PhysGAN:为自动驾驶生成物理世界弹性对抗示例

孔泽伦、郭俊峰、李昂和刘丛
达拉斯德克萨斯大学

马里兰大学帕克分校

摘要

尽管深度神经网络(DNNs)正被广泛用于基于视觉的自动驾驶系统，但人们发现它们容易受到敌对攻击，即在测试期间输入的小幅度扰动会导致输出发生巨大变化。虽然大多数最近的攻击方法针对数字世界的各种场景，但是不清楚它们在物理世界中如何执行，更重要的是，在这些方法下产生的干扰将覆盖整个驾驶场景，包括那些固定的背景图像，例如天空，使得它们不适用于物理世界的实现。我们提出了 PhysGAN，它以连续的方式生成了误导自动驾驶系统的物理世界弹性对抗示例。我们通过广泛的数字和现实世界评估展示了 PhysGAN 的有效性和健壮性。我们将 PhysGAN 与一组最先进的基线方法进行了比较，这进一步证明了我们方法的稳健性和有效性。我们还表明，PhysGAN 优于最先进的基线方法。据我们所知，PhysGAN 可能是第一个为攻击常见的自动驾驶场景生成现实和物理世界弹性对抗示例的技术。

1. 介绍

虽然深度神经网络(DNNs)已经建立了基于视觉的自动驾驶系统的基础，但它们仍然容易受到恶意攻击和表现出错误的致命行为。最近关于对抗性机器学习研究的工作表明，dnn 在集中于分类问题的干扰下相当容易受到有意的对抗性输入的影响[4，12，19，22，25]。为了解决自动驾驶系统中的安全问题，提出了自动生成对抗示例的技术，该技术添加了小数量的 pertur-

*现在在 DeepMind

图 1:与原始标志(左上)在视觉上无法区分的对抗性路边广告标志(右上)及其在现实世界中的部署(下图)的图示。

评估基于 DNN 的自动驾驶系统的鲁棒性的输入 bations[12，8，27]。

然而，这些技术主要集中于生成数字对抗示例，例如，改变图像像素，这在现实世界中永远不会发生[12]。它们可能不适用于真实的驾驶场景，因为在这种技术下产生的扰动将覆盖整个场景，包括固定的背景图像，例如天空。最近，一些作品迈出了研究物理世界攻击/测试静态物理对象[2，17]，人体对象[24，7]，交通标志[23，18，8]的第一步。虽然它们在目标场景和某些假设下显示有效，但它们专注于研究静态的物理世界场景(如停车标志的单个快照[8，23])，这妨碍了它们在实践中的应用，因为真实世界的驾驶是一个连续的过程，会遇到动态变化(如视角和距离)。此外，它们生成的对抗性例子在视觉上是不真实的(例如，粘贴在停车标志上的驾驶员可注意到的黑白标签很容易注意到攻击目的[8])。这些方法中的大多数也关注于不同于我们所研究的转向模型的分类模型，转向模型是回归模型。还要注意，直接扩展现有的数字 per-

这个分析存疑，相关工作没有这么弱，有点夸张

干扰生成技术(例如，FGSM)对于物理世界设置可能是无效的，即，仅将目标路边标志输入到这种技术中会输出相应的对立示例。由于产生扰动的过程没有考虑物理世界中任何潜在的背景图像(例如天空),因此导致的攻击效率可能会显著降低(在我们的评估中也证明了这一点),这些图像会在驾驶过程中被任何摄像机捕捉到。

我们的目标是生成一个真实的单个敌对示例，该示例可以被物理打印出来，以替换相应的原始路边物体，如图 1 所示。由于目标车辆持续观察这种对抗打印输出，这里的主要挑战是如何生成单个对抗示例，该示例可以在驾驶过程中的每一帧持续误导转向模型。此外，对于一个实际的物理世界部署，任何生成的对立示例都应该与其原始符号(已经在物理世界中部署的符号)在视觉上无法区分。

为了应对这些挑战，我们提出了一种新的基于 GAN 的框架，称为 PhysGAN 1，它通过观察驾驶过程中捕获的多个帧来生成单个敌对示例，同时保持对特定物理世界条件的弹性。我们的架构包含一个编码器(即目标自动驾驶模型的 CNN 部分)，它在驾驶过程中从帧中提取特征，并将它们转换为一个向量，作为生成器的输入。通过考虑从帧中提取的所有因素，这种设计确保生成器能够生成具有攻击效果的通用示例。如果没有这个编码器，效率会大大降低。为了生成一个可以不断误导导向模型的对立示例，PhysGAN 将一个 3D 张量作为输入。这增强了生成的样本对某些物理世界动态的弹性，因为使用视频切片更有可能捕捉到这种动态。

我们通过使用一组最先进的转向模型和数据集进行广泛的数字和现实世界实验，证明了 PhysGAN 的有效性和鲁棒性。数字实验结果表明，PhysGAN 对各种转向模型和场景都是有效的，能够将平均转向角误导高达 21.85 度。物理案例研究进一步证明，PhysGAN 在生成物理世界的敌对实例方面具有足够的弹性，能够将平均转向角误差高达 19.17 度。通过与一套全面的基线方法进行比较，也证明了这种功效。

