双重注意抑制攻击:在物理世界中产生对抗性伪装

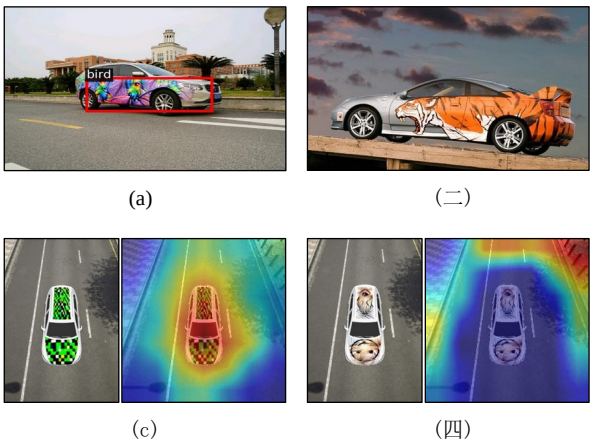
王、刘爱山、尹、刘顺昌、唐和*
软件开发环境国家重点实验室，
北京航空航天大学，中国北京
- jk 北航 scse，刘爱山，yzx835，liusc，sytang，xlliu}@buaa.edu.cn

摘要

深度学习模型容易受到对立例子的影响。作为一种对实用深度学习系统更具威胁性的类型，物理对立范例近年来受到了广泛的研究关注。然而，在没有利用诸如模型不可知和人类特定模式等内在特征的情况下，现有作品在物理世界中产生微弱的对抗性扰动，这不足以跨不同模型进行攻击，并且在视觉上表现出可疑的外观。受注意反映识别过程的内在特征这一观点的启发，提出双重注意抑制(DAS)攻击，通过抑制模型和人的注意来产生视觉自然的、具有强可转移性的物理攻击伪装。至于攻击，我们通过将模型共享的相似注意模式从目标区域转移到非目标区域来产生可转移的敌对标记。同时，基于人类的视觉注意力总是集中在显著的项目(例如，可疑的扭曲)，我们回避人类特有的自下而上的注意，以生成与场景上下文相关的视觉自然的伪装。我们在数字和物理世界中对最新模型(例如 Yolo-V5)进行了广泛的分类和检测实验，并显著证明了我们的方法优于最先进的方法。[1]

1. 介绍

深度神经网络(DNNs)已经在广泛的应用领域取得了显著的性能，例如计算机视觉 24]，自然语言[39]，和声学[34]，*等等，但他们容易受到对立的例子[41]*.这些精心设计的扰动对人类来说是察觉不到的，但很容易导致 DNNs 做出错误的预测，这对数字和物理世界中的深度学习应用构成了强大的安全挑战[22,13,31].



图一。(a)显示先前工作生成的伪装的可疑外观(即，刚果爱国者联盟[19]).(b)是现实世界中普遍存在的喷漆汽车。(c)显示了由现有作品(即 CAMOU [)产生的反面例子(分类为 pop 瓶)48])及其对应的注意力地图。
(d) 显示了由我们的 d as 和它的注意力分散图生成的对抗性例子(分类为西施犬)。

在过去的几年中，已经提出了一长串的工作来在不同的场景和不同的设置下执行对抗性攻击[26,7,2].虽然对深度学习提出了挑战，但对立的例子对于理解 DNNs 的行为也是有价值的，这可以提供对盲点的观察，并有助于建立健壮的模型[20,42,28,47].一般来说，对抗性攻击可以分为两类:数字攻击，它通过干扰数字空间中的输入数据来攻击 DNNs 以及通过修改 vi-来攻击 dnn 的物理攻击

*通讯作者 Corresponding author

1 您可以在中找到我们的代码 https://github.com/nlsde-safety-team/DualAttentionAttack。

物理世界中真实物体的一些特征。与数字世界中的攻击相反[23,45,21,48]，由于复杂的物理限制和条件(例如，照明、距离、摄像机等)，物理世界中的对抗性攻击更具挑战性。)，这将削弱所产生的对抗性扰动的攻击能力[12].在本文中，我们主要关注更具挑战性的物理世界攻击任务，这对于在实践中部署深度学习应用也更有意义。

