愚弄自动监控摄像机:攻击人员检测的对抗性补丁

西蒙·泰斯*
simen.thys@student.kuleuven.be

维贝·范·兰斯特*
wiebe.vanranst@kuleuven.be

图恩·戈德梅
toon.goedeme@kuleuven.be

库·鲁汶
比利时鲁汶大学内耶科技学院。

*作者对本文有同等贡献。

摘要

在过去几年中，对机器学习模型的对抗性攻击引起了越来越多的兴趣。通过对卷积神经网络的输入进行细微的改变，网络的输出可以被改变以输出完全不同的结果。第一种攻击通过稍微改变输入图像的像素值来欺骗分类器输出错误的类。其他方法试图学习可以应用于对象的“补丁”,以欺骗检测器和分类器。这些方法中的一些还表明这些攻击在现实世界中是可行的，即通过修改对象并用摄像机拍摄它。然而，所有这些方法都针对几乎不包含类内变化的类(例如停车标志)。然后，使用对象的已知结构在其顶部生成一个对抗性补丁。

在这篇文章中，我们提出了一种方法来生成具有大量类内变化的目标，即人的通用补丁。目标是生成一个补丁，能够成功地隐藏一个人从一个人检测器。例如，一种可能被恶意用来干扰监控系统的攻击，入侵者可以在不被发现的情况下偷偷摸摸地四处走动，方法是在他们的身体前面拿一个小纸板，对准监控摄像机。

从我们的结果可以看出，我们的系统能够显著降低人检测器的准确度。我们的方法在现实生活中也能很好地工作，在现实生活中，补丁是由摄像机拍摄的。据我们所知，我们是第一个尝试这种攻击的目标具有高水平的类内多样性，如人。

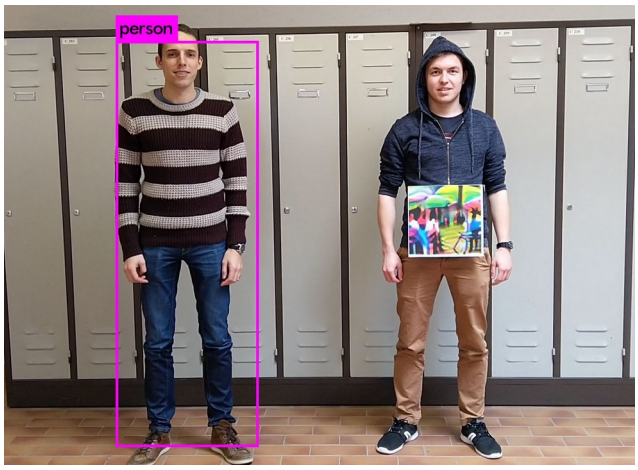

图 1:我们创建了一个对抗性的补丁，它能够成功地隐藏人，使其不被人探测器发现。左图:成功检测到没有补丁的个人。右图:拿着补丁的人被忽略。

1. 介绍

卷积神经网络(CNN)的兴起在计算机视觉领域取得了巨大成功。CNN 学习图像的数据驱动的端到端管道已被证明在广泛的计算机视觉任务中获得了最佳结果。由于这些结构的深度，神经网络能够在网络的底部(数据进入的地方)学习非常基本的过滤器，在顶部学习非常抽象的高级特征。为了做到这一点，典型的 CNN 包含数百万个学习参数。虽然这种方法产生了非常精确的模型，但可解释性却大大降低了。理解为什么网络会把一个人的图像归类为

人很辛苦。这个网络通过观察许多其他人的照片来了解一个人的长相。通过评估该模型，我们可以通过将该模型与人类标注的图像进行比较来确定该模型对于人检测工作得有多好。然而，以这种方式评估模型只能告诉我们检测器在特定测试集上的表现如何。该测试集通常不包含旨在以错误的方式操纵模型的示例，也不包含特别旨在欺骗模型的示例。这对于不太可能受到攻击的应用来说没什么问题，例如老年人跌倒检测，但对于安全系统来说却是一个真正的问题。安全系统的人员检测模型中的一个漏洞可能被用来绕过监视摄像机，该摄像机被用来在建筑物中进行非法闯入防范。

