基于引导扩散模型的自然对抗补丁生成方法

何 琨[1,2],佘计思[1,2],张子君[1,2*],陈 晶[1,2,3],汪欣欣[1,2],杜瑞颖[1,2,4]

(1. 武汉大学国家网络安全学院,湖北武汉 430072;2. 武汉大学空天信息安全与可信计算教育部重点实验室,湖北武汉 430072;
3. 武汉大学日照信息技术研究院,山东日照 276800;4. 地球空间信息技术协同创新中心,湖北武汉 430079)

摘 要: 近年来,物理世界中的对抗补丁攻击因其对深度学习模型安全的影响而引起了广泛关注. 现有的工作主要集中在生成在物理世界中攻击性能良好的对抗补丁,没有考虑到对抗补丁图案与自然图像的差别,因此生成的对抗补丁往往不自然且容易被观察者发现. 为了解决这个问题,本文提出了一种基于引导的扩散模型的自然对抗补丁生成方法. 具体而言,本文通过解析目标检测器的输出构建预测对抗补丁攻击成功率的预测器,利用该预测器的梯度作为条件引导预训练的扩散模型的逆扩散过程,从而生成自然度更高且保持高攻击成功率的对抗补丁. 本文在数字世界和物理世界中进行了广泛的实验,评估了对抗补丁针对各种目标检测模型的攻击效果以及对抗补丁的自然度. 实验结果表明,通过将所构建的攻击成功率预测器与扩散模型相结合,本文的方法能够生成比现有方案更自然的对抗补丁,同时保持攻击性能.

关键词: 目标检测;对抗补丁;扩散模型;对抗样本;对抗攻击;深度学习

基金项目: 国家重点研发计划项目(No. 2022YFB3102100);中央高校基本科研业务费专项资金(No.2042022kf1034);国家自然科学基金(No.62206203,No.62076187);湖北省重点研发计划项目(No.2022BAA039);山东省重点研发计划项目(No.2022CXPT055)

中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2024)02-0564-10

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230481

A Guided Diffusion-based Approach to Natural Adversarial Patch Generation

HE Kun[1,2], SHE Ji-si[1,2], ZHANG Zi-jun[1,2*], CHEN Jing[1,2,3], WANG Xin-xin[1,2], DU Rui-ying[1,2,4]

(1. *School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan*, *Hubei* 430072, *China*;
2. *Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education*, *Wuhan University*, *Wuhan*, *Hubei* 430072, *China*;
3. *Rizhao Institute of Information Technology*, *Wuhan University*, *Rizhao*, *Shandong* 276800, *China*;
4. *Collaborative Innovation Center of Geospatial Technology*, *Wuhan*, *Hubei* 430079, *China*)

Abstract: Adversarial patch attacks in the physical world have gained a lot of attention in recent years due to their safety implications. Existing work has mostly focused on generating adversarial patches that can attack certain models in the physical world, but the resulting patterns are often unnatural and easy to identify. To tackle this problem, we propose a guided diffusion-based approach to natural adversarial patch generation. Specifically, we construct a predictor for attack success rate (ASR) prediction by parsing the output of the target detector, such that the reverse process of a pre-trained diffusion model can be guided by the gradient of the classifier to generate adversarial patches with improved naturalness and high ASR. We conduct extensive experiments in both the digital and the physical worlds to evaluate the attack effectiveness against various object detection models, as well as the naturalness of generated patches. The experimental results show that by combining the ASR predictor with a pre-trained diffusion model, our method is able to produce more natural adversarial patches than the state-of-art approaches while remaining highly effective.

Key words: object detection; adversarial patch; diffusion model; adversarial example; adversarial attack; deep learning

收稿日期:2023-05-30;修回日期:2024-01-07;责任编辑:覃怀银

*通讯作者:张子君

Foundation Item(s): National Key Research and Development Program of China (No.2022YFB3102100); Fundamental Research Funds for the Central Universities (No. 2042022kf1034); National Natural Science Foundation of China (No.62206203, No.62076187); Key Research and Development Program of Hubei Province (No.2022BAA039); Key Research and Development Program of Shandong Province (No.2022CXPT055)

1 引言

随着深度学习的飞速发展,计算机视觉在过去十年中取得了巨大的进步[1~3]. 目标检测作为计算机视觉的关键技术之一,已广泛应用于安全关键的场景中,如人脸检测[4]、自动驾驶[5]和医学影像[6]. 在2014年,Szegedy等人发现深度神经网络容易受到对抗样本(adversarial examples)的攻击[7]. 对抗攻击是指向深度神经网络的输入图像样本添加精心设计但人眼难以察觉的扰动来生成对抗样本,从而诱使深度学习模型分类错误. 现有的目标检测模型也会受到对抗样本的威胁[8].

对抗样本攻击通常可以分为两类:数字世界的攻击和物理世界的攻击. 数字世界的攻击指攻击者可以在数字空间中直接修改输入图像[7,9~12],而在物理世界的攻击中,攻击者改变摄像机捕捉的目标物体的视觉特征,以欺骗深度学习模型[13],如目标检测模型.