据我们所知，PhysGAN 是第一个为攻击普通自治系统而产生现实的和对物理世界有弹性的对抗范例的技术

1 https://github . com/kongzelun/phys gan . git

转向系统。我们的贡献可以归纳为以下三个方面。

- 我们提出了一种新的基于 GAN 的框架 Phys- GAN，它可以生成对应于任何路边流量的物理世界弹性通用示例
  fic/广告标志，并使用生成的视觉上难以辨认的对立示例误导自动驾驶转向模型。
- 我们提出了一种 GAN 架构，使用 3D 张量作为优化生成器的输入，解决了物理世界部署中的一个关键技术挑战
  在整个驾驶过程中使用一个对立的例子来不断误导驾驶。
- 我们进行了广泛的数字和物理世界的评估，用几个指标，这表明了优越
  PhysGAN 的攻击性能优于最先进的方法。我们相信 PhysGAN 可以为自动驾驶的未来安全研究做出贡献。

2. 相关作品

对抗性攻击。最近已经提出了许多工作来产生在白盒环境中攻击的对抗例子[21]，其中对手知道网络的参数。快速梯度符号方法(FGSM) [11]代表了这些方法中的先驱，它沿着每个像素处梯度符号的方向执行一步梯度更新。FGSM 在[13]中进一步扩展为通过最大化目标类别的概率的目标攻击策略，这被称为 OTCM 攻击。还提出了基于优化的方法[26，14，4，5，29]。GAN 是最近在[10]中介绍的，由两个神经网络系统在零和游戏框架中相互竞争来实现。GAN 在人脸生成[16]和操作[30]方面都取得了视觉上吸引人的效果。

[29]提出了 AdvGAN，它利用 GAN 产生在分类问题上具有高攻击成功率的对抗例子。这些方法侧重于将干扰应用于整个输入，并且只考虑数字世界的攻击场景。很难将它们应用于现实世界，因为不可能使用一些生成的扰动来代替现实世界的背景(例如天空)。产生身体对抗的例子。据我们所知，只有最近的一套作品[15，8]开始致力于产生物理攻击。[15]侧重于对静态物理对立例子的理解。[8]明确地将扰动设计为在不同的真实世界条件下有效。他们的方法主要集中在动态距离和视角下的物理路标分类。不幸的是，这些工作集中于静态攻击场景(例如，最大化对抗效果

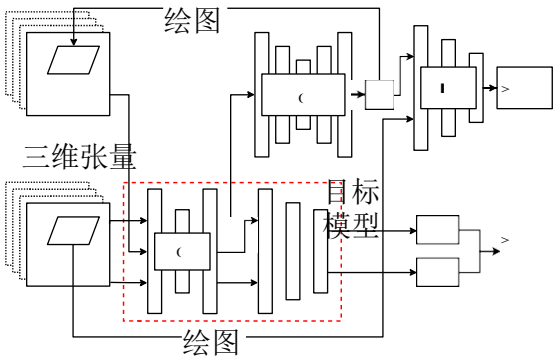


图 PhysGAN 框架概述。

物理示例的单个快照)，因此不需要解决一对多挑战。

与这些作品不同的是，PhysGAN 只能根据路边的交通/广告标志来生成对物理世界有弹性的对立例子；除了路牌以外的区域不会产生干扰。PhysGAN 解决了不断攻击转向模型的一对多挑战，并生成了真实的对抗性示例，这些示例对各种物理世界条件具有弹性，并且在视觉上与原始路边标志无法区分。

3. 我们的方法:PhysGAN

PhysGAN 的目标是生成与任何普通路边物体(例如，路边交通或广告标志)在视觉上无法区分的敌对示例，以通过用敌对示例物理地替换路边广告牌来持续误导驶过的自动驾驶车辆的转向角模型(目标模型)。当自动驾驶车辆驶过路边标志时，转向角模型会被愚弄而做出错误的决定。

3.1. 问题定义

我们在这一节中定义我们问题和符号。设 X = {Xi} 是视频切片集，使得 X ⊆ Rn w h，其中 n 是视频切片中的帧数，w 和
他的宽度和高度分别是一个框架。设 Y = {Yi}为地面真实转向角集合，Y ⊆ Rn。假设(Xi，易)是数据集中的第 I 个样本
由视频切片∈ X 和易∈ Y 组成，每个
其元素表示地面真实转向角
对应于它的框架。预训练的目标操纵模型(例如，Nvidia Dave-2、Udacity Cg23 和兰博)从视频切片集 X 到学习映射 f : X → Y
在训练阶段，地面真实转向角设置为 Y。
举一个例子(，易)，PhysGAN 的目标是制作一个对抗性的路边标志 Sadv，其目的是误导目标 a σ 自主驾驶模型 f 为 f (Xi) /=
易和最大化 伊|法(Xi)——易|。为了实现这个目标，
PhysGAN 需要生成一个对抗性的路边标志
Sadv 将取代数字或物理世界中的原始路边标志 Sorig。根据ℓ2-norm 距离度量，敌对的路边标志 Sadv 被假定为接近于原始的路边标志 Sorig，这意味着敌对的
路边标志副词 和原始路边标志 原始 都是-
视觉上最难分辨。

3.2. 物理世界挑战

对一个物体的物理攻击必须能够在不断变化的条件下生存，并在欺骗转向角模型时保持有效。我们围绕常见的驾车场景(例如，车辆驶向路边标志)对这些条件进行决策。

“一对多”的挑战。一个关键的技术挑战是解决“一对多”的挑战，即在整个驾驶过程中生成单个对立样本来连续误导车辆的转向角度决策。在生成敌对样本时考虑多个帧是一个挑战，因为车辆到电路板的距离、视角，甚至每个帧上的细微像素都可能不同。一个有效的对抗样本必须能够在所有框架中展现出最大的整体攻击效果。为了实现这个目标，敌对样本需要对每一帧中表现出的变化的条件有弹性。为了解决这一问题，PhysGAN 应用了一种基于 GAN 的新型框架，并将整个驾车视频片段(而不是单个帧)视为生成过程中的输入(参见第。3.5).

