用于愚弄物理世界中的人探测器的对抗性纹理

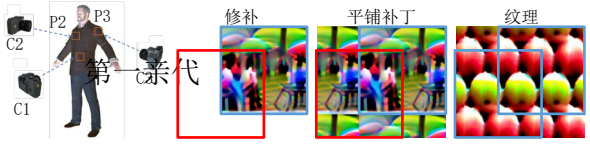
胡思远黄晓佩朱 1 富春孙 1 张博 1 肖林胡 1，3，4*
1 人工智能研究所计算机科学与技术系，
中国北京清华大学智能技术与系统国家重点实验室
2 中国北京清华大学集成电路学院 3 中国北京清华大学 DG/麦戈文脑研究所 4 中国北京中国脑研究所(CIBR)
zxp18 @mails.tsinghua.edu.cn huzhanha17
xlhu @mail.tsinghua.edu.cn siyuanhuang，fcsun，dcszb

摘要

如今，装有人工智能系统的相机可以捕捉和分析图像，以自动检测人。然而，AI 系统在接收真实世界中故意设计的模式时会出错，即，phys-对立的例子。先前的工作已经表明，在衣服上印刷敌对的补丁以躲避基于 DNN 的个人探测器是可能的。然而，当视角(即摄像机对物体的角度)发生变化时，这些敌对的例子可能会导致攻击成功率的灾难性下降。为了执行多角度攻击，我们提出了对抗性纹理(AdvTexture)。广告织物可以覆盖任意形状的衣服，使得穿着这种衣服的人可以从不同的视角隐藏而不被人察觉。我们提出了一种生成方法，称为基于环形裁剪的可扩展生成攻击(TC-EGA)，来处理具有重复结构的 AdvTexture。我们用 adv tex ure 打印了几块布料，然后在物理世界制作了 T 恤衫、裙子和连衣裙。实验表明，这些衣服可以骗过现实世界中的人体探测器。

1. 介绍

最近的工作表明，深度神经网络(DNNs)容易受到数字世界中通过向原始图像添加细微噪声而制作的对立示例的攻击 6,9,11,19,25 - 27,34]，并且 dnn 可以被物理世界中的人造物体攻击[1,4,10,32].这些人造物体被称为物理对抗的例子。最近，一些基于补丁攻击的方法 [32] 已经被提出来躲避人员探测器[15,16,35,37,38,40].具体来说，Thys 等人



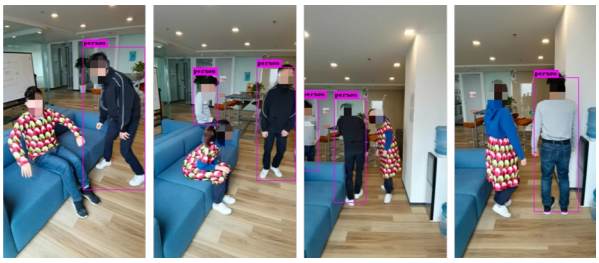
(一) (二) (三) (四)

图一。不同视角下的攻击图解。
(a)当设置为不同视角(C1、C2、C3)时，摄像机捕捉衣服的不同部分(P1、P2、P3)。(b-d)方框是摄像机可能捕捉到的区域。蓝色的表示最有效的攻击区域，而红色的则不太有效。

艾尔。[35]提出在纸板上贴一个补丁。通过将纸板放在摄像机前，人探测器无法探测到人。徐等[38]提出了一件印有对抗性补丁的对抗性 t 恤。穿 T 恤衫的人还可以躲避人体探测器。这些工作对广泛部署的基于深度学习的安全系统造成了相当大的威胁。它敦促研究人员重新评估这些系统的安全性和可靠性。

然而，上面提到的人检测器攻击方法只有在对抗性补丁面对摄像机时才有效。显然，一件衣服上的单个敌对补丁很难从多个视角攻击探测器，因为相机可能只捕捉到严重变形补丁的一部分(图。1a 和图。1b).我们称之为片段缺失问题。一种简单的扩展是用多个补丁覆盖衣服(例如，将补丁紧密地贴在衣服上；参见图。1c).然而，它不能完全解决片段丢失问题，因为摄像机将捕获属于不同面片单元的几个片段，使得攻击效率低下。另一个简单的解决方案是建立一个人的 3D 模型

*通讯作者。



图二。Adv- Texture 攻击 YOLOv2 时对抗效果的可视化。连衣裙、T 恤衫和裙子都是由覆盖着 AdvTexture 的大型聚酯布料裁剪而成。穿着这些衣服的人没有被探测器探测到。

身体和特定的一件衣服，以不同的视角呈现，如以前的作品[1]做了。然而，衣服是非刚性的，并且当前的 3D 渲染技术难以模拟真实世界中衣服的自然变形。例如，王等[36]在 3D 人类网格的平坦区域(正面和背面)上渲染 3D 徽标，但是当应用于看不见的网格时，攻击成功率(ASR)下降。

为了解决这个问题，我们提出了使用通用纹理的思想。与基于补丁的攻击不同，AdvTexture 可以以任意大小生成，因此可以覆盖任何大小的任何布料。我们要求纹理的任何局部都具有对抗效果(图。1d).然后，当衣服被 AdvTexture 覆盖时，摄像机捕捉到的每个局部区域都可以攻击检测器，这就解决了缺段问题。

为此，我们提出了一种基于环形裁剪的可扩展生成攻击(TC-EGA)的两阶段生成方法来制作 AdvTexture。在第一阶段，我们训练一个全卷积网络(FCN) [24,33]作为生成器，通过采样随机变量作为输入来产生纹理。不同于 GAN [17,28]，我们在每一层都使用卷积运算，包括潜在变量。因此，潜变量是一个具有空间维度的张量，只要我们沿着空间维度扩展潜变量，就能够使生成器生成多种尺寸的纹理。在第二阶段，我们用一种裁剪技术——环形裁剪(TC)来搜索潜在变量的最佳局部模式。优化后，我们可以通过平铺局部模式来生成一个足够大的潜变量。我们把它输入到 FCN，最后得到 AdvTexture。

我们实现了 TC-EGA 来攻击各种人体检测器，并在物理世界中实现了 AdvTextures。图。2 显示了一些针对 YOLOv2 的攻击示例。我们的实验表明，由这种织物制成的衣服显著降低了不同探测器的探测性能。

2. 相关著作

关于对立例子的早期作品[11,19,34]聚焦于数字攻击。小的敌对噪声可以被添加到原始图像中，并使 DNNs 输出错误的预测，对 DNNs 造成严重的安全问题。

与数字对抗性攻击相比，物理对抗性攻击在特定场景下会带来更多风险。几种方法[1,4,10,32]已经被提出来从物理上攻击图像分类模型。谢里夫等人[32]设计了一副眼镜来攻击人脸识别系统。Atha- lye 等人[1]通过引入对变换的期望(EoT) [1]方法。布朗等人[4]通过在对象附近放置对立的补丁来欺骗图像分类器。Evtimov 等人[10]将黑白贴纸贴在路标上，误导了路标分类。

最近，几种方法[15,16,35,35 - 38]准备攻击位于 DNN 的人员探测系统。Thys 等人[35[英语泛读材料一种可以贴在纸板上并由人拿着的对抗性贴片。黄等[16]提出通用物理伪装攻击(UPC)通过在虚拟环境中模拟 3D 物体来欺骗检测器。徐等[38]通过引入薄板样条(TPS)设计了一款广告 t 恤 2,8]来模拟衣服的变形(如褶皱)。吴等[37]介绍了对一系列检测模型、不同数据集和对象的攻击的系统研究。王等[36]用预设的徽标掩盖了对立的补丁，并将其映射到 3D 模型中。胡等[15]使用生成性对抗网络(GAN) [3,17]来制作看起来更自然的对抗性补丁。

部分作品[16,36,38]据报道，当视角增加时，攻击成功率下降。据王等[36]，当相机剧烈旋转时，部分补丁将不会被捕获。这可能导致低估威胁，而摄像机可以放置在现实世界场景中的任何地方。