在针对目标检测模型的物理世界攻击中,研究者们提出了一种较为常用的对抗补丁攻击方法,攻击者生成一张具有对抗性的图片,称为对抗补丁,将对抗补丁贴在到目标物体上(如眼睛框[14]、汽车车牌[8,15,16]、汽车车身[17]、人的衣服[18~21]),以使目标检测模型无法对该目标进行正确分类或定位. 例如使用对抗补丁来隐藏军用飞机或军舰,使其不受无人机的目标检测系统发现[22,23]. 目前部分研究者继续改进对抗补丁的攻击性能并增强其鲁棒性[19,24],也有另一部分研究者逐渐关注到对抗补丁的自然度和真实性问题,如将感知色彩距离与已有方法融合[25],在语义空间生成对抗补丁[26~28],将对抗损失与其他损失(如内容损失、风格损失和平滑度损失)一起优化[29,30],文献[31,32]提出基于生成对抗网络(Generative Adversarial Nets, GAN)的生成方法,通过在GAN学习到的图像流形上进行优化来生成对抗补丁.

然而这些方案部分仅仅只关注对抗补丁的攻击性能,没有关注对抗补丁的自然度,其生成的对抗补丁往往图案花哨,与自然图像之间存在明显差距,容易被人类观察者发现,因此不太可能迁移到由人类观察者审核的目标检测系统环境下,如军事背景. 另一部分方案虽然关注到自然度问题,但其要么需要优化额外的损失,或需要仔细调整训练过程中不稳定的超参数,要么有较明显的失真,反而降低了自身的隐蔽性,且基于GAN的对抗补丁是从GAN的隐空间进行采样的,缺乏可控性. 因此,这些方法都远远不够生成足够有效的对抗补丁.

为了解决这些问题,本文提出一种基于引导扩散模型的自然对抗补丁生成方法,最终生成的对抗补丁能够在自然性和攻击效果之间达到平衡. 受到条件扩散模型[33,34]的启发,本文构建了一个预测对抗补丁攻击成功率的预测器,通过该预测器的梯度引导预训练的扩散模型的逆扩散过程,最终在保证攻击效果的同时,生成更接近自然世界图片的对抗补丁,从而实现在物理世界中更有效的攻击.

2 基础知识

2.1 物理世界攻击

由于现实世界环境有更多复杂的条件和限制(如光照影响、位置、摄像机因素等),物理世界的攻击往往比数字世界中的攻击更具挑战性,这些环境因素可能会影响对抗补丁的像素值. 因此,文献[18,35]采用一组变换来模拟这些环境因素,使生成的对抗补丁在物理世界更具鲁棒性.

2.2 扩散模型

近年来,扩散式生成模型(扩散模型)在图像生成方面表现出优于GAN的性能[33,36]. 扩散模型是一种广泛应用于高质量图像生成的似然模型. 这些模型通过逐渐从潜在变量(通常是随机噪声)中移除噪声来生成样本[33,34,36]. 扩散模型由扩散过程和逆扩散过程(图像生成过程)组成,扩散过程可以看作一条固定的马尔科夫链,从原始数据$\boldsymbol{x}_0$开始,在T个扩散步骤内逐步加入噪声,以获得最终数据$\boldsymbol{x}_T$. 逆扩散过程即通过马尔可夫链从$\boldsymbol{x}_T$逐渐恢复到$\boldsymbol{x}_0$,扩散模型的逆扩散过程的马尔科夫链可表示为$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t):=N\left(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t\right),\boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t,t\right)\right)$,其中均值$\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t\right)$由参数为$\theta$的神经网络来拟合. 扩散模型训练的目的是使逆扩散过程生成的数据分布逼近原始数据$\boldsymbol{x}_0$的分布,从而生成更加真实和高质量的图像.

文献[33]提出条件扩散模型,该模型在逆扩散过程中可以被分类器引导,从而生成指定类别的图像. 具体来说,预训练的扩散模型$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$可以通过在逆扩散过程的采样均值上添加由分类器$p(y|\boldsymbol{x}_t)$的梯度计算出的条件来引导,从而生成在原来的基础上满足条件y的样本,称之为条件扩散模型$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,y)$. 条件扩散模

型的推导过程如下：

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,y)\propto p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\,p(y|\boldsymbol{x}_t)^s$$

$$\log p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,y)=\log\left[p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\,p(y|\boldsymbol{x}_t)^s\right]+C_1$$

$$\approx\log p(\boldsymbol{z})+C_2$$

其中，$\boldsymbol{z}\sim N(\boldsymbol{z};\boldsymbol{\mu}+s\boldsymbol{\Sigma}\boldsymbol{g},\boldsymbol{\Sigma})$，$\boldsymbol{\mu}\sim\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)$为预训练扩散模型的均值，$\boldsymbol{\Sigma}\sim\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t,t)$为预训练扩散模型的方差，$C_1$和$C_1$是常数，$s$为引导规模参数，$\boldsymbol{g}$是由用于引导扩散模型的分类器计算而来的梯度，即$\boldsymbol{g}=\nabla_{\boldsymbol{x}_t}\log p(y|\boldsymbol{x}_t)$.

3 本文方法

本文提出一种基于攻击成功率预测器引导扩散模型的自然对抗补丁生成方法，生成框架如图1所示. 首先选择一个预训练的扩散模型，该模型可以生成自然图像(例如真实世界的风景图或符合某种风格的图像). 随后，将当前扩散步骤生成的图像$\boldsymbol{x}_t$作为对抗补丁，将对抗补丁$\boldsymbol{x}_t$经过变换和贴图后的结果输入目标检测模型，分析对抗补丁$\boldsymbol{x}_t$攻击该目标检测模型的攻击成功率，构建一个攻击成功率预测器$p(y|\boldsymbol{x}_t)$，用于预测$\boldsymbol{x}_t$的攻击效果. 最后，利用攻击成功率预测器的梯度$\nabla_{\boldsymbol{x}_t}\log p(y|\boldsymbol{x}_t)$作为条件来引导扩散模型的逆扩散过程(即图像生成过程)，从而使该扩散模型可以按预期生成能够有效攻击目标检测模型的自然对抗补丁.