有限的操作区域。与大多数将扰动添加到整个输入图像的数字世界对抗方法不同，专注于物理世界场景的技术被限制为仅将扰动添加到图像的片段，即对应于原始物理对象的片段区域。此外，静态图像背景的基本假设在物理攻击中不成立，因为背景可以在驾驶过程中不断变化。

3.3. PhysGAN 概述

图 2 示出了 PhysGAN 的总体架构，其主要包括四个组件:编码器 E、生成器 G、鉴别器 D 和目标自治

驱动模型 f。编码器 E 表示卷积-
目标自动驾驶模型 f 的可选层
将 3D 张量作为输入，并用于提取视频的特征(原始的和扰动的)。为了解决仅生成持续影响驱动过程的单个示例的挑战，我们引入了一个新的想法，即在基于 GAN 的框架中将 3D 张量视为输入。2D 张量通常表示图像，而 3D

张量用于表示一小部分视频，通常包含数百帧。

如图 2 所示，提取的原始视频切片 Xorig 的特征被用作馈送到生成器的输入，以生成敌对的路边标志 Sadv。这样做允许我们考虑不同的原始视频切片 Xorig 可能对生成的对抗性路边标志 Sadv 具有不同影响的事实，从而确保生成器
g .生成对应于某个原始视频切片 Xorig 的最佳对抗性路边标志 Sadv。广告
sarial 路侧标志 Sadv 和原始路侧标志 Sorig 被发送到鉴别器 D，该鉴别器用于区分对抗性路侧标志 Sadv 和原始路侧标志 Sorig。鉴别器 D 代表损失函数，其测量敌对的视觉区别
路边标志 Sadv 和原始路边标志 Sorig，并且还鼓励生成器生成与原始标志在视觉上不可区分的示例。

3.4. 与目标模型一起训练 GAN

为了确保对抗性路边标志 Sadv 对目标自动驾驶产生对抗性影响

模型 f，我们引入以下损失函数:

$$\text{低频} = \beta \exp\left(-\beta \cdot \text{lf}(f(X), f(X))\right) \quad (1)$$

其中 β 是锐度参数，lf 表示用于训练目标自动驾驶模型 f 的损失函数，
如 MSE-loss 或ℓ 1-loss，Xorig 表示原始视频切片 Xorig，Xadv 表示对抗性视频切片 Xadv，它是通过将对抗性路边标志 Sadv 映射到原始视频的每一帧中而生成的
切片 Xorig。通过最小化 Lf ， 之间的距离
预测和基本事实将被最大化，
这确保了对抗的有效性。

为了计算 Lf，我们通过用生成的对抗性路边标记 Sadv 替换原始路边标记 Sorig 来获得对抗性视频片段 Xadv。注意，所生成的对抗性路侧标志 Sadv 是矩形图像，并且视频切片中的原始路侧标志 Sorig
可以呈现任意的四边形形状，其可以在不同的框架中变化。我们利用经典的透视映射方法[1]来解决这种不匹配。我们首先获得原始路侧标志 Sorig 在每一帧内的四个坐标，然后将生成的敌对路侧标志 Sadv 映射到每一帧内相应的四边形区域上(细节可在补充材料中找到)。

PhysGAN 的最终目标表示为:

$$L = L_{GAN} + \lambda L_f, \quad (2)$$

其中 λ 表示平衡两项之间折衷的系数，LGAN 是经典的 GAN 损耗，

PhysGAN 算法 1 优化需要:I -迭代次数;
要求:f -目标模型，参数固定;
1:而我< I:做
2:Sadv = G(E(Xorig));
3:LGAN = log D(Sorig)+log(1d(Sadv));
4: //固定 G 的参数
5: 做反向传播优化 argmax lgan;
6: Sadv = G(E(Xorig))
7:LGAN = log D(Sorig)+log(1d(Sadv));
8: //固定 D 的参数
9:对于输入视频片段中的每一帧，执行 per-
使用对立的路侧标志 Sadv 替换原始路侧标志 Sorig。
10: 做反向传播优化 argmin lgan;
11:LADV =β exp(β lf(f(Xorig)));
12: 做反向传播优化 argmin ladv;
13:结束时间

这可以表示为

$$L_{GAN} = E_{Sorig \sim pS}[\log D(Sorig)] + E_{Sadv \sim pS}[\log(1d(Sadv))]。 \quad (3)$$

为了解释这个目标函数，L 开始 鼓励
对立的路边标志 Sadv 在视觉上与
原始路边标志 Sorig，而 Lf 杠杆作用于
生成最大化 at 的对抗视频切片 Xadv
粘性有效性。我们得到编码器 E，发电机
g 和鉴别器 D，通过求解:

参数最小值最大值 L. (4)

