一种通用防御物理空间补丁对抗攻击方法

翔 云[1,2], 韩瑞鑫[1,2], 陈作辉[1,2], 李香玉[1,2], 徐东伟[1,2]

[1]浙江工业大学网络安全研究院 杭州 中国 310023
[2]浙江工业大学信息工程学院 杭州 中国 310023

摘要 目标检测算法具有优异的性能, 在工业上已经得到广泛应用。然而, 最近研究表明目标检测算法容易遭受对抗攻击, 对抗样本会使得模型的性能大幅下降。攻击者在数字空间中在图片上贴一个对抗补丁, 或者在物理空间中手持一张打印的对抗补丁, 都可以使得待检测的对象从目标检测器中“消失”。补丁对抗攻击在物理空间中可以攻击自动驾驶汽车和躲避智能摄像头, 对深度学习模型的应用造成了重大安全隐患。在物理空间中攻击目标检测器的对抗补丁具有鲜明特点, 它们色彩鲜艳、变化剧烈, 因此包含大量高频信息。基于这个特点, 我们提出了一种遮罩防御方法。我们先把待检测的图片分割成若干个像素块, 再用快速傅里叶变换和二值化处理求这些像素块中高频信息的含量, 依次对含有较多高频信息的像素块使用遮罩, 最后用目标检测器验证。此防御方法能够在物理空间中快速定位补丁的位置并破坏补丁的攻击效果, 使得目标检测器可以检测到被攻击者隐藏的对象。本方法与模型无关, 也和生成对抗补丁的方法无关, 能够通用防御物理空间中的补丁对抗攻击。我们在物理空间中使用了两个应用广泛的目标检测器做防御补丁对抗攻击实验, 在三个数据集中都能以超过 94%的防御成功率防御攻击, 比对比方法中最好的高出 6%, 实验结果证明了我们的方法的有效性。

关键词 深度学习; 补丁攻击; 物理攻击; 对抗防御
中图法分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2023.03.11

A general defense method for physical space patch adversarial attacks

XIANG Yun[1,2], HAN Ruixin[1,2], CHEN Zuohui[1,2], LI Xiangyu[1,2], XU Dongwei[1,2]

[1]Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China
[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract The deep learning based object detection algorithms have been widely used in many modern industry areas. However, recent research progress suggests that they are quite vulnerable to various adversarial attacks, which can greatly reduce the performance of deep learning models. Attaching an adversarial patch in the digital or physical space can make the attacked object “disappear” from the object detector. Therefore, patches generated by the adversarial attacks can cause major security risks to the deep learning models, e.g., automatic driving and intelligent camera evasion etc. Fortunately, those adversarial patches typically have very distinct features, e.g., rich colors, drastic changes, and hence, plenty of high frequency information. In this work, by tanking advantage of these features of the patches, we propose a mask based defense method for patch attack that utilizes fast Fourier transform, which can quickly locate the adversarial patch in the physical space. Specifically, we first divide the testing images into multiple pixel blocks. Then we use fast Fourier transformation and binaryzation to extract and process the high frequency information in each block. The blocks containing more high frequency components are masked. Finally, the masked images are re-processed using the original detection algorithm. In that case, the detected patch is consequently located and neutralized, and the hidden objects can be detected afterwards. This defense method is not related to the model or adversarial patch generation methods. It can be used as a general method against all similar adversarial patches. In our experiment, we use two widely used object detection algorithms in physical space to evaluate the performance of our technique. The experimental results show a defense success rate of more than 94% in three commonly used data sets, which is 6% higher than the state of art method. This results demonstrate the effectiveness of our techniques in real-world scenarios.

Key words deep learning; patch attack; physical attack; adversarial defense

通讯作者: 徐东伟, 副教授, 博士, Email: dongweixu@zjut.edu.cn
本课题得到国家自然科学基金(No. 61903334), 浙江省重点研发-尖兵项目(No. 2022C01018), 浙江省自然科学基金(No. LY21F030016)资助。
收稿日期: 2021-12-11; 修改日期: 2022-03-21; 定稿日期: 2023-01-04

1 引言

深度神经网络(Deep Neural Network, DNN)因为具有强大的特征提取能力, 在自然语言处理、车牌识别、目标检测等众多领域扮演重要角色, 并取得重大成功。然而, 随着 DNN 技术不断发展, 其安全问题日益受到研究人员关注。研究表明, 攻击者在样本数据中加入不易察觉的、特别设计的扰动, 可以使得深度学习模型做出错误的分类预测。例如在熊猫图像中添加扰动, 导致分类器错误的预测为长臂猿[1]。研究表明, 数字空间设计的扰动可以用于物理空间。将带有扰动的对抗贴纸贴在停车标志上, 致使自动驾驶汽车误认为限速 45 码标志[2]; 带上一副边框贴有对抗贴纸的眼镜, 成功攻击人脸识别算法[3]; Duan 等人[4]创造性的提出了一种伪装对抗攻击, 将对抗补丁伪装成自然界存在的物体的样子, 成功欺骗了最先进的图像分类器。

相比图像分类算法, 目标检测模型更加复杂, 也更难攻击成功, 不过也因此引起攻击者的兴趣, 成为被频繁攻击的目标。已经有一些攻击方法可以成功的愚弄目标检测器, 使得目标对象从目标检测中消失。Liu 等人[5]在图片的任意位置添加一个补丁, 可以使得目标在 YOLO[6-8]和 Faster R-CNN[9]等目标检测器中消失。不仅仅局限在数字空间, 它们中的一些可以在物理空间中实施, 在物理空间它们也可以取得成功。Thys 等人[10]将补丁对抗攻击应用在物理空间, 正如他们展示的, 通过把精心设计的补丁放置在人的中心位置, 可以欺骗当今最流行的 YOLO 目标检测器, 使得目标检测器发现不了贴了补丁的人。Mark 等人[11]在视野中放置一张对抗贴纸, 使得视野中所有的对象都成功地从目标检测器中逃脱了。我们相信在未来的研究中会出现更多优秀地攻击算法。

补丁对抗攻击愈发威胁到深度学习模型的安全性, 一个重要的问题出现了, 即我们如何防御已知的和将来可能会出现的补丁对抗攻击。目前已经提出了几种针对补丁对抗攻击的防御方法, 如 Hayes 等人[12]提出的数字水印和 Naseer 等人[13]提出的局部梯度平滑方法, 然而这些防御方法很容易被白盒攻击打破[14]。Xu 等人[15]通过类激活映射找到图片中对分类影响最大的像素, 再通过语义不一致性和预测激活不一致性可以找到补丁, 有效防御攻击。Liang 等人[16]提出基于 Grad-CAM 的防御方法,找到对分类影响最大的区域, 然后用灰色覆盖, 可以很好的防御补丁对抗攻击。他们还提出通过改进补丁生成算法, 可使得基于 Grad-CAM 的防御方法失去防御效果。然而目前还缺少一种在物理空间中快速高效地防御补丁对抗攻击的通用方法。