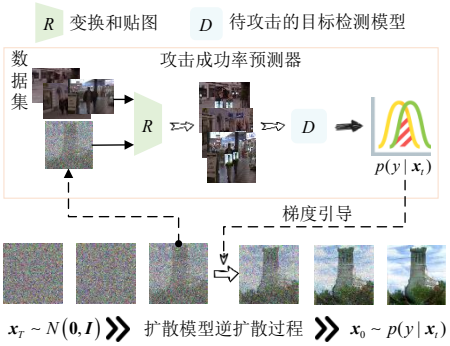

图1 基于引导扩散模型的自然对抗补丁生成方法框架

3.1 对抗补丁变换和贴图

本文的攻击目标是生成一个自然的对抗补丁，将其印在衣服上，以使穿戴者无法被人体目标检测模型检测到. 本文的攻击方案也可以应用于其他检测目标和其他目标检测模型.

首先用人体目标检测模型检测行人数据集中的k张干净图像，以获取图像上“人”的位置，以便在“人”中心位置贴上对抗补丁. 由第2.1节，本文引入一组随机变换R，对对抗补丁进行变换，这些变换包括随机遮挡、随机旋转和平移、亮度和对比度的随机变化、随机缩放. 将对抗补丁$\boldsymbol{x}_t$经过变换R之后，贴到k张干净图像$\boldsymbol{o}=(o_1,o_2,\cdots,o_k)$中“人”的中心位置，从而构成$k$张对抗图像$\boldsymbol{a}=(a_1,a_2,\cdots,a_k)$，此过程称为贴图步骤$Ap(\boldsymbol{o},\boldsymbol{x}_t)$.

3.2 攻击成功率预测器

将k张未贴对抗补丁的干净图像$\boldsymbol{o}$输入待攻击的人体目标检测模型D，得到的m个检测目标$\{(\boldsymbol{b}_i,c_i,\text{score}_i),i=1,2,\cdots,m\}$，将这些检测框作为干净检测框，其中$\boldsymbol{b}_i$是第$i$个检测目标的检测框坐标，$c_i$是第$i$个检测目标的类别标签，$\text{score}_i$为第$i$个检测目标的置信度得分. 本文以攻击人体目标检测模型为例，因此类别标签均为“人”，当选择其他类别作为目标类别时，c_i即为该目标类别标签.

将k张对抗图像$\boldsymbol{a}$同样输入待攻击的人体目标检测模型D，得到的n个检测目标$\{(\boldsymbol{b}_i^{\text{adv}},c_i^{\text{adv}},\text{score}_i^{\text{adv}}),i=1,2,\cdots,n\}$，其中$\boldsymbol{b}_i^{\text{adv}}$是第$i$个检测目标的检测框坐标，$c_i^{\text{adv}}$是第$i$个检测目标的类别标签，$\text{score}_i^{\text{adv}}$为第$i$个检测目标的置信度得分.

为了引导扩散模型生成对抗补丁，本文构建攻击成功率预测器，用该预测器的梯度进行引导. 攻击成功率预测器$p(y|\boldsymbol{x}_t)$能够预测对抗补丁$\boldsymbol{x}_t$的攻击成功率，其中标签$y=1$代表攻击成功率，$y=0$代表攻击失败率. 本文用对抗补丁$\boldsymbol{x}_t$的攻击成功率代表攻击成功的概率密度$p(y=1|\boldsymbol{x}_t)$，用对抗补丁$\boldsymbol{x}_t$的攻击失败率近似代表攻击失败的概率密度$p(y=0|\boldsymbol{x}_t)$.

在人体目标检测领域中，当且仅当一个检测框为真正例(True Positive，TP)时该检测框才会被作为检测结果保留. 因此，对抗补丁$\boldsymbol{x}_t$对上述k张干净图像中m个检测目标的攻击成功率应该定义为$1-\text{TP}/m$，其中TP为真正例检测框的数量. 但若以此方法直接计算攻击成功率，即$p(y=1|\boldsymbol{x}_t)=1-\text{TP}/m$，那么$p(y=1|\boldsymbol{x}_t)$将不可导，因此无法求出梯度$\nabla_{\boldsymbol{x}_t}\log p(y|\boldsymbol{x}_t)$.

因此本文选择用一种近似的方法计算攻击成功率. 分析目标检测模型的输出，对于k张干净图像上的第i个目标，若$\boldsymbol{b}_i$与$\boldsymbol{b}_i^{\text{adv}}$之间的交并比(Intersection Of Union，IOU)大于IOU阈值(一般为0.5)且$c_i\neq c_i^{\text{adv}}$，目标检测模型将此检测框标记为真正例，并输出所有真正例作为检测结果. 对于第一个条件$\boldsymbol{b}_i$与$\boldsymbol{b}_i^{\text{adv}}$之间的交并比大于IOU阈值，本文将满足此条件记为事件A，即事件A形式化表示为$\text{IOU}_i>0.5,i=1,2,\cdots,m$，其中$\text{IOU}_i$为$\boldsymbol{b}_i$与$\boldsymbol{b}_i^{\text{adv}}$之间的交并比. 在某些对抗图像中，存在多个检测框($\boldsymbol{b}_i^{\text{adv}}$)与干净图像上同一个干净检测框对应的情况，对于这些情况，具有最高交并比的检测框将被标

记为真正例.