在本文中，我们强调了对人员检测系统的这种攻击的风险。我们创造了一个小的(大约 40 厘米 40 厘米)“广告系列补丁”被用来作为一个斗篷
装置隐藏人从物体探测器。一场示威-
图中显示了这种情况 1。

本文的其余部分结构如下 2 回顾了对抗性攻击的相关工作。截面 3 讨论我们如何生成这些补丁。在第节 4
我们在 Inria 数据集上对我们的贴片进行了定量评估，并在手持贴片时拍摄的真实视频片段上进行了定性评估。我们在第一节中得出结论 5。源代码可从以下网址获得:https://gitlab.com/EAVISE/adversarial-yolo

2. 相关著作

随着 CNN 越来越受欢迎，对 CNN 的敌对攻击在过去几年中也越来越受欢迎。在本节中，我们将回顾这类攻击的历史。我们首先讨论对分类器的数字攻击，然后讨论对人脸识别和对象检测的真实攻击。然后，我们简单地讨论一下目标检测器 YOLOv2，它在本文中是我们攻击的目标。

对分类任务的对抗性攻击早在 2014 年 Bigio et at。[2]显示了对抗性攻击的存在。之后，Szegedy 等人[19]成功地为分类模型生成了对抗性攻击。他们使用一种方法，能够欺骗网络对图像进行错误分类，同时只稍微改变图像的像素值，这样人眼就看不到这种变化。随后，Goodfellow 等人[9]创建一个更快的梯度符号方法，使其更实用(更快)地在图像上生成对抗性攻击。而不是像[19]，他们在更大的一组图像中找到能够对网络进行攻击的单个图像。在[14]，Moosavi-Dezfooli 等人提出了一种能够通过改变图像来产生攻击的算法
比以前更少也更快。他们使用超平面来模拟输入图像的不同输出类之间的边界。Carlini 等人[4]提出了另一种对抗性攻击，使用优化方法，与已经提到的攻击相比，它们在准确性和图像差异(使用不同的标准)方面都有所改进。在[3 布朗等人创造了一种方法，它不是改变像素值，而是生成可以数字化放置在图像上的补丁，以欺骗分类器。他们不是使用一个图像，而是使用多种图像来建立内部的竞争。在[8] Evtimov 等人提出了一种用于分类的真实攻击。他们的目标是停车标志分类的任务，由于停车标志可能出现的不同姿态，这被证明是具有挑战性的。他们生成一个标签，可以贴在停车标志上，使其无法识别。Athalye 等人[1]提出了一种优化 3D 模型纹理的方法。不同姿态的图像被显示给优化器，以建立对不同姿态和光照变化的鲁棒性。然后使用 3D 打印机打印出最终的物体。穆沙维-德兹夫利的作品 13]提出了一种生成单个通用图像的方法，该图像可用作不同图像上的相反系列扰动。通用对抗图像也显示了对不同检测器的鲁棒性。

用于人脸识别的真实世界对抗性攻击真实世界对抗性攻击的示例在[17].Sharif 等人演示了如何使用印刷眼镜来欺骗面部识别系统。为了保证鲁棒性，眼镜需要在各种不同的姿势下工作。为了做到这一点，他们优化了眼镜上的印刷，这样他们就可以处理大量的图像，而不仅仅是单一的图像。它们还包括非印刷适性评分(NPS)，以确保图像中使用的颜色可以由打印机显示。