为了更好地说明本文提出的防御方法, 我们定义了两个新名词——对抗区域和良性区域。对抗区域指的是对抗补丁存在的那块区域, 是攻击有效的根源; 良性区域则相反, 指的是不包含对抗补丁地区域, 没有攻击效果。把全域像素看作一个集合, 则对抗区域与良性区互补。如果可以破坏对抗区域, 良性区域的像素修改很少, 检测器就能发现被隐藏的对象。最理想的是对抗区域完全被破坏, 良性区域保持不变。

现有技术在物理空间中对目标检测器发起攻击所生成的对抗补丁具有明显的特征。因为在实际应用过程中对抗补丁放在背景中或被攻击对象上, 只占被捕捉到图像的很小一部分。为了克服打印过程中的打印色差、图像传感器不能完美的捕捉到对抗补丁等问题, 保证对抗补丁具有攻击效果, 这些补丁都是相邻像素变化剧烈且色彩鲜艳, 因此引入了大量的高频信息[17-19]。局部梯度平滑、JPEG 压缩和位深压缩等防御方法修改全域像素, 良性区域也被修改, 而且不能完全破坏对抗区域。这导致降低了模型正常样本数据集上的表现, 不能很好的防御补丁对抗攻击。根据这个特征, 我们提出了基于快速傅里叶变换的通用防御方法。使用遮罩遮住补丁, 能够完全破坏补丁而且不改变良性区域的像素, 因此会有更好的防御效果。不失一般性, 我们选择人作为防御对象。实验证明, 在物理空间可以找到对抗补丁并高成功率防御攻击, 使得目标检测器可以找出被隐藏的人。

综上所述, 本文的主要创新点包括:

(1) 提出了基于快速傅里叶变换的防御方法, 有效防御物理空间中针对目标检测器的补丁对抗攻击, 并且防御算法对原模型的影响小。

(2) 我们用两个最先进的目标检测器在四个数据集上对防御方法进行评估, 展示了防御方法对原模型的影响小于同类方法, 且防御效果优于同类方法。

2 相关工作

深度学习对抗攻击指的是, 攻击者在良性样本上添加精心设计的扰动得到对抗样本, 使得深度学习模型做出错误的预测。防御对抗攻击指的是, 防御者通过去除对抗样本中扰动, 防止深度学习模型因扰动的存在而做出错误的预测。

2.1 对抗攻击

深度学习中的对抗攻击分为目标攻击和非目标攻击。目标攻击是指通过添加扰动欺骗深度神经网络, 使得预测结果为期望的错误。非目标攻击是指通过添加扰动欺骗深度神经网络, 使得预测结果不正确即可。对抗攻击可以应用在数字空间, 也可以用在物理空间中。在数字空间中, 攻击者可以随意修改图片中任何一个像素。不同于数字空间中的攻击, 在物理空间中, 攻击者没有权限访问和修改被攻击者使用图像传感器捕获到的图像, 只能够在要攻击的物体上预先添加扰动, 然后图像传感器捕捉已经添加扰动的图像。对抗攻击和防御是一个相互博弈的过程, 永远也不会结束。

对抗样本是深度神经网络安全研究的重要组成部分, 目前主要针对图像分类和目标检测等计算机视觉任务。Goodfellow 等人[20]提出的快速梯度符号法(fast gradient sign method, FGSM), 通过沿梯度反方向添加扰动, 使得分类器做出错误的预测。FGSM 虽然速度快, 但是容易陷入局部最优。一些研究人员在 FGSM 的基础上提出了多种改进方法, 不仅提高了攻击成功率, 也缓解了陷入局部最优的问题。例如 IFGSM[21]和 PI-FGSM[22]。Xin 等人[5]在图片的任意位置添加一个补丁, 可以使得目标从 YOLO 和 Faster R-CNN 检测器中逃离。

由于光的强度、拍摄角度和距离远近都会导致攻击失效, 更难在物理空间中发起对抗攻击。不过也有不少研究人员将对抗攻击应用在物理空间中。Mahmood 等人[3]设计了一种对抗贴纸, 贴在眼镜框上, 成功躲避了人脸识别模型。Chen 等人[23]在车牌上添加黑色色素块, 可以成功欺骗车牌识别系统。Kevin 等人[2]提出 RP2 攻击方法, 导致分类器将贴有对抗补丁的停止标志误分类为限速 45 码标志。

在物理空间中除了图像分类算法受到攻击, 许多研究人员也成功攻击目标检测算法。攻击者通过打印预先生成好的对抗补丁, 然后将补丁放在待检测对象的身上或者所在背景中, 使得持有补丁的对象在目标检测器中“消失”了, 不能被检测到。如图 1, (a)没有对抗补丁, 目标检测器可以发现图像中的人, (b)中的人手持对抗补丁的人成功躲过目标检测器。Thys 等人[10]首先提出生成对抗补丁, 通过手持打印的补丁, 成功欺骗 YOLOv2 目标检测模型, 使得手持对抗补丁的人在摄像头中消失, 逃避目标检测算法。Mark 等人[11]在背景中放置一张对抗贴纸, 导致 YOLOv3 目标检测模型不能发现视野中的任何物体。Wu 等人[24]将补丁印刷在衣服上, 成功攻击目标检测器。我们提出的防御算法就是防御此类针对目标检测器的补丁对抗攻击。

(a) (b)

图 1 在物理空间中攻击目标检测器

Figure 1 Attack the object detector in physical world

除了图像邻域, 语音识别系统也引起了研究人员的兴趣。Kreuk 等人[25]在声学特征(MFCC)上利用快速梯度符号法攻击端到端的说话人验证模型, 取得了很好的成功。Li 等人[26]通过构建一个单独的对抗转换网络, 直接将原始输入转化为对抗输入, 有效的攻击说话人识别模型。Carlin 和 Wagner[27]提出了一种直接修改原始音频的优化方法攻击语音识别模型, 该攻击属于目标攻击且攻击成功率可达到 99%。

2.2 防御

现在防御对抗攻击主要集中在数字空间, Chiang 等人[14]提出一种通过区间限制传播(IBP)对抗补丁[28-29]防御方法。Xiang 等人[30]提出 PatchGuard 方法防御针对图像的补丁对抗攻击, 在特征聚合中引入掩蔽算法, 去除导致图像分类错误的异常大的潜在损坏特征。Naseer 等人[13]提出了一种局部梯度平滑方法, 在把图片送入模型推理之前, 通过估计调整噪声区域的梯度, 能够防御对抗攻击潜在的影响。