对于第二个条件 $c_i \neq c_i^{\text{adv}}$,由于当 $\text{score}_i^{\text{adv}} > 0.5$ 时,此检测框将被分类为“人”这一类别.因此,$c_i \neq c_i^{\text{adv}}$ 等价于 $\text{score}_i^{\text{adv}} > 0.5$. 本文将满足第二个条件记为事件 B,即事件 B 形式化表示为 $\text{score}_i^{\text{adv}} > 0.5, i = 1, 2, \cdots, m$. 当同时满足以上两个条件,即事件 A 和事件 B 同时发生时,该目标被记为真正例,此时对抗补丁攻击该目标失败,计算攻击失败的目标个数占总目标个数的比例,即可得到该对抗补丁的攻击失败率,反之,用1减去攻击失败率即可得到该对抗补丁的攻击成功率.

对于一阶段目标检测模型,事件 A 和事件 B 是相互独立的. 为了使计算攻击成功率的方法可导,本文使用sigmoid函数来近似事件 A 和事件 B 发生的概率,计算公式见式(1)、式(2).

$$p(A) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{T_1}\left(\text{IOU}_i - 0.5\right)\right) \tag{1}$$

$$p(B) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{T_2}\left(\text{score}_i^{\text{adv}} - 0.5\right)\right) \tag{2}$$

其中,T_1 和 T_2 是温度参数,调节 T_1 和 T_2 能够控制上述近似概率更接近于真实概率. 最终本文的预测器定义见式(3).

$$p\left(y|\boldsymbol{x}_t\right) = \begin{cases} p(A)\,p(B), & y = 0 \\ 1 - p(A)\,p(B), & y = 1 \end{cases} \tag{3}$$

其中,标签 $y=1$ 代表攻击成功率,$y=0$ 为攻击失败率.

一个对抗补丁的真正攻击成功率需要在大量对抗补丁样本上统计出来,这个过程开销太大,难以预先大批量生成多样化的对抗补丁,无法进行精确计算. 又由于直接计算真正例和攻击成功率的过程是不可导的,无法直接计算其梯度并用于引导扩散模型,因此本文加入sigmoid函数使其可微,即可求出梯度. 本文构造攻击成功率预测器近似计算条件概率,便于利用扩散模型进行条件引导.

3.3　基于预测器引导的扩散模型

本文构建一个基于攻击成功率预测器引导的扩散模型来生成对抗补丁. 由第2.2节可知,$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 为一个预训练好的扩散模型,$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$ 是本文想要得到的满足条件 y 的扩散模型,其生成的图片在原来的基础上满足条件 y. 在本文方案中,条件 y 即“攻击成功或失败”,其中标签 $y=1$ 代表攻击成功,$y=0$ 代表攻击失败. 在本文中用于条件引导的分类器 $p(y|\boldsymbol{x}_t)$ 即本文所构建的攻击成功率预测器,在预测器的引导下,本文得到的条件扩散模型采样生成的样本具有对抗攻击的效果,同时图案近似于扩散模型的训练样本.

在本文方案中,用式(3)中构建的攻击成功率预测器来计算引导梯度 $\boldsymbol{g} = \nabla_{\boldsymbol{x}_t}\log p\left(y|\boldsymbol{x}_t\right)$. 为了生成攻击成功的图像,选择标签 $y=1$,最终本文提出的基于引导扩散模型的自然对抗补丁生成算法如算法1所示.

算法1　基于引导扩散模型的自然对抗补丁生成算法

输入:扩散步长 T,温度参数 T_1 和 T_2,引导规模参数 s,补丁变换 R,目标检测系统 D,预训练的扩散模型 $(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$,行人数据集 $\boldsymbol{o}$

输出:自然对抗补丁 $\boldsymbol{x}_0$

$\boldsymbol{x}_T \leftarrow N(\boldsymbol{0}, \boldsymbol{I})$

FOR $t = T$ TO 1 DO:

$\boldsymbol{a} \leftarrow Ap(\boldsymbol{o}, \boldsymbol{x}_t)$

$\left\{\left(\boldsymbol{b}_i, c_i, \text{score}_i\right), i = 1, 2, \cdots, m\right\} \leftarrow D(\boldsymbol{o})$

$\left\{\left(\boldsymbol{b}_i^{\text{adv}}, c_i^{\text{adv}}, \text{score}_i^{\text{adv}}\right), i = 1, 2, \cdots, n\right\} \leftarrow D(\boldsymbol{a})$

FOR $i = 1$ TO m DO:

$\text{IOU}_i = \text{IOU}\left(\boldsymbol{b}_i^{\text{adv}}, \boldsymbol{b}_i\right)$

$p(A) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{T_1}\left(\text{IOU}_i - 0.5\right)\right)$

$p(B) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{T_2}\left(\text{score}_i^{\text{adv}} - 0.5\right)\right)$

$p\left(y = 1|\boldsymbol{x}_t\right) = 1 - p(A)\,p(B)$

$\boldsymbol{g} = \nabla_{\boldsymbol{x}_t}\log p(y = 1|\boldsymbol{x}_t)$

$\boldsymbol{\mu}, \boldsymbol{\Sigma} \leftarrow \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t, t\right)$

$\boldsymbol{x}_{t-1} \leftarrow N\left(\boldsymbol{\mu} + s\boldsymbol{\Sigma}\boldsymbol{g}, \boldsymbol{\Sigma}\right)$