尽管已经采用了几种尝试来执行物理攻击31,19,30]，现有的工作总是忽略了固有的特征，如模型不可知和人类特有的模式，因此它们的攻击能力仍然不能令人满意。特别地，这些限制可以总结为(1)现有的方法忽略了模型之间的共同模式，并且使用模型特定的线索(例如，特定模型的梯度和权重)生成对抗性的干扰，这不能攻击不同的目标模型。换句话说，对抗性干扰的可转移性很弱，这削弱了他们在现实世界中的攻击能力；(2)当前的方法产生了具有视觉可疑外观的对抗性干扰，这与人类感知不一致，甚至吸引了人类的注意力。例如，涂在敌对伪装上的[19]，分类器将汽车误分类为鸟。然而，如图所示1(a)伪装明显包含不自然和可疑的与鸟相关的特征(例如鸟头)，这吸引了人类的注意。

为了解决上述问题，本文提出了双重注意抑制(DAS)攻击，通过抑制模型和人类的注意。关于攻击的可转移性，受生物观察的启发，当刺激特征被抵消时，不同个体之间的大脑活动共享相似的模式[46](即选定的注意力[27])，我们通过抑制不同模型之间共享的注意力模式来执行对抗性攻击。具体来说，我们通过连通图将模型共享的相似注意力从目标区域转移到非目标区域。因此，由于没有注意目标区域中的对象，目标模型将被错误分类。由于我们生成的敌对伪装捕获模型不可知的结构，它可以在不同的模型之间转移，这提高了可移植性。

至于视觉的自然性，心理学家发现，人类视觉的自下而上的注意力会提醒人们注意突出的物体(如失真)6].现有的方法生成具有视觉上可疑外观的物理对抗实例，其显示了人类感知的显著特征。因此，我们试图通过生成包含与场景上下文高度语义相关的对抗性伪装来回避这种人类特有的视觉注意。结果，生成的伪装更加可疑人类感知的自然。数字1(c)敌对伪装是由CAMOU生成的吗[48]这是对人类视觉的怀疑。相比之下，我们生成的对抗性伪装产生了更自然的外观，如图所示1(d)。

据我们所知，我们是第一个探索模型间共享注意力特征，并通过抑制模型和人类注意力在物理世界中产生敌对伪装的人。在数字世界和物理世界中对分类和检测任务进行的大量实验表明，我们的方法优于其他最先进的方法。

2. 相关作品

对立的例子是精心设计的干扰，人类察觉不到，但可能误导DNNs [41,22].在过去的几年中，已经提出了一长串的工作来开发对抗性攻击策略[25,13,30,43,11,29,48,19].一般来说，有几种不同的方法来对对抗性攻击方法进行分类，例如，有针对性的或无针对性的攻击，白盒或黑盒攻击等。根据对抗性扰动产生的领域，对抗性攻击可以分为数字攻击和物理攻击。

数字攻击对数字像素域中的输入数据产生不利的扰动。Szegedy等人[41]首先介绍了对抗性的例子，并使用L-BFGS方法来产生它们。通过利用目标模型的梯度，Goodfellow等人提出了快速梯度符号方法(FGSM)[22]这会很快产生反面例子。此外，Madry等人[1]建议投影梯度体面(PGD)，这是目前最强的一阶攻击方法。基于目标模型的梯度，已经提出了一系列攻击方法[25,8,45,9].尽管这些攻击在数字世界中取得了实质性的成果，但是当它们被引入到物理世界中时，它们的攻击能力会显著退化。

另一方面，物理攻击旨在通过改变物理世界中真实物体的视觉特征来产生普遍的扰动。为了实现这一目标，一些作品首先在数字世界中产生对抗性扰动，然后通过在真实对象上绘制对抗性伪装或直接创建受扰动的对象来进行物理攻击。通过构造一个渲染函数，Athalye等人[2]在物理世界中生成3D对抗性物体来攻击分类器。Eykholt等人[13]引入的NPS [33]引入到考虑制造误差的损失函数中，使得它们可以对交通标志识别产生强烈的恶意攻击。最近，黄等[19提出了通用物理伪装At- tack (UPC),它通过联合愚弄

区域提议网络和分类器。另一种工作试图通过生成对抗性补丁来进行物理对抗性攻击 3]，它将噪声限制在一个小的和局部的小块，而没有扰动约束[30,31].