目前防御对抗攻击的方法主要分为两大类: a)在把样本送入模型之前对样本进行处理, 如 LGS[13]对图片进行局部平滑处理, 从而破坏对抗补丁; 用遮罩覆盖对抗补丁, 除去补丁[31]。b)修改 DNN 网络, 如在目标类中增加补丁类, 通过对抗训练找到图片中的补丁[32], Xu 等人[17]通过类激活映射的深度神经网络激活可视化方法找到并抠出对分类影响最大的区域, 再与正常样本求差异性, 可以找到潜在的对抗补丁, 最后通过图像修复算法恢复原始图像, 得到正确的分类。这两类防御方法都是开环, 没有构成闭环。我们的防御方法不同于这两类, 模型的输出会反馈给输入。我们通过快速傅里叶变化找到高频信息最多的像素块, 把原图像中该像素块对应位置替换成灰色遮罩, 再送入目标检测, 然后从模型输出可得知检测结果, 最后判断是否需要用灰色遮罩替换

高频信息其次的像素块, 形成闭环, 能够快速找到存在对抗补丁。

除了在图像领域, 研究人员也研究了语音领域的对抗防御。Yang 等人[33]提出利用语音数据的时间依赖性来检测对抗样本, 该方法可以有效的防御对抗攻击。

3 方法

3.1 图像中的快速傅里叶变换

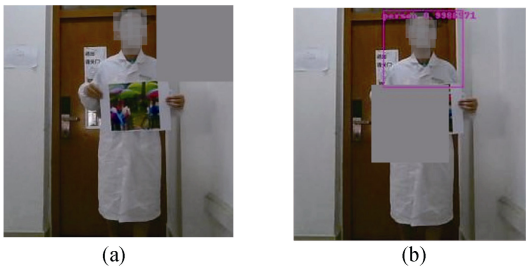
因为对抗攻击发起者只能修改样本中特定区域中的像素, 而不可以修改全域像素。为了保证扰动能够被打印并且在物理空间中仍然有效果, 攻击者会引入色彩艳丽、变化剧烈的扰动, 所以对抗补丁会引入大量高频信息。对抗样本和良性样本有着很大的区别, 因此我们采用二维快速傅里叶变换(2D-FFT)区分正常样本和对抗样本, 一张尺寸为 M×N 的图像经过二维快速傅里叶变换后得到。二维快速傅里叶变换公式如式 1 所示。

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)e^{-j2\Pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{1}$$

式中 M 和 N 分别表示图像的总行数和总列数, x、y 分别表示第 x 行和第 y 列, f(x,y)表示第 x 行第 y 列的像素值。

3.2 方法概要

为了防护目标检测模型, 需要找到一个方法, 准确找到对抗补丁, 然后消除补丁的影响。相当于在目标检测器或分类器前面构造一个过滤器, 用于找到并去除待检测样本中的补丁。最直观的方法就是用遮罩遮盖对抗补丁, 如果对抗区域被覆盖, 检测器就能检测到原本被隐藏的对象。如图 2 所示, 图(a)中遮罩没有破坏对抗区域, 攻击对抗补丁的攻击效果不会受影响,目标检测器不能检测到被隐藏的人。图(b)中遮罩破坏了绝大部分对抗区域, 使得对抗补丁失去攻击效果, 检测器可以检测到被隐藏的人。由



(a) (b)

图 2 在不同位置应用遮罩, 再送入目标检测器

Figure 2 Apply mask at different positions and feed it to the object detector

于不知道补丁所在位置, 如果在每一个有可能出现对抗补丁的位置应用遮罩, 再用检测器验证, 必定会花费大量时间。我们的研究就是如何快速准确的找到补丁的位置, 并破坏补丁, 再将图像送入检测器, 这样就能正确检测到图像中被隐藏的对象。

根据物理空间中攻击目标检测算法生成的对抗补丁包含变化剧烈、含有大量高频信息的特征, 我们提出基于快速傅里叶变换的快速遮罩防御方法。首先, 这种方法直接对图像进行处理, 与模型无关。其次, 这类在物理空间中攻击目标检测器的对抗补丁都具有此特征[17-19]。因此本方法具有普适性, 能够防御这一类攻击。该方法不仅可以检测到图像中是否存在对抗补丁, 还能够定位到对抗补丁的位置并破坏补丁。受 YOLO 的启发, 首先我们将输入样本切分为 $n\times n$ 大小的像素方块, 然后将每一个像素方块都进行快速傅里叶变换, 再二值化处理。从二值化后的图像中可以观察到, 变化越剧烈的图像得到的结果中高频信息(白色像素)占比越大。我们按照白色占比由高到低将所有的像素方块排序, 并记录像素块的位置。我们认为如果存在对抗补丁, 那么一定出现在靠前的位置。然后取前 K 个(实验中我们记为 TOP K), 对原像素方块用灰色(像素值为(128, 128, 128))遮罩覆盖, 得到应用遮罩的样本。然后把此样本送入目标检测器, 若应用遮罩前检测器没有发现人, 应用遮罩后的样本发现了人, 则说明该位置存在对抗补丁。我们的防御算法流程如图 3 所示。

我们的防御方法主要分为三步: ①首先我们将样本切分为若干个方形像素块。②接着进行快速傅里叶变换和二值化, 得到二值图像, 接着再计算所有二值图像中白色像素的占比, 表示像素块包含高频信息占比。根据白色像素占比从高到低将像素块排序, 排在前面的更有可能是对抗补丁。③最后用灰色遮罩覆盖疑似补丁, 送入目标检测器。遮罩覆盖了对抗补丁, 破坏了攻击效果, 因此检测器就能发现被隐藏的人。加入我们的防御方法的目标检测算法如式 2。实验证明, 在物理空间中, 不论对抗补丁放在被待检测对象身上还是放置在背景中, 我们的方法都可以有效防御这类对抗攻击。

$$result=F(x\odot M,step,BlockSize,K) \tag{2}$$

式中 x 是输入图片; $step$ 表示切分的步长, 输入图像被切分成像素块的大小为 $BlockSize\times BlockSize$; M 表示遮罩, 像素值为(128, 128, 128), 大小与切分的像素块相同; K 表示只取前 K 个高频占比最大的像素块并应用遮罩; $F(\bullet)$是目标检测器, $result$ 是检测器的输出, 包含检测对象的置信度和位置信息。