END FOR

RETURN $\boldsymbol{x}_0$

3.4　对抗补丁自然度

本文的工作旨在生成既能避开目标检测模型,又能避开人眼的对抗补丁,使之更合理地用于物理世界的攻击. 所谓的“躲避人眼”实际上是一种主观上约束,即对抗补丁应该更接近现实世界中存在的模式,自然度应该更高. 图像的自然度指的是图像所呈现的内容与真实世界中的自然场景或对象相符合的程度,即图像中的视觉元素和属性与真实世界中的相似性,包括颜色、光照、纹理、形状、结构等方面的特征. 图像自然度越高,则代表图像可以更好地模拟真实世界中的场景和对象. 图像的自然度评价主要是一个主观的过程,涉及人类主观感知,但在实际的评价过程中,可以通过主观评价和客观评价相结合的方式进行.

图像自然度主观评价可以通过人类观察者对图像的直接感受及打分来进行[31]. 因此,本文统计人类观察者对对抗补丁选择的百分比作为对抗补丁的自然度主观评价. 主观评价可以真实地反映人类观察者的主观视觉感受,评价结果准确、可靠.

图像自然度客观评价可以通过计算图像的统计特征、纹理信息、颜色分布等来量化图像的自然度[37],建立数学模型来模拟人眼视觉感知. 由于本文的方案中不使用参考图像,因此选择使用无参考的图像质量评估方法进行对抗补丁自然度的客观评价. 本文的客观评估方法包括基于自然场景统计的失真通用无参考图像质量评估模型[38](Blind/Referenceless Image Spatial Quality Evaluator, BRISQUE)、基于自然场景图像计算的默认模型进行比较的无参考图像质量评估模型[39](Naturalness Image Quality Evaluator, NIQE)和基于感知的非参考图像质量评估器[40](Perceptual Image Quality Evaluator, PIQE).

4 实验结果与分析

4.1 实验准备

4.1.1 数据集、目标模型与评价指标

(1)数据集. 本文的对抗补丁生成过程的数据集选择INRIA行人数据集,该数据集包含614个训练图像和288个测试图像. 实验中,将所有图像尺寸调整为416×416. 由于各种目标检测模型的输出略有不同,为了公平比较,本文选择每个目标检测模型在干净图像上预测的具有最大IOU的检测框作为干净检测框.

(2)目标模型. 本文实验选择了YOLOv2[41],YOLOv3[42],YOLOv3 tiny,YOLOv4,YOLOv4 tiny,YOLOv5作为待攻击目标检测模型,这些模型均是在包含“人”类的COCO数据集上进行训练的. 所有目标检测模型的输入图像尺寸均为416×416,NMS的置信度阈值和IOU阈值分别设置为0.4和0.6.

(3)扩散模型. 本文选择在ImageNet数据集上预训练的扩散模型来生成对抗补丁,扩散模型生成图像的尺寸为256×256.

(4)攻击性能评价指标. 本文选择mAP@0.5作为评估指标来评估本文方案的性能,mAP@0.5为当IOU阈值设置为0.5时目标检测模型测试阶段的平均精度(mean Average Precision, mAP),该指标在目标检测领域中被广泛使用. 本文选此评价指标是为了进行公正比较,不需要考虑置信度阈值.

(5)自然度评价指标. 对于对抗补丁自然度主观评价,本文实验中以随机顺序向25位观察者展示所有对抗补丁和自然图片,观察者投票选择自己认为自然的对抗补丁,统计每个对抗补丁被投票选择的百分比,作为对抗补丁的自然度主观评分.

(6)对抗补丁自然度客观评价,实验中采用基于概率模型的无参考的图像质量评估方法计算每个对抗补丁的图像质量,作为对抗补丁的几种自然度客观评分,具体评价方法为第3.4节中提到的方法.

4.1.2 实验细节、实验环境及对比方案

(1)实验细节. 除非另有说明,实验中默认设置$T_1=0.5, T_2=0.5, s=2\,000$,批大小设置为16,扩散步长为1 000,随机种子选择5 423.

(2)实验环境. 本文代码使用Pytorch实现,在两个NVIDIA Tesla V100-SXM2-32 GB GPU上运行.

(3)对比方案. 为了进行全面的比较,本文选择了文献[18~21,31]、随机图像及自然图像,在INRIA行人测试数据集上进行对比.

4.2 攻击效果实验

4.2.1 数字世界实验

本文将目标检测模型在干净图像上的检测框作为干净检测框,也就是说,当不部署攻击时,目标检测模型在测试数据集上的mAP@0.5为100 %. 当部署攻击时,mAP@0.5越小表示对抗补丁攻击效果更好. 表2记录了在数字世界中,本文方案攻击不同目标检测模型时的mAP@0.5,表1记录了对比方案的mAP@0.5. 图2记录了本文中所有的对抗补丁,其中P8为随机噪声图像,P9为自然图像.

从表1和表2可以看出,本文方案的攻击效果接近于基于GAN的对抗补丁生成方案[31],并且攻击效果更接近于基于优化的对抗补丁生成方案[18]. 基于优化的对抗补丁生成方案[18]及其他对比方案虽然攻击效果更优,但这种方案并没有考虑对抗补丁的自然度,其生成的对抗补丁图案扭曲,易被观察者检测出来.

从表2还可以看出,本文生成的对抗补丁可以迁移到不同的目标检测模型上.

