针对图像分类的鲁棒物理域对抗伪装

段晔鑫[1,4],贺正芸[2],张 颂[3],詹达之[4],王田丰[4],林庚右[4],张 锦[1],潘志松[4*]
(1. 陆军军事交通学院镇江校区,江苏镇江 212003;2. 湖南工业大学轨道交通学院,湖南株洲 412007;
3. 北京电子科技学院网络空间安全系,北京 100071;4. 陆军工程大学指挥控制工程学院,江苏南京 210007)

摘 要: 深度学习模型对对抗样本表现出脆弱性. 作为一种对现实世界深度系统更具威胁性的攻击形式,物理域对抗样本近年来受到了广泛的研究关注. 现有方法大多利用局部对抗贴片噪声在物理域实现对图像分类模型的攻击,然而二维贴片在三维空间的攻击效果将由于视角变化而不可避免地下降. 为了解决这一问题,所提Adv-Camou方法利用空间组合变换来实时生成任意视角及变换背景的训练样本,并最小化预测类与目标类交叉熵损失,使模型输出指定错误类别. 此外,所建立的仿真三维场景能公平且可重复地评估不同的攻击. 实验结果表明,Adv-Camou生成的一体式对抗伪装可在全视角欺骗智能图像分类器,在三维仿真场景比多贴片拼接纹理平均有目标攻击成功率高出25%以上,对Clarifai商用分类系统黑盒有目标攻击成功率达42%,此外3D打印模型实验在现实世界平均攻击成功率约为66%,展现出先进的攻击性能.

关键词: 对抗样本;对抗伪装;对抗攻击;图像分类;深度神经网络
基金项目: 国家自然科学基金(No.62076251)
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2024)03-0863-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221301

Robust Physical Adversarial Camouflages for Image Classifiers

DUAN Ye-xin[1,4], HE Zheng-yun[2], ZHANG Song[3], ZHAN Da-zhi[4], WANG Tian-feng[4], LIN Geng-you[4], ZHANG Jin[1], PAN Zhi-song[4*]
(1. *Zhenjiang Campus, Army Military Transportation University, Zhenjiang, Jiangsu* 212003, *China*;
2. *Railway Transportation College, Hunan University of Technology, Zhuzhou, Hunan* 412007, *China*;
3. *Department of Cyberspace Security, Beijing Electronic Science and Technology Institute, Beijing* 100071, *China*;
4. *Command and Control Engineering College, Army Engineering University, Nanjing, Jiangsu* 210007, *China*)

Abstract: Deep learning models are vulnerable to adversarial examples. As a more threatening type for practical deep learning systems, physical adversarial examples have received extensive research attention in recent years. Most of the existing methods use the local adversarial patch noise to attack the image classification model in the physical world. However, the attack effect of 2D patches in 3D space would inevitably decline due to the change in the view angle. To address this issue, the proposed Adv-Camou method uses spatial combination transformation to generate training examples of arbitrary viewpoints and transformed backgrounds in real time. Moreover, the cross-entropy loss between the prediction class and target class is minimized to make the model output the specified incorrect class. In addition, the established 3D scene can evaluate different attacks fairly and reproducibly. The experimental results show that the coated adversarial camouflage generated by the Adv-Camou method can fool image classifiers from arbitrary viewpoints. In the 3D simulation scene, the average targeted attack success rate of Adv-Camou is more than 25% higher than that of piecing together patches. The success rate of black-box targeted attacks on the Clarifai commercial classification system reaches 42%. In addition, the average attack success rate of 3D printing model experiments in the real world is about 66%, which significantly demonstrates that our method outperforms state-of-the-art methods.

Key words: adversarial example; adversarial camouflage; adversarial attack; image classification; deep neural network

收稿日期:2022-11-11;修回日期:2023-09-02;责任编辑:李勇锋
*通讯作者:潘志松

Foundation Item(s): National Natural Science Foundation of China (No.62076251)

1 引言

深度神经网络在诸多计算机视觉任务中都达到了先进水平,并被部署于现实应用,如无人驾驶系统[1].然而研究表明[2~4],对输入样本添加不妨碍人类识别的对抗噪声后,能导致智能识别模型以意想不到的方式失常.此种添加精心扰动的图像样本被称为对抗样本[2],所带来的潜在危险性使其成为当前的研究热点.

对抗样本攻击根据可获知的目标模型知识量,可分为白盒和黑盒攻击.白盒攻击方式是假定攻击者能访问模型的内部结构和参数,可通过梯度反传更新对抗噪声来获得对抗样本.但实际中由于攻击者通常不能访问目标模型的内部结构和参数,属于黑盒模型,因此需要采用黑盒方式进行攻击.黑盒攻击方法主要有两种,分别是基于查询的和基于迁移的.基于查询的黑盒攻击方法[4]要求大量的查询,计算成本高,且现实中难以满足自由查询的条件,客观上造成基于查询的方法可行性低.研究发现对抗样本具有跨模型迁移性[2],即通过白盒方式攻击一个模型所生成的对抗样本,可以欺骗另一个相同任务的黑盒模型,为实现黑盒攻击提供了重要途径.

对抗样本攻击按场景可分为数字域攻击和物理域攻击[5].数字域攻击通常利用不易察觉的对抗噪声欺骗深度模型,但容易因环境噪声等因素而遭到破坏,因而其性能难以在物理域有效保持[6];物理域攻击所利用的可视噪声能在三维物理域保持一定的鲁棒欺骗性.然而,生成鲁棒的物理域对抗样本,也存在多方面的挑战,首先物理域对抗样本的对抗噪声在制作(比如打印)过程中会有损失,其次对抗噪声容易被环境噪声、光照等破坏,第三,对抗噪声经由摄像头采集也存在信息损失,因为摄像头无法完美地拍摄图像,第四是摄像头的角度和距离也会影响攻击效果.