3.5. 用 PhysGAN 攻击

我们假设目标自动驾驶模型 f 是预先训练好的，目标自动驾驶模型 f 的参数是固定的，PhysGAN 的生成器 G 只能在训练时访问目标自动驾驶模型 f 的参数。我们训练 PhysGAN 的算法在算法 1 中示出，该算法包括
两个阶段。如算法 1 所示，第一阶段是训练鉴频器 D，稍后用于形成 LGAN 的一部分(第 2 - 5 行)；第二阶段是用两个损失函数 Lf 和 LGAN 训练生成器 G，这促使生成器 G 生成视觉上不可分辨的对抗样本，并使生成的样本
ple 对目标自动驾驶模型是不利的
分别为 f(第 6 - 11 行)。编码器 E，也就是
CNN 部分目标自动驾驶车型 f，aims
从所有观察到的帧中提取特征
驱动并将它们转换成输入到发生器的矢量。这种设计保证了发电机可以发电

通过考虑从视频切片中提取的所有有用特征，给出了攻击效果在视觉上不可区分的例子。对于物理世界部署，攻击者应打印与目标路边标志相同大小的敌对示例，以确保视觉上的不可分辨性。

4. 实验

我们使用广泛研究的基于 CNN 的操纵模型和流行的数据集，通过数字和物理世界评估来评估 PhysGAN。

4.1. 实验设置

转向模型。我们在几个流行的和广泛研究的[6，28，3]基于 CNN 的转向模型上评估 PhysGAN，如 NVIDIA Dave-2 2、Udacity Cg23 3 和 Udac-city 兰博 4。值得注意的是，由于原始模型应用了用单个图像训练的 2D CNN，我们将 2D CNN 改编成 3D CNN，并用一组 20 帧视频切片训练 3D CNN。

数据集。在我们的数字实验中使用的数据集包括(1) Udacity 自动驾驶汽车挑战数据集 5，它包含由驾驶汽车的仪表板安装的相机捕获的 101396 幅训练图像和由人类驾驶员为每幅图像应用的同时方向盘角度；(2) DAVE-2 测试数据集[20] 6，包含 45，568 张图片，用于测试 NVIDIA DAVE-2 模型；(3) Kitti [9]数据集，其包含来自六个不同场景的 14，999 个图像，由配备有四个摄像机的 VW Passat 旅行车捕获；以及(4)用于物理世界评估的定制数据集，其包含超过 20000 帧用于在物理情况下训练 PhysGAN。

对于物理世界的实验，我们首先执行颜色增强以提高图像对比度，使对立的例子在变化的光照条件下更加鲁棒。然后，我们打印出每个评估方法下生成的示例，并将其粘贴到选定的路边物体上。我们通过该对象驾驶车辆，并使用捕获的驾驶视频进行离线分析。为了了解 PhysGAN 在实际的自动驾驶车辆上的表现，我们还进行了在线驾驶测试，模拟了面对这种敌对的路边物体时的真实驾驶过程。

视频切片选择标准。我们的驾驶场景选择标准是路边交通或广告标志应该完全出现在驾驶视频的第一帧中

2 https://dev blogs . NVIDIA . com/深度学习-自动驾驶-汽车/

3 https://github . com/uda city/self-driving-car/tree/master/steering-models/community-models/cg23

4 https://github . com/uda city/自驾-汽车/树/主人/转向-模型/社区-模型/兰博

5 https://medium . com/uda city/challenge-2-using-deep-learning-to-predict-steering-angles-f 42004 a 36 ff 3

6 https://github . com/sully Chen/driving-datasets

景色	形象	大小	部	最大
戴夫-直道 1	20	455 × 256	21 × 22	41 × 49
戴夫-曲线 1	20	455 × 256	29 × 32	51 × 49
乌达城-直线 1	20	640 × 480	48 × 29	66 × 35
uda city-曲线 1	20	640 × 480	51 × 51	155 × 156
Kitti-straight1	20	455 × 1392	56 × 74	121 × 162
Kitti-straight2	20	455 × 1392	80 × 46	247 × 100
基蒂曲线 1	20	455 × 1392	64 × 74	173 × 223

表 1:实验中评估的场景。

具有超过 400 个像素，并且在最后一帧中部分消失。我们从上述数据集中选择 7 个场景，并对所有选择的场景进行评估。每个数据集中所选的场景包括直道和弯道场景。由于所有这些数据集都不包含路边标志的坐标，我们必须在所选场景的每一帧中标记标志的四个角。我们使用 Adobe After Effects 7 的运动跟踪器功能来自动跟踪连续帧中四个角的标志运动。表 1 显示了我们选择的场景的属性。

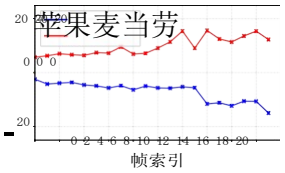
基线方法。我们将 PhysGAN 与几种基线方法进行了比较:

- *原始标志。第一个基线是简单地测试原始路边标志。这种比较很重要，因为它验证了转向角误差是否是由于*
  PhysGAN 但不是原来的标志。我们在数字和物理评估中都包括了这个基线。
- *FGSM。FGSM [11]非常强大，它旨在通过利用*
  渐变。在我们的问题上下文中，我们直接应用 FGSM 来生成给定的捕获输入帧的扰动。我们在数字评估中仅包括 FGSM，因为不可能在物理世界中应用覆盖整个图像帧(如天空)的生成扰动。
- *物理科学。为了在物理世界环境中应用 FGSM，我们开发了一种称为 Phys- FGSM 的新方法作为附加基线，它基于*
  并且仅对输入图像中的目标路边标志产生扰动。这样做可以让我们打印出混乱的图像，并粘贴到相应的标志上。我们在数字评估和物理评估中都包括 PhysFGSM。因为视频切片包含多个帧，所以物理切片基于中间帧产生扰动。
- *RP2。我们还将 PhysGAN 与物理世界基线 RP2 [8]进行了比较，RP2 是一种为单个输入场景生成扰动的优化方法。*
  最初的 RP2 方法侧重于分类