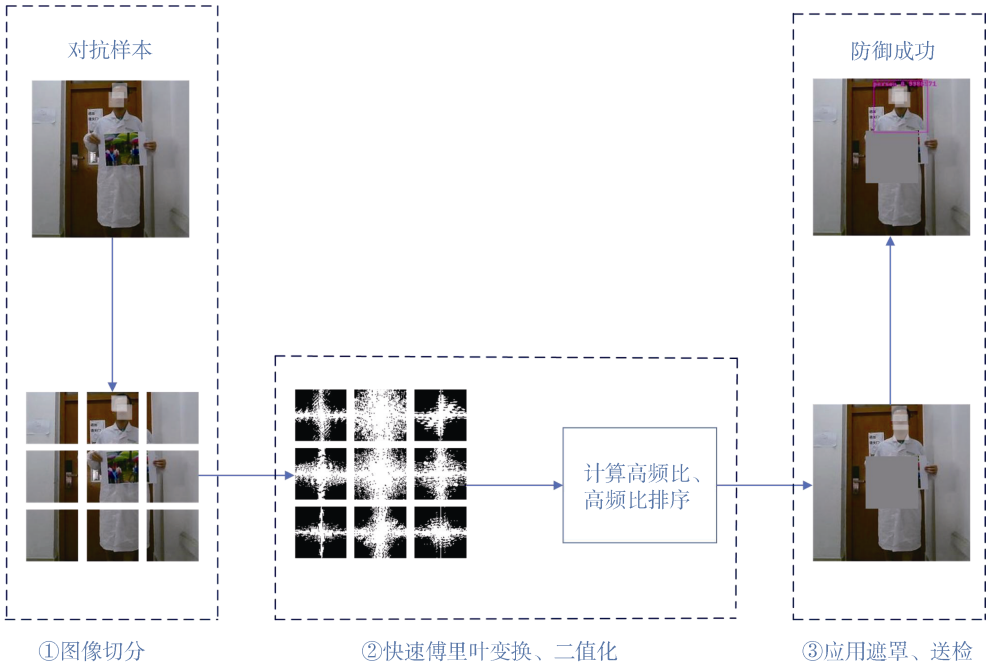


图 3　基于快速傅里叶变换的通用物理空间补丁对抗攻击防御方法示意图

Figure 3　General physical space patch attack defense method based on fast Fourier transform

3.3　防御方法系统阐述

YOLO 是目前最先进的目标检测器, 经过训练后可以找出感兴趣的对象。这类检测器的鲁棒性能非常地强, 即使目标的一部分被遮盖住了, 甚至是输入仅仅包含对象的一部分, 都可以以高置信度检测到对象。因此用遮罩覆盖对象的一部分对检测结果的影响可以忽略不计。

3.3.1　图像切分

如果将整张待检测的样本图像直接快速傅里叶变换, 可以知道该样本变化剧烈程度, 但是不能准确定位到变化最剧烈的区域。所以先将输入样本切分成若干个像素块, 再对得到的像素块进行快速傅里叶变换。因为变化越剧烈越有可能是对抗补丁, 因此只要找到变化最剧烈的像素块, 再找到该像素块在输入样本中的位置就能确定疑似补丁的位置。由于攻击者的补丁有可能出现在样本的任何一个位置, 所以我们需要检测所有可能出现补丁的位置, 因此计算量会很大。如样本图片上是 416×416 大小, 像素块取 60×60, 总共有超过 127k 种可能情况。如果设置划分区域的大小为 70×70, 步长为 30, 总共得到 144 个像素块, 极大地减小了计算量并缩短时间。

3.3.2　快速傅里叶变换和二值化处理

将划分得到的像素块按照从左到右, 从上往下的顺序依次进行快速傅里叶变换和二值化处理。我们可以从二值图像中明显地发现, 变化越剧烈的像素块得到的二值图像中白色像素占比越高, 即含有的高频信息越多。如图 4 所示, 像素块 1 包含对抗补丁大部分, 像素块 2 只包含小部分, 像素块 3 完全不包含。分别经过快速傅里叶变换和二值化, 得到体现高频信息含量的二值图像。可以看到, 像素块包含对抗补丁面积越大, 得到的二值图像中白色像素越多, 即高频信息占比高的像素块更有可能是对抗补丁, 我们优先验证这些像素块。

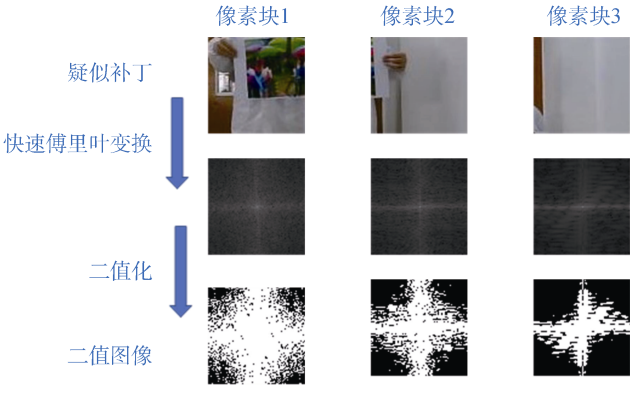


图 4　包含补丁面积不同的像素块经过快速傅里叶变换和二值化得到的结果

Figure 4　The result of fast Fourier transform and binarization of pixel blocks containing different patch areas

3.3.3　排序、像素块选择

把每一个输入样本分割成指定大小的像素块, 这些像素块经过快速傅里叶变换和二值化处理后可以得到一系列二值图像, 我们需要根据这些像素块

的二值图像确定对抗补丁的位置。二值图像中白色像素表示像素块中包含的高频信息, 依次计算每一个像素块对应的二值图像中白色像素的占比。为了方便计算, 我们先将二值图像归一化, 再计算占比, 公式如式(3):

$$Rate(i)=\frac{1}{M*N}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y) \tag{3}$$

式中 i 表示第 i 个像素方块, M 是像素方块的长, N 是像素方块的宽, f(x,y)表示第 x 行, 第 y 列像素的值, $f(x,y)\in[0,1]$。

图 4 中三个疑似补丁由此计算高频占比如表 1 所示。求出待检样本所有的像素方块高频占比, 由高到低排序, 并且记录像素块在原图像中的位置。3 个包含补丁面积不同的像素块经过快速傅里叶变换和二值化处理后求得高频信息占比, 从表中可以看到像素块 1 中高频比为 72.08%, 像素块 2 中高频比为 47.42%, 像素块 3 中高频比为 28.92%。可以看到, 包含补丁面积越大的像素块高频比越高。

表 1　图 4 中包含补丁面积不同的像素块的高频信息占比

Table 1　the proportion of high-frequency information in the three pixel blocks with different patch areas in Figure 4