相关研究已取得一些很有意义成果,如局部对抗贴片[6~8]、停车标识攻击[5]以及小的对抗物体如海龟[9].但是这些工作依然存在一些不足:(1)基于局部对抗贴片的攻击[4,6,7],由于对抗贴片是在二维空间训练生成,即在训练过程贴片并不随图像中的物体进行相应的空间转换,而只在图像上执行二维图像转换,导致对抗贴片随物体在三维空间中变化姿态和视角时,攻击效果急剧下降;(2)对抗噪声只用于平面物体,如停车标识[5],对于任意视角的物体攻击容易失效;(3)缺少一个物理实验环境来公平且可重复地评估测试攻击效果.

为了解决这些问题,本文提出 Adv-Camou 攻击方法生成对抗伪装,首先利用空间组合分布变换来实时渲染目标物几何及环境变化,并且对抗伪装在训练过程中跟随三维目标物姿态变化进行相应变换,消除对抗性视角盲点,通过最小化预测类与目标类交叉熵损失,使模型输出指定错误类别,得到全视角对抗伪装.此外,本文利用 Unity 引擎建立了三维虚拟场景以进行变化场景仿真实验,并对 Clarifai 商用图像分类系统进行了攻击测试,进一步利用 3D 打印技术在现实世界对伪装目标物的攻击性进行了评估验证,实验结果证明所生成的对抗伪装具有出色的攻击欺骗性.

总体而言,本文主要贡献有:(1)所提 Adv-Camou 可生成欺骗图像分类器的三维全视角一体式对抗伪装,表现出显著优于多对抗贴片拼接纹理的攻击性能;(2)建立了三维仿真环境以公平且可重复地评估不同视角、距离和亮度等条件下的攻击;(3)Adv-Camou 可用于任意物体的对抗伪装纹理生成,结果可 3D 打印,或将对抗伪装纹理贴于目标物体,在现实世界表现出良好的泛化性.

2 相关工作

2.1 数字域对抗样本

数字域对抗样本通常是对原输入图像添加不易察觉的对抗噪声而获得.自 Szegedy 等人[2]首次发现对抗样本以来,已有多种图像对抗样本方法被提出,如快速梯度方法[10],通过计算单次梯度,以很低的计算代价生成大量的对抗样本,但表现出较低的攻击成功率[10].Kurakin 等人[11]将单步快速梯度法拓展为迭代方法,显著提升了白盒攻击成功率,但容易过拟到所攻击的白盒模型参数,表现为较弱的跨模型迁移攻击性.一系列增强方法被提出用来缓解过拟合问题,避免陷入局部极值,提升对抗样本的迁移性,如引入动量因子[12]、多输入增强对抗攻击[13]、平移无关法增强对抗噪声[14]和对网络进行特征级微扰[15]等.然而,由于数字域对抗样本的对抗噪声较弱,因此在物理域容易受视角、环境噪声和亮度等变化因素影响而遭到破坏,导致攻击失效.

2.2 物理域对抗样本

物理域对抗样本通常对目标物添加较强的对抗噪声,但噪声对人眼也较为显著,此类噪声被称为可视对抗噪声.Eykholt 等人[5]针对交通标识的识别任务,发现在停车标识不同位置粘贴贴纸会有不同的攻击效果,通过对敏感脆弱区域应用贴纸攻击,成功导致交通标识被错误分类.Brown 等人[6]生成了场景无关的局部对抗贴片,可以在二维图像的任意位置导致分类模型输出指定类别标签,但由于对抗贴片在二维空间训练生成,即对抗贴片在训练过程中不随图像中物体进行相应的空间变换,导致在三维空间测试时,贴片攻击效果随视角变大而急剧下降.Wang 等人[7]生成了同时抑制

模型和人眼注意力的无目标攻击贴片. Athalye 等人[9]成功将二维平面对抗样本拓展到三维空间,生成了对抗海龟和垒球,但是生成的对抗物体只能在较小的相机距离欺骗图像分类器,难以保持鲁棒攻击性.

3　图像分类对抗伪装生成方法

所提 Adv-Camou 对抗伪装生成过程主要包括三维训练样本实时生成和对抗伪装纹理生成及更新. 下面首先介绍对抗伪装生成流程框架,然后对三维环境条件参数和伪装生成过程作详细介绍.

3.1　对抗伪装生成流程

由于自动驾驶是安全攸关的研究热点,车辆为其重点关注对象,因此本文选取车为实验目标物,如图 1 所示,对抗伪装生成主要包括两个步骤.

(1)第一步,实时生成三维训练样本. 以目标物三维模型文件及相应纹理图片文件作为输入,将纹理映射到目标物的渲染图中,并根据给定的三维环境条件参数分布,引入一系列图像转换如旋转、平移和距离变化等几何变换,以及亮度、噪声和背景变化等环境变换,以近似模拟真实环境的分布,实时获得不同视角、亮度以及不同环境背景的训练样本,为生成对抗纹理提供多样化训练样本.

(2)第二步,攻击分类模型(源模型)以更新对抗伪装纹理. 分类模型在对抗纹理训练过程中对伪装纹理车进行类别预测,通过最小化预测类与指定目标类的交叉熵损失,更新对抗噪声,减小伪装车辆目标与指定目标类别(如拖拉机)的类间距离,使模型将目标车识别为指定错误类. 对抗伪装在训练过程中随车辆姿态变化进行相应的空间变换,消除对抗性视角盲点. 更新对抗噪声后的对抗伪装作为新的迭代步输入,直到达到预设的迭代步数,输出能在物理域保持鲁棒对抗性的三维全视角对抗伪装.