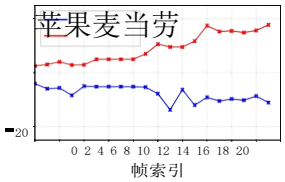
[7]https://www.adobe.com/products/aftereffects.html

	戴夫		乌达城		凯蒂		
	直道 1	曲线 1	直道 1	曲线 1	直道 1	直道 2	曲线 1
苹果							
麦当劳							

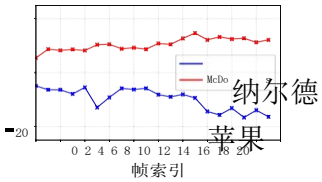
表 2:原始和生成的对抗性片段以及各种场景下对应的图像帧。



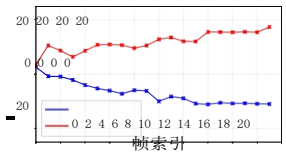
(a) 戴夫·斯特里特 1



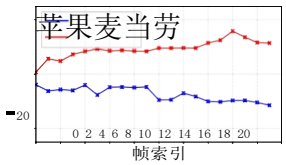
(b) 戴夫曲线 1



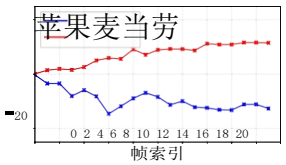
(c) Udacity 直线 1



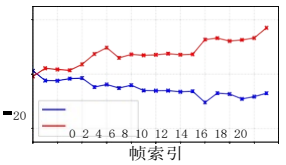
(d) Udacity 曲线 1



(e) Kitti 直线 1



(f) Kitti 直道 2



(g) 基蒂曲线 1

图 3:苹果麦当劳Dave-2 转向模型上转向角度误差随时间变化的图示。 苹果麦当劳

问题，所以我们通过用回归损失代替分类损失来扩展它以适用于操纵模块。

- *随机噪声。我们还打印了一张受随机噪声干扰的图像，并将其粘贴在路边的标志上。*

评估指标。在我们的实验中，我们使用两个指标来评估 PhysGAN 的功效:转向角均方误差(表示为转向角 MSE)和最大转向角误差(MSAE)。转向角 MSE 测量预测的转向角和地面真实值之间的误差的平方的平均值，MSAE 表示在属于一个视频切片的所有帧中观察到的最大转向角误差。大的转向角 MSE 和 MSAE 意味着更好的攻击效能。

此外，我们还进行了在线驾驶测试案例研究，根据实时计算得出的各种评估方法下的转向角误差，手动控制每一帧中的转向角(大约)。我们在这里使用指标“到达路缘的时间”来衡量攻击效能，它衡量实际自动驾驶车辆开到路缘上所需的时间。请注意，所有结果与地面真实转向角相关。

4.2. 结果

我们首先报告了 PhysGAN 在数字和物理世界场景中的总体功效。补充文件中给出了一套完整的结果。

数字场景的结果。表 2 显示了每个场景的代表帧，其中符号被替换为从 PhysGAN 生成的对立示例(使用目标导向模型 NVIDIA Dave-2)。表 2 的每一列代表一个特定的场景。据观察，Phys- GAN 可以生成相当真实的敌对样本，在视觉上无法与原始对象区分开来。原始视频片段中的焦油化路边标志被我们选择的麦当劳和 Apple Watch 标志取代，修改后的视频片段用于所有实验。这是因为原始视频切片中的路边标志的分辨率较低，这使得我们很难验证生成的路边标志在视觉上是否可区分。