3. 方法

在这一节中，我们首先给出问题的定义，然后详细阐述我们提出的框架。

3.1. 问题定义

给定深度神经网络 F θ 和具有地面真实标签 y 的输入干净图像 I，数字世界中的对立示例 Iadv 可以使模型进行如下错误预测:

$$f\theta(iadv) /= y \quad s\|. \quad t \quad .\|I \quad iadv < \epsilon, \quad (1)$$

其中距离度量用于量化足够小的两个输入 I 和 Iadv 之间的距离。

在物理世界中，让(M，T)表示具有网格张量 M、纹理张量 T 和地面真实 y 的 3D 真实对象。深度学习系统的输入图像 I 是真实对象(M，T)在环境条件 c C(例如，相机视图、距离、照明等)下的渲染结果。)从渲染器由 I = ((M，T)，c)。为了执行物理攻击，我们通过用具有不同物理属性(例如，颜色、形状)的对抗性纹理张量 Tadv 替换原始 T 来生成 Iadv = ((M，Tadv)，c)。因此这个定义我们的问题可以描述为:

$$F\theta(Ia \quad y \quad s. \quad t. \quad t \quad tadv < \epsilon, \quad (2)$$
$$dv)$$

在这里，我们确保了 ϵ.在物理世界中生成的对抗性伪装的自然性

在本文中，我们主要讨论物理世界中的对抗性攻击，并生成一种对抗性伪装(即纹理)，当它被绘制或覆盖在真实对象上时，能够欺骗真实的深度学习系统。

3.2. 框架概述

为了生成视觉自然、可移植性强的物理对抗伪装，提出了双重注意抑制(DAS)框架，同时抑制模型和人的注意。总体框架可以在图中找到 2。

关于攻击的可转移性，受生物学观察的启发，我们抑制了模型间共享的相似注意模式。具体来说，我们一般-
通过分散模特的注意力来伪装敌人
从目标区域到非目标区域(例如背景)的转换。通过
连通图。由于不同的深度模型对同一物体产生相似的注意模式，我们生成的对抗性伪装可以捕获模型不可知的结构并转移到不同的模型。

至于视觉自然性，我们的目标是避免人类视觉中特定的自下而上的注意 6]通过生成视觉上自然的伪装。通过引入种子内容补丁 PO(其与场景上下文具有很强的感知相关性),在这种情况下生成的敌对伪装对于人类感知来说可能更加不可疑和自然。因为人类在做预测时更关注物体的形状 29]，我们进一步保留种子内容块的形状信息，以改善人类的注意力相关性。因此，回避了人类特有的注意机制，导致更多的自然伪装。

3.3. 模型注意力分散

生物学家发现，相同的刺激特征(即选择性注意力)会在不同的个体中产生相似的大脑活动模式 46](即神经元超感知的相似特征)。由于人工神经网络是从人类中枢神经系统实现的[16]，我们也有理由假设 DNNs 可能具有相同的特征，即不同的模型在做出相同的预测时对相同的对象具有相似的注意模式。基于上述观察，我们考虑通过捕获模型不可知的注意结构来提高对抗性伪装的迁移能力。