用于目标检测的真实世界对抗性攻击陈等 5]提出了针对目标检测的真实攻击。他们的目标是在更快的 R-CNN 检测器中检测停车标志 16].喜欢[1]，他们使用对变换的期望(EOT)的概念(对图像进行各种变换)来建立对不同姿态的鲁棒性。我们最近发现的欺骗现实世界中物体探测器的工作是 Eykholt 等人的工作[18].在影片中，他们再次瞄准停车标志并使用 YOLOv2 [15]检测器进行白盒攻击，他们在停车标志的整个红色区域填充一个图案。他们还在 Faster-RCNN 上进行评估，发现他们的攻击也转移到其他检测器。

与这项工作相比，所有针对对象检测器的攻击都集中在具有固定视觉模式的对象上，如交通标志，并且没有考虑类内变化。据我们所知，以前的工作没有提出

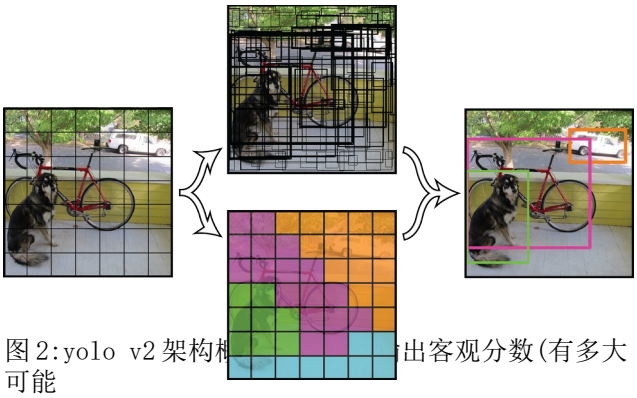

图 2:yolo v2 架构[...]出客观分数(有多大可能

该检测包含一个对象)，如中上图所示，以及一个类得分(哪个类在边界框中)，如中下图所示。图片来源:https://github.com/pjreddie/darknet/维基/YOLO:-实时目标检测

一种适用于不同类别(如人)的检测方法。

目标检测在本文中我们的目标是流行的YOLOv2 [15]物体探测器。YOLO 属于更大类别的单次拍摄物体探测器(连同像 SSD 这样的探测器[12])，其中边界框、对象分数和类分数是通过在网络上进行单次传递来直接预测的。YOLOv2 是完全卷积的，输入图像被传递到网络，在网络中，各层将其缩小到分辨率比原始输入分辨率小 32 倍的输出网格。此输出格网中的每个像元包含五个预测(称为“锚点”)，其边界框包含不同的纵横比。每个锚点包含一个向量-

tor

[xoffset, yoffset, w, h, pobj, pcls1, pcls2, ..., pclsn]。xoffset 和 *yoffset 是边界框 com-的中心位置*与当前锚点相比，w 和 h 是边界框的宽度和高度，pobj 是该锚点包含对象的概率，pcls1 到 pclsn 是使用交叉熵损失学习的对象的类得分。数字 2 显示了该体系结构的概述。

3. 生成对抗 per- son 检测器的对抗性补丁

这项工作的目标是创建一个系统，能够生成可打印的敌对补丁，可用于欺骗人员探测器。如前所述，陈等人[5]和 Eykholt 等人[18]已经表明，在现实世界中，对目标探测器的敌对攻击是可能的。在他们的工作中，他们瞄准停止标志，在这项工作中，我们关注不同于站牌的统一外观，人可以有更多的变化。使用优化过程(在图像像素上),我们试图找到一个补丁，在一个大数据集上，有效地降低人检测的准确性。在本节中，我们将深入解释生成这些对抗性补丁的过程。

我们的优化目标由三部分组成:

- *不可印刷性分数[*17]，这是一个因素，代表我们的补丁中的颜色可以由普通打印机表现得多好。给出者:

$$Lnps = \sum_{ppatch \in p} \text{部}_{cprint \in C} \; | \; p\ patch\ c\ print \; |$$

