敌对伪装:用自然风格隐藏物理世界的攻击

段冉杰马王
1 温伯恩科技大学 2 墨尔本大学 3 上海交通大学

摘要

众所周知，深度神经网络(DNNs)容易受到对立示例的攻击。现有的工作主要集中在通过小的和察觉不到的扰动创建的数字对抗例子，或者通过人类观察者容易识别的大的和不太真实的失真创建的物理世界对抗例子。在这篇文章中，我们提出了一种新的方法，称为对抗伪装(AdvCam)，将物理世界的对抗例子加工和伪装成对人类观察者来说似乎合法的自然风格。具体来说，AdvCam将大的对抗性扰动转化为定制的风格，然后“隐藏”在目标对象或非目标背景中。实验评估表明，在数字和物理世界场景中，AdvCam制作的敌对示例具有良好的伪装性和高度的隐蔽性，同时在欺骗最先进的DNN图像分类器方面仍然有效。因此，AdvCam是一种灵活的方法，可以帮助设计隐形攻击来评估DNNs的鲁棒性。AdvCam还可以用来保护私人信息不被深度学习系统检测到。密码

1. 介绍

深度神经网络(DNNs)是一系列功能强大的模型，已广泛用于各种人工智能系统，并在许多应用领域取得了巨大成功，如图像分类[13]、语音识别[26]、自然语言处理[33]和自动驾驶[6]。然而，众所周知，DNNs容易受到敌对示例(或攻击)的攻击，这些攻击是通过在正常示例上添加精心设计的扰动而制造的[25，12，20，2，27，28]。这引起了对以下问题的严重关切
安全关键型应用[23，8，10，15，21]。例如，如图1所示，添加了精心制作的扰动，类似于停车标志表面的污渍、积雪和变色。一辆配备了最先进的分类器的自动驾驶汽车以非常高的可信度将修改后的停车标志检测为其他物体(我们稍后将检查这种扰动是如何产生的)。

通讯地址:马(xingjun.ma@unimelb.edu.au)

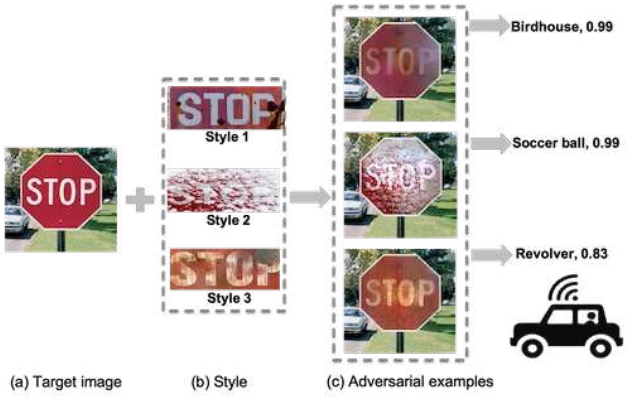

图1:由提议的AdvCam攻击精心制作的伪装敌对示例。给定(a)中的目标图像，对手可以从(b)中选择不同的伪装风格，以在(c)中制造看起来自然发生的敌对例子，但可以欺骗DNN分类器以高置信度做出不正确的预测。

对抗性攻击可以应用于两种不同的设置中:1)数字设置，其中攻击者可以将输入的数字图像直接输入到DNN分类器中；以及2)物理世界设置，其中DNN分类器只接受来自摄像机的输入，并且攻击者只能向摄像机呈现敌对的图像。有三个特性可以用来表征对抗性攻击:1)对抗性强度，它代表欺骗DNNs的能力；2)敌对的隐蔽性，即敌对的干扰是否能被人类观察者察觉；以及3)伪装灵活性，这是攻击者能够控制敌对形象出现的程度。

大多数攻击方法都是针对数字环境开发的，例如投影梯度下降(PGD) [22]、Carlini和Wagner (CW)攻击[4]以及使用生成式对抗网络(Ad- vGAN)制作的对抗示例[30]。对于数字攻击，小扰动通常就足够了。然而，物理世界的攻击需要大的甚至无限制的扰动[17，1]，因为小的扰动太细微，在复杂的物理世界环境中无法被摄像机捕捉到。已经存在了

(a) RP2(b)adv cam(c)adv patch(d)adv cam

图 2:成功的物理世界攻击的例子。
AdvCam 指的是我们提出的对抗性伪装。

超越小扰动的几种攻击方法，如对抗性补丁(AdvPatch) [3]和鲁棒物理扰动(RP2) [10]。表 1 总结了现有攻击的特性，其中⋆⋆表示比⋆.好总而言之，小扰动可以实现隐身，这种小扰动只在数码环境下有用。此外，现有的攻击需要精确的扰动大小来实现隐蔽性，而对于视觉不可感知性和对抗强度来说，很难决定合适的扰动大小，尤其是在物理环境中。此外，当前方法的生成过程是难以控制的，例如，攻击者不能决定对抗性例子的出现。因此，这些方法的伪装灵活性相当有限。针对大扰动的灵活的(然而坚固和隐蔽的)伪装机制仍然是需要解决的公开问题。