视觉注意力技术[49]已经被研究了很长时间，以提高对深度学习行为的解释和理解，例如 CAM [49]，Grad-CAM [37]，以及

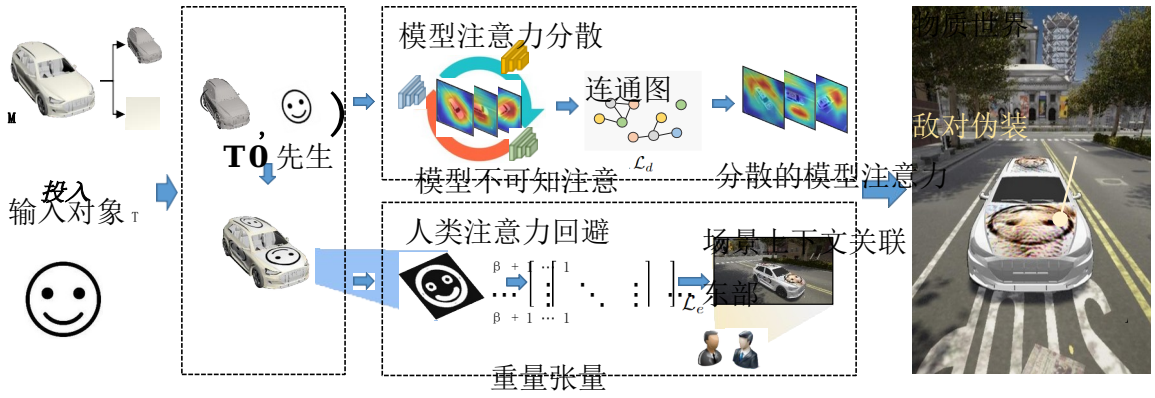
Grad-CAM++ [5].在进行预测时，模型将大部分注意力放在目标对象上，而不是均值上
无光泽的部分。直觉上，要成功攻击一个模型，我们直接把模型的注意力从显著的物体上转移开。换句话说，我们将显著区域上的模型共享的相似注意图转移到其他区域，并迫使注意权重均匀地分布在整个图像中。因此，该模型可能无法聚焦于目标对象并做出错误的预测。

具体地说，给定一个对象(M，T)，一个要优化的对立性质张量 Tadv，以及某个标签 y，我们通过 R 得到 Iadv，然后用注意模块 A 计算注意图 Sy

$$Sy = A(Iadv, y)。(3)$$

更准确地说，注意力模块 A 是

$$A(I, y) = \sum_k \sum_y \sum_{ij} a_{ij} relu\left(\frac{\partial py}{\partial Ak}\right)$$



图二。我们的 DAS 方法的框架。首先，通过充分利用模型相似的注意模式，利用损失函数 d 迫使“热”区域远离目标对象来转移固有的注意特性。然后，通过将对手的外观与上下文场景相关联并保留种子内容图像的形状信息来生成视觉上自然的对手伪装，从而规避人类特有的视觉注意机制。

其中 g ky 是特定类别 y 的梯度权重，并且激活图 k，py 是类 y，Ak 的分数

像素值是否到位 第 k 个特征图的(I，j)，relu()表示 relu 函数。注意，注意力模块可以是任意的深度学习模型，而不是目标模型。

给定由等式 2 计算的注意力地图 Sy3 我们的目标是转移注意力区域，并迫使模型聚焦于非目标区域。直观上，注意力图的像素值表示该区域对模型预测的贡献程度。为了降低显著对象的注意力权重并分散这些注意力区域，我们利用了连通图，该图包含图中任意一对节点之间的路径。在图像中，每个像素的关注权重高于特定阈值的区域可以被视为连通区域。为了使用连通图转移模型的注意力，我们考虑以下两个任务：(1)通过将连通图分成多个子图来降低整体连通性；(2)降低连通子图中每个节点的权重。为了实现这些目标，我们建议注意力分散损失为

$$= \frac{1}{k} \quad 1 \frac{Gk, \; s.t. \; G}{N \; Nk} \quad \subseteq \text{的}, \qquad (5)$$

其中 Gk 是对应于 Sy 中第 k 个连通图的区域中的像素值之和，N 是 Sy 的总像素数，Nk 是 Gk 的总像素数。通过最小化 d，注意力图中的显著区域变得更小(即，注意力分散)，并且显著区域的像素值变得更低(即，不再被“加热”)，从而导致“注意力分散”的注意力图。

3.4. 人类注意力回避

为了克服复杂环境带来的问题-物质世界的环境条件，大多数物质的攻击会产生相对巨大的对抗性扰动[11]. 由于自下而上的人类注意力机制总是提醒人们注意显著的物体(例如，失真)[6]，在这种情况下，由于显著的干扰，对立的例子总是能够吸引人类的注意，在物理世界中表现出可疑的外观和较低的隐秘性。

在本文中，我们的目标是通过抑制人类的视觉机制来产生更多的视觉自然的伪装。直觉上，我们期望生成的伪装与被攻击的环境有相似的视觉感觉(例如，车辆上漂亮的油漆比无意义的扭曲更容易被人接受)。因此，生成的对抗性伪装可以与人类感知高度相关，而人类感知是不可疑的。