图 3 示出了在每个帧场景中沿着时间线的转向角误差的结果，其中不利图像的尺寸随着时间几乎单调增加。图 3 中的每个子图表示特定的场景，其中

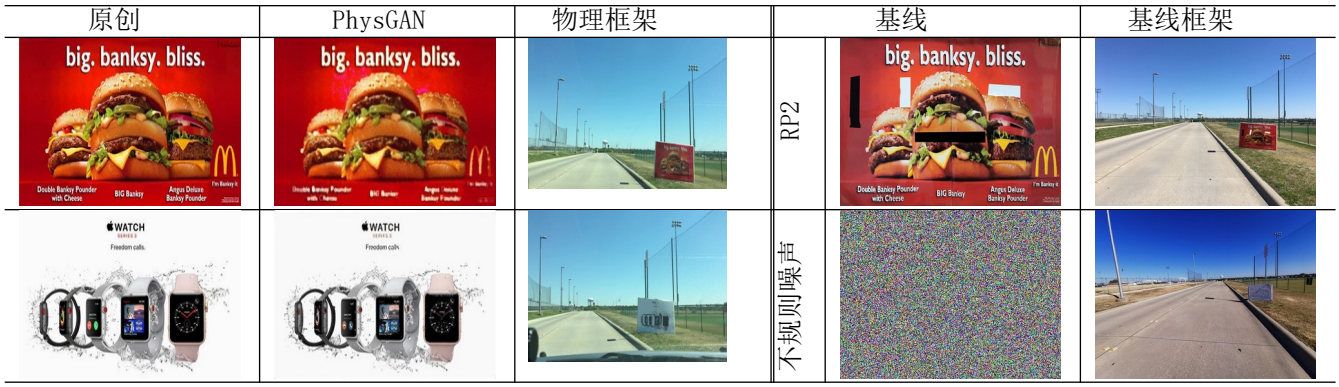


表 3:不同方法下的物理世界对抗场景的图解。

框架编号	一	2	3	四	5	6	七	8	9	10	11	12	13	14	15	16	17	18	19	20
原始苹果标志	0.36	−0.51	0.82	0.45	0.10	−0.16	0.84	−1.38	−2.16	-0.86	0.60	-1.11	0.21	-0.49	-0.55	-0.56	0.10	-0.51	0.49	-1.00
PhysGAN(苹果)	0.17	0.89	1.68	7.94	1.93	4.79	2.87	6.34	2.08	3.54	9.06	8.37	5.93	12.51	13.43	11.37	12.75	11.74	13.63	13.44
最初的麦当劳标志	−0.17	−0.42	−1.49	−1.34	−0.51	−0.08	0.60	−0.35	0.70	-0.75	-0.43	-0.35	0.59	-0.89	1.49	0.61	0.94	-0.99	1.13	-0.00
PhysGAN(麦当劳)	−1.24	−1.37	−0.02	−0.30	−2.48	−0.17	−1.06	−0.80	−0.01	-5.37	-1.60	-2.62	−2.45	-4.68	-11.71	-10.85	-9.83	-8.74	−11.35	−19.17

表 4:物理世界实验下的每帧转向角误差。第 2 行和第 4 行(第 3 行和第 5 行)示出了当原始标志(由 PhysGAN 生成的相应敌对标志)展开时的转向角误差。

x 轴表示沿时间轴的帧索引，y 轴表示转向角误差。我们清楚地观察到，PhysGAN 会导致几乎所有帧都出现明显的角度误差，甚至对于横向样本与背景相比相对较小的早期帧也是如此。

现实世界场景的结果。我们如下执行物理世界实验。我们首先录制驾驶车辆驶向原始路边标志的训练视频，并使用这些视频来训练 Dave-2 模型。然后，我们按照与数字世界评估相同的配置来训练 PhysGAN，以生成对抗样本。然后将生成的对立样本打印并粘贴在原始路边标志上。然后，我们记录了相同的驾车经过过程的测试视频，但使用的是对抗性样本。然后通过分析这些测试视频获得转向角误差。具体来说，对于训练和测试视频片段，我们从 70 英尺远开始记录，并在车辆实际经过广告标志时停止记录。用于培训视频，行车速度设为 10mph 捕捉 suf-
虚构的图像。测试视频的速度设置为 20 英里每小时，以反映普通校园内的驾驶速度限制。物理案例研究在直道上进行
出于安全考虑。我们实验中使用的路边广告牌尺寸为 48′ ×72′ 。

表 3 显示了原始标志和在不同方法下生成的相应对抗示例，以及每个示例的摄像机捕捉的场景。为了清楚地解释结果，我们列出了由于 PhysGAN 和使用原始符号导致的每帧转向角误差
在表 4 中(其他比较结果详见第二节。4.3).如表 4 所示，PhysGAN 能够生成单个可打印的物理世界弹性对抗示例，该示例可能在整个驾驶过程中误导连续帧的驾驶模型。这里一个有趣的观察是转向角误差趋向于随着帧索引的增加而增加。这是因为，在帧索引较大的情况下，敌对样本的大小在整个帧中占据相对较大的空间，因此能够对操纵模型产生更负面的影响。此外，我们观察到，对于原始路边标志，在所有帧下转向角误差几乎可以忽略不计。

4.3. 与基线方法的比较

数字基线。对于每个转向模型，我们将我们的方法与其他四个基线进行比较，包括 FGSM、PhysFGSM、随机噪声和原始符号。表 5 显示了七个不同场景的结果。这些结果表明:(1)虽然 FGSM 实现了最高的攻击效果，但它需要对整个场景应用扰动，这不适用于物理世界；(2)我们的方法的攻击效果比 PhysFGSM 好得多，这意味着一旦考虑物理世界的实现约束，PhysGAN 将优于现有方法的直接扩展。
(3)由于在随机噪声和原始符号下的角度误差是微不足道的，所以每个转向模型是相当鲁棒的。

物理基线。对于物理世界的场景，我们将 PhysGAN 与 PhysFGSM、随机噪声和原始符号进行比较。结果如表 6 所示。我们观察到随机噪声和原始符号都具有可忽略的 im-1