	像素块 1	像素块 2	像素块 3
高频比	72.08%	47.42%	28.92%

3.3.4　送检、判断是否存在对抗补丁

包含高频信息更多的像素块会排在更前面, 代表该像素块位置更有可能存在对抗补丁。用灰色遮罩覆盖该疑似补丁的像素块, 然后送入目标检测器。目标检测器 YOLO 如果检测到了人, 会返回一个矩形框, 表示人在图像中的位置。如果应用遮罩前没有检测到人, 应用遮罩后检测到了人, 则说明遮罩在原样本中对应的位置存在对抗补丁。我们还需要判断检测到的人是否是真的人。如果遮罩中心在矩形框内部, 我们认为检测到的人是真的人; 如果像素块中心不在矩形框内部, 当人的置信度超过阈值 th(实验中我们取 th=0.7), 我们也认检测到的人是真的人。这两种情况都是防御成功。整个防御方法的伪代码如算法 1 所示。

伪代码中第 1 行是将输入样本分割为指定大小的像素块。第 2 行到第 7 行依次是将像素块进行快速傅里叶变换、二值化处理, 再求高频占比。第 8 行表示按照高频占比将切分得到的像素块排序。第 9 行表示只取出前 K 个高频占比较的像素块。第 10 和第 11 行分别是对取出来的前 K 个像素块在原始图像对应的位置用灰色遮罩替换, 和将应用遮罩的样本送给模型, 得到目标检测器输出的类别和矩形框, 框表示对象在样本中的位置信息。第 10 行到第 18 行表示判断是否有补丁。如果检测到的结果没有人, 则对下一个位置应用遮罩。如果检测到有人, 判断遮罩中心是否在检测器给出的方框内, 若在方框内或者在方框外但置信度大于阈值, 则表示防御成功。

算法 1　防御算法

输入: 一张样本图像 x, 取高频信息最高的像素块数 k, 划分块大小 $BlockSize$, 阈值 th, 目标检测器 F(·)

输出: 是否存在对抗补丁

```
1.BlockList=Divide(x,BlockSize,step)
2.FOR block in BlockList:
3.    FBlock=2DFFT(block)
4.    BinBlock=Binarization(FBlock)
5.    HFRate=Rate(BinBlock)
6.    BinBlockList.append(HFRate)
7.    END FOR
8.RateList=BinBlockList.sort()
9.FOR i in range(k):
10.   MaskOnImg=ApplyMask(img, RateList[i])
11.    result,box=F(mask_on_img)
12.    IF result is person:
13.        IF RateList[i] in box:
14.        return result
15.        ELSE IF P(person)>th
16.            return result
17.        END IF
18     END IF
19.END FOR
20.return null
```

4　实验

4.1　方案概述

我们使用当下流行且表现出色的 YOLOv2 和 YOLOv3 两种目标检测算法作为防护模型。物理空间中对目标检测算法的攻击方法可根据对抗补丁作用位置不同分为两大类, 分别是作用在待检测对象身上和作用在待检测对象所在背景中, 这两类攻击方法都可以使得目标检测器不能发现补丁作用下的人。为了体现防御方法的普适性, 分别选用两类攻击方法中的一个典型攻击方法, 直接打印文章[10-11]中的对抗补丁作为本实验用的补丁。这三个攻击方法所生成的对抗补丁分别是要贴在待检测对象身上[10]和待检测对象所在背景中[11], 两个对抗补丁都具有

高攻击成功率高鲁棒性, 且含有这类补丁的特征, 能够代表物理空间中攻击目标检测器的补丁对抗攻击方法。数据集在 4.2.1 节介绍。当测到样本中存在对抗补丁, 我们提前结束该样本的检测, 如果检测完包含高频信息较多的前 K 个像素块, 我们也结束此样本的检测, 然后进行下一个样本检测。我们一共选取了 4 种对比方法, 分别是局部梯度平滑(LGS)[13]、JPEG 压缩、总方差最小化(TMV)[34]和位深压缩(BR)。其中 LGS 的具体做法是先估计梯度域中的噪声位置, 然后在将样本送入深度神经网络之前进行正则化处理以估计噪声区域的梯度, 从而破坏图像中变化剧烈的区域; JPEG 压缩通过使用离散余弦变换(DCT)去除高频分量, 从而破坏对抗样本中具有攻击效果的扰动; 位深压缩是减少像素存储时的位数, 从而降低图像分辨率。

4.2 实验设置

4.2.1 数据集

在物理空间中针对目标检测器的补丁对抗攻击还没有公开数据集, 因此我们在物理空间中使用杰锐微通公司的 HF899 摄像头拍摄了四组数据集, 分别是良性样本、对抗样本 1、对抗样本 2、对抗样本 3, 每组 1000 张图片。然后筛选出良性样本中所有能够被目标检测器检测到的样本、对抗样本 1、对抗样本 2 和对抗样本 3 中所有不能被目标检测器检测到的样本。如果检测器不能检测到良性样本中的对象, 评估防御方法对良性样本的影响失去意义。如果未使用防御方法前检测器可以检测到样本中的对象, 那么评估防御成功率也没有意义, 所以在制作数据集的过程中我们会进行筛选, 所有的筛选工作都是由目标检测器自身完成。数据集由筛选后的四组样本构成。其中良性样本共计 869 张图片; 对抗样本 1 数据集筛选得到的对抗样本数据集 A, 共计 532 个对抗样本; 对抗样本 2 数据集筛选得到的对抗样本数据集 B, 共计 418 个对抗样本; 对抗样本 3 数据集筛选得到的对抗样本数据集 C, 共计 733 个对抗样本。我们直接使用文章[10]中的补丁作为对抗样本 1 和对抗样本 2 中的对抗补丁, 该补丁必须要贴在被攻击对象的中间位置, 该对抗样本 YOLOv2 目标检测器有效。直接使用作者[11]提供的补丁作为对抗样本 3 中的对抗补丁, 补丁可以放置在被攻击对象中间置, 也可以放置在背景中, 该对抗样本对 YOLOv3 目标检测器有效。

4.2.2 模型和类别

我们选用 YOLOv2 和 YOLOv3 模型作为本次实验目标检测器。我们选用人作为攻击和防御的目标。对抗样本数据集 A 和对抗样本数据集 B 攻击目标检测器 YOLOv2, 对抗样本数据集 C 攻击目标检测器 YOLOv3。

4.2.3 参数设置

我们设置 YOLOv2 在对抗样本数据集 A、对抗样本数据集 B 和良性样本数据集中的区域划分的步长为 30, 分割像素块的大小为 60×60 和 70×70。YOLOv3 在对抗样本数据集 C 和良性样本数据集中区域划分的步长为 30, 分割像素块的大小为 85×85。两个目标检测器都是取包含高频信息最高的前 10 个像素块用作遮罩。