表1 数字世界中对比方案的攻击效果

对抗补丁标号	对比方案	攻击的目标	mAP@0.5/%
P10	基于优化对抗补丁[18]	YOLOv2	2.13
P11	通用对抗补丁[19]	Faster RCNN	61.87
P13	隐身斗篷[21]	YOLOv2	26.00
P14	对抗性T恤[20]	YOLOv2	10.70
P12	基于GAN对抗补丁[31]	YOLOv3	34.90

4.2.2 物理世界实验

本文在室内和室外各种环境中进行了物理世界实验. 安排穿印有对抗补丁T恤的人和穿着正常衣物的人一起站在摄像机前1~4 m的位置,随后用一部HUAWEI nova 2S手机拍摄一组他们的照片作为对抗图像,将拍摄的对抗图像输入目标检测模型,得到检测结果. 本实验中对抗补丁训练和测试均选择YOLOv3 tiny

图2　本文方案及对比方案所生成的对抗补丁

作为目标检测模型. 表3中记录了目标检测模型检测到的穿着正常衣物的人和穿印有对抗补丁的T恤的人的百分比. 图3中展示了不同环境下的对抗图像. 如表3所示,在室内环境中,对抗补丁将目标检测模型的检测率从100 %降低到约15 %,在室外环境中降低到43.75 %.

4.3　自然度实验

本文对生成的对抗补丁的自然度进行主观评价和客观评价. 分别统计本文生成的对抗补丁、对比方案的对抗补丁及一张自然图像的自然度主观评分和三种自然度客观评分,统计结果记录在表4中. 从表4中可以看出,本文生成的对抗补丁P1的自然度高于其他方案,更接近自然图像,更加逼真.

表3　物理世界中对抗补丁的攻击效果　　单位:%

场景	正常衣物	印对抗补丁的T恤
卧室	100	15.38
楼道	100	14.29
暗处	75	30.00
街道	100	43.75
院子	100	53.84
客厅	100	27.78

图3　不同环境下的对抗图像

表2　数字世界中本文方案的攻击效果　　单位:%

对抗补丁标号	训练时目标模型	测试时目标模型					
		YOLOv2	YOLOv3	YOLOv3 tiny	YOLOv4	YOLOv4 tiny	YOLOv5
P1	YOLOv2	**8.96**	42.82	27.55	25.77	21.96	19.10
P2	YOLOv3	41.21	**33.94**	37.31	34.49	45.78	34.72
P3	YOLOv3 tiny	25.28	52.59	**8.37**	42.86	33.28	27.07
P4	YOLOv4	51.32	64.36	48.41	**15.18**	54.36	39.90
P5	YOLOv4 tiny	29.43	59.95	16.57	36.77	**14.99**	24.79
P6	YOLOv5	54.58	56.84	54.35	34.99	48.18	**32.92**
P8	随机噪声	75.03	73.75	78.91	76.71	75.74	57.26

4.4　消融实验

4.4.1　引导梯度消融实验

为了探究本文方案的有效性,将算法1中用来引导生成对抗性补丁的梯度 $\boldsymbol{g}=\nabla_{x_t}\log p(y=1|\boldsymbol{x}_t)$ 替换为 $\boldsymbol{g}=\nabla_{x_t}\log-\text{loss}(\boldsymbol{x}_t)$,其中 $\text{loss}(\boldsymbol{x}_t)$ 是基于优化的对抗补丁生成方案[18]和基于GAN的对抗补丁生成方案[31]的损失函数,是由对抗图像中具有最大类别置信度得分的检测框计算而来的,旨在最小化目标检测模型输出的类别置信度. 表5中记录了使用两种不同引导梯度时生成的对抗补丁,实验结果表明,在达到相同攻击效果的情况下,本文方案生成的补丁更加自然.

表 4 自然度实验

对抗补丁和自然图像		对抗补丁							自然图像
		P2	P3	P4	P5	P12	P10	P13	P9
自然度主观评分/%		80	48	32	64	75	8	16	88
自然度客观评分	BRISQUE	9.715 0	43.458 2	33.164 9	31.405 9	27.536 3	44.008 5	43.458 2	14.543 8
	NIQE	18.878 8	18.872 1	18.875 5	18.875 2	18.873 9	5.438 8	18.872 1	3.512 1
	PIQE	26.967 7	48.060 0	38.941 3	48.724 8	31.757 9	63.870 4	76.225 0	34.265 2
对抗补丁来源		本文	本文	本文	本文	基于GAN对抗补丁[31]	基于优化对抗补丁[18]	隐身斗篷[21]	扩散模型[36]

表 5 使用不同引导梯度时生成的对抗补丁

对抗补丁	引导梯度	mAP@0.5/%
P2	$\nabla_{\boldsymbol{x}_t}\log p(y=1\|\boldsymbol{x}_t)$	33.94
P7	$\nabla_{\boldsymbol{x}_t}\log-\text{loss}(\boldsymbol{x}_t)$	33.27

4.4.2 温度参数的实验

温度参数 T_1 和 T_2 会影响自然度和攻击效果. 图 4 中记录取不同的温度参数值时生成的对抗补丁的攻击效果和自然度主观评分. 图 4 中对应的对抗补丁如图 5 所示. 图 4 中各个对抗补丁训练时的温度参数 T_1 和 T_2 的值如表 6 所示,所有对抗补丁均在 YOLOv3 上训练和测试. 从图 4 和表 6 中可以看出,当温度参数减小时,mAP@0.5 降低,即攻击效果更好,但同时自然度降低. 随着温度参数 T_1 和 T_2 的减小,$1/T_1$ 和 $1/T_2$ 增大,$p(y=1|\boldsymbol{x}_t)$ 更接近于真实攻击成功率,梯度 $\nabla_{\boldsymbol{x}_t}\log p(y=1|\boldsymbol{x}_t)$ 更大. 因此,生成的对抗补丁具有更强的攻击效果,但同时扩散模型会更关注攻击成功率预测器的梯度,生成的图片也更不自然. 调整温度参数 T_1 和 T_2 可以在自然度和攻击效果之间达到权衡.