3. 方法

我们的目标是生成任意大小的纹理，当这些纹理被印在布上时，从布上提取的任何补丁在对抗性攻击中都是有效的。我们首先介绍了一个对抗性补丁生成器，然后描述了基于补丁生成器的 TC- EGA。

3.1. 对抗性补丁生成器

设 τ 表示覆盖有纹理的整块布料， τ \u 表示提取的小块。我们假设 τ \u 遵循一个分布 padv，当它的对抗效果更显著时，概率 padv(τ \u\u)更高。我们使用能量函数 U(τ ∞)来建模

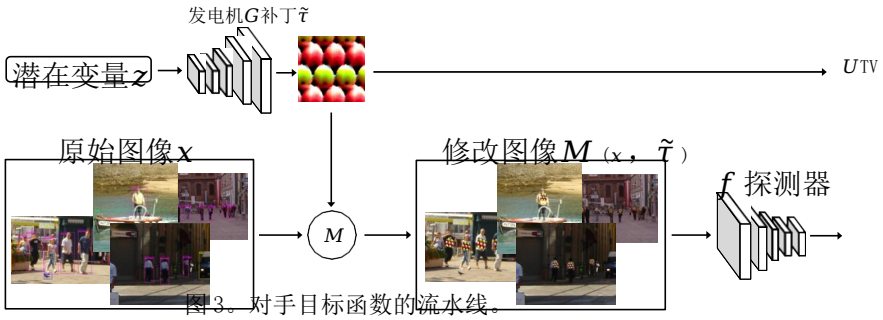


图3。对手目标函数的流水线。

这样的分布:

$$padv(\tau\backslash\backslash) = \frac{\text{鄂族}}{\quad}, \quad (1)$$

其中 zu =∫ τ\\ $e^{U(\tau\infty)d\tau}$ 叫做配分函数。

但是，由于配分函数的原因，很难直接从 padv(τ\\) 采样。因此，我们使用一个参数-terized generator gΦ:z→ τ\\,以逼近 padv(τ\\)，其中 z (0，I)。我们将 qΦ(τ\734;)定义为 τ\734;= gΦ(z)的分布，可以写成

$$q_\phi(\tau\infty) = \int \delta(\tau - g_\phi(z))p_z(z)\,dz, \quad (2)$$

其中，pz 是标准正态分布(0，I)的概率密度函数(PDF)，δ()是狄拉克 δ 函数。为了更精确地表示 padv(τ∞),我们调整 GΦ 以最小化 KL 散度 KL(qΦ(τ\\) padv(τ\\))。借助 Deep InfoMax (DIM) [14]我们有以下定理:

定理 1 最小化 KL(qΦ (τ˜) 副词 p (τ\\))相当于

$$\text{最小}_{q\Phi\Phi,\ \omega} \ (\tau\backslash\backslash)[U(\tau\backslash\backslash)] - IJSD(\tau\backslash\backslash, z),$$

在哪里

$$JSD_{\phi,\omega}(\tau, z) = E(\tau, z)\infty q\tau, z(\tau, z)[-sp(t\omega(\tau, z))] - E\tau\backslash q\phi(\tau\backslash), z'\infty pz(z')[sp(t\omega(\tau\backslash, z'))], \quad (3)(4)$$

qτ\\，表示 τ\\和 z 的联合分布，sp(t)= 1
log(1 + e)是 softplus 函数。Tω 是由神经网络建模的标量函数，其参数 ω 必须为
与参数 φ 一起优化。

证明见补充材料。

等式中的目标函数。（3)由两个术语组成。第一项 Eτ∿qΦ(τ∿)[U(τ∿)]被称为对手目标函数，因为最小化它改善了对手目标函数
生成的补丁的 ial 有效性。第二学期

JSD(τ，z)被称为信息目标函数，因为最小化它等价于最大化 z 和 τ 的互信息 14]，它需要不同的潜变量来生成不同的补丁。

3.1.1 对手目标函数

对手目标函数 Eτ∿qΦ(τ∿)[U(τ∿)]可以是 es-~ 通过采样 z 并生成 τ 来估计:

$$\frac{1}{N}\sum_{i=1}^{\text{常通}}[U(G_\phi(\text{子}))], \quad (5)$$

其中 zi 是从(0，I)采样的潜在变量，N 表示样本总数。

现在我们需要设置适当的能量函数，使得降低能量会导致人员检测器的检测失败。我们注意到检测器输出多个边界框，每个框有一个置信度得分，当

接收图像。置信度得分低于预定阈值的盒子将被过滤掉。因此我们选择信心的期望

作为能量函数 U(τ\\)的一部分的盒子上的分数。那么最小化高级对象功能将降低 con-
箱子的可信度分数，使得箱子很容易被过滤掉。

具体来说，我们在每一步随机生成补丁，

并应用一组物理变换，例如根据期望过变换(EoT)随机调整比例、对比度、亮度和附加噪声 32,35].

我们还加入了随机薄板自旋(TPS) [8,38]

变形作为附加的随机变换。然后，我们根据来自训练集的图像 x 上的预测框，将补丁随机附加到人身上。我们用 M (x，τ∞)来表示上述过程，并得到
修改后的图像将被发送到目标设备
泰克多。因此，这部分能量函数定义为

$$Uobj = Ex, M[f(M(x, \tau\backslash\backslash))], \quad (6)$$

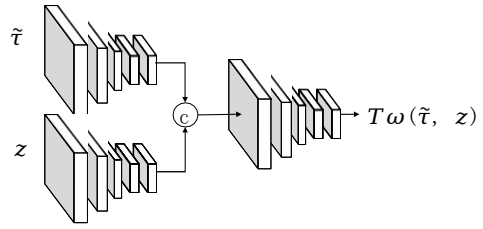


图 4。辅助网络的架构

输入 τ 和 z，输出标量值 Tω （τ，z)。手术 *Tω。它有两个 图中的 c 代表连接。*

其中 f 表示由目标检测器预测的盒子的置信度得分。

我们使用总方差(TV)损失的可微变量[32]作为能量函数的另一部分，以促使面片更平滑:

$$UTV = \sum_{i,j} \tau_{i,j}\tau_{i,j} + \tau_{i,j}\tau_{i,j+1} \quad (7)$$

我们一起形成了能量函数

$$u(\tau \backslash\backslash) = \frac{1}{\beta}(Uobj + \alpha UTV), \quad (8)$$

其中，α 和 β 是系数。参见图。3 为了插图。当最小化对手目标函数时，能量函数的每个部分将一起被最小化。

3.1.2 信息目标函数

如等式中所述。（4)，我们用一个辅助网络 Tω 来增加 z 和 τ 的互信息。我们在图 2 中说明了 Tω 的架构。4。情商。（4)有两个术语，和 估计他们每个人都需要随机抽样。继之前的工作之后[14]，为了估计第一项，我们首先从(0，I)中采样 z，然后由 Gφ(z)生成 τ \\ 每个训练步骤。为了估计第二项，我们保持 τ \ 和重采样 z。

在训练过程中，我们同时最小化敌对目标函数和信息目标函数。因此，分布 qφ 可以近似为 *padv，这意味着生成的面片 τ \\可以是 ad-* 与目标探测器垂直。

3.2. 基于环形裁剪的可扩展广义攻击

以秒计。3.1 中，我们已经描述了训练敌对补丁 τ \\'的生成器的方法。在本节中，我们使用 TC-EGA 生成 AdvTextures τ 基于 adv sar-ial 补丁生成器。我们利用一个特定的网络架构和一个样本技术来将对立的补丁扩展到对立的纹理。TC-EGA 有两个阶段。在第一阶段，我们训练一个全卷积网络(FCN) [24,33]