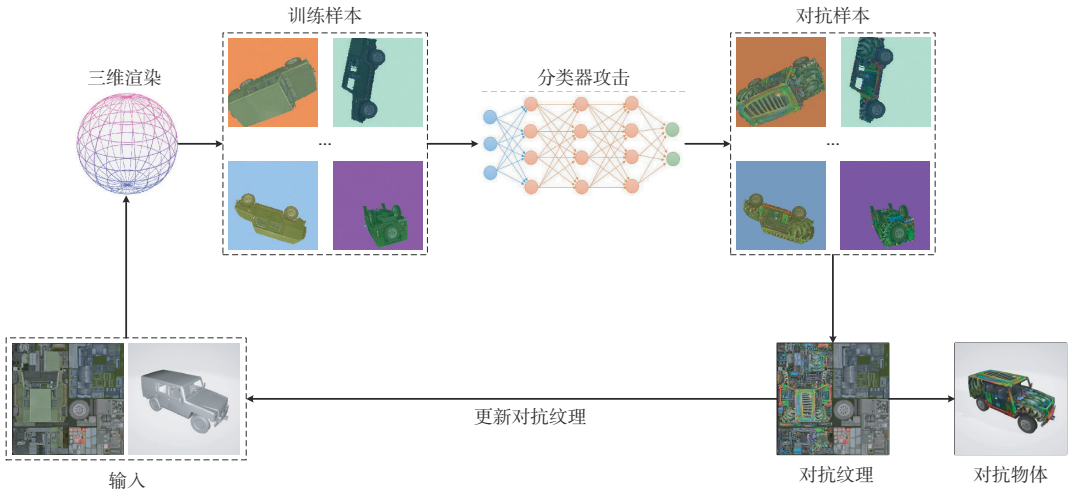

图 1　生成对抗伪装流程

3.2　三维渲染

一个 3D 物体由网格张量 $\boldsymbol{m}$ 覆盖纹理张量 $\boldsymbol{c}$ 生成,记为 $(\boldsymbol{m},\boldsymbol{c})$. 令 e 表示环境条件参数,E 为环境条件参数分布,包含视角、相机距离、光照、噪声、背景变化等,参数分布见表 1,均为独立连续随机变量,符合均匀分布(除了高斯噪声). 通过亮度和通道像素的加值和乘值,对真实场景光照变化和打印误差建模近似. 在环境条件 e 下,经图像渲染操作 $\mathcal{R}$ 得到样本 x 为:

$$x=\mathcal{R}((\boldsymbol{m},\boldsymbol{c}),e) \tag{1}$$

通过对纹理 $\boldsymbol{c}$ 添加对抗噪声,得到对抗伪装纹理 $\boldsymbol{c}_{\text{adv}}$,所生成的对抗样本为:

$$x_{\text{adv}}=\mathcal{R}((\boldsymbol{m},\boldsymbol{c}_{\text{adv}}),e) \tag{2}$$

其中 $(\boldsymbol{m},\boldsymbol{c}_{\text{adv}})$ 为所得到的 3D 对抗物体.

3.3　对抗伪装纹理生成

对于一个干净的输入图像 x,它能被图像分类模型 f 正确分类为真实标签 y,对 x 添加对抗噪声后得到对抗样本 x_{adv}. 对抗样本按照攻击者的期望输出可分为无目标对抗样本和有目标对抗样本,无目标对抗样本满足 $f(x_{\text{adv}})\neq y$,即对抗样本不被分类为标签 y 即为攻击成

表 1　物理域环境条件参数分布

环境条件	最小	最大
相机距离	1.5	4.0
X/Y 平移	−0.05	0.05
旋转角度	任意	
背景色	(0.1,0.1,0.1)	(1.0,1.0,1.0)
亮/暗(加值)	−0.15	0.15
亮/暗(乘值)	0.5	2.0
通道像素(加值)	−0.15	0.15
通道像素(乘值)	0.7	1.3
高斯噪声(标准差)	0.0	0.1

功,而有目标对抗样本满足$f(x_{adv})=y^*$且$y^*\neq y$,即要求对抗样本被误分类为指定目标类y^*.

由于ImageNet数据集[16]中车的类别过于细粒度化,即不同车的类间距离小,一种车容易被误识别为另一种车,所以不指定错分类别的无目标攻击相对无趣. 另外用于训练分类器的真实图片数据和三维仿真数据之间存在域差,可能导致随机误分类,有目标攻击可以排除此因素对结果的影响,因此本文研究更具挑战性的有目标攻击,使目标车被误分类为与原类别有较大差异的指定目标类,如“拖拉机”和“红绿灯”.

研究表明,集成模型攻击法相比单模型攻击能得到更强的对抗样本,具有更强的跨模型黑盒迁移攻击性,因为如果对抗样本可以欺骗多个模型,那么它们更可能成功攻击其他未知模型,并且发现logits集成优于损失集成和softmax集成方案[12]. K个分类模型logits集成可表示为:

$$l(x_{adv})=\sum_{k=1}^{K}w_k l_k(x_{adv}) \tag{3}$$

其中$l_k(x_{adv})$为第k个模型的logits输出,w_k为第k个模型的权重系数,$w_k\geqslant 0$且$\sum_{k=1}^{K}w_k=1$.

对于有目标攻击,需要最小化预测值与目标类标签y^*的交叉熵损失函数$J(l(x_{adv}),y^*)$. 不同于二维物体,人眼对三维物体识别的主要是基于物体的形状轮廓,而纹理几乎不影响人眼的识别结果. 鉴于此,实验不约束对抗纹理与原始纹理$\boldsymbol{c}$之间的距离,结合式(2),以对抗伪装纹理$\boldsymbol{c}_{adv}$为优化变量,通过最小化交叉熵损失函数,更新对抗伪装纹理$\boldsymbol{c}_{adv}$. 针对K个模型集成的攻击优化问题可以表示为:

$$\arg\min_{\boldsymbol{c}_{adv}} J(l(x_{adv}),y^*) \tag{4}$$

其中$J(l(x_{adv}),y^*)$为$-\mathbf{1}_{y^*}\cdot\log(\mathrm{softmax}(l(x_{adv})))$,$-\mathbf{1}_{y^*}$为$y^*$的one-hot编码.