表 1:现有攻击和我们的 AdvCam 总结。

攻击数字物理隐形灵活性 PGD	√	×	⋆⋆	⋆
AdvPatch	×	√	⋆	⋆
RP2			⋆	⋆
AdvCam	√	√	⋆⋆	⋆⋆

为了解决这个问题，本文提出了一种新颖的 ad-通用伪装方法(AdvCam)，使用风格转换技术将对立的示例制作成自然的风格。图像的风格是一个抽象的概念，通常指其视觉外观，如颜色和纹理，与其结构信息形成对比[16]。在 AdvCam 中，攻击者可以根据不同的攻击场景定制伪装风格和攻击区域。例如，图 1 显示了我们的 AdvCam 攻击精心制作的几个敌对交通标志。AdvCam 与现有物理攻击的快速对比如图 2 所示。虽然图 2 中的所有示例都成功攻击了 DNNs，但我们可以看到，与 RP2 创建的人工涂鸦相比，AdvCam 能够在停车标志上生成带有自然污渍的恶意干扰，或者与 AdvPatch 生成的带有突出图案的补丁相比，adv cam 能够生成伪装的产品标签。我们提出的 AdvCam 能够生成高度隐蔽的对抗实例，同时对各种物理世界条件具有鲁棒性。

总之，AdvCam 不是扰动限制型 at- tack，因此不会固有地受到现有扰动限制型技术通常要求的有限扰动量的影响。我们定义了一种灵活的机制来诱导自然外观中出现的扰动，这是一种与以前的攻击完全不同的范式。正是这种工作原理的内在差异使得 AdvCam 能够产生比现有方法更真实的图像。

我们在本文中的主要贡献是:

- 我们提出了一种灵活的对抗伪装方法 AdvCam，来构造和伪装对抗实例。
- *AdvCam 允许产生大扰动、可定制的攻击区域和伪装风格。它对于评估 DNNs 在物理世界攻击下的大扰动的脆弱性是非常灵活和有用的。*
- 在数字世界和物理世界场景上的实验表明，AdvCam 伪装的敌对示例具有很高的隐身性，同时在欺骗最先进的 DNN 图像分类器方面仍然有效。

2. 相关著作

2.1. 对抗性攻击

对抗性攻击是通过最大化目标模型(要攻击的模型)的分类误差来生成对抗性的例子[25]。有针对性和无针对性的攻击。针对性攻击就是忽悠网络误分类

敌对的例子进入攻击者期望的类别，而无目标攻击是欺骗网络将敌对的例子误分类到任何不正确的类别[12]。
对抗性攻击既可以在数字环境中直接应用于目标模型，也可以在现实世界环境中应用，在现实世界中，对抗性例子的重新捕获的照片被提供给目标模型[17]。

2.1.1 数字攻击

对抗的例子可以按照对抗梯度的方向，通过一个或多个扰动步骤来制作。这包括经典的快速梯度符号法(FGSM) [12]，基本迭代法(BIM) [17]，最强一阶方法投影梯度下降(PGD) [22]，以及针对可转移攻击的跳过梯度法(SGD) [29]。这些攻击可以是有目标的，也可以是无目标的，它们的扰动被一个小的
范数球 Qp≤，L2 和 L∞是最常用的范数。基于优化的攻击，例如
Carlini 和 Wagner (CW)攻击[4]和 elastic-net (EAD)攻击[7]，直接将扰动最小化为对抗性损失的一部分。

还存在几种无限制的对抗性攻击。这些攻击搜索对图像的可替换成分(属性)的修改，例如颜色[14]、纹理和物理参数[34，18]，同时保留图像的关键成分。然而，这些攻击要么产生大的非自然失真，要么依赖于具有语义注释的训练数据。此外，这些攻击不能生成复杂的对抗模式，因此对于复杂的真实场景是非常有限的。对抗性例子也可以由生成性对抗性网络(GANs)产生[30，24]。然而，由于生成过程难以控制，因此很难用GANs对给定的测试映像进行有针对性的攻击。

现有的大多数攻击通过制造小扰动或修改目标图像的语义属性来实现隐蔽性。然而，一种灵活的伪装机制，可以有效地隐藏具有自然风格的敌对实例，仍然是文献中所缺少的。在本文中，我们通过提出这样一种方法来解决这个问题。