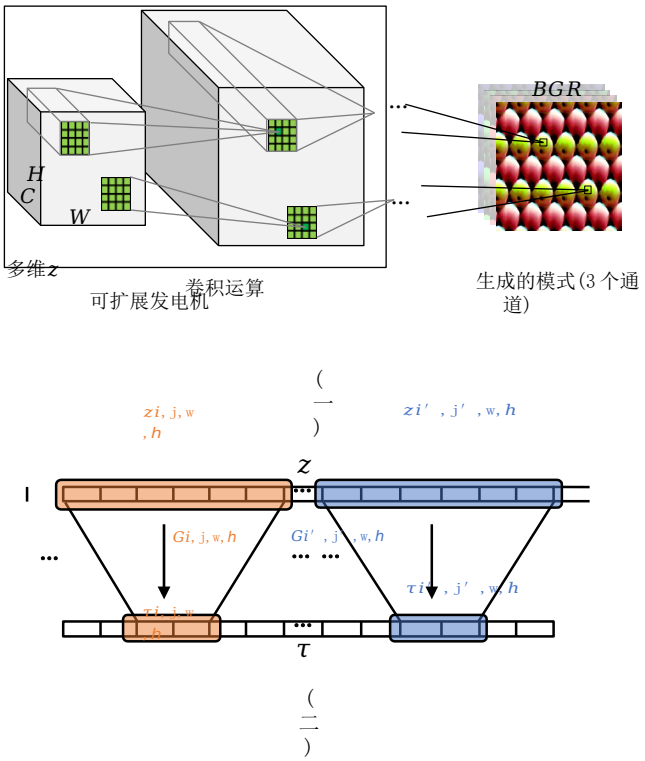


图 5。(a)FCN 发电机的图示。生成器网络的所有层都是具有零填充的卷积层，包括第一层。(b)当输入是 zi，j，w，h 时，从位置 I，j 提取的每个面片 τi，j，w，h 可以被视为子生成器 Gi，j，w，h 的输出

以帮助从对立纹理的分布中取样。在第二阶段，我们寻找最佳的潜在代表，以产生最有效的对抗性纹理。

3.2.1 第一步:训练一个可扩展的发电机

我们的目标是训练一个生成器，使它可以通过输入一个随机的 z 值来生成任意大小的面片。关键点是通过构造 FCN 赋予生成器平移不变量属性，其中所有层都是具有零填充的卷积层，包括输入潜在变量的第一层(见图。5a).潜在变量是 B C H W 张量，其中 B 是批量大小，C 是通道数量，H、W 分别是高度和宽度。

这里我们展示了使用 FCN 的原因。我们假设整体纹理 τ 是由具有隐藏变量 z (0，I)的全局生成元 G：z τ 生成的。我们-用 τi，j，w，h 标注提取的面片，其中心位于在整个纹理的位置(I，j)，并且具有(w，h)的形状。此外，面片 τi，j，w，h 可视为子生成器 Gi，j，w，h 的输出 :ziz，jz，wz，hz τi，j，w，h，其中 ziz，jz，wz，hz 是由所有依赖于 τi，j，w，h 的元素(见图。5b).假设 τi，j，w，h 服从分布 I，j，w，h，我们有下面的定理和推论。

定理 2 设 τ 1 = G1(z1)， τ 2 = G2(z2)，Z1∞ 1 *z2 2， τ 1 1， τ 2 2. 如果 1 等同于 2 和 G1 相当于 G2，那么 1 等同于 2.*

推论 2.1 Gi， j， w， h 和 i,j,w,h 与 I， j 无关， 即 Gi， j， w， h = Gw， h 和 i,j,w,h = w,h，如果 g 是 FCN，输入 z∞ (0， I)。

校样见补充资料。因此，只要训练子生成器 Gw，h 来近似 w，h 到 padv 的分布，从
具有形状(w，h)的整体纹理也近似如下

低 p 副词，即它具有对抗效力。此外，
由于卷积的平移不变性
操作、子发电机 Gw、h 和全局发电机

除了潜变量 z 的空间形状 H 和 W 不同之外，可以共享相同的架构和参数。结果，我们只需要训练一个小生成器。

请注意，隐藏变量 z 的高度 H 和宽度 W 不能太小，否则输出将太小，无法在空间形状(W，H)中裁剪面片。我们用 Hmin 和 Wmin 表示最小空间尺寸。在训练期间，我们在形状 B 中取样一个小 z
并在每个训练步骤中生成相应的补丁。此后，我们可以通过用任何 H ≥ Hmin 和随机化 z 来产生任意大小的不同纹理
W ≥ Wmin。

3.2.2 第二阶段:找到最佳潜在模式

经过训练后，生成器可以通过对潜在变量进行采样来生成不同的纹理。为了找到对抗攻击的最佳纹理，我们建议更进一步，即在冻结生成器参数的情况下优化潜在变量。然而，由于纹理没有特定的形状，并且潜在变量的大小需要足够大以产生大的纹理布料，所以直接优化潜在变量是困难的。

受拓扑中支持上下左右延拓的 torus 展开的启发[12](图。6a)，我们引入了环形裁剪(TC)技术，该技术旨在将局部模式 zlocal 优化为一个单元，使得最终的潜在变量 z 可以通过平铺多个相同的单元来产生。详细地说，zlocal 可以被参数化为形状 B C L L 中的张量，形状 B C L L 具有形状超参数 L，形状超参数 L 可以被视为一个两个
拓扑学中的三维环面 T2。6a).因此，可以从 zlocal 中裁剪出任意形状的潜变量
以递归的方式(图。6b)，可以看作是在圆环面上的裁剪。我们用 Croptorus 来表示这种作物操作。

在优化过程中，我们随机抽样潜在的
变量以形状为例 B C Hmin Wmin 这样

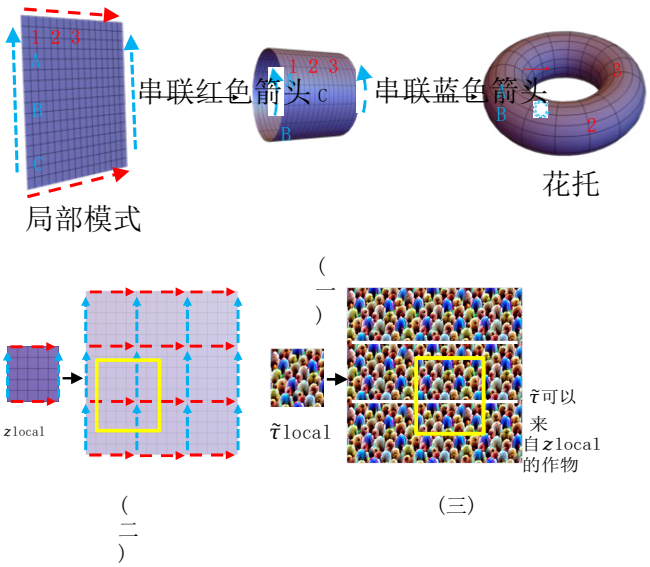


图 6。环形裁剪插图。(a)通过首先连接其水平边缘(红色箭头)，然后连接垂直边缘(蓝色箭头)，局部图案可以折叠成环面。(b)可以通过并排平铺局部图案来创建任意形状的潜在变量，因此在接合处裁剪的变量等同于在环面上裁剪的变量，意味着图案仍然是连续的。(c)这种裁剪技术也适用于像素空间。参见第节。4.3 对于这个变种。

裁剪技术。因为我们在这个阶段只考虑对手的有效性，所以我们通过 zsample 生成补丁并最小化对手损失(等式)。 (5)).优化后，通过平铺 zlocal 可以产生一个任意大小的潜变量。

4. 实验设置

4.1. 学科

我们招募了三名受试者(平均年龄:24.0；范围:
21 26;两个男性和一个女性)来收集身体测试组。招募和研究程序获得批准
清华大学心理伦理委员会，中国北京。

4.2. 资料组

我们采用了 Inria 个人数据集[7]作为我们的训练集。这是一个用于行人检测的数据集，由 614 幅用于训练的图像和 288 幅用于测试的图像组成。我们评估-对 Inria 测试集进行基于补丁的攻击。对于物理-
根据评估，我们制作了不同质地的衣服。三名受试者穿着不同的对手服装，在摄像机前慢慢转了一圈
它固定在离地面 1.38 米的地方。除非另有规定，否则摄像机和人之间的距离固定为 2 米。我们为每个小组录制了两个视频
对象和每一件敌对的衣服。其中一个视频是在室内(实验室)录制的，另一个是录制的

室外(砖砌走道)。然后我们从每个视频中提取了 32 帧。我们录制了 3 ^ 2 = 6 个视频，并为每件对立的衣服收集了 6 ^ 32 = 192 帧。我们手动标记它们来构建一个测试集。