生成鲁棒三维对抗伪装需优化对任意角度、变化距离和其他环境条件下的整体损失,即为训练优化以下目标函数:

$$\arg\min_{\boldsymbol{c}_{adv}} \mathrm{E}_{x\sim X,e\sim E}J(l(x_{adv}),y^*) \tag{5}$$

其中X是由渲染器实时生成的任意视角和变化环境条件的图像训练集,E是由渲染器模拟的环境条件分布,E是变换期望(EOT)技术[9],它在优化过程中模拟对抗扰动. 训练过程中分类器f获得的“真”输入x_{adv}是渲染操作后带有环境条件e的输入对象$(\boldsymbol{m},\boldsymbol{c}_{adv})$. 最后生成的伪装像素值被限制在图像有效像素[0,1]范围内. 集成模型的攻击过程如算法1所示,经过迭代更新获得3D目标物的对抗伪装纹理$\boldsymbol{c}_{adv}$.

算法1 集成模型Adv-Camou伪装生成算法

输入:3D对象$(\boldsymbol{m},\boldsymbol{c})$,环境条件参数$e\in E$,神经渲染器$\mathcal{R}$,目标类标签$y^*$,$K$个分类器$f_1,f_2,\cdots,f_K$,相应logits为$l_1,l_2,\cdots,l_K$,集成权重为$w_1,w_2,\cdots,w_K$,最大迭代步$N$

输出:对抗伪装纹理张量$\boldsymbol{c}_{adv}$

1.$\boldsymbol{c}_{adv}^0\leftarrow\boldsymbol{c}$

2. FOR $t=0$ TO $N-1$ DO

3.生成训练数据:$x_{adv}^t\leftarrow\mathcal{R}((\boldsymbol{m},\boldsymbol{c}_{adv}^t),e)$

4.由输入x_{adv}^t计算出K个模型的logits值并集成:

$$l(x_{adv}^t)=\sum_{k=1}^{K}w_k l_k(x_{adv}^t)$$

5.通过攻击所有分类更新伪装纹理$\boldsymbol{c}_{adv}^t$:

$$\arg\min_{\boldsymbol{c}_{adv}^t}\mathrm{E}_{x\sim X,e\sim E}J(l(x_{adv}^t),y^*)$$

$$\boldsymbol{c}_{adv}^t=\mathrm{Clip}(\boldsymbol{c}_{adv}^t,0,1)$$

6.END FOR

7.RETURN: $\boldsymbol{c}_{adv}=\boldsymbol{c}_{adv}^N$

4 实验结果与分析

4.1 实验设置

源模型 实验选取四个在ImageNet数据集上预训练的分类模型进行集成来生成对抗样本:Inception-v3(Inc-v3)[17]、Inception-v4(Inc-v4)、Inception-ResNet-v2(IncRes-v2)[18]和ResNet-v2-152(Res-152)[19].

目标模型 为了全面评估所生成的对抗伪装的攻击效果,除了以源模型作为目标模型以进行白盒攻击测试,另外选取三个模型来评估跨模型黑盒攻击效果:Densenet-169(Dense-169)[20]、Xception-71(Xcep-71)[21]和NASNetLarge(NASNet)[22].

评价指标 对于测试数据,以有目标攻击成功率作为评价指标,即当原输入x被正确分类为标签y,且对抗样本x_{adv}被误分类为指定目标类标签y^*才视为攻击成功,这相比于无目标攻击更有难度,可表示为:

$$\frac{\sum\{f(x_{adv})=y^*\wedge f(x)=y\}}{\sum\{f(x)=y\}} \tag{6}$$

基线方法 现有工作生成的物理域对抗样本主要利用局部对抗贴片附于物体表面或在物体邻近区域,实验选取原车纹理(Original)、目标类实物贴图纹理(Natural)、双注意力抑制(Dual Attention Suppression,DAS)贴片纹理[8]以及RP2对抗贴片纹理[4]、Adv-Patch对抗贴片纹理[7]作为基线对比. 如图2,为了公平对比,贴片图案分布在车顶、引擎盖和车侧等各个面,Adv-Camou对抗伪装纹理覆盖整个车体,而轮胎、窗户、车灯等不易或不宜有对抗噪声的部位维持原纹理不变. 图2右列第一行目标类为拖拉机,第二行目标类为红绿灯. DAS为无目标攻击.

4.2 物理仿真环境

实验使用Unity引擎构建了一个逼真的3D模拟场

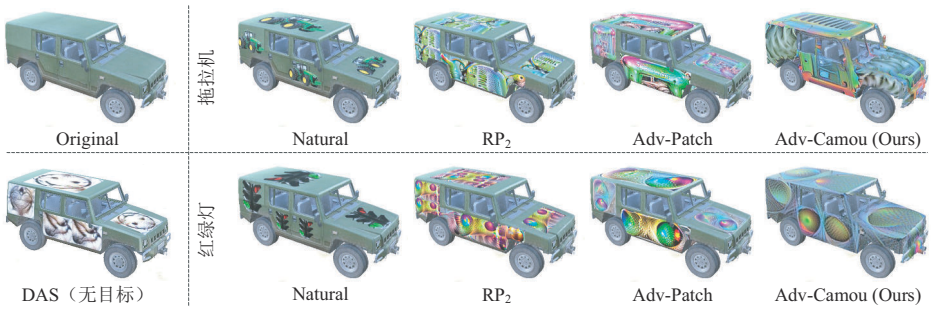

图2　不同纹理车

景,可以设置不同实验环境参数,以公平且可重复地评估不同的物理域攻击.