其中 ppatch 是补丁 P 中的一个像素，而 cprint 是一组可打印颜色 c 中的一种颜色。这种损失有利于图像中与我们的可打印颜色组中的颜色接近的颜色。

- *1tv 图像中的总变化，如[*17].这种损失确保了我们的优化器偏爱具有平滑色彩过渡的图像，并防止噪声图像。我们可以从面片 P 计算 Ltv，如下所示:

$$Ltv = \sigma((pi,\ j\ pi+1,\ j\ )2 + (pi, \text{q}j\ pi,\ j+1)2$$

我，j

如果相邻像素相似，得分较低，如果相邻像素不同，得分较高。

- *图像中的最大客观分数。我们补丁的目标是隐藏图像中的人物。为此，我们训练的目标是最小化检测器输出的对象或类分数。这一分数将在本节的后面进行深入讨论。*

从这三部分得出我们的总损失函数:

$$L = \alpha\, Lnps + \beta\, Ltv + Lobj$$

我们根据经验确定的系数 α 和 β 来计算三个损失的总和，并使用 Adam [10]算法。

我们优化器的目标是最小化总损失

长度在优化过程中，我们冻结了网络中的所有权重，只改变补丁中的值。在该过程开始时，补丁被初始化为随机值。

数字 3 概述了如何计算对象损失。遵循相同的过程来计算类别概率。在本节的剩余部分，我们将深入解释这是如何实现的。

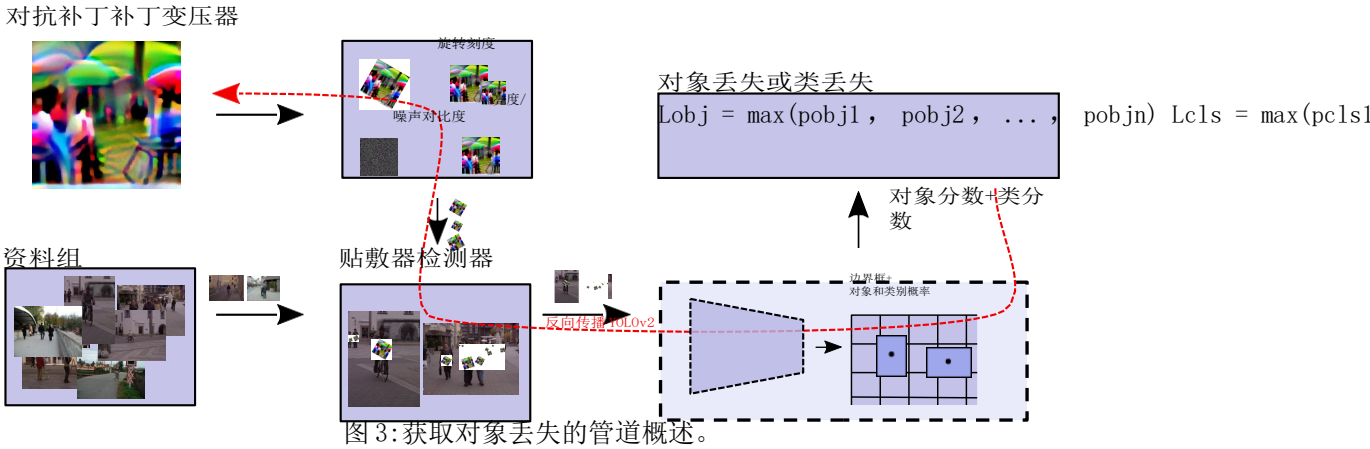

图 3:获取对象丢失的管道概述。

(a)所得到的学习补丁(b)通过最小化分类的优化过程生成的另一个补丁，该优化过程略微和客观性评分。

c)通过最小化- (d)最小化分类生成的补丁客观分数。不同的参数。仅得分。

图 4:使用不同方法的补丁示例。

3.1. 最小化检测器输出的概率

如第节所述 2YOLOv2 对象检测器输出一个单元格网格，每个单元格包含一系列锚点(默认为五个)。每个锚点包含边界框的位置、对象概率和类别分数。让探测器忽略我们排除的人-