4.3 实验结果

在物理空间中用良性样本数据集评估防御方法对原模型的影响, 用对抗样本数据集评估防御方法对补丁攻击的防御效果。防御成功率计算方法如下:

$$acc=\frac{\text{防御成功的样本数}}{\text{总样本数}}\times 100\% \tag{4}$$

两个目标检测器在良性样本和对抗样本数据集中的实验结果如表 2 所示。一共测试了包括我们的方法在内的 5 种防御方法分别在对抗样本和良性样本数据集中的表现, 分别是局部梯度平滑(LGS)[13]、JPEG 压缩、总方差最小化(TMV)、位深压缩(BR)和我们提出得基于快速傅里叶防御方法(OURS)。实验中用良性样本和对抗样本数据集 A 和对抗样本数据集 B 评估我们的方法防御针对目标检测器 YOLOv2 的补丁对抗攻击的有效性, 用良性样本和对抗样本数据集 C 评估我们的方法防御针对目标检测器 YOLOv3 的补丁对抗攻击的有效性。在三个数据集中, 我们都是取高频信息占比最高的前 10 个像素块对其应用作罩, 检测像素块在是否包含对抗补丁。从实验结果中可以看到, 加了防御方法后, 原模型在良性样本数据集中的精度会降低, 因此防御方法对原模型存在消极影响。添加我们的防御方法后, YOLOv2 和 YOLOv3 都可以检测出所有良性样本中的人。在对抗样本数据集 A 中, 防御成功率达到 96.62%; 在对抗样本数据集 B 中, 防御成功率达到 94.50%; 在对抗样本数据集 C 中, 防御成功率达到 95.09%。在对抗样本数据集 A 中, 我们的方法防御成功率比对比方法中效果最好的位深压缩(59.59%)高出 37.03%; 在对抗样本数据集 B 中, 我们的方法防御成功率比对比方法中效果最好的 JPEG(防御成功率等于 21.53%)高出 6.14%; 在对抗样本数据集 C 中, 我们的方法防御成功率比对比方法中效果最好的局部梯度平滑(防御成功率等于 88.95%)高出 6.14%。在良性样本数据集中, YOLOv2 和 YOLOv3

表 2 物理空间中防御针对两种典型目标检测器补丁对抗攻击性能总结

Table 2 Summary of the performance of the defense against two typical target detector patches in the physical space

	YOLOv2			YOLOv3	
	良性样本数据集	对抗样本数据集 A	对抗样本数据集 B	良性样本数据集	对抗样本数据集 B
未加防御	100%	0	0	100%	0
LGS[lambda=13]	98.39%	31.77%	45.45%	100%	88.68%
LGS[lambda=11]	98.39%	32.71%	52.93%	100%	88.95%
LGS[lambda=9]	98.85%	35.34%	61.48%	100%	87.86%
LGS[lambda=7]	99.19%	35.53%	65.55%	100%	86.63%
LGS[lambda=5]	99.77%	35.34%	72.01%	100%	81.31%
LGS[lambda=3]	99.77%	28.20%	70.33%	100%	72.17%
JPEG[quality=90]	99.88%	8.08%	72.73%	100%	26.74%
JPEG[quality=70]	99.88%	8.08%	70.81%	100%	24.56%
JPEG[quality=50]	99.65%	21.99%	72.97%	100%	29.20%
JPEG[quality=30]	99.42%	31.39%	69.38%	100%	74.22%
JPEG[quality=10]	75.72%	21.24%	41.15%	84%	38.47%
TMV[weights=1]	86.77%	25.94%	35.89%	97.93%	41.06%
TMV[weights=5]	88.84%	28.20%	34.69%	97.93%	40.79%
TMV[weights=10]	87.23%	26.32%	35.89%	97.81%	40.65%
TMV[weights=20]	87.11%	25.94%	36.36%	97.81%	42.43%
TMV[weights=30]	86.42%	25.38%	34.69%	97.93%	42.16%
BR[depth=1]	100%	13.72%	71.29%	100%	19.78%
BR[depth=2]	98.27%	59.59%	50.48%	99.78%	61.26%
BR[depth=3]	79.75%	19.17%	3.11%	97.93%	22.37%
OURS[top 10]	100%	96.62%	94.50%	100%	95.09%

都能检测到所有样本中的人, 且防御后原模型精度降低小于同类方法。

4.4 像素块分割大小对实验结果的影响

实验过程中我们发现输入样本被切分为不同大小像素块, 防御效果会不同。因为遮罩大小等于图像被划分的大小, 所以遮罩尺寸直接决定覆盖对抗补丁的程面积, 从而影响对抗区域被破坏的程度。我们在 3 个对抗样本数据集中, 将样本图片划分为不同的大小, 以测试其对防御效果的影响。YOLOv2 目标检测器在对抗样本数据集 A 和对抗样本数据集 B 上测试, YOLOv3 目标检测器在对抗样本数据集 C 上测试。我们设置步长为 30, TOP-K 表示只取高频信息占比最高的前 K 个像素块用作遮罩, 检测是否为对抗补丁。实验结果如表 3 所示。*BlockSize* 表示划分块的大小, 依次取 40×40、50×50、60×60、70×70。

从表 3 中我们可以看到, 可以发现当划分块的大小固定, K 值越大, 防御成功率越高, 这是因为考虑到了更多可能包含补丁的像素块。但是这并不意味着越大越好, 因为划分块很大意味着遮罩很大, 会将图像中的对象覆盖。如果遮罩足够大, 将人的大部分甚至全身都覆盖了, 检测器就发现不了样本中的人。还可以发现当 K 值固定时, *BlockSize* 不同防御

表 3 将样本切分为不同大小的像素块的防御成功率

Table 3 The defense success rate of dividing the sample into pixel blocks of different sizes

像素块大小(*BlockSize*)		40×40	50×50	60×60	70×70
对抗样本数据集 A	TOP 1	11.47%	11.47%	15.23%	19.36%
	TOP 3	23.68%	25.19%	29.70%	34.97%
	TOP 5	35.90%	41.93%	51.13%	53.95%
	TOP 10	64.29%	74.81%	96.62%	86.10%
对抗样本数据集 B	TOP 1	67.22%	57.42%	55.50%	51.44%
	TOP 3	79.90%	77.03%	77.51%	76.79%
	TOP 5	83.49%	82.30%	84.93%	86.37%
	TOP 10	86.84%	88.76%	90.67%	94.50%
对抗样本数据集 C	TOP 1	66.44%	65.21%	77.22%	85.00%
	TOP 3	75.31%	74.62%	83.90%	89.90%
	TOP 5	75.58%	77.90%	84.31%	91.95%
	TOP 10	77.63%	79.67%	84.31%	91.95%