4.4.3 引导规模参数的实验

引导规模参数 s 也会影响自然度和攻击效果. 当 $s>1$ 且 s 增加时,由于 $s\nabla_{\boldsymbol{x}_t}\log p\left(y|\boldsymbol{x}_t\right)=\nabla_{\boldsymbol{x}_t}\log p(y|\boldsymbol{x}_t)^s$,此分布被指数放大了,变得比 $p\left(y|\boldsymbol{x}_t\right)$ 更尖锐[33]. 换句话说,s 值越大,基于引导的扩散模型就越关注攻击成功率预测器的梯度,从而使生成的对抗补丁攻击效果更好,但同时自然度降低. 图 6 中展示了取不同的引导规模参数 s 值时生成的对抗补丁的攻击效果和自然度主观评分. 图 7 记录了图 6 中对应的对抗补丁. 所有对抗补丁均在 YOLOv3 上训练和测试. 从图 6 中可以看出,当 s 增大时,mAP@0.5 降低,即攻击效果更好,但同时自然度降低. 因此,调整引导规模参数 s 也可以平衡对抗补丁的自然度和攻击效果.

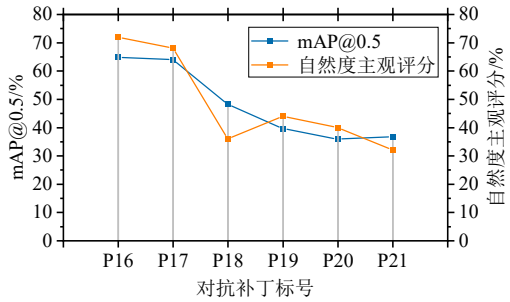

图 4 温度参数对对抗补丁自然度和攻击效果的影响

图 5 不同温度参数下生成的对抗补丁

表 6 不同对抗补丁在训练时的温度参数值

对抗补丁标号	P16	P17	P18	P19	P20	P21
T_1	1	1	2	2	2.5	3
T_2	1	2	1	2	2.5	3

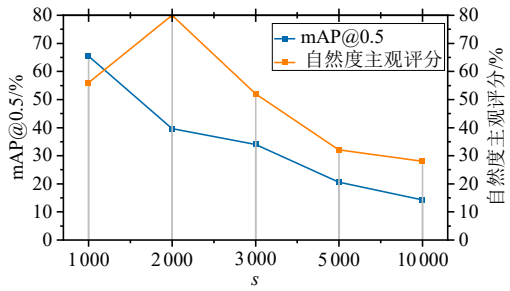

图 6 引导规模参数对抗补丁自然度和攻击效果的影响

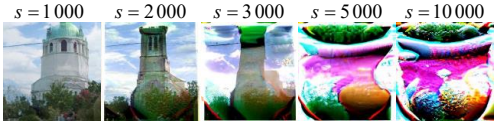

图 7 不同引导规模参数下生成的对抗补丁

5 总结与展望

针对目标检测系统生成的对抗补丁在一方面可以保护个人隐私,在另一方面也会严重威胁到人们的生命财产安全. 现有方案生成的对抗补丁在自然度方面不够有效,本文提出了一种基于引导扩散模型的自然对抗补丁生成方法,用于攻击目标检测模型,充分运用了扩散模型生成高质量图像的优势,能够生成既有攻击效果又更加自然的对抗补丁. 实验结果表明,本文方案生成的对抗补丁在攻击效果和自然度之间取得了良好的平衡. 未来工作的方向是提高生成对抗补丁的多样性,使攻击者能够生成个性化的对抗补丁,并且将本文的攻击扩展到其他类别上,用于隐藏任何类别的对象.

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[3] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2818-2826.

[4] DENG J K, GUO J, VERVERAS E, et al. RetinaFace: single-shot multi-level face localisation in the wild[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 5202-5211.

[5] CHEN C Y, SEFF A, KORNHAUSER A, et al. DeepDriving: learning affordance for direct perception in autonomous driving[C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 2722-2730.

[6] GE Y F, ZHANG Q, SUN Y T, et al. Grayscale medical image segmentation method based on 2D&3D object detection with deep learning[J]. BMC Medical Imaging, 2022, 22(1): 33.

[7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations. Banff: OpenReview.net, 2014: 1-10.

[8] SONG D, EYKHOLT K, EVTIMOV I, et al. Physical adversarial examples for object detectors[C]//12th USENIX Workshop on Offensive Technologies. Berkeley: USENIX Association, 2018: 1-10.

[9] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations. San Diego: OpenReview.net, 2015: 1-11.

[10] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2574-2582.

[11] CARLINI N, WAGNER D A, CARLINI N, et al. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2017: 39-57.

[12] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018: 1-23.

[13] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017: 1-14.

[14] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016: 1528-1540.

[15] BOSE A J, AARABI P. Adversarial attacks on face detectors using neural net based constrained optimization[C]// 2018 IEEE 20th International Workshop on Multimedia Signal Processing. Piscataway: IEEE, 2018: 1-6.