用三种不同的方法:我们可以最小化类人的分类概率(例如图中的小块)4d，最小化客观性分数(图 4c)，或两者的组合(图 4b 和 4a).我们尝试了所有的方法。最小化班级分数倾向于将班级成员切换到不同的班级。在我们使用在 MS COCO 数据集上训练的 YOLO 检测器的实验中[11]，我们发现生成的补丁被检测为 COCO 数据集中的另一个类。数字 4a 和 4b 是在图形的情况下，取类与对象概率的乘积的一个例子 4a 学习过的补丁最终看起来像一只泰迪熊，它在视觉上也是重新组装的。班上的“泰迪熊”似乎压倒了

类“人”。然而，因为该补丁开始类似于另一个类，所以该补丁不太可转移到在不包含该类的数据集上训练的其他模型。

我们提出的另一种最小化客观分数的方法没有这个问题。虽然我们只是在优化过程中把它放在人的上面，但是得到的补丁对于某个职业来说没有其他方法那么具体。数字 4c 显示了此类补丁的示例。

3.2. 准备培训数据

与之前针对停车标志所做的工作相比[5,18]，为阶级人士制造对立的补丁是多
更具挑战性:

- 人们的外貌变化更大:衣服、肤色、身材、姿势...相比之下，停止标志总是有相同的八角形，通常是红色的。
- 人可以出现在很多不同的情境中。停车标志大多出现在街道一侧的相同环境中。
- 一个人的外貌会因人而异——

取决于一个人是背对镜头还是面向镜头。

- 一个人身上没有一个固定的地方可以贴上补丁。在停车标志上，很容易计算出补丁的确切位置。

在本节中，我们将解释如何应对这些挑战。首先，不是像在[]中那样人工修改目标对象的单个图像并进行不同的变换 5,18]，我们用的是不同人的真实影像。我们的工作流程如下:我们首先在图像数据集上运行目标人物检测器。这产生了边界框，根据检测器显示图像中人出现的位置。在相对于这些边界框的固定位置上，我们然后在不同的变换下将我们的补丁的当前版本应用于图像(这将在第节中解释)3.3).然后，生成的图像(与其他图像一起分批)被输入检测器。我们测量仍然被检测到的人的分数，我们用它来计算损失函数。使用整个网络上的反向传播，优化器然后进一步改变补丁中的像素，以便更好地欺骗检测器。

这种方法的一个有趣的副作用是，我们不局限于带注释的数据集。任何视频或图像集合都可以输入到目标检测器中，以生成边界框。这使得我们的系统也可以进行更有针对性的攻击。当我们从目标环境中获得数据时，我们可以简单地使用该镜头来生成特定于该场景的补丁。这可能比一般数据集执行得更好。

在我们的测试中，我们使用 Inria 的图像[6]数据集。这些图像更适合全身行人，更适合我们的监控摄像机应用。我们承认像 COCO 女士这样更具挑战性的数据集[11]和 Pascal VOC [7]是可用的，但是它们包含了太多的人出现的变化(例如一只手被标注为 person)，这使得我们很难将补丁放在一个一致的位置上。

3.3. 使补丁更加健壮

在本文中，我们的目标是必须在现实世界中使用的补丁。这意味着它们首先被打印出来，然后被摄像机拍摄下来。当你这样做时，许多因素影响补片的外观:照明可能改变，补片可能稍微旋转，补片相对于人的大小可能改变，相机可能增加噪声或稍微模糊补片，视角可能不同。。。为了尽可能地考虑到这一点，我们在将补丁应用到图像之前对其进行了一些变换。我们进行以下随机变换:

- 贴片每向旋转 20 度。

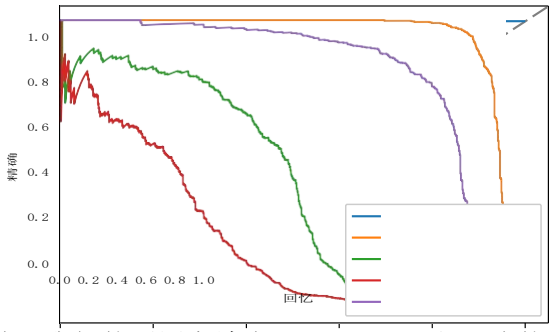

图 5:我们的不同方法(OBJ-CLS，OBJ 和 CLS)的 PR 曲线，与随机补丁(噪声)和原始图像(干净)相比较。

- 补丁随机放大和缩小
- 随机噪声被放在面片的顶部。
- 补丁的亮度和对比度是随机变化的

在整个过程中，重要的是要注意，必须保持对所有朝向补片的操作计算向后梯度的可能性。

4. 结果

在本节中，我们将评估补丁的有效性。我们通过将补丁应用于 Inria 测试集来评估我们的补丁，使用的过程与我们在训练期间使用的过程相同，包括随机变换。在我们的实验中，我们试图最小化一些可能隐藏人的不同参数。作为对照，我们还将我们的结果与包含随机噪声的补丁进行比较，该补丁以与我们的随机补丁完全相同的方式进行评估。数字 5 显示了我们不同补丁的结果。OBJ-CLS 的目标是最小化目标分数和类别分数的乘积，在 OBJ 只最小化目标分数，在 CLS 只最小化类别分数。噪声是随机噪声的控制补丁，干净是没有应用补丁的基线。(因为边界框是通过在数据集上运行相同的检测器生成的，所以我们得到了完美的结果。)从该 PR 曲线中，我们可以清楚地看到生成的贴片(OBJ-CLS、OBJ 和 CLS)与作为对照的随机贴片相比的影响。我们还可以看到，与使用类得分相比，最小化对象得分(OBJ)具有最大的影响(最低平均精度(AP))。