成功率也不同。取包含高频信息最高的前 10 个像素块时, 在对抗样本数据集 A 中, 当 *BlockSize* 等于 60, 防御成功率为 96.2%; 在对抗样本数据集 B 中, 当 *BlockSize* 等于 70, 防御成功率为 94.50%; 在对抗样本数据集 C 中, 当 *BlockSize* 等于 70, 防御成功率为 91.95%。

4.5 复杂背景对方法的影响

我们的方法采用先分割再通过快速傅里叶变换评估每个像素块含有高频信息多少, 对每个样本只取前 K 个变化最剧烈的像素块验证。因此如果背景复杂, 包含变化剧烈的图案, 那么经过快速傅里叶变换和排序处理后排在前面位置的像素块可能是背景而不是对抗补丁, 会对防御结果造成影响。但是只要对抗补丁在前 K 个像素块里, 我们的防御方法依然有效。为了研究防御方法在复杂背景中的防御效果, 我们采用贴纸的方式在对抗样本中数据集中添加烟花和鲜花模拟复杂背景。我们在对抗样本数据集A中添加烟花和鲜花, 记为对抗样本数据集D; 在对抗样本数据集 B 中添加烟花和鲜花记为对抗样本数据集E; 在对抗样本数据集C中添加烟花和鲜花记为对抗样本数据集 F 良性样本数据集添加烟花和鲜花记为良性样本数据集 G。如图 5 所示: 其中(a)是良性样本数据集添加复杂背景; (b)是对抗样本数据集A添加复杂背景; (c)是对抗样本数据集B添加复杂背景; (d)是对抗样本数据集 C 添加复杂背景。模型和类别同 4.2.2, 参数设置同 4.2.3。

(a) (b) (c) (d)

图 5 复杂背景数据集

Figure 5 Datasets with complex backgrounds

实验结果如表 4 所示, 从表中可以看到当 K=10 时, 加了防御方法后的 YOLOv2 在对抗样本数据集 D 和在对抗样本数据集 E 中防御成功率分别为 63.16%和 89.47%, 当 K=20 时, 防御成功率分别提升到 94.93%和 95.93%。当 K=10 时, 加了防御方法后的 YOLOv3 在对抗样本数据集 F 中防御成功率为 88.81%, 当 K=20 时, 防御成功率提升到 94.27%。当 K=10 时, 加了防御方法后的 YOLOv2 在良性样本数据集 G 中检测成功率为 100%。当 K=10 时, 加了防御方法后的 YOLOv3 在良性样本数据集 G 中的检测成功率为 100%。在 3 个复杂背景的对抗样本数据集中防御成功率相比正常背景对抗样本数据集都有所降低, 均超过 94%, 均高于对比方法在正常背景的对抗样本数据集中的防御成功率。在复杂背景的良性样本数据集中, 加了防御算法的两个目标检测模型均能达到 100%的检测成功率。

复杂背景对我们的防御方法存在一定的影响, 经快速傅里叶变换后排在前面的像素块有可能是背景而不是对抗补丁。但只要补丁出现在前 K 个像素块中, 就能破坏对抗区域, 检测到被隐藏的人。因此在复杂背景的样本中, 只要增大 K 值, 依然可以以高成功率防御补丁对抗攻击。

表 4 复杂背景下的防御成功率

Table 4 The defense success rate in complex background

	YOLOv2			YOLOv3	
	良性样本数据集 G	对抗样本数据集 D	对抗样本数据集 E	良性样本数据集 G	对抗样本数据集 F
未加防御	100%	0	0	100%	0
OURS[top 10]	100%	63.16%	89.47%	100%	88.81%
OURS[top 15]	100%	86.28%	93.30%	100%	93.32%
OURS[top 18]	100%	92.67%	95.69%s	100%	94.27%
OURS[top 20]	100%	94.93%	95.93%	100%	94.27%

5 讨论

我们提出的基于快速傅里叶变换思想的防御方法, 能够有效防御物理空间中针对目标检测器的补丁对抗攻击, 对原模型的影响小。实验中我们用两个最先进的目标检测器在三个数据集上对防御方法进行评估, 展示了防御方法对原模型的影响小于同类方法, 且防御效果优于同类方法。

我们提出的防御方法有两个参数可以设置, 分别是图像切分时像素块的大小 *BlockSize* 和检测时取高频信息最高像素块的数量 *K*。*BlockSize* 的选择直接决定防御的效果, 原因如下: 如果分割的像素块太小, 会导致补丁只有小部分被破坏, 依然具有攻击性; 如果分割的像素块太大, 有可能待检测对象会被遮罩覆盖, 即使完全破坏了对抗补丁目标检测器也不能检测到人。因此本防御方法在目标在图像中的尺寸大小差异不大, 即待检测者距离图像传感器的距离变化不大的场合中效果最佳。*K* 值需根据防御成功率要求和背景中高频信息选择, 若需要高防御成功率则设置为大值, 若背景高频信息多则也需要设置为大值。如果参数 *BlockSize* 可以根据目标在图像中的尺寸大小自适应调节, 就可以打破当前的局限性, 得到更宽松的应用条件, 这也是我们今后努力的方向。

6 结论

为了防护目标检测算法不受补丁对抗攻击的消极影响, 能够安全地部署, 我们提出了基于快速傅里叶变换的通用补丁对抗攻击防御方法。既考虑到每一个有可能出现补丁的位置, 又不需要用模型检测每一种可能性, 可以节约计算成本和时间。而且我们的方法可以完全破坏对抗区域且不会对良性区域造成影响。此防御方法不受补丁生成算法的影响, 可以在物理空间中有效防御不同的补丁对抗攻击。我们的方法对原模型的不利影响更小, 而且防御成功率更高。在三个对抗样本数据集中, 两个目标检测器都可以检测到所有良性样本中的对象, 而且防御成功率都超过 94%。

参考文献

[1] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. https://arxiv.org/abs/1412.6572

[2] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Visual Classification[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1625-1634.

[3] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528-1540.

[4] 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*, 2020: 997-1005.

[5] Liu X, Yang H R, Liu Z W, et al. DPatch: An Adversarial Patch Attack on Object Detectors[EB/OL]. 2018: arXiv: 1806.02299. https://arxiv.org/abs/1806.02299

[6] Redmon J, Divvala S, Girshick R, et al. You only Look Once: Unified, Real-Time Object Detection[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779-788.

[7] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6517-6525.

