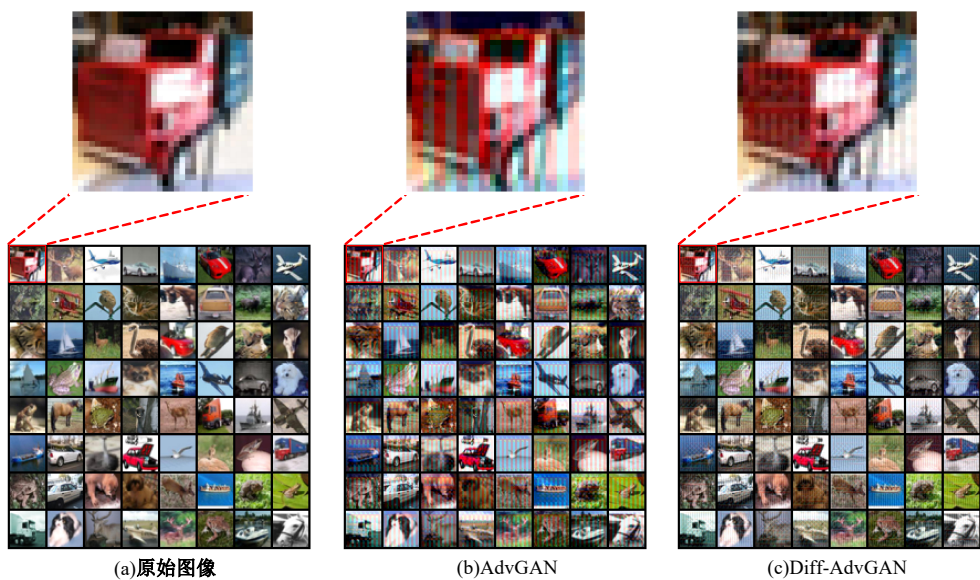


图 8 CIFAR-10 数据集上攻击 ResNet32 生成对抗样本

Fig. 8 Attack ResNet32 to generate adversarial examples on the CIFAR-10 dataset

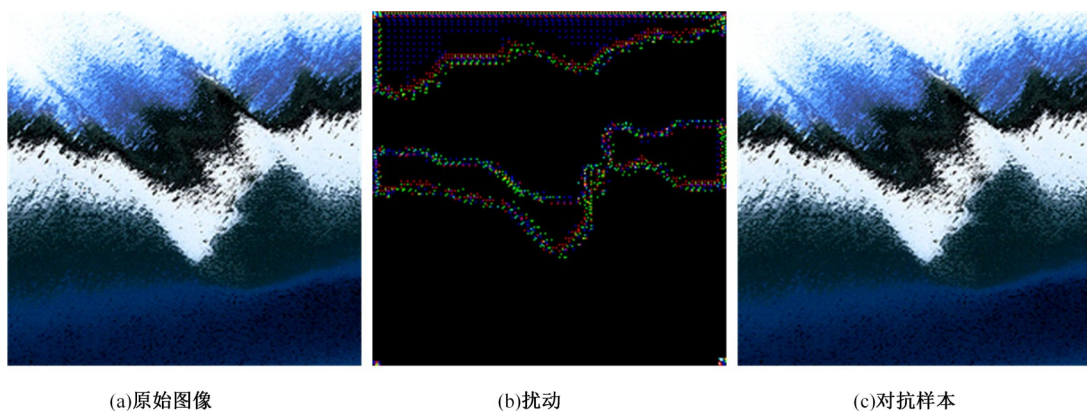


图 9 Imagenet 数据集上攻击 VGG16_BN 模型生成对抗样本

Fig. 9 Attack VGG16_BN model to generate adversarial examples on the Imagenet dataset

3 结束语

本文提出了一种改进的图像对抗样本生成方法 Diff-AdvGAN,通过自适应空间特征融合模块获取图像的全局特征和细节信息,使生成器在图像关键区域生成扰动,提升了攻击效果。同时,构造了一个随机微分引导模块,利用其反向过程增强扰动的可控性,提升了扰动的稀疏性并降低了幅度,生成了真实性高的对抗样本。实验结果表明,Diff-AdvGAN 方法生成的关键扰动可有效误导目标模型,且稀疏性高、幅度小,生成的对抗样本更加真实。

参考文献:

- [1] Kim I, Baek W, Kim S. Spatially attentive output layer for image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, 2020: 9533-9542.
- [2] Boutros F, Damer N, Kirchbuchner F, et al. Elastic-face: Elastic margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 1578-1587.
- [3] Hausler S, Garg S, Xu M, et al. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, 2021: 14141-14152.
- [4] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations, Banff, AB, Canada, 2014: 23-41.
- [5] 吉顺慧, 胡黎明, 张鹏程, 等. 基于稀疏扰动的对抗样本生成方法[J]. 软件学报, 2023, 34(9): 4003-4017.
Ji Shun-hui, Hu Li-ming, Zhang Peng-cheng, et al. Adversarial example generation method based on sparse perturbation[J]. Journal of Software, 2023, 34(9): 4003-4017.
- [6] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J/OL]. [2025-05-01]. https://www.researchgate.net/publication/269935591_Explaining_and_Harnessing_Adversarial_Examples
- [7] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[C]//5th International Conference on Learning Representations, Toulon, France, 2017: 1-14.
- [8] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 2017: 39-57.
- [9] Baluja S, Fischer I. Adversarial transformation networks: learning to generate adversarial examples[J/OL]. [2025-05-01]. <https://doi.org/10.48550/arXiv.1703.09387>
- [10] Zhao Z L, Dua D, Singh S. Generating natural adversarial examples[C]//6th International Conference on Learning Representations, Vancouver, BC, Canada, 2018: 1-15.
- [11] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 2017: 214-223.
- [12] Zhang W J. Generating adversarial examples in one shot with image-to-image translation GAN[J]. IEEE Access, 2019, 7: 151103-151119.
- [13] Xiao C W, Li B, Zhu J Y, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 3905-3911.
- [14] 黄帅娜, 李玉祥, 毛岳恒, 等. 基于集成 advGAN 的黑盒迁移对抗攻击[J]. 吉林大学学报: 工学版, 2022, 52(10): 2391-2398.
Huang Shuai-na, Li Yu-xiang, Mao Yue-heng, et al. Black-box transferable adversarial attacks based on ensemble advGAN[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(10): 2391-2398.
- [15] 刘悦文, 孙子文. 抵御对抗攻击的生成对抗网络 IWSN 入侵检测模型[J]. 江苏大学学报: 自然科学版, 2025, 46(5): 562-569.
Liu Yue-wen, Sun Zi-wen. Intrusion detection model based on generative adversarial networks in IWSN against adversarial attacks[J]. Journal of Jiangsu University (Natural Science Edition), 2025, 46(5): 562-569.
- [16] Jandial S, Mangla P, Varshney S, et al. AdvGAN++: harnessing latent layers for adversary generation[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop, Los Alamitos, CA, USA, 2019: 2045-2048.
- [17] Singh J, Gould S, Zheng L. High-fidelity guided im-

- age synthesis with latent diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023: 5997-6006.
- [18] Podell D, English Z, Lacey K, et al. SDXL: improving latent diffusion models for high-resolution image synthesis[J/OL]. [2024-05-03]. <https://doi.org/10.48550/arXiv.2307.01952>
- [19] Dai X L, Liang K S, Xiao B. AdvDiff: generating unrestricted adversarial examples using diffusion models[J/OL]. [2024-05-03]. <https://doi.org/10.48550/arXiv.2307.12499>
- [20] Chen X Q, Gao X T, Zhao J J, et al. Advdiffuser: Natural adversarial example synthesis with diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 4562-4572.
- [21] Liu D C, Wang X J, Peng C L, et al. Adv-diffusion: Imperceptible adversarial face identity attack via latent diffusion model[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(4): 3585-3593.
- [22] Liu S T, Huang D, Wang Y H. Learning spatial fusion for single-shot object detection[J/OL]. [2024-05-03]. <https://doi.org/10.48550/arXiv.1911.09516>
- [23] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
- [24] Song J M, Meng C L, Ermon S. Denoising diffusion implicit models[J/OL]. [2024-05-03]. <https://doi.org/10.48550/arXiv.2010.02502>
- [25] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: attacks and defenses[C]//6th International Conference on Learning Representations, Vancouver, BC, Canada, 2018: 1-20.
- [26] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//6th International Conference on Learning Representations, Vancouver, BC, Canada, 2018: 1-23.