虚拟场景　如图3,所建立的虚拟场景为典型的城市环境,包括建筑、街道、路牌和绿植等. 仿真引擎使实验能够在不同的环境条件下进行实验,如不同视角、相机距离、光照、背景、遮挡等.

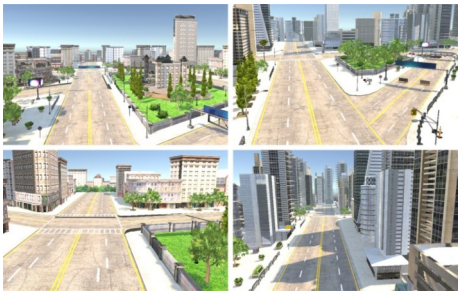

图3　Unity城市场景

相机设置　如图4,相机视角和距离在环绕车辆周围的半椭球内360°自由布置,相比局部对抗贴片限制视角,这更符合现实世界摄像头分布的随机性,从而更贴近实际地评估物理域攻击效果.

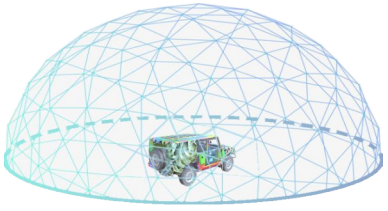

图4　自由视角示意图

4.3　仿真场景实验

将所得对抗伪装攻击结果与原纹理、实物贴图纹理以及对抗贴片纹理在任意视角和不同光照下对比. 图2显示了DAS无目标纹理以及拖拉机、红绿灯目标类纹理. 在模拟场景的不同位置,为每一种纹理车渲染240(120×2)张自由视角两种亮度的图像. 其中原纹理车均能被正确识别.

表2给出了在白盒和跨模型黑盒攻击场景下任意视角和不同光照条件下的攻击成功率. 由于在现实环境通常无法访问目标模型的内部结构和参数,为黑盒场景,因此重点关注黑盒迁移攻击结果更有意义,即基于白盒源模型生成的对抗样本对其他目标模型的跨模型攻击结果. 受页面空间限制,表2选取Inc-v3模型为白盒攻击测试代表模型. 可以看出,Adv-Camou方法所得一体式对抗伪装纹理的平均有目标攻击成功率比基线方法高出25%以上,也优于DAS无目标攻击法,在三维场景中能保持较强的鲁棒性. Adv-Camou取得了更高的黑盒攻击成功率,表明生成的对抗伪装更好地学习到了多模型的决策边界信息,从而具有良好的跨模型攻击性能,验证了所提方法的有效性.

对比表2有目标攻击结果,不同亮度下,目标类为拖拉机的攻击成功率在明亮环境下较高,而目标类为红绿灯的攻击成功率在阴暗环境下较高. 分析ImageNet训练集中拖拉机和红绿灯的数据,发现拖拉机训练数据绝大部分在明亮环境下拍摄,而红绿灯有相当一部分是在阴暗环境或夜间拍摄,红绿灯在阴暗环境下特征更明显,可能导致阴暗环境下红绿灯目标类攻击成功率高于明亮环境下. 另外,拖拉机目标类攻击成功率整体显著高于红绿灯,分析原因是吉普车的车轮也属于拖拉机的重要特征,所以吉普车与拖拉机的类间距离较小,因此较易攻击成功.

图5和图6分别显示了明亮以及阴暗场景下,多种视角不同纹理车样本Inc-v3模型分类结果. 由图可见二维贴片对三维场景的泛化性较弱,主要因为视角大导致贴片变形从而攻击效果下降,另外多个局部贴片拼合攻击效果也不如一体式对抗纹理. 此外,Adv-Camou所得对抗纹理在变化距离能保持较强对抗性. 图5和图6中Adv-Camou纹理车Inc-v3模型的top-2分类结果及置信分数见表3.

实验发现了一个有趣的现象,拖拉机目标类的对抗纹理车可能被分类为割草机、犁或收割机. 分析ImageNet训练集中这三个类别的数据,发现这三个类别的部分图片与拖拉机有非常相似的特征,甚至犁和收割机图片中就是由拖拉机牵引作业. 而红绿灯目标类容易被误分类为桶和三轮车,可能是因为圆形红绿灯抽象图案与圆桶或轮子相似.

进一步地,为验证目标物在部分遮挡情况下对抗

表 2 任意视角和不同光照下的攻击成功率(*代表白盒攻击) 单位:%

模型	Inc-v3				Dense-169				Xcep-71				NASNet			
目标类	拖拉机		红绿灯		拖拉机		红绿灯		拖拉机		红绿灯		拖拉机		红绿灯	
亮度	亮	暗	亮	暗	亮	暗	亮	暗	亮	暗	亮	暗	亮	暗	亮	暗
Natural	3.3	3.3	0.0	4.2	5.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0
RP_2	66.7*	50.0*	4.2*	5.8*	21.7	0.0	0.0	0.8	26.7	8.3	0.0	0.0	33.3	16.7	0.0	0.0
Adv-Patch	60.0*	54.2*	6.7*	1.7*	53.3	8.3	0.8	4.2	16.7	4.2	14.2	11.7	31.7	19.2	10.0	9.2
Adv-Camou	**89.2***	**75.0***	**31.7***	**30.0***	**71.7**	**15.8**	**5.8**	**6.7**	**49.2**	**30.8**	**21.7**	**58.3**	**86.7**	**76.7**	**30.0**	**39.2**
DAS	32.9 亮		55.8 暗		20.0 亮		72.9 暗		55.0 亮		67.9 暗		19.2 亮		52.9 暗	