特别地，我们首先合并种子内容补丁 PO，其包含与场景上下文的强语义关联。然后，我们通过 TO = ψ(PO，T)在车辆(M，T)上绘制种子内容补丁。具体来说，ψ()是首先传递 2D 的变换运算符种子内容补丁成 3D 张量，然后通过张量加法给汽车上色。

因为人类在聚焦时更关注形状-观察物体并做出预测[29]，我们旨在通过更好地预先提供种子内容补丁的形状来进一步改善人类注意力相关性。具体地，我们使用边缘提取器 Φ[,获得边缘补片 Pedge = Φ(PO)4]来自 seed 内容修补程序。需要注意的是，Pedge 在每个像素中有 0-1 的值。然后，我们简单地把边缘块变换成一个掩模张量

e 与 T0 具有相同的维数。

利用掩模张量 E，我们可以区分边缘和非边缘区域，并限制添加到边缘区域的扰动。因此，注意力回避损失 e 可以表述为

$$le = (\beta E+1) \odot (Tadv \quad T0)2, \quad (6)$$

其中 β E + 1 是权张量，1 是这样一个张量，其中每个元素都是 1，并且它的维数与 E 相同，并且表示逐元素乘法。

为了进一步提高伪装的自然度，我们引入了平滑损失[13]通过减少相邻像素之间的差值平方。对于渲染的敌对图像 Iadv，平滑损失可以表示为:

$$ls = \sigma (xi, xi+1, j)2 + (xi, xi, j+1)2, \quad (7)$$

其中 Xi，j 为 Iadv 在坐标(I，j)处的像素值。

综上所述，在这种情况下生成的伪装将在像素和感知层面上与场景上下文视觉相关，从而逃避人类的感知注意。

3.5. 整体优化流程

总的来说，我们通过联合优化模型注意力分散损失 d、人类注意力回避损失 e 和平滑损失 s 来生成对抗性伪装。

具体来说，我们首先将目标模型从显著的物体转移到无意义的部分(如背景)；然后，我们通过增强与场景上下文的强感知相关性来回避人类特有的注意机制。因此，我们可以通过最小化以下公式来生成可转移的和视觉上自然的敌对伪装

$$\min Ld + \lambda Le + Ls, \quad (8)$$

其中 λ 控制项 e 的贡献。

为了平衡攻击能力和外表的自然-ness，我们在分类任务中将 λ 设定为 105，在检测任务中将 λ 设定为 103，并根据我们的实验结果将 β 设定为 8。整个训练算法可以描述为算法 1。

4. 实验

在本节中，我们首先概述实验设置，然后通过在数字和物理世界中的全面评估来说明我们提出的攻击框架的有效性。

算法 1 双重注意力抑制(DAS)攻击

输入:环境参数集 C = c1，c2，...cr、3D 真实对象(M，T)、种子内容补丁 P0、神经渲染器、注意力模型和类标签 y

输出:对抗性纹理张量 Tadv

T0← ψ(P0，T)

Pedge Φ(P0)将
Pedge 变换为 E，
初始 Tadv 为 T0
对于多少个纪元呢
从 C 中选择微型批次环境条件
对于 m = r/minibatch 步骤 do
Iadv ((M，Tadv)，cm)
叙利亚(Iadv，y)
通过等式计算 d、e 和 s(5,6,7)通过等式优化 Tadv(8)
结束于
结束于

4.1. 实验设置

虚拟环境。为了执行物理世界 at- tack，我们选择卡拉[10]作为我们的 3D 虚拟仿真环境，这是自动驾驶研究常用的开源模拟器。基于 Un- real Engine 4，CARLA 提供了许多高分辨率的开放数字资产，例如城市布局、建筑和车辆，以模拟与现实世界几乎相同的数字世界。

评估指标。为了评估我们提出的方法的性能，我们选择广泛使用的准确度作为分类任务的度量；对于检测任务，我们采用 P@O.5 以下[48]，它反映了 IoU 和精度信息。

比较方法。我们在 3D 攻击和物理攻击文献中选择了几个最先进的作品，包括 UPC [19]，CAMOU [48]，和 MeshAdv [44].为了更好的分析，我们以不同的方式实现 MeshAdv。我们使用 ResNet-50 作为分类的基础模型，使用 Yolo-V4 作为检测的基础模型。我们在补充材料中提供了关于这些方法的更多信息。