[8] Farhadi A, Redmon J. YOLOv3: An incremental improvement[C]. *Computer Vision and Pattern Recognition. Berlin/Heidelberg, Germany: Springer*, 2018: 1804.02767.

[9] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.

[10] Thys S, Van Ranst W, Goedemé T. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 49-55.

[11] Lee M, Kolter Z. On Physical Adversarial Patches for Object Detection[EB/OL]. 2019: arXiv: 1906.11897. https://arxiv.org/abs/1906.11897

[12] Hayes J. On Visible Adversarial Perturbations & Digital Watermarking[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018: 1678-16787.

[13] Naseer M, Khan S, Porikli F. Local Gradients Smoothing: Defense Against Localized Adversarial Attacks[C]. *2019 IEEE Winter Conference on Applications of Computer Vision*, 2019: 1300-1307.

[14] Chiang P Y, Ni R K, Abdelkader A, et al. Certified Defenses for Adversarial Patches[EB/OL]. 2020: arXiv: 2003.06693. https://arxiv.org/abs/2003.06693

[15] Xu Z R, Yu F X, Chen X. DoPa: A Comprehensive CNN Detection Methodology Against Physical Adversarial Attacks[EB/OL]. 2019: arXiv: 1905.08790. https://arxiv.org/abs/1905.08790

[16] Liang B, Li J C, Huang J J. We can always Catch You: Detecting Adversarial Patched Objects WITH or WITHOUT Signature[EB/OL]. 2021: arXiv: 2106.05261. https://arxiv.org/abs/2106.05261

[17] Xu Z R, Yu F X, Chen X. LanCe: A Comprehensive and Lightweight CNN Defense Methodology Against Physical Adversarial Attacks on Embedded Multimedia Applications[C]. *2020 25th Asia and South Pacific Design Automation Conference*, 2020: 470-475.

[18] Xiang C, Mittal P. PatchGuard++: Efficient Provable Attack Detection Against Adversarial Patches[EB/OL]. 2021: arXiv: 2104.12609. https://arxiv.org/abs/2104.12609

[19] Zhou G Z, Gao H C, Chen P, et al. Information Distribution Based Defense Against Physical Attacks on Object Detection[C]. *2020 IEEE International Conference on Multimedia & Expo Workshops*, 2020: 1-6.

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. *3rd International Conference on Learning Representations*, *ICLR 2015 - Conference Track Proceedings*, 2015.

[21] Kurakin A, Goodfellow I, Bengio S. Adversarial Examples in the Physical World[EB/OL]. 2016: arXiv: 1607.02533. https://arxiv.org/abs/1607.02533

[22] Gao L L, Zhang Q L, Song J K, et al. Patch-Wise Attack for Fooling Deep Neural Network[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 307-322.

[23] Chen J Y, Shen S J, Su M M, et al. Black-Box Adversarial Attack on License Plate Recognition System[J]. *Acta Automatica Sinica*, 2021, 47(1): 121-135.
(陈晋音, 沈诗婧, 苏蒙蒙, 等. 车牌识别系统的黑盒对抗攻击[J]. *自动化学报*, 2021, 47(1): 121-135.)

[24] Xu K D, Zhang G Y, Liu S J, et al. Adversarial T-Shirt! Evading Person Detectors in a Physical World[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 665-681.

[25] Kreuk F, Adi Y, Cisse M, et al. Fooling End-to-End Speaker Verification with Adversarial Examples[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 1962-1966.

[26] Li J G, Zhang X F, Xu J Z, et al. Learning to Fool the Speaker Recognition[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2020: 2937-2941.

[27] Carlini N, Wagner D. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text[C]. *2018 IEEE Security and Privacy Workshops*, 2018: 1-7.

[28] Gowal S, Dvijotham K, Stanforth R, et al. Scalable Verified Training for Provably Robust Image Classification[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 4841-4850.

[29] Mirman M, Gehr T, Vechev M. Differentiable abstract interpretation for provably robust neural networks[C]. *International Conference on Machine Learning. PMLR*, 2018: 3578-3586.

[30] Xiang C, Bhagoji A N, Sehwag V, et al. PatchGuard: A Provably Robust Defense Against Adversarial Patches via Small Receptive Fields and Masking[EB/OL]. 2020: arXiv: 2005.10884. https://arxiv.org/abs/2005.10884

[31] McCoyd M, Park W, Chen S, et al. Minority Reports Defense: Defending Against Adversarial Patches[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 564-582.

[32] Ji N, Feng Y F, Xie H D, et al. Adversarial YOLO: Defense Human Detection Patch Attacks via Detecting Adversarial Patches[EB/OL]. 2021: arXiv: 2103.08860. https://arxiv.org/abs/2103.08860

[33] Yang Z L, Li B, Chen P Y, et al. Characterizing Audio Adversarial Examples Using Temporal Dependency[EB/OL]. 2018: arXiv: 1809.10875. https://arxiv.org/abs/1809.10875

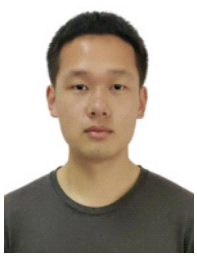
[34] Guo C, Rana M, Cisse M, et al. Countering Adversarial Images Using Input Transformations[EB/OL]. 2017: arXiv: 1711.00117. https://arxiv.org/abs/1711.00117

翔云 于 2014 年在密西根大学电子信息专业获得博士学位。现任浙江工业大学讲师。研究领域为人工智能安全、嵌入式系统。研究兴趣包括: 机器学习、网络空间安全。Email: xiangyun@zjut.edu.cn

李香玉 于 2018 年在三峡大学电子信息科学与技术专业获得学士学位。现在浙江工业大学电子信息专业攻读专业硕士学位。研究方向为可信 AI 安全, 研究兴趣包括: AI 安全, 人工智能鲁棒性。Email: 2084921287@qq.com

韩瑞鑫 于 2020 年在宜春学院自动化专业获得学士学位。现在浙江工业大学电子信息专业攻读硕士学位。研究领域为 AI 安全。研究兴趣包括: 深度学习、对抗攻击和对抗防御。Email: mr_hanruixin@163.com

徐东伟 于 2014 年在北京交通大学交通安全工程专业获得博士学位。现任浙江工业大学副教授。研究领域为交通信息处理、交通复杂网络、机器学习。研究兴趣包括: 人工智能、信号分析。Email: dongweixu@zjut.edu.cn



陈作辉 于 2019 年在浙江工业大学自动化专业获得学士学位。现在浙江工业大学控制科学与工程专业攻读博士学位。研究领域为 AI 安全。研究兴趣包括: 计算机视觉、AI 应用。Email: czuohui@gmail.com