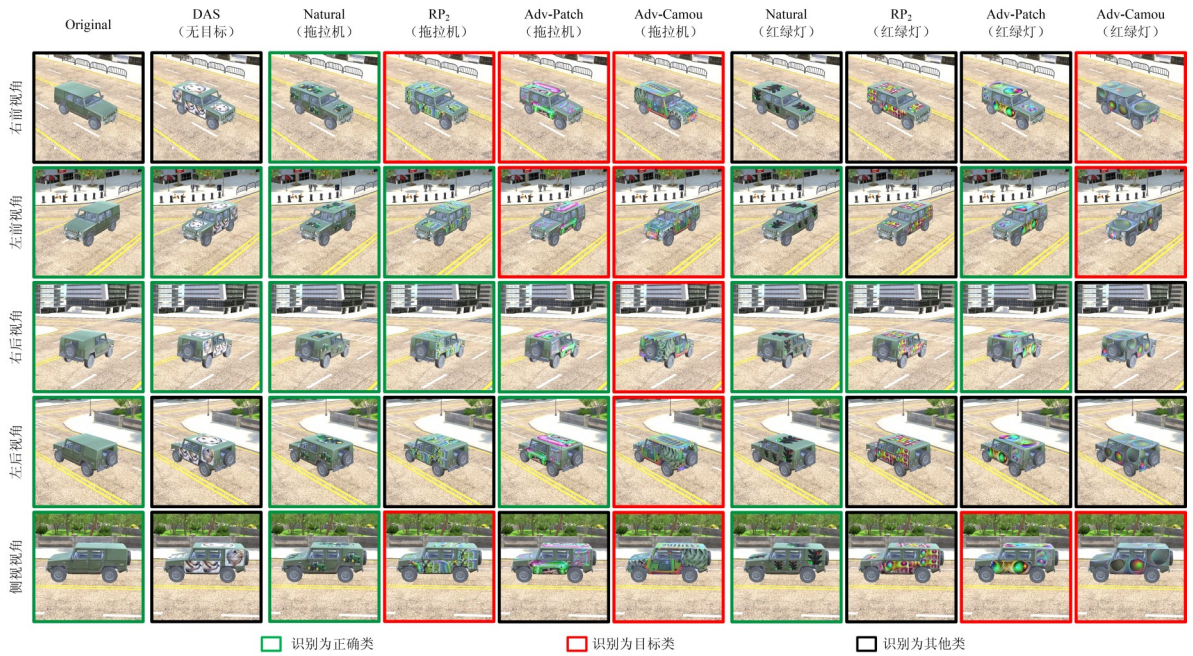

图 5 明亮场景多种视角下不同纹理车

图 6 阴暗场景多种视角下不同纹理车

表3　对抗伪装纹理车对Inc-v3模型有目标攻击top-2结果及置信分数

亮度	视角	拖拉机目标类		红绿灯目标类	
亮	右前	拖拉机 (0.46)	割草机 (0.23)	红绿灯 (0.15)	垃圾桶 (0.12)
	左前	拖拉机 (0.69)	犁 (0.17)	红绿灯 (0.37)	桶 (0.11)
	右后	拖拉机 (0.74)	犁 (0.06)	桶 (0.42)	垃圾桶 (0.03)
	左后	拖拉机 (0.43)	割草机 (0.18)	桶 (0.13)	三轮车 (0.10)
	侧视	拖拉机 (0.78)	收割机 (0.16)	红绿灯 (0.44)	三轮车 (0.09)
暗	右前	拖拉机 (0.75)	割草机 (0.05)	红绿灯 (0.35)	出租车 (0.13)
	左前	拖拉机 (0.41)	割草机 (0.09)	三轮车 (0.21)	红绿灯 (0.10)
	右后	拖拉机 (0.80)	割草机 (0.02)	红绿灯 (0.41)	桶 (0.04)
	左后	拖拉机 (0.77)	犁 (0.06)	红绿灯 (0.25)	哑铃 (0.13)
	侧视	拖拉机 (0.85)	收割机 (0.05)	红绿灯 (0.42)	手推车 (0.05)

伪装纹理的攻击性能,针对吉普车不同部位设置遮挡物,并对原纹理车、拖拉机目标类和红绿灯目标类伪装纹理车分别取60张自由视角图像.实验结果如表4所示,可以看出遮挡情况相比未遮挡情况攻击效果反而有所提升,原因可能是虽然部分对抗纹理被遮挡,但遮挡也可能导致吉普车重要特征(如原纹理车轮)信息的损失,使得模型判别结果相对远离原类别空间,并使得模型更加关注其余对抗纹理区域,从而提升了欺骗效果.图7为不同部位遮挡结果示例.

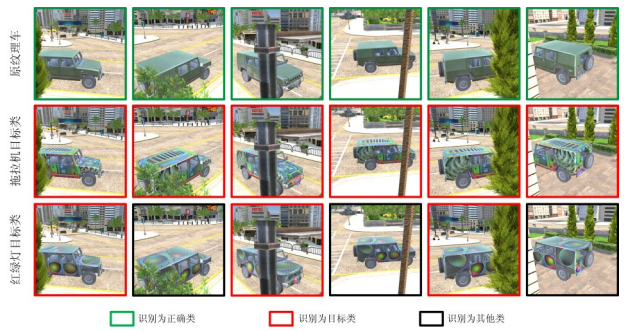

图7　对抗纹理车不同遮挡结果示例

表4　对抗纹理车遮挡任意视角的攻击成功率(*代表白盒攻击)　　单位:%

模型	Inc-v3		Dense-169		Xcep-71		NASNet	
目标类	拖拉机	红绿灯	拖拉机	红绿灯	拖拉机	红绿灯	拖拉机	红绿灯
Adv-Camou	91.7*	53.3*	73.3	10.0	71.7	13.3	85.0	46.7

4.4　商用分类系统实验

为了检验生成的对抗伪装对现实世界分类应用的攻击效果,实验对Clarifai的通用图像分类系统进行了黑盒攻击测试,以验证所提方法的有效性和先进性.Clarifai是国际领先的商用深度学习人工智能平台,提供常用的图片和视频识别服务.