4.3. 基线方法

我们评估了 Thys 等人提出的对抗性补丁 35]和徐等[38]，并分别用 AdvPatch 和 AdvTshirt 命名。我们从他们的原始论文中复制了这些模式。我们还平铺了 AdvPatch 和 AdvTshirt，以形成具有重复图案的纹理。这两种变体称为 AdvPatchTile 和 AdvTshirtTile。此外，我们评估了一个纹理与重复随机颜色，这是随机表示

此外，TC-EGA 有多个组成部分，其中一些可以单独应用于制作敌对特征。为了研究每个组件的性能，我们设计了三种 TC-EGA，如下所述。

可扩展生成攻击(EGA)我们训练了一个 FCN 作为 TC-EGA 的第一级，没有优化最佳潜变量。在评估过程中，最终纹理可以由任意大小的潜在变量生成，并从标准正态分布中采样。

环形裁剪攻击(TCA)我们直接优化纹理，而不是训练 FCN 来生成纹理。具体来说，我们初始化了一个局部纹理模式
300×300 像素，按大小随机抽取一个面片
150×150，在每个优化步骤中通过环形裁剪。

随机剪切攻击(RCA)我们直接优化一个大小固定的大块。我们初始化了大的补丁，并根据大小随机裁剪了一个小补丁
优化期间 150 150。这种方法被称为随机裁剪攻击(RCA)。我们在 RCA2 和 RCA6 实施了两次攻击，大块补丁的大小
分别是 300 300 和 900 900。

4.4. 实施细节

我们精心制作的 AdvTexture 主要是为了愚弄 YOLOv2 [29]，YOLOv3 [30]，更快的 R-CNN [31]并屏蔽 R-CNN [13].
检测器在 MS COCO 数据集上进行了预训练[22].他们的输出被过滤，只输出 person 类。

对于每个目标检测器，我们首先使用非最大抑制(NMS)阈值 0.4 从训练集中提取图像上的预测边界框。我们
选择置信度大于 1 的盒子-
t 阈值(YOLOv2 和 YOLOv3 为 0.5，fast 和 Mask R-CNN 为 0.75)。我们还过滤掉了面积小于整个图像的 0.16%的盒子

(a)随机(b) AdvPatchTile (c) TC-EGA

图 7。不同纹理的可视化。(a)具有重复随机颜色的纹理。(b)由平铺一个横向片形成的纹理[35]反复。(c)TC-EGA 攻击 YOLOv2 产生的纹理。

为了更快和屏蔽 R-CNN。然后，正如我们在第二章中所描述的。3.1.1 在优化过程中，我们将提取的图像贴在人体上，并将修改后的图像输入到检测器中。

此外，我们应用了 Adam [18]优化器优化两个阶段的参数。超参数如下所示。(1)第一阶段:初始学习率来训练发电机是 0.001。生成器是 7 层 FCN，其输入是大小为 B128 9 9 的潜在变量 z。相应输出的大小是 B 3 324 324，其中第二维代表 RGB 通道。(2)第二阶段:我们优化了一个本地 la-
tent 变量 zlocal，大小为 1 128 4 4，然后通过环形裁剪技术产生大小为 B 128 9 9 的 z 样本。优化后的学习率为 0.03。

为了物理实现 AdvTexture，我们打印了
通过数字纺织品印花在聚酯布料上形成纹理。后来，我们聘请了一位专业的裁缝来制作广告服装，包括 t 恤、裙子和连衣裙。

5. 结果

图。7 展示了一些用不同方法得到的纹理，更多的可以在补充材料中找到。

5.1. 数字世界中基于补丁的攻击

我们首先评估了数字世界中基于补丁的攻击形式。具体来说，在评估除 AdvPatch 和 AdvTexture 之外的大多数方法时，我们从纹理中随机提取面片。我们用重采样面片来表示这种面片。然后，我们将补丁附加到 Inria 测试集的图像上，就像制作对抗性补丁一样。我们在最初的测试中使用了目标检测器提出的包围盒
置信阈值为 0.5 的图像作为地面真实。我们计算了该方法的平均精度
在修改后的测试图像上设置包围盒来测量对抗效果。请注意，AP 越低，攻击越强。

标签。1 给出了 YOLOv2 在不同条件下的 AP。clean 表示原始测试集上的 AP。因为

方法	美国联合通讯社(Associated Press)	可膨胀的	重新取样
干净的	1.000		
随意	0.963	✓	✓
AdvPatch [35]	**0.352**	✗	✗
AdvPatchTile	0.827	✓	✓
AdvTshirt	0.744	✗	✗
[38]	*		
AdvTshirtTile	0.844	✓	✓
RCA6			

表 1。在 Inria 测试集上不同攻击下 YOLOv2 的攻击点。可扩展表示方法是否可以产生任意大小的纹理。重采样表示是否随机提取面片。

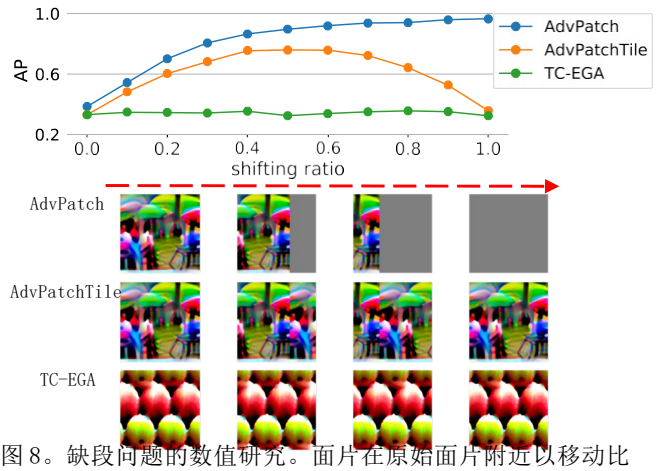
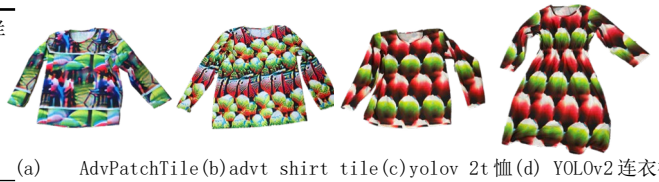


图 8。缺段问题的数值研究。面片在原始面片附近以移动比率进行裁剪。以 AdvPatch 为例，当裁剪的面片恰好是原始面片时，偏移比率为 0.0。变速比是 1.0 当原始面片完全移出裁剪范围时。

我们使用探测器对原始图像的预测作为地面实况，AP 为 1.000。AdvPatch 将 YOLOv2 的 AP 降至 0.352[1]。

与 AdvPatch 相比，可扩展的变体 Adv- PatchTile 将 AP 从 0.352 增加到 0.827。由于 Ad- vTshirt 是在不同的数据集(其作者的私人数据集)上训练的，所以它只获得了 0.744 的 AP。同样，Ad- vTshirtTile 将 AP 增加到 0.844。我们把它归因于-缺失线段问题的解决。与其相比

1 我们根据他们发布的代码 https://gitlab.com/EAVISE/adversarial-yolo.复制了一个对抗性补丁，复制的补丁获得了 0.378 的 AP。在所有的实验中，我们都使用了从他们的论文中复制的补丁。We reproduced an adversarial patch according to their released code



(a) AdvPatchTile(b)advt shirt tile(c)yolov 2t恤(d) YOLOv2连衣裙

图 9。现实世界的敌对服装。

变异，TC-EGA 得到最低的 AP 0.362，也是所有重采样斑块中最低的。AdvPatch 使 AP 略低于 TC-EGA。然而，它不可扩展，因此不适合多视角攻击。此外，EGA 将 AP 降至 0.470，TCA 用 AP 0.664 创建了可扩展补丁。它是低于 AdvPatchTile，这表明环形裁剪技术的有效性。此外，RCA6 比 RCA2 差得多，这表明在优化大补丁时存在困难。