2.1.2 物理世界的攻击

一项研究表明，通过使用手机摄像头打印和重新拍摄，数字对抗的例子仍然可以有效[17]。然而，后续工作发现，由于视点移动、相机噪声和其他自然转换，这种攻击在物理世界条件下不容易实现[1]。因此，强烈的物理世界攻击需要大的扰动，以及包括照明、旋转、透视投影等在内的变换分布的特定适应。AdvPatch攻击允许较大的扰动，不受缩放或旋转的影响，因此可直接用作物理攻击[3]。敌对贴纸和涂鸦也被用于攻击，如交通标志分类器和图像网络分类器在现实世界的场景[10]。其他现实世界的攻击包括敌对的眼镜框架[23]、车辆[35]或t恤[32]，它们可以欺骗人脸识别系统或物体探测器。所有这些物理世界的攻击产生了巨大的扰动来增加对抗的力量，这不可避免地导致巨大的和不现实的扭曲。这大大降低了它们的隐蔽性，如图2所示。

2.2. 神经类型转移

神经风格转移是从纹理转移问题发展而来的，其目标是将源图像的纹理转移到目标图像，同时保留目标图像的结构信息。传统的纹理转换方法主要集中在非参数算法上，即从源纹理中重新采样像素[9]。然而，这些方法受到像素替换的基本限制，即它们不能处理复杂的样式。神经类型转移显示出显著的图像风格化的结果[11]。其中，图像的内容和风格信息可以从卷积神经网络(CNN)学习的特征表示中分离出来。然后，将风格图像的风格信息重新组合到目标图像中，实现风格转换。这项技术已经吸引了几个后续工作的不同方面的改进[5，19]。在本文中，我们将利用这些技术来伪装通用示例。

3. 我们的敌对伪装方法

在这一节中，我们首先概述了对抗性攻击问题和我们提出的伪装方法。然后，我们介绍了我们提出的方法使用的损失函数，以及对物理世界条件的适应。

3.1. 概观

给定具有类别标签y的测试图像$x \in R^m$，DNN分类器$F : R^m \rightarrow \{1, k\}$将图像像素映射到离散标签集和目标类$y_{adv}$，对抗性攻击是为目标图像找到一个对立的例子x'

x通过解决以下优化问题:

$$\begin{aligned} &\text{最小化 } D(x, x') + \lambda L_{adv}(x') \\ &\text{使得 } x' \in [0, 255]^m, \end{aligned} \tag{1}$$

其中D(x，x′)是定义对抗示例的隐蔽性的距离度量，Ladv是通用损失，[0，255]表示有效像素值，λ是调整对抗强度的参数。注意

在隐秘和对抗力量之间有一个权衡。在整个实验中，我们将所有其他参数固定为常数[19]，仅调整对抗强度参数λ。

我们的目标是开发一种机制，将具有较大扰动的对立示例加工和标记成定制的样式，并且攻击者可以灵活地定义攻击区域或样式。我们使用风格转移技术来达到伪装的目的，使用对抗性攻击技术来达到对抗性的强度。最终的损失

是对抗强度的对抗损失Ladv、风格生成的风格损失Ls、保持源图像内容的内容损失Lc和生成局部平滑区域的平滑损失Lm的组合。我们将这种最终损失称为敌对伪装损失:

$$L = (L_s + L_c + L_m) + \lambda L_{adv}, \tag{2}$$

其中括号中的三个损失函数一起用于伪装的目的。我们的方法概述如图3所示。攻击者定义目标图像、目标攻击区域和预期的目标样式。然后，我们提出的AdvCam在预期区域产生具有预期风格的对抗性扰动，如右侧所示

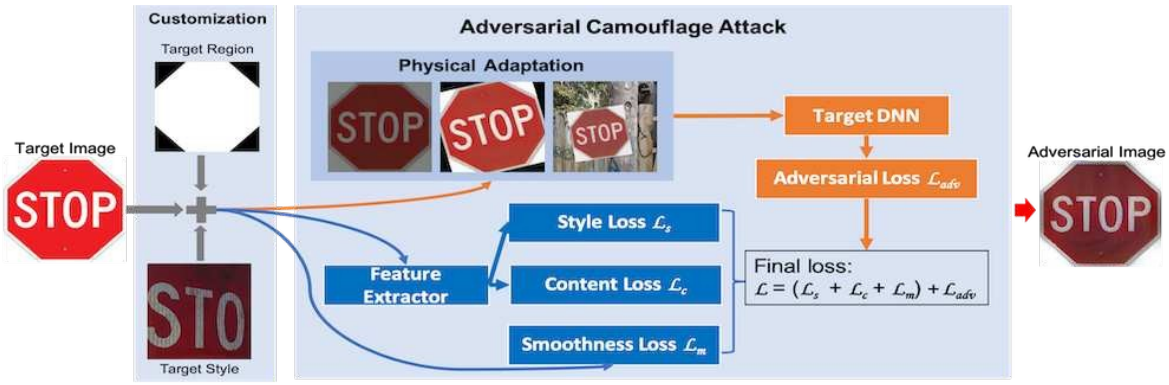

图 3:建议方法的概述。

图 3。使广告示例对各种环境条件(包括光照、旋转等)具有鲁棒性。我们
为生成的 x" in 增加额外的身体适应训练
每一步。

3.2. 敌对伪装损失

3.2.1 风格丧失

对于传统攻击，隐蔽性度量由 D(x，x')= L2 x x' 定义，其中是 Lp 范数，通常使用和 L∞来表示。这是为了约束 pertur-
bations 是“小”的。对于我们提出的伪装
隐秘性由对立示例 x' 和样式参考图像 xs 之间的样式度量来定义。两幅图像之间的风格距离可以由它们的不同来定义
风格表现中的差异:

$$D_{s\,l\,l} = \sum_{l \in l} \| G(f(xs)) - G(f(x')) \|_2^2, \tag{3}$$

其中 f 是可以不同于目标模型的特征提取器(例如公共 DNN 模型)，G 是在 f 的一组样式层提取的深度特征的 Gram 矩阵。由于不同的风格可以在不同的层学习，我们使用网络的所有卷积层作为风格层。为了在目标图像的预期区域中产生风格化的扰动，我们表示定义

攻击区域和非攻击区域分别由 M 和 M 表示。在生成过程的每一步之后，我们屏蔽广告
sarial image x' 乘 M，使只有攻击区域可修改，而非攻击区域不变(x 被 M 屏蔽)。

内容损失

上述风格损失可以在参考风格中形成对立的图像，然而，对立图像的内容可能看起来与原始图像的内容非常不同。原始图像的内容可以通过

内容保存损失:

$$L_{c\,l\,l} = \sum_{l \in l} \| f(x) - f(x') \|_2^2, \tag{4}$$

其中 Cl 表示用于提取内容表示的内容层集合。这是为了确保敌对图像具有与原始图像非常相似的内容
深层表示空间中的图像。我们使用特征提取器网络的更深层作为内容层。

注意，当攻击仅发生在不包含任何特定内容的小区域中时，内容损失是可选的。然而，如果攻击区域包含语义内容，则内容丢失有助于减少敌对图像与其原始版本之间的语义差异。

平滑度损失

可以通过减少相邻像素之间的变化来提高对立图像的平滑度。对于敌对图像 x'，平滑度损失被定义为:

$$Lm = \sum ((xi, xi+1, j)' - (xi, xi, j+1)')^2 1^2, \tag{5}$$

其中 x' 是图像 x' 的坐标(I，j)处的像素。显然，这将促使图像具有低方差
(例如，平滑)局部补丁。我们注意到，如果目标图像和风格图像的表面都已经是平滑的，则平滑度损失在改善隐秘性方面是有限的。但我们仍然建议在物理设置中加入它，正如 Sharif 等人指出的[23]，光滑项对于提高物理环境中对抗性例子的鲁棒性是有用的。

3.2.4 对抗性损失

对于对抗性损失，我们使用以下交叉熵损失:

$$L_{副词} = \begin{cases} -\log(py_{副词}(x')), & \text{用于目标攻击} \\ -\log(py(x')), & \text{对于无目标攻击,} \end{cases} \quad (6)$$

其中 pyadv()是目标模型 F 相对于 yadv 类的概率输出(softmax on logits)。我们注意到
所提出的伪装攻击不限于敌方损失的特定形式，并且可以与现有的攻击方法结合使用。

3.3. 适应物理世界的条件

为了使 AdvCam 生成的对手在物理上可实现，我们在生成伪装对手的过程中对物理条件进行建模。由于物理世界环境经常涉及到条件波动，如视点移动、相机噪声和其他自然变化[1]，我们使用一系列适应措施来适应这些变化的条件。特别是，我们采用了一种类似于期望转换(EOT) [1]的技术，但是没有期望。在谢的工作[31]中，他们还采用了 EOT 来提高对立举例的可迁移性。然而，我们的目标是提高各种物理条件下的通用示例的适应性。因此，我们考虑了用于模拟物理世界条件波动的变换，包括旋转、缩放、颜色偏移(模拟照明变化)和随机背景。

$$x' = \text{部}(x), \quad \text{最大}_{T \in} \ \lambda \ \text{Ladv}(o + T') + (\text{Ls} + \text{Lc} + \text{Lm}) \quad (7)$$

其中 o 代表我们
样本在物理世界中，T 表示旋转、调整大小和颜色偏移的随机变换。

原则上，“视觉”是人类观察者和 DNN 模型的主要感知。通过根据原始图像 x 和背景图像 o 来调整样式参考图像 xs，所提出的伪装攻击可以制造出能够欺骗人类观察者和 DNN 模型的高度隐蔽的对抗示例。

4. 实验评估

在本节中，我们首先概述实验设置。然后，我们通过消融研究分析 AdvCam 攻击。随后，我们通过人类感知研究评估了 Ad- vCam 对数字攻击的伪装性能，并给出了几个 AdvCam 制作的高隐身性对抗实例。我们最后进一步执行物理世界攻击。