目标模型。我们选择不同的常用模型架构进行实验。具体来说，Inception-V3 [40]，VGG-19 [38]，ResNet-152 [15]，以及 DenseNet [18]被用于分类任务；Yolo-V5 [35]，固态硬盘[32]，更快的 R-CNN [36]，并屏蔽 R- CNN [14]被用于检测任务。对于所有模型，我们使用 ImageNet 和 COCO 上的预训练版本。

实施细节。我们根据经验为分类任务设定 λ = 105，为探测任务设定 λ = 5103，我们设定 β = 8。我们采用 Adam 优化器，学习率为 0.01，权重衰减为 104，最大值为

五世之母。我们采用种子内容补丁(例如，笑脸图像)作为 3D 对象的外观

在训练过程中。我们所有的代码都是用 PyTorch 实现的。我们在 NVIDIA Tesla V100-SXM2-16GB GPU 集群上进行培训和测试。在现实世界的攻击场景中，对手对部署的模型(即架构、重量等)只有有限的了解和访问。).考虑到这一点，我们主要关注黑盒设置中的攻击，这对于物理世界的攻击更有意义和适用。

4.2. 数字世界攻击

在这一部分中，我们评估了在黑盒设置下，在数字世界中的车辆分类和检测任务中，我们生成的对抗性伪装的性能。

我们在模拟环境中随机选择 155 个点来放置车辆，并使用虚拟摄像机在每个点使用不同的设置(即角度和距离)捕捉 100 个图像。具体来说，我们使用不同的距离值(5° 、10° 、15° 和 20°)，四个摄像机俯仰角值(22.5° 、45° 、67.5° 、90°)和八个摄像机偏航角值(南、北、东、西、东南、西南、东北、西北)。然后，我们收集 15，500 幅不同设置组合的模拟图像，并选择 12，500 幅图像作为训练集，3，000 幅图像作为测试集。为了进行公平的比较，我们使用 ResNet-50(用于分类)和 Yolo-V4(用于检测)的主干作为训练中的注意力模块。如表中所示 1 和桌子 2，我们可以得出如下几个结论:

(1) 我们的对抗性伪装在不同模型上实现了明显更好的分类和检测任务性能(在 ResNet-152 上最大下降 41.02%，在更快的 R- CNN 上最大下降 23.93%)。

(2) 我们发现，对于检测任务，UPC 比其他基线工作得更差。我们推测原因可能是 UPC 主要是为物理攻击设计的，因此在数字世界中表现出较差的攻击能力。相比之下，我们的 DAS 攻击利用了固有特性，在数字世界中仍然具有良好的攻击能力。

(3) 与其他主干模型相比，SSD 表现出明显更好的稳健性(即精度下降更低)。原因可能是 SSD 中的一些模块不太容易受到恶意攻击，这可以用来进一步提高模型的鲁棒性。我们把它作为未来的工作。

4.3. 物理世界攻击

对于物理世界的攻击，我们进行了几次实验来验证我们生成的对抗性伪装的实际有效性。由于资金和条件的限制，我们印刷了我们的敌对伪装

方法	准确度(%)			
	盗梦空间-第三版	VGG-19	ResNet-152	DenseNet
生	74.36	40.62	73.51	71.91
MeshAdv	42.31	32.44	35.33	58.04
卡穆	47.51	31.46	48.93	57.56
单件生产成本(Unit Production Costs)	42.40	38.00	48.18	65.87
我们的	**39.86**	**30.18**	**32.49**	**55.42**

表 1。数字世界中分类任务的结果。

	P@0.5 (%)			
	Yolo-V5	SSD	更快的 R-CNN	口罩 R-CNN
生的	92.07	81.54	86.04	89.24
MeshAdv	72.45	66.44	71.84	80.84
卡穆	74.01	73.81	69.64	76.44
UPC	82.41	74.58	76.94	81.97
我们的	**72.58**	**65.81**	**62.11**	**70.21**