我们进一步研究了片段缺失问题，通过评估在移位位置裁剪的补丁的对抗有效性(见图。8).当移位比率增加时，基于补丁的攻击 AdvPatch 变得不那么有效。平铺补丁缓解了这个问题，但是仍然有问题。TC- EGA 生成的纹理在移动期间是健壮的。

数字世界其他探测器被 TC-EGA 攻击的结果见补充资料。

5.2. 物理世界中的攻击

图。9 展示了用不同方法制作的衣服，更多信息可在补充材料中找到。

我们首先在 YOLOv2 上比较了不同的方法。由于检测器预测的盒子可以通过特定的置信阈值进行过滤，我们在图中绘制了召回置信曲线。10 并在传说中展示了自己的 AP。请记住，召回表示成功检索到的箱子的比例。这些盒子由置信阈值填充。因此，对于每个特定的置信阈值，较低的回忆表示较好的对抗效果。来自图。10AdvPatch 和 AdvTshirt 的平铺变体比原始方法更有效。TC-EGA 在所有方法中以最低的回忆-信任曲线和最低的 AP 表现最好。

此外，我们使用了另一个度量来评估攻击的有效性。具体来说，对于每个输入图像，我们收集目标检测器的预测边界框，其置信度得分大于某个置信度阈值。只要这些框之一与基本事实框的交集(IoU)大于 0.5，则认为检测器已正确检测到。我们将攻击成功率定义为没有被正确预测的测试图像。自从

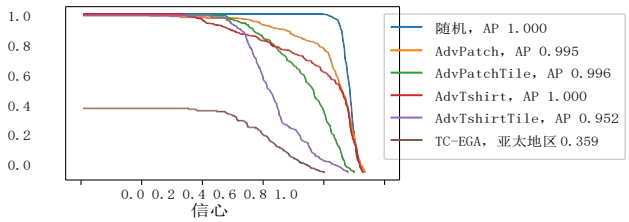


图 10。物理上的召回 v.s 置信曲线和 AP_____对抗测试集。目标网络是 YOLOv2。

服装随机 t 恤裙子连衣裙 mASR 0.092			
0.771	0.287	0.893	

表二。不同敌对服装的面具。

ASR 与置信阈值相关，我们计算了多阈值下 ASR 的平均值，即 mASR。阈值分别为 0.1，0.2，...，0.9 英寸实验。

图。11 从多个视角展示 mASRs。与随机纹理相比，AdvPatch 和 AdvTshirt 在人面对摄像机(视图-ing 角度在图中为 0♀或 360♀)。然而，这两种方法的 mASRs 随着视角的增大而减小
增加，这表明了段缺失问题。两种方法的平铺变体在多视角下具有一些对抗效果，而 mASRs
几乎每个视角都低于 0.5。TC- EGA 几乎在每个视角都优于其他方法。在视角 0♀和 180♀时，mASR 约为 1.0，表明当置信度阈值大于 0.1 时，人总是能够避开检测器。当视角接近 90° 或 270° 时，效果较差，因为此时摄像机捕捉的区域较小
视角。

我们研究了衣服类型和人与相机之间的距离的影响。从选项卡。2 当这种质地应用于不同种类的衣服时，其通用效果是不同的。当应用于较大的衣服(例如，连衣裙)时，攻击更有效，因为更多的纹理区域被相机捕获。此外，敌对服装在室内和室外场景中的 mASRs 相当(见补充材料)。当远离摄像机时，它们的效果下降(见补充材料)。

标签。3 提出了对抗各种探测器的伪装方法。从表中可以看出，TC-EGA 获得了比随机高得多的 mASR。此外，当敌对服装通过不同的探测器时，敌对有效性仍然存在。有关迁移研究的详细信息，请参见补充材料。此外，

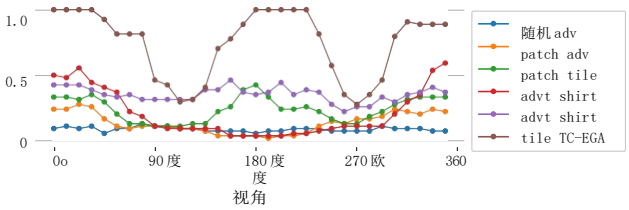


图 11。不同视角下的袭击事件。

检测器 YOLOv2	YOLOv3	FasterRCNN		
随意	0.08 7	0.000 1	0.000	0.000
TC-EGA	0.74 3	0.701 1	0.930	0.855

1 在将输入发送到 YOLOv3 之前，我们将输入的大小缩放了 50%。原因见补充资料。

表 3。物理世界中不同探测器的 mASR。

我们在补充视频中提供了一个视频演示。

6. 结论

我们提出了一种方法来制作 AdvTextures，以实现对人检测系统的物理攻击。主要思想是首先训练一个可扩展的生成器，通过在潜在空间中随机输入来生成 AdvTexture，然后搜索潜在变量的最佳局部模式进行攻击。AdvTexture 的有效性通过优化潜在输入来提高。我们通过将 AdvTexture 印在一块大布上，并制作不同的 t 恤、裙子和连衣裙，来实现 adv texture。根据我们在现实世界中的实验评估，这些衣服在穿着者转身或改变姿势时是有效的。

局限性虽然针对一个检测器的精心制作的纹理在某种程度上也可以攻击另一个检测器，但可移植性不是很好。模型集成可以用来提高可移植性。

潜在的负面影响对抗性研究可能会导致现实世界社区中潜在的不必要的应用，例如相机安全问题。已经提出了许多基于先前暴露的漏洞的防御方法[11,20,39]，提高了我们社区的安全水平，有益地说明了攻击研究的价值。

确认

这项工作得到了国家自然科学基金(编号 U19B2034，62061136001，61836014)和清华-丰田联合研究基金的支持。

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。国际机器学习会议，284-293 页。PMLR，2018。1,2

[2] 弗雷德·l·布克斯坦。主变形:薄板样条和变形的分解。IEEE 模式分析和机器智能汇刊，11(6):567–585，1989。2

[3] 安德鲁·布洛克，杰夫·多纳休和卡伦·西蒙扬。用于高保真自然图像合成的大规模 GAN 训练。参加 2019 年 5 月 6 日至 9 日在美国路易斯安那州新奥尔良举行的 2019 年 ICLR 第七届学习代表国际会议。OpenReview.net，2019。2

[4] 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。arXiv 预印本 arXiv:1712.09665，2017。1,2

[5] 蔡兆伟和努诺赛洛斯。级联 r-cnn:高质量对象检测和实例分割。2019 年 IEEE 模式分析与机器智能汇刊。15

[6] 尼古拉斯·卡里尼和戴维·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。1

[7] Navneet Dalal 和 Bill Triggs。用于人体检测的方向梯度直方图。2005 年 IEEE 计算机学会计算机视觉和模式识别会议(CVPR'05)，第 1 卷，第 886–893 页。Ieee，2005 年。5

[8] 吉安卢卡·多纳托和塞尔日·贝隆吉。近似薄板样条映射。在欧洲计算机视觉会议上，第 21-31 页。斯普林格，2002 年。2,3

[9] 董、、廖、庞天宇、、、、、。用动力推动对抗性攻击。IEEE 计算机视觉和模式识别会议论文集，第 9185–9193 页，2018 年。1,15

[10] Kevin Eykholt、Ivan Evtimov、Earlence Fernandes、、Amir、Xiao、Atul Prakash、Tadayoshi Kohno 和 Dawn Song。对深度学习视觉分类的鲁棒物理世界攻击。IEEE 计算机视觉和模式识别会议论文集，第 1625-1634 页，2018 年。1,2

[11] 伊恩·古德菲勒，黄邦贤·史伦斯和克里斯蒂安·塞格迪。解释和利用对立的例子。2015 年国际学习代表会议。1,2,8

[12] 艾伦·哈奇。代数拓扑。剑桥大学出版社，2002 年。5

[13] 明凯·何、乔治娅·格基奥萨里、彼得·多拉·r 和罗斯·吉尔-希克。屏蔽 r-cnn。IEEE 计算机视觉国际会议论文集，第 2961-2969 页，2017 年。6,13