4.1. 实验装置

4.1.1 基线攻击

我们将 AdvCam 攻击与两种现有的代表性方法进行了比较:PGD [22]和对抗性补丁(Adv-
补丁)[3]。PGD 代表了数字攻击

最强一阶攻击。AdvPatch 代表可以直接应用于物理的无限制攻击

世界背景。其他一些物理世界的攻击，如 RP2，需要根据具体情况进行手工设计，因此在大规模生产中受到限制。我们从攻击成功率和视觉效果两个方面对这些方法进行了比较。对于 AdvCam 攻击，我们使用与[19]中相同的网络层来提取样式和内容(必要时)表示。

4.1.2 威胁模型

我们在数字和现实环境中测试了有针对性和无针对性的攻击。威胁模型采用灰箱设置:源网络和目标网络都是 VGG-19 网络，但在 ImageNet 上单独训练。对于物理世界的测试，我们首先使用谷歌 Pixel2 智能手机拍摄目标对象的照片，然后在攻击源网络的数字设置中制作一个敌对图像，之后我们打印出该敌对图像来替换或放置在目标对象上，然后我们使用同一智能手机从不同的视角和距离重新拍摄该对象。物理词攻击成功率由目标网络在重新拍摄的对抗图像上的预测准确度来衡量。

4.2. 消融研究

这里，我们在 ImageNet 上进行了一系列的实验
为了分析我们的 AdvCam 攻击的以下方面:1)伪装区域的形状和位置，2)伪装损失(例如样式损失、内容损失和平滑度损失)，

以及 3)对抗强度参数 λ 和区域大小。

4.2.1 伪装区域:形状和位置

这里，我们展示了伪装区域在形状和位置方面的选择如何影响精心制作的对抗示例的对抗强度。给定选定的攻击区域的形状和大小，我们以 1000 为间隔将强度参数 λ 从 1000 增加到 10000，直到攻击成功。该范围是根据大量实验选择的，[1000，10000]是找到具有高对抗强度和隐蔽性的广告的有效范围。图 4 显示了用不同区域制作的伪装敌对示例。我们发现，无论伪装区域是在目标对象的中心还是远离目标对象的中心，攻击都可以以很高的置信度成功。我们将在这一小节的最后一部分展示攻击可以伪装成一个甚至远离目标物体的区域，秘密隐藏在背景中。

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

图 4:在两次有针对性的攻击中，区域形状和大小的烧蚀:“背包”→“坦克”(顶排)和“雨披”→“交通灯”。(a):具有预期风格的原始干净图像(左下角)。(b) - (e):左:选定的攻击区域，右:精心制作的伪装对抗示例，在图像底部给出了 top-1 预测(“预测类别，置信度”)。

(a) 原件(b) Ls (c) Ls+Lc (d) All

图 5:烧蚀 3 次伪装损失:(a):右下角有意图伪装风格的原始图像；(b) - (d):使用不同损失函数的伪装敌对例子。

4.2.2 伪装损失(Ls，Lc，Lm)

图 5 示出了三组伪装有或没有两个可选增强的对抗示例
ments(内容保持 Lc 和平滑度增强 Lm)。当合并一个增强时，它的损失函数通过下面的等式被直接添加到最终对象。(2)(Ladv 中的 λ 被设置为 2000)。正如可以观察到的，内容保存可以帮助保存原始内容
帐篷，如“交通标志”示例(第三列)，而平滑度增强有助于生成平滑的物体表面。这些增强是可选的，因为对于一些通用示例，它们仅略微改善了视觉外观，例如，在“台灯”示例(第三行)中的内容保持或在所有示例中的平滑度增强。

4.2.3 对抗强度(λ)和区域大小

我们从 50 个类别中随机选择 2000 张 ImageNet 测试图像，使用不同的 λ ∈ [1000，10000]进行有针对性和无针对性的伪装攻击。针对-

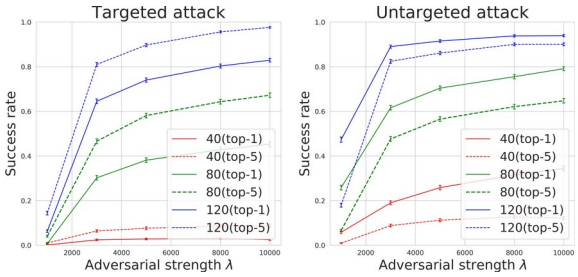

图 6:对抗强度 λ 和区域大小的消融:无针对性攻击(左)和有针对性攻击(右)的成功率。

tack，目标类是随机选择的，与真实类不同。为了测试 λ 在不同攻击区域下的影响，这里我们还改变了区域的大小，从 40*40 到 120*120。图 6 展示了前 1/5 的成功率，这是通过目标类是否在前 1/5 的类中来衡量的。如图，在区域固定的情况下，较大的对抗强度 λ 最多可以提高 20%的成功率；而当 λ 固定时，更大的攻击区域可以提高成功率高达 40%。与目标攻击相比，无目标攻击更容易成功。在最高成功率和最高成功率之间，最高成功率更难实现(实线比虚线低)。不同对抗强度和区域大小的标准误差在 0.07%和 1.13%之间。请注意，在这些实验中，攻击区域的伪装类型和位置是随机选择的，这可能会降低伪装效果和成功率。但是，这并不影响更大的 λ 和区域大小有助于制造更强攻击的结论。