表二。数字世界中的探测任务的结果。

通过惠普彩色激光打印机 Pro MFP M281fdw 打印机，将它们贴在具有不同背景的玩具车模型上，以模拟真实的车辆绘画。为了进行公平的比较，我们使用华为 P40 手机在各种环境条件下(即左、右、前、后 8 个方向及其相应的交叉方向，3 个角度
0° ♀、45° ♀、90° ♀，2 个距离长和短距离以及 3 种不同的环境)拍摄了 144 张汽车模型的照片。我们生成的对抗性伪装的可视化可以在图中找到 3。

评估结果见表 3 和桌子 4。与其他方法相比，DAS 表现出竞争性的可转移攻击能力，明显优于比较基准(例如，在 Inception-V3 上为 31.94%，在 VGG-19 上为 27.78%，在 ResNet- 152 上为 29.86%，在 DenseNet 上为 34.03%)。此外，UPC 的评估结果比数字世界有明显的改善，这与我们的分析是一致的。然而，SSD 在物理世界中表现出较低的鲁棒性，值得进一步研究。此外，Yolo-V5 显示了惊人的 P@0.5 值，这可能是因为 Yolo-V5 是专门为物理世界中的应用程序设计的。尽管面对这种强模型，与其他方法相比，我们的 DAS 方法仍然表现出一定的攻击能力。

综上所述，实验结果证明了我们的敌对阵营在物理世界中具有很强的迁移攻击能力。

4.4. 模型注意力分析

在这一部分，我们通过定性和定量的研究对模型注意力进行了详细的分析，以验证模型注意力分散在 DAS 攻击中的有效性。

首先，我们通过可视化不同模型对同一即时消息的注意区域来进行定性研究

图 3。攻击玩具车的结果。它们分别被识别为汽车、凉鞋、汽车、鼠标。

年龄。如图所示 4(a)不同的 dnn 对同一幅图像表现出相似的注意模式。换句话说，不同的模型将它们的注意力放在相似的区域，表明注意力在模型之间是共享的，并且可以被认为是模型不可知的特征。

然后，我们通过计算结构相似性指数(SSIM)进行定量研究 50]，这是一个众所周知的质量度量，用于测量两个图像之间的相似性[17].具体来说，我们在不同的模型上生成特定图像(即熊猫)的注意力图，并计算不同模型(即 Inception-V3、VGG-19、ResNet-152 和 DenseNet)上每对注意力图之间的 SSIM 值。如图 2 所示

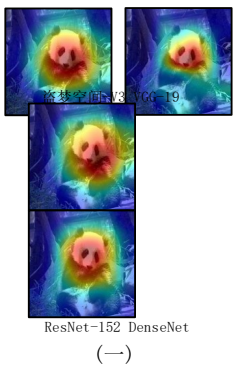
方法准确度(%)	盗梦空间-第三版	VGG-19	ResNet-152	DenseNet
生的	58.33	40.28	41.67	46.53
MeshAdv	40.28	34.03	38.89	36.11
卡穆	40.28	29.17	31.25	45.14
单件生产成本	35.41	33.33	33.33	41.67

表 3。在分类任务中物理世界的结果。

	P@0.5 (%)			
	Yolo-V5	SSD 更快的	R-CNN 口罩	R-CNN
生的	100.00	90.28	68.06	93.75
MeshAdv	100.00	61.11	56.25	63.19
卡穆	99.31	61.11	61.81	63.19
单件生产成本	100.00	63.19	52.08	61.81