[14] 德文·赫杰姆、亚历克斯·费多罗夫、萨缪尔·拉沃伊-马奇尔登、卡兰·格雷瓦尔、菲尔·巴赫曼、亚当·特里施勒和约舒阿·本吉奥。通过互信息估计和最大化学习深度表示。2019 年国际学习代表会议。3,4,12

[15] 胡宇智团，孔博汉，丹尼尔·斯坦利·谭，俊，华凯龙，郑文煌。物体探测器的自然物理对抗补丁。IEEE/CVF 计算机视觉国际会议论文集，第 7848–7857 页，2021。1,2

[16] 黄，，高，，周，谢慈航，阿兰·L·尤耶，邹长青，。对目标探测器的通用物理伪装攻击。IEEE/CVF 计算机视觉和模式识别会议论文集，第 720–729 页，2020 年。1,2

[17] 泰罗·卡拉斯、萨穆利·莱恩和蒂莫·艾拉。一种基于风格的生成对抗网络生成器体系结构。IEEE/CVF 计算机视觉和模式识别会议论文集，第 4401-4410 页，2019 年。2

[18] 迪德里克·P·金马和吉米·巴。亚当:一种随机优化方法。arXiv 预印本 arXiv:1412.6980，2014。6

[19] 阿列克谢·库拉金、伊恩·古德菲勒和萨米·本吉奥。物理世界中的普遍例子。在 2017 年国际学习代表大会上。1,2

[20] 廖、、、董、庞天宇、。使用高级表示引导的 denoiser 防御对抗性攻击。IEEE 计算机视觉和模式识别会议论文集，第 1778-1787 页，2018 年。8

[21] 密集物体探测的焦损失。IEEE 计算机视觉国际会议论文集，第 2980-2988 页，2017 年。15

[22] 宗-林逸、迈克尔·梅尔、塞尔日·贝隆吉、詹姆斯·海斯、彼得罗·佩罗娜、迪瓦·拉马南、彼得·多拉·r 和 C·劳伦斯·兹尼克。微软 coco: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。6

[23] 刘燕佩，，陈，，宋晓明。探究可转移的对立例子和黑盒攻击。在 2017 年 4 月 24 日至 26 日在法国 ICLR 土伦举行的第五届学习表征国际会议上，会议记录。OpenReview.net，2017。15

[24] 乔纳森·朗，埃文·谢尔哈默和特雷弗·达雷尔。语义分割的全卷积网络。IEEE 计算机视觉和模式识别会议论文集，第 3431–3440 页，2015。2,4

[25] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi 和 Pascal Frossard。Deepfool: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。1

[26] Anh Nguyen，Jason Yosinski 和 Jeff Clune。深度神经网络很容易被忽悠:对无法识别的图像的高置信度预测。IEEE 计算机视觉和模式识别会议论文集，第 427–436 页，2015 年。1

[27] 尼古拉斯·帕伯诺特、帕特里克·麦克丹尼尔、萨默什·贾、马特·弗雷德里克松、Z·伯凯·切利克和阿南瑟拉姆·斯瓦米。对抗环境下深度学习的局限性。2016 年

*IEEE 欧洲安全和隐私研讨会(Eu- roS&P)，第 372 - 387 页。IEEE，2016。*1

[28] 亚历克·拉德福德，卢克·梅斯，和索史密斯·钦塔拉。具有深度卷积生成对抗网络的无监督表示学习。arXiv 预印本 arXiv:1511.06434，2015。2

[29] 约瑟夫·雷蒙德和阿里·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。6,13

[30] 约瑟夫·雷蒙德和阿里·法尔哈迪。Yolov3:增量改进。arXiv 预印本 arXiv:1804.02767，2018。6,13

[31] 邵青·任、明凯·何、罗斯·吉斯克和孙健。更快的 r-cnn:用区域建议网络实现实时目标检测。IEEE 模式分析与机器智能汇刊，39(6):1137 - 1149，2016。6,13

[32] 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。2016 年 acm sigsac 计算机和通信安全会议论文集，第 1528-1540 页，2016 年。1,2,3,4

[33] 约斯特·托拜厄斯·斯普林根贝格、阿列克谢·多索维茨基、托马斯·布罗克斯和马丁·里德米勒。追求简单：全卷积网。arXiv 预印本 arXiv:1412.6806，2014。2,4

[34] 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。在 2014 年国际学习代表大会上。1,2

[35] 西蒙·提斯，维贝·范·兰斯特和图恩·戈德默。愚弄自动监控摄像机:攻击人员检测的对抗性补丁。《IEEE/CVF 计算机视觉和模式识别研讨会论文集》，第 0-0 页，2019 年。1,2,3,6,7

[36] 、周景阳、、、、常、钱德拉吉特·巴贾杰和王。3d 对抗性标识能遮掩人类吗？arXiv 预印本 arXiv:2006.14655，2020。2

[37] 吴祖轩、林世南、拉里·戴维斯和汤姆·戈德斯坦。制造隐身衣:真实世界对目标探测器的恶意攻击。在欧洲计算机视觉会议上，第 1-17 页。斯普林格，2020。1,2

[38] 、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。在欧洲计算机视觉会议上，第 665-681 页。斯普林格，2020。1,2,3,6,7,13

[39] 、徐、、齐。特征压缩:检测深层神经网络中的对立实例。arXiv 预印本 arXiv:1704.01155，2017。8

[40] 朱晓佩，小李，，，王，-胡奥林。在现实世界中使用小灯泡愚弄热红外行人探测器。2021 年在 AAAI 举行的第三十五届 AAAI 人工智能大会上。1

A. 证明

A.1. 定理 1 的证明

KL 散度 KL(qϕ(τ˜) padv(τ˜))可以分为两项:

$$\mathrm{KL}(q_\phi(\tau\tilde{}) \| p_{adv}(\tau\tilde{})) = \int_{\tau\tilde{}} q_\phi(\tau\tilde{}) \log \frac{q_\phi(\tau\tilde{})}{p_{adv}(\tau\tilde{})} d\tau\tilde{}$$

$$= \int_{\tau\tilde{}} q_\phi(\tau\tilde{}) \log q_\phi(\tau\tilde{}) d\tau\tilde{} - \int_{\tau\tilde{}} q_\phi(\tau\tilde{}) \log p_{adv}(\tau\tilde{}) d\tau\tilde{}, \quad (9)$$

其中第一项是 qϕ 的负熵，即，hϕ(τ˜)。我们引入互信息来帮助计算熵:

$$I_\phi(\tau\tilde{}, z) = \int_{\tau\tilde{}, z} p(\tau\tilde{}, z) \log \frac{p(\tau\tilde{}, z)}{q_\phi(\tau\tilde{})p_z(z)} d\tau\tilde{}, dz, \quad (10)$$

其中 p(τ˜, z)是 τ˜= gϕ(z)和 z 的联合分布.由于 p(τ˜, z)= p(τ˜ z)pz(z)和 qϕ(τ˜)是边际分布

$$I_\phi(\tau\tilde{}, z) = \int_{\tau\tilde{}, z} p(\tau\tilde{}, z) \log \frac{p(\tau\tilde{}, z)}{p_z(z)} d\tau\tilde{}, dz - \int_{\tau\tilde{}, z} p(\tau\tilde{}, z) \log q_\phi(\tau\tilde{}) d\tau\tilde{}dz$$

$$= \int_{\tau\tilde{}, z} p(\tau\tilde{} z)p_z(z) \log p(\tau\tilde{} z) d\tau\tilde{}, dz - \int_{\tau\tilde{}} \log q_\phi(\tau\tilde{}) \int_z p(\tau, z) dz d\tau\tilde{}$$

$$= p_z(z) \int_{\tau\tilde{}} p(\tau\tilde{} z) \log p(\tau\tilde{} z) d\tau\tilde{} dz - \int_{\tau\tilde{}} q_\phi(\tau\tilde{}) \log q_\phi(\tau\tilde{}) d\tau\tilde{}$$

$$= -h_\phi(\tau\tilde{} z) + h_\phi(\tau\tilde{}), \quad (11)$$

其中 Hϕ(τ˜ z)称为条件熵。所以，Eq 的第一项。（9)可以用 Iϕ(τ, z)hϕ(τ˜ z).代替因为 τ˜∼qϕ 由 z 决定，即 p(τ˜ z)= δ(τ˜− gϕ(z))，所以我们有