4.3. 数字攻击

4.3.1 攻击设置

我们从 ImageNet ILSVRC2012 测试集的 5 个类别中随机选择 150 个干净的图像。然后，我们应用

三种方法(PGD、AdvPatch 和我们的 AdvCam)为每个干净的图像制作有针对性的对立示例。所选的源和目标类对是:“ashcan”→
《知更鸟》《背包》→《水罐》《大炮》→《大众电梯》，
“球衣”→“单杠”，“邮箱”→“娱乐中心”。对于 PGD 和 AdvCam，我们攻击主要目标-
通过手动选择获得目标区域，而对于高级补丁，我们进一步选择目标区域内的圆形攻击区域。对于 PGD，我们使用最大扰动
ϱ = 16/255(表示为 PGD-16)。对于 AdvCam，我们从与相同的类别中随机选择第二个图像
样式图像，并逐渐将 λ 从 1000 增加到 10000，直到找到一个对立的示例。为了公平比较，我们过滤掉失败的对立例子。最后，我们为 PGD、Ad- vPatch 和 AdvCam 分别收集了 132、101、122 个对抗性例子。图 7 显示了我们用来进行人类感知研究的三种方法的一些精心制作的对立例子。

(a) 原始(b) PGD-16 (c) AdvPatch (d) AdvCam

图 7:PGD-16、AdvPatch 和我们的 AdvCam 攻击精心制作的原始和敌对图像。

4.3.2 人类感知研究结果

我们在 Amazon Mechani- cal Turk (AMT)上建立了一项人类感知研究，要求人类评估者选择显示的图像是“自然和真实”还是“不自然或不真实”。为了模拟真实世界场景中的敌对例子，我们以随机顺序向用户呈现三种类型的敌对图像，并且是单独呈现，而不是成对呈现。最后，我们从 130 名参与者中收集了 1953 个选项。AdvPatch 被选为“自然和真实”
19.0±1.68% 的时间，PGD 被选中的时间为 77.3±1.53%，而我们的 AdvCam 被选中的时间为 80.7±1.53%。我们将这些统计数据总结为“隐秘”
三种方法的得分如图 8 所示。这证实了我们提出的 AdvCam 攻击能够制造像小扰动 PGD-16 方法一样隐蔽的敌对示例，尽管 AdvCam 攻击的扰动在大小上是不受限制的。

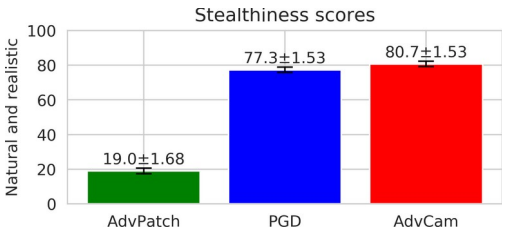

图 AdvPatch、PGD-16 和 AdvCam 的隐蔽性。

4.3.3 定制示例

在这里，我们展示了 AdvCam 如何制造极其隐蔽的伪装，尤其是无目标的伪装。图 9 展示了几个这样的例子。第一行显示了目标上的伪装旗示例，第二行显示了非目标上的伪装旗，这些伪装旗是通过攻击精心选择的背景区域而制作的。对于目标上的，AdvCam 在目标物体的表面产生自然和隐形的扰动。对于脱靶的，在第一个脱靶的例子中(第二行左边的两张图片)，我们将攻击隐藏在价格标签中，以欺骗 VGG-19 分类器将左轮手枪错误分类为卫生纸。在第二个例子中(第二行中间的两个图像)，AdvCam 通过在背景中添加花朵，成功地将一只布伦海姆猎犬伪装成熊皮。在第三个示例中(第二行右侧的两幅图像)，我们将攻击伪装成背景中的海报，这导致停在墙前的一辆小型货车被误认为是交通灯。这些例子不仅证明了我们的 AdvCam 攻击的隐蔽性和灵活性，还表明对深度学习系统的威胁无处不在，在许多情况下，甚至可能很难被人类观察者察觉。