转向模型	方法	戴夫		乌达城		凯蒂		
		直道 1	曲线 1	直道 1	曲线 1	直道 1	直道 2	曲线 1
Nvidia Dave-2	PhysGAN	106.69 / 15.61	66.79 / 11.63	81.76 / 17.04	114.13 / 14.64	108.76 / 17.72	150.00 / 17.34	95.87 / 15.83
	FGSM	115.91 / 17.41	199.27 / 19.47	141.23 / 16.17	192.19 / 21.23	156.16 / 17.84	217.52 / 19.50	103.38 / 14.54
	物理科学	15.88 / 6.42	4.73 / 4.87	13.91 / 5.74	3.08 / 2.89	15.17 / 8.04	8.67 / 4.54	13.12 / 7.24
	不规则噪声	3.00 / 2.01	2.25 / 2.37	2.36 / 2.60	1.77 / 3.10	3.15 / 3.16	1.60 / 0.96	5.92 / 4.41
	原始标志	4.17 / 3.15	4.35 / 2.40	3.84 / 1.79	1.09 / 0.72	4.20 / 2.98	3.06 / 1.23	2.86 / 1.30
Udacity Cg23	PhysGAN	91.85 / 13.80	113.41 / 14.78	50.61 / 10.43	78.56 / 15.46	46.53 / 11.72	62.64 / 11.64	71.09 / 18.14
	FGSM	203.34 / 19.70	157.98 / 14.67	171.92 / 19.89	96.74 / 17.75	136.08 / 14.00	162.35 / 18.53	89.75 / 16.71
	物理科学	58.53 / 11.86	36.44 / 10.68	30.72 / 9.41	46.74 / 8.88	28.89 / 11.37	22.63 / 7.61	61.23 / 10.95
	不规则噪声	5.32 / 3.67	3.75 / 2.72	4.05 / 2.52	4.20 / 2.26	5.31 / 4.49	6.54 / 1.98	6.10 / 3.68
	原始标志	4.17 / 3.15	4.35 / 2.40	3.84 / 1.79	4.09 / 2.72	4.20 / 2.98	3.06 / 1.23	2.30 / 1.86
兰博乌达城	PhysGAN	61.87 / 11.28	113.78 / 15.29	87.68 / 13.90	42.71 / 12.55	56.41 / 12.42	58.67 / 10.42	145.66 / 21.85
	FGSM	209.81 / 21.78	147.28 / 16.43	151.14 / 15.28	166.50 / 16.27	169.17 / 18.57	126.14 / 14.19	175.28 / 19.36
	物理科学	16.43 / 8.95	14.24 / 8.34	5.32 / 3.73	14.82 / 6.11	16.58 / 7.78	13.89 / 7.93	29.58 / 19.18
	不规则噪声	1.90 / 2.55	3.49 / 5.79	6.06 / 5.00	1.92 / 3.98	3.82 / 5.42	2.09 / 3.05	1.52 / 1.91
	原始标志	3.93 / 2.01	6.30 / 4.46	1.80 / 1.28	6.54 / 2.52	5.06 / 3.52	5.75 / 4.03	4.95 / 2.07

表 5:所有评估方法下的转向角 MSE(左)和 MSAE(右)。虽然 FGSM 产生最大的攻击，但是它修改了整个图像观察，并且不适用于真实世界。在所有物理世界的攻击方法中，我们的方法 PhysGAN 产生了最好的性能。

	PhysGAN	RP2	不规则噪声	原始标志
Nvidia Dave-2	73.94 / 13.63	23.48 / 6.52	2.48 / 1.02	2.12 / 1.56
Udacity Cg23	99.23 / 14.56	25.15 / 7.86	2.56 / 2.11	2.15 / 1.73
兰博乌达城	87.56 / 17.60	32.54 / 7.51	1.51 / 1.15	3.12 / 2.48

表 PhysGAN、RP2、随机噪声和原始符号下的转向角 MSE(左)和 MSAE(右)。

转向模型协定，表明预先训练的转向模型(没有被攻击)在现实世界环境中足够强大。如表 6 所示，Phys-GAN 的性能明显优于 RP2，在所有转向模式下都可以实现非常高的攻击效率，这可能会导致现实世界中的危险驾驶行为。

	PhysGAN	RP2	不规则噪声	原创
停车时间	10s	-	-	-
中心距离	1.5 米	1.09 米	0.29 米	0.47 米

表 7:在线驾驶测试结果。第二行显示到达路缘的时间结果，第三行显示车辆偏离正确路径(即直线行驶)的最大距离。

4.4. 在线驾驶案例研究

上述评估不符合政策，驾驶轨迹不受敌对标志的影响。在本节中，我们进一步进行政策评估，即模拟实际驾驶场景的在线驾驶案例研究，以了解 PhysGAN 如何影响实际自动驾驶汽车的转向决策。在这些案例研究中，我们在每一帧中根据下计算的转向角误差实时手动控制转向

每种方法都采用 Nvidia Dave-2 的转向模式。我们要求人类驾驶员以 5 英里/小时的速度驾驶车辆 1 秒钟，以反映一个帧和相应的手动转向 ac-

tion。我们注意到这个在线评估设置是

真正的自主车辆，并提供攻击系统的适当评估。我们不使用虚拟模拟器进行评估，因为它们通常会导致模拟到真实的转换问题。因此，模拟器上的评估结果不会反映模型在现实世界中的能力。如表 7 所示，在这两个指标下，PhysGAN 优于其他基线。此外，只有 PhysGAN 下生成的敌对标志会引导车辆驶上路边石，这需要 10 秒钟(考虑到安全问题，驾驶速度非常低)。这一在线驾驶案例研究进一步证明了自动驾驶汽车由于 PhysGAN 而可能采取的危险转向动作，表明了它在应用于实际自动驾驶汽车时的有效性。