$$h_\phi(\tau\tilde{} z) = \int_{\tau\tilde{}} \delta(\tau\tilde{} - g_\phi(z)) \log \delta(\tau\tilde{} - g_\phi(z)) d\tau\tilde{} \frac{p_z(z)}{p_z(z)} dz$$

$$= -\int_{\tau''} \delta(\tau'') \log \delta(\tau'') d\tau'' \int_z p_z(z) dz \quad (12)$$

$$= -\int_{\tau''} \delta(\tau'') \log \delta(\tau'') d\tau'', \quad (13)$$

这表明 Hϕ(τ˜ z)是常数[2]。因此，我们忽略等式中的这一项。（11).此外，在第二届任期内情商。（9)，自 p

$$q_\phi(\tau\tilde{}) \log p_{adv}(\tau\tilde{}) d\tau\tilde{} = \int_{\tau\tilde{}} q_\phi(\tau\tilde{}) \log \frac{q_\phi(\tau\tilde{})}{p_{adv}(\tau\tilde{})} d\tau\tilde{} - \int_{\tau\tilde{}} q_\phi(\tau\tilde{}) \log q_\phi(\tau\tilde{}) d\tau\tilde{}$$

$$= \int_{\tau} q_\phi(\tau|) U(\tau|) d\tau| + \int_{\tau} q_\phi(\tau|) \log ZU \, d\tau|$$

$$= E_{\tau \sim q_\phi(\tau)}[U(\tau)] + \log ZU, \tag{14}$$

其中配分函数 $ZU = \int_{\tau} e^{u(\tau|)} d\tau$ 是一个常数。

因此，最小化 Eq。（9)相当于

$$\min I_\phi(\tau, z) + E_{\tau \sim q_\phi(\tau)}[U(\tau)]。 \tag{15}$$

2 事实上，对于离散分布，它是零；对于连续分布，它是无穷大。In

换句话说，我们需要同时最大化 Iϕ(τ↷, z)和最小化 Eτ↷)qϕ(τ↷)[U(τ↷)]。根据 Deep InfoMax (DIM) [14]，最大化 Iϕ(τ, z)等价于最大化 Jensen-Shannon 互信息(MI)估计器，

$$\phi, \omega \quad (\tau, z) = E^{-}_{\tau, z(\tau, z)\infty q} [sp(t\omega(\tau, z))] - E\tau\sim q\phi \quad (\tau \backslash\backslash, z' \infty pz(z')) [sp(t\omega(\tau \backslash\backslash, z'))], \quad (16)$$

其中 qτ\\, z 表示 τ\\, z 的联合分布，sp(t) = log(1 + et)为 softplus 函数。Tω 是由神经网络建模的标量函数，其参数 ω 必须与参数 ϕ 一起优化。因此，我们用 JSD(τ˜, z)代替 Iϕ(τ∞, z),同时优化 ϕ 和 ω。

鉴于上述情况，最小化 KL(qϕ(τ˜) padv(τ˜等于

$$\min \text{IJSD}(\tau, z) + E\tau \quad (\tau\sim)【U(\tau\sim)】。(17)$$

ϕ, ω ϕ, ω ϕ

A.2. 定理 2 的证明

因为 G1 等价于 G2，所以 τ1 与 τ2 具有相同的维数。我们用 k 来表示维数，设 τk 是 τ1 的第 k 个元素， τk 是 τ2 的第 k 个元素。由于 1 与 2 相同，即概率密度函数(PDF) pZ1 (z) = pZ2 (z)，我们有

$$\begin{aligned} &\Pr(\tau k < hk, k = 1, 2, \ldots, K) \\ &= \int_{G1(z)k<hk,\ k=1,2,\ldots,K} pZ1(z)dz \\ &= \int_{G2(z)k<hk,\ k=1,2,\ldots,K} \\ &= \Pr(\tau k < hk, k = 1, 2, \ldots, K), \quad (18) \end{aligned}$$

在 hk k=1,2，...，K 是任意实数的列表。因此， 1 的累积分布函数(CDF)等于 2 的 CDF，这证明了 1 与 2.是相同的

A.3. 推论 2.1 的证明

假设 FCN 有 L 层，我们将 Conv(l)，核(L)和 Act(l)分别定义为第 L 层的卷积函数，卷积核和逐元素激活函数。设核(l)的空间大小为 a(l)和 b(l)。我们用 o(l)表示第 l 层激活函数之前的值，用 v(l)表示特征图。我们进一步定义 v(0)为输入 z，定义 v(L)为输出 τ。因此，对于 l ∈ 1，2，...，L ，我们有

$$o(l) = conv(l)(v(l\ 1)) = v(l\ 1) \div 内核(l), \quad (19)$$
$$v(l) = Act(l)(o(l)), \quad (20)$$

其中运算代表卷积。我们用 v(l)表示 $_{我, j, w, h}^{(l)}$ 作为一个大小为 w h 的矩形区域中心分别在 v(l)和 o(l)中的位置 I，j。忽略边界条件，对于所有 l，I，j，I′ ，j′ ，w，h，根据卷积运算的性质，我们有

$$_{I, j, w(l), h(l)}^{(l)} = 法案(l)_{I, j, w(L1), h(L1)}^{(L1)} \int 内核, \quad (21)$$

$$= v(l\ 1) \int 内核, \quad (22)$$

$$五(l)_{I, j, w(l), h(l)} = 和 \int 内核, \quad (23)$$

$$o(l)_{I', j'} \quad (1)$$

$$I'_{(1)}(i', j', w(1), h(1)) \quad I'_{(1)}(i', j', w(L1), h(L1)) \quad I'_{(1)}(i', j', w(1), h(1)), \quad (24)$$

其中 $w(L\ 1) = w(L) + a(L)1$，$h(L\ 1) = h(L) + b(L)1$，$w(L) = w$，$h(L) = h$，因此，我们可以定义一个函数

$$I_{(1)}(i, j, w, h) \quad I_{(1)}(i, j, w, h) \quad I_{(0)}(i, j, w(0), h(0)) \quad I_{(1)}(i, j, w(1), h(1))$$

当 l = 0 时，G(l) 显然等价于 G(l)，因为它们都是相同的函数。此外，分布 (l) I′, j′, w(l), h(l) 也等同于 v(l)，因为 v(0) 的每个元素都是独立的，并且是同分布的。

对于 l > 0，我们假设 G(L1) 相当于 G(L1)，以及 v(L1) 的分布与相同 (L1) I′, j′, w(L1), h(L1) 对于所有的 I，j，I′，j′，w，h。（21）到（24），对于所有 v(0) (0) I′, j′, w(0), h(0)

(l) 我, j, w, h (0) I, j, w(0), h(0) (l) (L1) ∫内核) I, j, w(L1), h(L1)

(l) (l 1) 我, j, w, h (0) I, j, w(0), h(0))) ∫内核(l))

= Act(l)(G(L1) (五(0) I′, j′, w, h I′, j′, w(0), h(0) ∫内核)

= Act(l)(v(L1) (l) I′, j′, w, h (0) I′, j′, w(0), h(0)

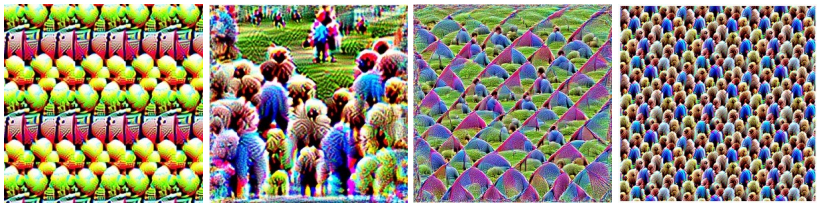
因此，G(l) 相当于 G(l)。此外，由于方程中的卷积。（21）和（23）是等效的，则 o(l) 的分布与 o(l) 的相同根据定理 2。此外，由于主动功能 Act(l) 是逐元素的，即，它对于 I，j 和 v(l) 的分布 I′，j′ 是等价的也与的相同 (l) I′, j′, w(l), h(l) 根据定理 2。