[16] CHEN S T, CORNELIUS C, MARTIN J, et al. ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector[C]//European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2019: 52-68.

[17] WANG J K, LIU A S, YIN Z X, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 8561-8570.

[18] THYS S, VAN RANST W, GOEDEME T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//2019 IEEE/CVF Conference on

Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2019: 49-55.

[19] HUANG L F, GAO C Y, ZHOU Y Y, et al. Universal physical camouflage attacks on object detectors[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 717-726.

[20] XU K D, ZHANG G Y, LIU S J, et al. Adversarial T-shirt! evading person detectors in a physical world[C]// Computer Vision - ECCV 2020. Cham: Springer, 2020: 665-681.

[21] WU Z X, LIM S N, DAVIS L S, et al. Making an invisibility cloak: Real world adversarial attacks on object detectors[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 1-17.

[22] AURDAL L, LKKEN K H, KLAUSEN R A, et al. Adversarial camouflage for naval vessels[C]//Artificial Intelligence and Machine Learning in Defense Applications. Bellingham: SPIE, 2019, 11169: 163-174.

[23] ADHIKARI A, den HOLLANDER R, TOLIOS I, et al. Adversarial patch camouflage against aerial detection[C]// Artificial Intelligence and Machine Learning in Defense Applications II. Bellingham: SPIE, 2020: 115430F.

[24] LEI X C, CAI X, LU C, et al. Using frequency attention to make adversarial patch powerful against person detector[J]. IEEE Access, 2023, 11: 27217-27225.

[25] ZHAO Z Y, LIU Z R, LARSON M. Towards large yet imperceptible adversarial image perturbations with perceptual color distance[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1036-1045.

[26] LIU H T D, TAO M, LI C L, et al. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer[C]//5th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019: 1-14.

[27] HOSSEINI H, POOVENDRAN R. Semantic adversarial examples[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2018: 1695-1700.

[28] HU Z H, HUANG S Y, ZHU X P, et al. Adversarial texture for fooling person detectors in the physical world[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 13297-13306.

[29] DUAN R J, MA X J, WANG Y S, et al. Adversarial camouflage: Hiding physical-world attacks with natural styles [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 997-1005.

[30] GUESMI A, BILASCO I M, SHAFIQUE M, et al. AdvART: Adversarial art for camouflaged object detection attacks[EB/OL]. (2023-03-03) [2023-05-20]. https://arxiv.org/abs/2303.01734.

[31] HU Y C T, CHEN J C, KUNG B H, et al. Naturalistic physical adversarial patch for object detectors[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 7828-7837.

[32] DOAN B G, XUE M H, MA S Q, et al. TnT attacks! universal naturalistic adversarial patches against deep neural network systems[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3816-3830.

[33] DHARIWAL P, NICHOL A Q. Diffusion models beat GANs on image synthesis[C]//Advances in Neural Information Processing Systems 34. La Jolla: Curran Associates, 2021: 8780-8794.

[34] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[C]//Advances in Neural Information Processing Systems 32. La Jolla: Curran Associates, 2019: 11895-11907.

[35] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]//Proceedings of the 35th International Conference on Machine Learning. San Diego: PMLR, 2018: 284-293.

[36] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems 33. La Jolla: Curran Associates, 2020: 6840-6851.

[37] TAN J, JI N, XIE H D, et al. Legitimate adversarial patches: evading human eyes and detection models in the physical world[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 5307-5315.

[38] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on Image Processing, 2012, 21(12): 4695-4708.

[39] MITTAL A, SOUNDARARAJAN R, BOVIK A C. Making a "completely blind" image quality analyzer[J]. IEEE Signal Processing Letters, 2013, 20(3): 209-212.

[40] VENKATANATH N, PRANEETH D, BH M C, et al. Blind image quality evaluation using perception based

features[C]//2015 Twenty First National Conference on Communications. Piscataway: IEEE, 2015: 1-6.
[41] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6517-6525.
[42] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-04-08) [2023-05-20]. http://arxiv.org/abs/1804.02767.

作者简介

何　琨　男,1986年10月出生于湖北省武汉市. 博士. 现为武汉大学副教授、博士生导师. 主要研究方向为应用密码学、网络安全、云计算安全、人工智能安全和区块链安全等. 中国电子学会会员编号:E190156480M.
E-mail: hekun@whu.edu.cn

佘计思　女,1999年7月出生于湖北省随州市. 现为武汉大学在读硕士生. 主要研究方向为人工智能安全、目标检测.
E-mail: shejisi@whu.edu.cn

张子君　男,1989年4月出生于湖北省武汉市. 博士. 现为武汉大学副教授. 主要研究方向为神经网络优化算法、正则化、网络架构、表示学习和强化学习等.
E-mail: zijunzhang@whu.edu.cn

陈　晶　男,1981年3月出生于湖北省武汉市. 博士. 现为武汉大学教授、博士生导师. 主要研究方向为网络安全、人工智能安全、分布式系统安全和区块链等.
E-mail: chenjing@whu.edu.cn

汪欣欣　女,1995年8月出生于湖北省随州市. 现为武汉大学在读博士生. 主要研究方向为目标检测、对抗学习和后门学习等.
E-mail: xinelwang@whu.edu.cn

杜瑞颖　女,1964年10月出生于河南省新乡市. 博士. 现为武汉大学教授、博士生导师. 主要研究方向为网络安全、隐私保护、云安全和移动安全等.
E-mail: duraying@whu.edu.cn