Clarifai的类别标签与ImageNet数据集类别标签存在差别,且置信分数并不是softmax分类结果,即加和不为1.由于无法获知Clarifai系统模型结构和参数,且所用模型与基于ImageNet训练的源模型存在较大差异,攻击具有更大挑战性.

实验将100张目标类为拖拉机的对抗伪装车辆样本输入到Clarifai系统,有42%的图像能以高置信分数被识别为拖拉机,进一步证明了所提方法的有效性.图8和图9给出了明亮和阴暗环境下不同视角的攻击结果示例图.

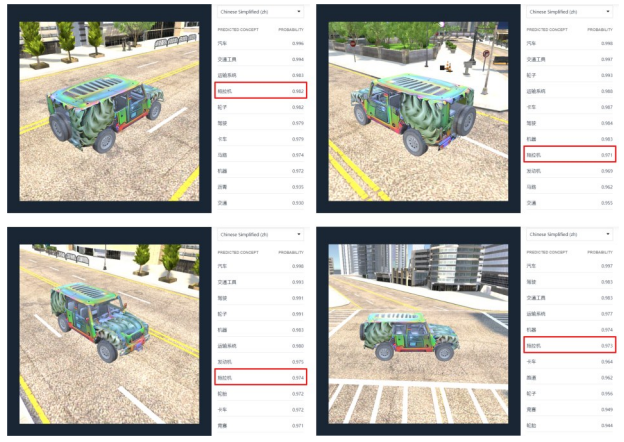

图8　明亮环境下对抗车辆Clarifai分类结果

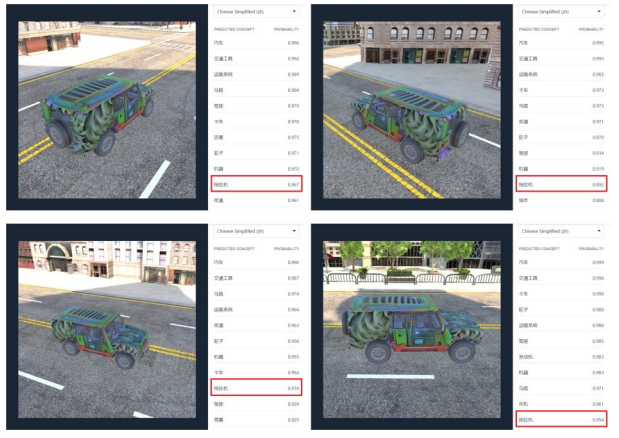

图9　阴暗环境下对抗车辆Clarifai分类结果

4.5　现实场景实验

为了进一步验证对抗攻击效果,以3D打印的拖拉机目标类的对抗车辆实物为研究对象开展实验分析,评估对抗伪装纹理在现实世界的鲁棒性.实验使用HONOR20手机进行拍照和录制视频.与仿真场景实验类似,在自由视角对3D打印车辆拍摄.对抗伪装纹理车被放置在一个旋转的转盘上,评估360°环绕视角下的对抗攻击性.

5个视频总帧数为2 126帧,分别为405、409、427、435和446帧,光照亮度有不同变化,使测试数据更加多样化并具有代表意义,以全面评估有对抗伪装的有效性.针对

Inc-v3模型有目标攻击成功率分别为71%(289/405)、65%(266/409)、72%(308/427)、66%(288/435)和58%(259/446).实验结果显示Adv-Camou方法所得对抗伪装能在现实场景取得较高的有目标攻击成功率,证明了所得对抗样本对现实世界的泛化能力.图10展示了固定间隔采样的100张图像帧,并给出了Inc-v3模型的分类结果.

实验发现在现实世界明亮环境比阴暗环境能取得更高有目标攻击成功率,这可能是因为光线较暗条件下,摄像机难以捕捉车身对抗纹理细节,导致攻击效果下降.

4.6 其他目标物实验

对其他目标生成对抗伪装来进一步验证Adv-Camou法.选取集装箱和圆桶为方形物和圆形物的代表,在Unity中分别生成120张不同视角的图像.表5为Adv-Camou对抗纹理集装箱和圆桶任意视角的攻击成功率,可以看出,对抗伪装纹理集装箱和圆桶以较高成功率被误分为指定目标类,表明Adv-Camou可用于除车之外的其他物体.由实验结果可知,相比于吉普车目标物,集装箱和圆桶对拖拉机目标类攻击成功率有所下降,分析原因是集装箱和圆桶没有"车轮"这种拖拉机所具有的特征,导致与拖拉机类别空间距离相距较远,攻击较难;然而,集装箱和圆桶的红绿灯目标类攻击成功率比吉普车目标物有明显提升,分析也是因为集装箱和圆桶没有"车轮"这种车辆显著特征,相比吉普车与红绿灯类间距离较近,攻击较易.图11为集装箱和圆桶对抗伪装样本示例.