确定用于检测的 PR 曲线上的良好工作点的典型方法是在

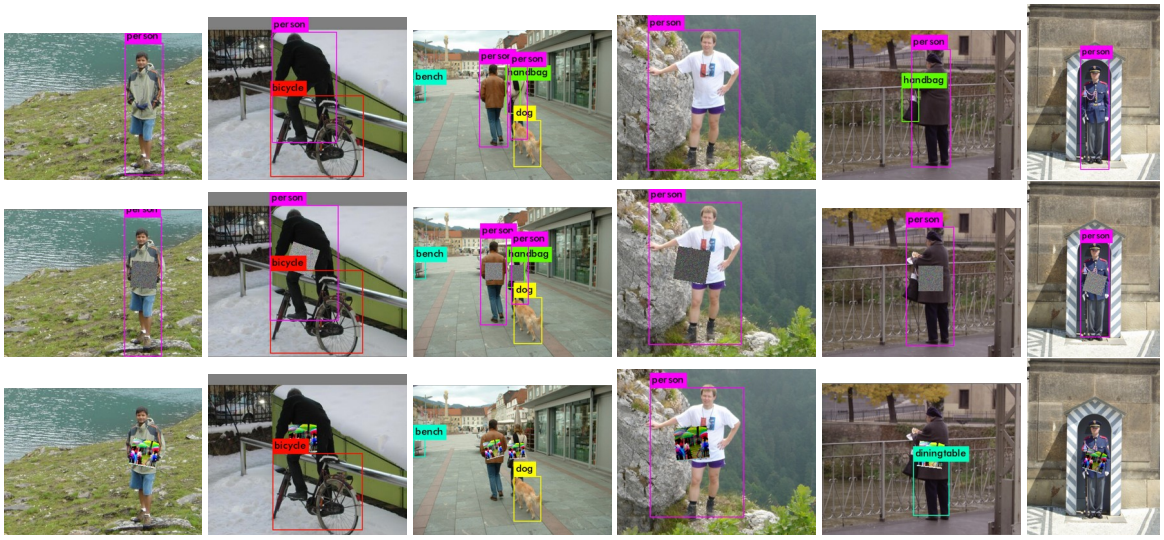

图 6:我们在 Inria 测试集上的输出示例。

进场召回 (%)
清洁 100
噪音 87.14
OBJ-CLS 39.31
目标文件 **26.46**
CLS 77.58

表 1:不同召回方法的比较。不同的方法规避警报的效果如何？

PR 曲线(图中的虚线 5)，看它和 PR 曲线的交点。如果我们对干净的 PR-曲线这样做，我们可以使用在该工作点得到的阈值(在我们的例子中为 0.4)作为参考，以查看我们的方法将降低检测器的召回多少。换句话说，我们会问这样一个问题:有多少由监控系统产生的警报是通过使用我们的补丁来规避的？桌子 1 使用图中的缩写显示了该分析的结果 5。由此我们可以清楚地看到，使用我们的补丁(OBJ-CLS、OBJ 和 CLS)显著降低了生成的警报数量。

数字 6 显示了应用于 Inria 测试集中某些图像的补丁示例。我们首先将 YOLOv2 检测器应用于没有补片的图像(第 1 行)、具有随机补片的图像(第 2 行)以及具有我们生成的最佳补片 OBJ 的图像(第 3 行)。在大多数情况下，我们的贴片能够成功地将人从探测器中隐藏起来。在不是这种情况的情况下，补片不与人的中心对齐。这可以通过以下事实来解释，即在优化期间，补片也仅位于由边界框确定的人的中心。

在图中 7 我们测试我们的印刷版本补丁在现实世界中有效。总的来说，该补丁似乎工作得很好。由于面片是在相对于边界框的固定位置上训练的，因此将面片保持在正确的位置上似乎非常重要。演示视频可在以下网址找到:https://youtu.be/MIbFvK2S9g8。

5. 结论

在这篇文章中，我们提出了一个系统来产生敌对补丁的人探测器可以打印出来，并在现实世界中使用。我们通过优化图像来最小化检测器输出中与人的出现相关的不同概率。在我们的实验中，我们比较了不同的方法，发现最小化对象丢失创建了最有效的补丁。

从我们用打印出来的补丁进行的真实世界测试中，我们还可以看到，我们的补丁在隐藏人免受物体探测器的攻击方面非常有效，这表明使用类似探测器的安全系统可能容易受到这种攻击。

我们相信，如果我们将这种技术与复杂的服装模拟相结合，我们可以设计出一种 t 恤印花，这种印花可以让自动监控摄像头几乎看不见一个人(使用 YOLO 探测器)。

6. 未来的工作

在未来，我们希望通过使其更加健壮来扩展这项工作。做到这一点的一种方式是通过对输入数据进行更多(仿射)变换或使用模拟数据(即，将补片作为纹理应用于人的 3D 模型上)。另一个可以做更多工作的领域是可移植性。我们当前的补丁不能很好地转移到

图 7:使用我们补丁的打印版本的真实镜头。

完全不同的架构，如更快的 R-CNN [16]，同时针对不同的架构进行优化可能会对此有所改进。

参考

[1] A.阿萨里，1 .英斯特朗，a .易勒雅斯和郭国光。综合有力的对立例子。arXiv 预印本 arXiv:1707.07397，2017。2

[2] B.放大图片作者:Michael b .
页（page 的缩写）拉斯科夫，贾辛托和花小蕾。测试时对机器学习的规避攻击。在关于数据库中的机器学习和知识发现的欧洲联合会议上，第 387-402 页。斯普林格，2013。2

[3] T.B. Brown、D. Mane、A. Roy、M. Abadi 和 J. Gilmer。敌对补丁。arXiv 预印本 arXiv:1712.09665，2017。2