5. 结论

我们提出了 PhysGAN，它为误导自主转向系统产生了物理世界弹性对抗的例子。我们提出了一种新的基于 GAN 的框架，用于生成在整个过程中不断误导驾驶模型的单个对立示例。所生成的对立范例在视觉上与原始路边物体不可区分。大量的数字和物理实验表明了 PhysGAN 的有效性和鲁棒性。我们希望我们的工作可以激励未来对自动驾驶的安全和鲁棒的机器学习的研究。

参考

[1] 透视映射，https://www。geometrictools.com/Documentation/PerspectiveMappings.pdf。四

[2] 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。arXiv 预印本 arXiv:1707.07397，2017。一

[3] 阿尔贝托·布罗吉、米歇尔·布佐尼、斯特凡诺·德伯蒂斯蒂、保罗·格里斯勒里、玛丽亚·基娅拉·拉吉、保罗·美第奇和彼得罗·弗萨里。自动驾驶技术的广泛测试。IEEE 智能交通系统汇刊，14(3):1403－1415，2013。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。2017 年 IEEE 安全与隐私研讨会(SP)，第 39－57 页。IEEE，2017。1，2

[5] 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。2

[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。IEEE 计算机视觉国际会议论文集，第 2722－2730 页，2015 年。5

[7] Gamaleldin Elsayed、Shreya Shankar、Brian Cheung、Nico- las Papernot、Alexey Kurakin、Ian Goodfellow 和 Jascha Sohl-Dickstein。欺骗计算机视觉和受时间限制的人类的对立例子。《神经信息处理系统进展》，3914－3924 页，2018 年。一

[8] 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。在 CVPR，第 1625-1634 页，2018。1，2，5

[9] 安德烈亚斯·盖革、菲利普·伦茨、克里斯托弗·斯蒂勒和拉克尔·乌尔塔森。视觉遇上机器人:kitti 数据集。国际机器人研究杂志，32(11):1231－1237，2013。5

[10] 伊恩·古德菲勒、让·普吉-阿巴迪、迈赫迪·米尔扎、徐炳、戴维·沃德-法利、谢尔吉尔·奥泽尔、亚伦·库维尔和约舒阿·本吉奥。生成对抗网络。在 NIPS 中，第 2672-2680 页。2014.2

[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 年在 ICLR。2，5

[12] 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。arXiv 预印本 arXiv:1607.02533，2016。一

[13] 阿列克谢·库拉金、伊恩·古德菲勒和萨米·本吉奥。大规模的对抗性机器学习。ICLR，2017。2

[14] 刘燕佩，，陈，，宋晓明。探究可转移的对立例子和黑盒攻击。ICLR，2017。2

[15] 陆家骏、侯赛因·西拜、埃文·法布里和大卫·福赛思。无需担心自主车辆中物体检测的对立例子。arXiv 预印本 arXiv:1707.03501，2017。2

[16] 陆咏仪，戴玉荣和邓志强。属性引导的人脸图像生成的条件循环方法。arXiv 预印本 arXiv:1705.09966，2017。2

[17] 扬·亨德里克·梅岑、穆马迪·柴塔尼亚·库马尔、托马斯·布罗克斯和福尔克·菲舍尔。针对语义图像分割的普适对抗性扰动。2017 年 IEEE 计算机视觉国际会议(ICCV)，第 2774－2783 页。IEEE，2017。一

[18] Andreas Mogelmose、Mohan Manubhai Trivedi 和 Thomas B Moeslund。用于智能驾驶员辅助系统的基于视觉的交通标志检测和分析:观点和调查。IEEE 智能交通系统汇刊，13(4):1484－1497，2012。一

[19] 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。在 CVPR，第 2574－2582 页，2016。一

[20] 潘，尤玉荣，，陆。自动驾驶的虚拟到真实强化学习。arXiv 预印本 arXiv:1704.03952，2017。5

[21] 尼古拉斯·帕伯诺特、帕特里克·麦克丹尼尔、伊恩·古德菲勒、萨默什·贾、Z·伯凯·切利克和阿南瑟拉姆·斯瓦米。针对机器学习的实用黑盒攻击。2017 年 ACM 亚洲计算机与通信安全会议论文集，第 506-519 页。ACM，2017。2

[22] 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 欧洲安全与隐私研讨会(Eu- roS&P)，第 372－387 页。IEEE，2016。一

[23] 皮埃尔·塞尔马内和扬·勒昆。基于多尺度卷积网络的交通标志识别。在 IJCNN，第 2809-2813 页，2011 年。一

[24] 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。ACM，2016。一

[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。神经网络的有趣特性。arXiv 预印本 arXiv:1312.6199，2013。一

[26] 托马斯·塔尼和刘易斯·格里芬。对立范例现象的边界倾斜透视。arXiv 预印本 arXiv:1608.07690，2016。2

[27] 田宇池、裴可欣、贾纳和白沙奇雷。Deeptest:深度神经网络驱动的自动驾驶汽车的自动化测试。在 ICSE，第 303-314 页。ACM，2018。一

[28] 田宇池、裴可欣、贾纳和白沙奇雷。Deeptest:深度神经网络驱动的自动驾驶汽车的自动化测试。《第 40 届国际软件工程会议论文集》，303-314 页。ACM，2018。5

[29] 肖、、、何华伦、刘、宋晓。用对抗网络生成对抗实例。IJCAI，2018。2

[30] 朱俊彦，菲利普克拉亨布 hl，埃利谢赫特曼，和阿列克谢埃夫罗斯。自然图像流形上的生成视觉操作。在 ECCV，第 597-613 页。斯普林格，2016。2