4.4. 物理世界的攻击

我们进一步设计了三个现实世界的攻击场景来测试我们的 AdvCam 攻击的伪装能力。为了进行比较，我们还执行了 AdvPatch 和 PGD 攻击。对于 PGD 攻击，我们在(16/255，32/255，64/255，128/255)中测试了不同的 ϱ，并展示了具有最小 ϱ. 的成功对抗例子我们将对手的图案打印在 A3 或 A4 纸上，然后在不同的视角和距离下拍 20 张照片。
第一种情况是将野生图案伪装成街道标志，这可能会给自动驾驶带来问题
汽车。图 10 中的第一行显示了 PGD-128 (ϱ = 128/255)、AdvPatch 和 AdvCam 设计的一些成功模式。如你所见，攻击是完美的-
被 AdvCam 标记到树的纹理中。虽然 PGD 在数字环境中具有高度的隐蔽性，但在物理环境中它需要大的扰动(ϱ = 128/255)。因此
对抗模式比 AdvCam 的隐蔽性差得多，
与 AdvPatch 相同。第二种情况是保护穿着运动衫的人的身份。我们模拟了这样一个

图 9:我们的 AdvCam 攻击精心制作的伪装敌对图像及其原始版本。

(a)adv_patch (b)PGD_128 (c)adv_cam

图 10:上图:被视为街道标志的对立木质纹理。下图:t恤上的敌对标志。

图 11:带有三种污渍的敌对交通标志。

使用伪装的时尚标志“皮卡丘”攻击球衣的场景(参见图 10 中的底行)。这三种攻击都是从“泽西岛”到“爱尔兰梗”的攻击。请注意,即使是在最大的干扰下,PGD 的攻击也失败了。这显示了 AdvCam 具有定制伪装风格的高度灵活性,提供了满足各种攻击场景的灵活的隐蔽创建。

我们还使用 AdvCam 以三种不同的自然风格对“理发店”进行了“路牌”攻击(见图 11).AdvCam 的模式平滑且自然,几乎不能被人类观察者检测到,但成功地以高置信度欺骗了分类器。总之,由于 AdvCam 产生的对手的高度隐蔽性,它对当前基于 DNNs 的系统构成了无处不在的威胁。因此,AdvCam 可以成为评估物理世界中使用的 DNNs 的鲁棒性的有用工具。

5. 结论和未来工作

在本文中,我们研究了通用示例的隐蔽性,提出了一种新的方法称为对抗伪装(AdvCam),它结合了自然风格转移和对抗攻击技术,将对抗示例加工和伪装成隐蔽的自然风格。AdvCam 是一种灵活的方法,有助于为 DNN 模型的鲁棒性评估精心策划隐形攻击。除了从攻击的角度来看,提出的 Adv- Cam 可以是一种有意义的伪装技术,以保护被人类观察者和基于 DNN 的设备检测到的物体或人。

目前提出的 AdvCam 仍然需要攻击者手动指定攻击区域和目标类型,我们计划在未来的工作中探索语义分割技术来自动实现这一点。此外,我们还将探索将 AdvCam 应用于其他计算机视觉任务,包括对象检测和分割。此外,针对伪装攻击的有效防御策略将是另一个重要且有希望的方向。

确认

杨云得到了澳大利亚研究委员会发现项目 DP180100212 的资助。我们也非常感谢与史文朋科技大学的文生博士进行的早期讨论。

参考

[1] 安尼施·阿萨莱，洛根·恩斯特罗姆，安德鲁·易勒雅斯和郭凯文。综合强有力的对抗性例子。2017 年在 ICLR。1，3，5

[2] 、王、、、舒、、。基于希尔伯特的对抗性例子的生成性辩护。在 ICCV，2019 年。一

[3] 汤姆·布朗、蒲公英·马内、奥克·罗伊、马丁·阿巴迪和贾斯汀·吉尔默。敌对补丁。在 NIPS 车间，2017。2，3，5

[4] 尼古拉斯·卡里尼和戴维·瓦格纳。评估神经网络的鲁棒性。在 2017 年 IEEE S&P 大会上。1，2

[5] 亚历克斯·詹班达。语义风格转移和将二位涂鸦变成精美的艺术品。2016 年在 ICLR。3

[6] 陈，阿里塞夫，阿兰科恩豪斯和肖。深度驾驶:学习自动驾驶中直接感知的启示。2015 年在 ICCV。一

[7] 平、亚什·夏尔马、、易金凤和谢卓瑞。Ead:通过对立的例子对深层神经网络的弹性网络攻击。在 2018 年的 AAAI。2

[8] 董、、吴宝元、、、、、。人脸识别中高效的基于决策的黑盒对抗性攻击。在 CVPR，第 7714－7722 页，2019。一

[9] 阿列克谢·埃夫罗斯和威廉·T·弗里曼。用于纹理合成和传递的图像拼接。在 2001 年的 PACMCGIT。3

[10] 伊万·埃夫蒂莫夫、凯文·艾克霍尔特、厄尔伦斯·费尔南德斯、小野泰代、李博、阿图尔·普拉卡什、阿米尔·拉赫马蒂和宋。对深度学习模型的强大的物理世界攻击。在 2018 年的 CVPR。1，2，3