[4] 名词（noun 的缩写）卡里尼和 d .瓦格纳。评估神经网络的鲁棒性。2017 年 IEEE 安全与隐私研讨会(SP)，第 39 - 57 页。IEEE，2017。2

[5] 南-T .陈、c .科尼利厄斯、j .马丁和 D. H .周。对更快的 r-cnn 目标检测器的攻击。arXiv 预印本 arXiv:1804.05810，2018。2,3,4,5

[6] 名词（noun 的缩写）达拉和 b .特里格斯。用于人体检测的方向梯度直方图。计算机视觉和模式识别国际会议(CVPR'05)，第 1 卷，第 886 - 893 页。IEEE 计算机学会，2005 年。5

[7] 米（meter 的缩写））埃灵厄姆，1 .范古尔，C. K .威廉姆斯，j .温，和
A.塞斯曼。pascal 视觉对象类(voc)挑战。国际计算机视觉杂志，88(2):303 - 338，2010。5

[8] 埃夫蒂莫夫、埃霍尔特、费尔南德斯、科尼奥、李、
A.普拉卡什，拉赫马蒂和宋。对深度学习模型的强大物理世界攻击。arXiv 预印本 arXiv:1707.08945，1:1，2017。2

[9] I. J .古德费勒 j .史伦斯和 c .塞格迪。解释和利用对立的例子。arXiv 预印本 arXiv:1412.6572，2014。2

[10] D.金玛和巴。亚当:一种随机优化方法。arXiv 预印本 arXiv:1412.6980，2014。3

[11] T.-林、梅尔、贝隆吉、海斯、佩罗娜、拉马南、多拉和兹尼克。Microsoft coco: Com-上下文中的公共对象。在欧洲计算机视觉会议上，第 740-755 页。斯普林格，2014。4,5

[12] W.刘、安盖洛夫、尔汉、塞格迪、里德
Y.傅和 A. C .伯格。Ssd:单次多盒探测器。在欧洲计算机视觉会议上，第 21-37 页。
斯普林格，2016。3

[13] 南-穆萨维-德兹富利先生、法齐先生、法齐先生和
页（page 的缩写）弗罗沙。普遍的对抗性干扰。IEEE 计算机视觉和模式识别会议论文集，第 1765-1773 页，2017 年。2

[14] 南-穆萨维-德兹富利、法齐和弗罗萨德。深度欺骗:一种简单而准确的欺骗深度神经网络的方法。IEEE 计算机视觉和模式识别会议论文集，第 2574-2582 页，2016 年。2

[15] J.雷德蒙和法尔哈迪。Yolo9000:更好、更快、更强。IEEE 计算机视觉和模式识别会议论文集，第 7263 - 7271 页，2017 年。2,3

[16] 南任，何国光，R. Girshick，和 J. Sun。更快的 r-cnn:用区域建议网络实现实时目标检测。《神经信息处理系统进展》，第 91 - 99 页，2015 年。2,7

[17] 米（meter 的缩写））Sharif、S. Bhagavatula、L. Bauer 和 M. K. Reiter。承认一项罪行:对最先进的人脸识别技术进行真实而隐秘的攻击。2016 年 ACM SIGSAC 计算机和通信安全会议论文集，第 1528-1540 页。ACM，2016。2,3

[18] D.宋、k .艾克霍尔特、I .叶夫提莫夫、e .费尔南德斯、b .李、
A.拉赫马蒂、f .特拉默、a .普拉卡什和 t .科尼奥。物体探测器的物理对抗例子。2018 年第 12 届 USENIX 进攻性技术研讨会(WOOT 18)。2,3,4,5

[19] C.塞格迪、w .扎伦巴、I .苏茨基弗、j .布鲁纳、d .埃汉、
I .古德费勒和 r .弗格斯。神经网络的有趣特性。arXiv 预印本 arXiv:1312.6199，2013。2