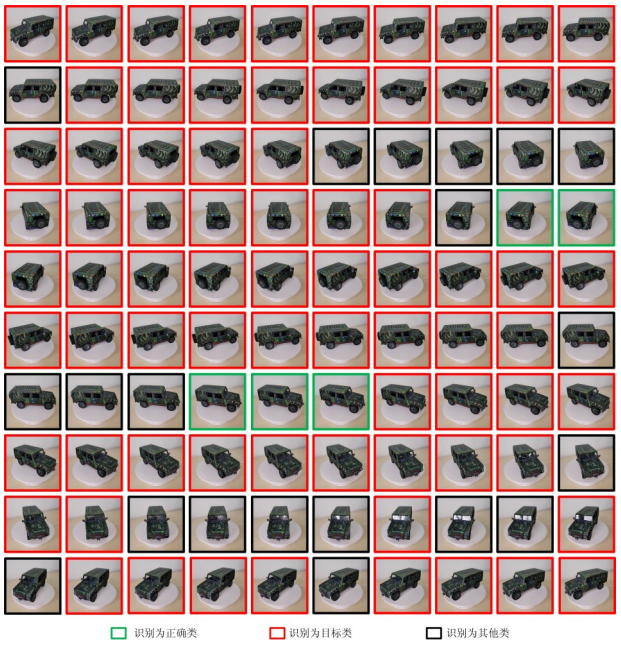

图10 对抗伪装车模型分类实验结果示例

表5 对抗纹理集装箱和圆桶任意视角的攻击成功率(*代表白盒攻击) 单位:%

模型	Inc-v3		Dense-169		Xcep-71		NASNet	
目标类	拖拉机	红绿灯	拖拉机	红绿灯	拖拉机	红绿灯	拖拉机	红绿灯
集装箱	51.7*	68.3*	12.5	16.7	33.3	61.7	64.1	76.7
圆桶	63.3*	66.7*	32.5	23.3	65.8	70.8	78.3	72.5

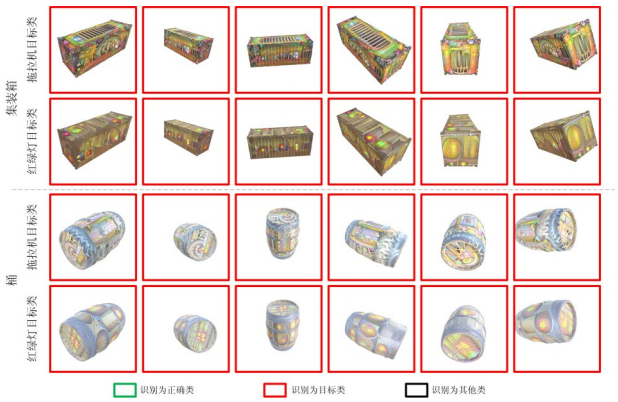

图11 对其他目标物的对抗伪装攻击结果示例

5 结论

本文研究了物理域对抗伪装生成,所提Adv-Camou方法生成的对抗伪装可三维全视角欺骗智能图像分类器输出指定的目标类别.在所建立的仿真场景评估测试了不同视角、亮度以及遮挡等条件下的攻击效果,实验结果表明所生成的全视角对抗伪装具有显著优于多对抗贴片拼接纹理的攻击欺骗性能.进一步地,通过对Clarifai商用图像分类系统测试、3D打印模型实验以及其他目标物伪装纹理实验,验证了所生成对抗伪装出色的黑盒迁移攻击性、对现实世界场景以及其他目标物的泛化性.研究成果有助于更有针对性地增强智能识别系统的安全性,另一方面有助于发展面向智能识别系统的隐私保护技术.未来将进一步对伪装纹理视觉自然度问题开展优化研究.

参考文献

[1] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 4909-4926.

[2] SZEGEDY C, ZAREMBA W, SUTSKEVERET I, et al. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations. Banff: ICLR, 2014: 2632-2640.

[3] DUAN Y, CHEN J, ZHOU X, et al. Learning coated adversarial camouflages for object detectors[C]//Proceedings of

the Thirty-First International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022: 891-897.

[4] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 1083-1092.

[5] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1625-1634.

[6] LU J, SIBAI H, FABRY E, et al. No need to worry about adversarial examples in object detection in autonomous vehicles[EB/OL]. (2017-07-12) [2022-10-12]. https://arxiv.org/abs/1707.03501.

[7] BROWN T B, MANE D, ROY A, et al. Adversarial patch [EB/OL]. (2017-12-27) [2022-10-12]. https://arxiv.org/abs/1712.09665.

[8] WANG J, LIU A, YIN Z, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 8565-8574.

[9] ATHALYE A, ENGSTROM L, et al. Synthesizing robust adversarial examples[C]//International Conference on Machine Learning. New York: ICML, 2018: 284-293.

[10] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 1-11.

[11] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial examples in the physical world[C]//5th International Conference on Learning Representations. Toulon: ICLR, 2017: 1-14.

[12] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9185-9193.

[13] XIE C, ZHANG Z, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2730-2739.

[14] DONG Y, PANG T, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4312-4321.

[15] DUAN Y, ZOU J, ZHOU X, et al. Adversarial attack via dual-stage network erosion[J]. Computers & Security, 2022, 122: 102888.

[16] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[17] SZEGEDY C, VANHOUCKE V, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2818-2826.

[18] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI Conference on Artificial Intelligence. New York: AAAI, 2017: 4278-4284.

[19] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]//European Conference on Computer Vision. Cham: Springer, 2016: 630-645.

[20] HUANG G, LIU Z, VAN D M, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 4700-4708.

[21] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2017: 1251-1258.

[22] LIU C, ZOPH B, et al. Progressive neural architecture search[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 19-34.

作者简介

段晔鑫　男,1987年生,江西余干人,2021年在陆军工程大学获得计算机科学与技术专业博士学位. 现为陆军军事交通学院讲师,主要研究方向为对抗机器学习.
E-mail: duanyexin0713@163.com

贺正芸　女,1976年生,湖南衡阳人,2005年、2023年分别在湖南大学、陆军工程大学获硕士学位和博士学位. 现为湖南工业大学讲师,主要研究方向为人工智能、机器视觉.
E-mail: zhengyun_he@126.com

潘志松　男,1973年生,江苏南京人. 2003年获南京航空航天大学博士学位,现为陆军工程大学教授、博士生导师. 主要研究方向为人工智能、模式识别.
E-mail: panzs@nuaa.edu.cn