通过运用数学归纳法，我们得出结论 G(l) 相当于 G(l)，以及 v(l) 的分布与 v(l) 的相同对于所有 l ∈ [1，2，...，L]。由于 v(L) = τ，w(l) = w，h(l) = h，我们导出推论 2.1。

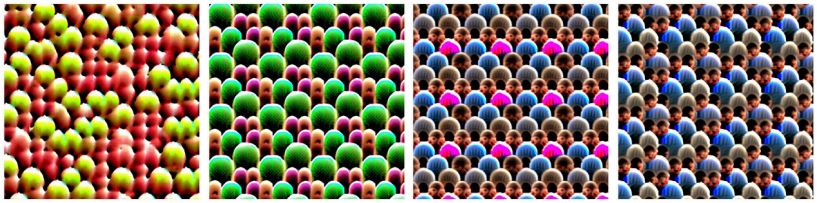
注意，FCN 的每个卷积层都需要零填充，以避免边界问题。

B. 对立的质地和对立的衣服

无花果。S1 和 S2 分别呈现额外的对立纹理和对立服装，由于页数限制，它们没有出现在主论文中（图 6 和 8）。除非另有说明，所有在主要论文和补充材料中出现的关于身体攻击的结果都是通过对抗性 T 恤衫获得的。



AdvTshirtTile (b) RCA2× (c) RCA6× TCA



(e) EGA (f)约洛夫 3 [30](g)faster CNN[31] (h) MaskRCNN [13]

图 S1。不同对立纹理的可视化，扩展了主文件中的图 6。(a)通过平铺对立斑块形成的纹理[38]反复。(b-e)通过不同方法产生的纹理攻击 YOLOv2 [29].(f-h)TC-EGA 产生的纹理分别攻击不同的探测器。

C. 攻击数字世界中不同检测器的结果

标签。S1 给出了在 Inria 测试集上 YOLOv3、FasterRCNN 和 MaskRCNN 的 AP。注意，原始测试图像上每个检测器的 AP 是 1.0。虽然这些 adv 纹理不如 AP 为 0.362 的 YOLOv2 有效

(d)YOLOv2 裙子(e)yolov 3(f)faster(g)MaskRCNN

图 S2。用不同方法制作的真实世界敌对服装，扩展了主文中的图 8。

目标检测器 yolo v3 faster CNN MaskRCNN AP 0.511
0.419 0.492

餐桌 S1。在 Inria 测试集上 TC-EGA 攻击不同检测器的接入点。

(参见选项卡。他们将干净图片的 AP 降低了一半。

D. 室内和室外条件的比较

我们比较了不同对抗性 t 恤在室内和室外场景下的攻击效果。我们使用了主论文第 4.2 节中描述的视频。我们从每个视频中提取了 32 帧，视角从 0º 到 3º。因此，我们为每个场景和每个检测器收集了 3 ^ 32 = 96 帧。结果显示在表中。S2。每件敌对服装的室内 mASR 与室外 mASR 相当。它表明敌对的衣服
在不同的场景下都有效。

目标 / 事件	YOLOv2	YOLOv3	FasterRCNN	MaskRCNN
室内的	0.771	0.764	0.912	0.832
户外的	0.714	0.638	0.948	0.878

餐桌 S2。在人和摄像机之间不同距离的攻击。

E. 攻击的有效性与到摄像机的距离有关

我们为每个穿着 YOLOv2 T 恤的人在室内和室外场景中录制了额外的视频。这些人仍然在相机前慢慢转了一圈，以不同的视角收集画面。我们将摄像机和人之间的距离改变为 1.6 米、2.0 米、2.6 米、3.4 米、4.4 米、5.6 米和 7.0 米。对于每个距离，我们收集
3(人)2(场景)32(每个视频的帧数)=总共 192 帧。图。S3 展示 YOLOv2 T 恤的 mASRs
各种距离。当人靠近摄像机时，mASR 最高(1.6 m，mASR 0.791)。当距离为 7.0 米时，它下降到 0.257

F. 攻击 YOLOv3

在本节中，我们提供了在发送到 YOLOv3 之前将输入大小缩放 50%的原因(参见表。主文中 4)。YOLOv3 有三个分支来预测不同比例的盒子。这些分支基于特征地图
并且在预测盒之前使用额外的块。因此，这些分支在受到对抗性攻击时是相对独立的。因为不同分支预测的盒子数量

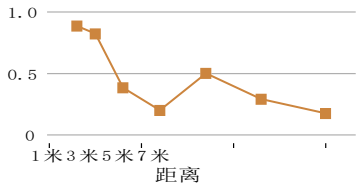


图 S3。在人和摄像机之间不同距离的攻击。

可能非常不同，攻击可能偏向于某个特定的分支。图。S4a 显示 Inria 训练数据集上每个分支的预测框的直方图，置信度阈值为 0.5。第一个分支预测大规模的盒子，第三个分支预测小规模的盒子。图。S4b 呈现相对于不同置信度阈值的预测框的分数。从图中可以看出，第二个分支预测了大多数盒子(当置信阈值为 62.8%时是 0.5)，表明产生的对抗模式可能偏向于攻击第二个分支。然而，在我们的记录的视频，人的规模在第二分支的预测框的范围之外(比较图 1 和 2)。S4a 和 S4c).因此，在将帧发送到 YOLOv3 之前，我们将输入的大小缩放了 50%。

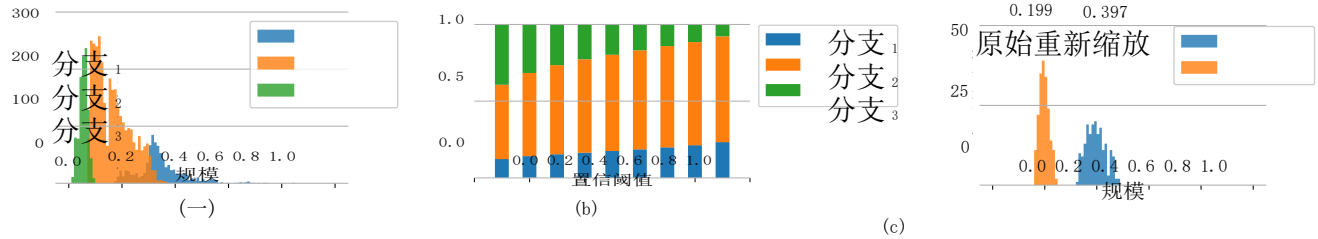


图 S4。(a)由 YOLOv3 的不同分支预测的盒尺度的 dist √分布。对于每个具有归一化尺寸 w h 的盒子，我们通过 w∫h 来定义尺度。(b)由不同分支相对于各种置信度阈值预测的盒子的分数。(c)原始和重新缩放的视频帧上的框的比例分布。红色实线表示原始视频帧的平均比例，红色虚线表示重新调整后的帧的平均比例(50%)。

G. 物理世界中的迁移研究

我们通过产生攻击特定检测器的敌对服装对几个检测器进行基于转移的攻击。标签。S3 给出了基于传输的攻击的 mASR。表中的每个数字都是在 192 帧上获得的，如主要论文的 4.2 节所述。YOLOv2 和 YOLOv3 的敌对服装在以下情况下仍然有效
它们分别被用来攻击 YOLOv3 和 YOLOv2。然而，这些衣服在攻击除 RetinaNet 之外的其他模特时得到了低屏蔽。更快的 RCNN 和 MaskRCNN 的敌对服装在用于攻击其他模型时仍然有效，尽管有时(例如攻击 YOLOv3)不如攻击自己有效。一个可能的解决方案是使用模型集合技术[9,23]，这是留给以后研究的。

目标 / 源 YOLOv2	YOLOv3	快速 CNN	马斯克 CNN	RetinaNet [21	级联掩蔽 CNN [5
YOLOv2 0.743	0.526	0.000	0.000	0.182	0.000
YOLOv3 0.518	0.701	0.014	0.037	0.453	0.009
快速 CNN 0.617	0.237	0.930	0.848	0.900	0.695
MaskRCNN 0.547	0.359	0.873	0.855	0.838	0.575

餐桌 S3。转移攻击的伪装。