[11] 利昂·A·加蒂斯、亚历山大·S·埃克和马蒂亚斯·贝奇。使用卷积神经网络的图像风格转换。2016 年在 CVPR。3，4

[12] 伊恩·J·古德费勒，黄邦贤·史伦斯，克里斯蒂安·塞格迪。解释和利用对立的例子。2014 年在 ICLR。1，2

[13] 何、、、任、。用于图像识别的深度残差学习。2016 年在 CVPR。一

[14] 侯赛因·侯赛因尼和 Radha Poovendran。语义对立的例子。在 2018 年 CVPR 研讨会上。3

[15] 林西江、马、陈、和蒋玉刚。对视频识别模型的黑盒对抗性攻击。在 ACM MM，2019。一

[16] 谢尔盖·卡拉耶夫、马修·特伦塔科斯特、海伦·汉、阿西姆·阿加瓦拉、特雷弗·达雷尔、亚伦·赫茨曼和霍尔格·温内默勒。识别图像风格。arXiv 预印本 arXiv:1311.3715，2013。2

[17] 阿列克谢·库拉金、伊恩·古德菲勒和萨米·本吉奥。物理世界中的对立例子。2016 年在 ICLR。1，2，3

[18] 刘雪娣、、李春良、Nowrouzezahrai 和亚历克雅各布森。超越像素规范球:使用解析微分渲染器的参数对手。在 2018 年的 ICLR。3

[19] 栾福军、西尔万·帕里斯、埃利·谢克曼和卡维塔·巴拉。深度照片风格转移。2017 年在 CVPR。3，5

[20] 马、、王、萨拉·埃尔法尼、苏丹西·维杰维克拉马、格兰特·舍内贝克、宋黎明、迈克尔·侯勒和。利用局部固有维数刻画对立子空间。在 2018 年的 ICLR。一

[21] 马、牛、、王、、、和。理解对基于深度学习的医学图像分析系统的对抗性攻击。在 arXiv:1907.10456，2019。一

[22] 亚历山大·马德雷、亚历山大·马克洛夫、路德维希·施密特、迪米特里斯·齐普拉斯和阿德里安·弗拉杜。抗对抗性攻击的深度学习模型。在 2018 年的 ICLR。1，2，5

[23] Mahmood Sharif、Sruti Bhagavatula、Lujo Bauer 和 Michael K Reiter。犯罪的附属品:对最先进的人脸识别技术的真实而隐秘的攻击。在 CCS，2016。1，3，4

[24] 宋洋，瑞舒，内特·库什曼和斯特凡诺·埃尔蒙。用一般模型构造无限制的对立范例。在 NIPS，2018。3

[25] 克里斯蒂安·塞格迪、沃伊切赫·扎伦巴、伊利亚·苏茨基弗、琼·布鲁纳、杜米特鲁·埃汉、伊恩·古德菲勒和罗布·弗格斯。神经网络的触发特性。2013 年在 ICLR。1，2

[26] 、王、、蒲、。用于自动语音识别的残差卷积 ctc 网络。arXiv 预印本 arXiv:1702.07793，2017。一

[27] 、王、、马、、易金凤、周博文和顾。对抗性训练的收敛性和鲁棒性。在 ICML，2019 年。一

[28] 、王、邹迪凡、易金凤、、马、、顾。提高对抗性的鲁棒性需要重新审视错误分类的例子。在 2020 年的 ICLR。一

[29] 吴，，王，，，，马。跳过连接很重要:关于由结果网生成的对立例子的可移植性。在 2020 年的 ICLR。2

[30] 肖、、、何华伦、刘、宋晓。用对抗网络生成对抗实例。在 IJCAI，2018。1，3

[31] 谢慈航、张志帅、、周、、周仁和艾伦·L·尤耶。利用输入多样性提高对立范例的可迁移性。在 CVPR，2019 年。5

[32] 、徐、、、范全福、、孙、陈红歌、、、等。对抗性 t 恤！在现实世界中躲避个人探测器。arXiv 预印本 arXiv:1910.11099，2019。3

[33] 曾敏，王，。对话生成中的狄利克雷潜变量分层递归编译码器。在 EMNLP，2019。一

[34] 曾晓辉，，王宇翔，，邱，谢灵曦，戴宇荣，邓志强，和艾伦．L．尤耶。超越图像空间的对抗性攻击。在 CVPR，2019 年。3

[35] 张旸、哈桑·弗鲁什、菲利普·大卫和龚柏青。伪装:学习物理车辆伪装，以便在野外敌对地攻击探测器。在 ICLR，2019 年。3