(Unit

表 4。物理世界中的探测任务的结果。



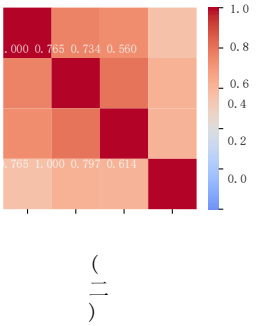


图 4。(a)是 4 个不同模型对某一特定图像的注意映射。(b)是根据 SSIM 值绘制的热图。

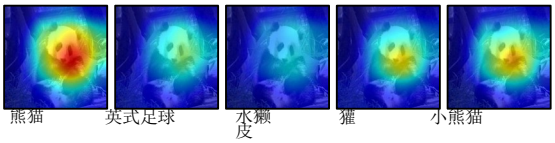


图 5。用 ResNet-152 实现同一幅图像上不同目标标记的熊猫注意力图的可视化。当不同的目标标签被提供给模型时，注意图显著不同。

抱朴说 4(b)不同的模型表现出相对较高的注意力地图相似性。

此外，我们通过改变模型预测(即，类别)来进一步可视化注意力地图。如图所示 5 当改变类别标签时，注意力图从显著的对象转移，并且在整个图像上变得更加稀疏。

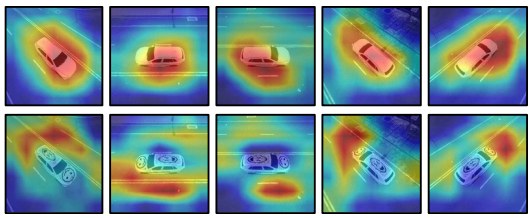


图 6。我们的 DAS 攻击前后的注意力地图。在我们的 DAS 攻击后，模特的注意力被分散了。

综上所述，我们可以得出如下几个结论:(1)不同的 dnn 对特定图像中的同一类别表现出相似的注意模式；(2)我们可以通过分散 DNN 的注意力来攻击它的错误预测。更多的实验结果可以在补充材料中找到。

4.5. 人类感知研究

为了评估我们生成的敌对伪装的自然性，我们在一个最常用的众包平台上进行了人类感知研究。我们使用不同的方法(例如 MeshAdv、CAMOU、UPC 和 Ours)来对抗性地扰动我们的 3D 汽车对象，并获得对抗性的纹理。然后，我们使用这些伪装来绘制汽车，并获得用于人类感知研究的渲染图像，如下所述:(1)识别。参与者

要求将上述方法生成的每个伪装分配到 8 个类别中的一个(地面真实类别，6 个类似于地面真实的类别，以及“我不能说出它是什么”)。至于 CAMOU，由于它缺乏语义信息，我们不考虑将其用于识别任务；
(2)自然。参与者被要求在全国范围内打分

伪装的强度从 1 到 10。特别是，我们收集了 106 名参与者的所有回答。

问题	百分比(%)		
	MeshAdv	卡穆	IPC 我们的
承认	36.6	-	27.4 49.6
自然	43.4	39.6	40.6 60.4

表 5。人类感知研究的结果。40

如表中所示 549.6%的参与者可以识别出我们的伪装的真实标签，这远远好于其他方法生成的标签。至于自然性任务，高达 60.4%的参与者认为我们的敌对伪装是自然的

远远超过其他公司(17%以上)。因此，我们可以得出结论，我们的敌对伪装在视觉上是最自然的，在感知上与人类的感知是一致的。[2]

4.6. 消融研究

在本节中，我们进行了几项消融研究，以进一步调查我们的两个主要损失项的贡献，即模型注意力分散损失和人类注意力回避损失。由于[]中充分研究了平滑损失这一事实 13]，我们将其设为固定期限。

不同损失项的影响。不同的损失项起不同的作用，我们进行烧蚀研究以进一步研究损失项的影响。我们认为，在我们的 DAS 方法中，注意力分散损失 d 模型主要提供了一种可转移的攻击能力，而人类的注意力回避提供了自然的表象。为了证明这些观点，我们通过计算不同的损失项组合来进行实验。具体来说，我们分别使用函数 d、e 和 d+λe(s 固定)来优化对抗性伪装。如表中所示 6 精度显示出显著下降(即，d 设置下的 36.53%下降到 e 设置下的 59.87%，d + e 设置下的 39.86%)。并且用良性图像生成的相应 SSIM 值分别是 0.6905、0.9987 和 0.7551，证明了我们的观点。此外，在我们的实验中可以观察到一个有趣的结果。在 e 环境下训练时，准确率在 VGG-19 和 DenseNet 上有明显提高，但在 Inception-V3 和 ResNet-152 上有所下降，这意味着常见纹理可能对 DNNs 造成不可知的影响，进一步证明了它们的脆弱性可怜虫。

	准确度(%)			
	盗梦空间-V3	VGG-19	ResNet-152	DenseNet
生的	74.36	40.62	73.51	7
致死	36.53	25.87	31.20	5
	59.87	50.00	47.87	75
$Ld + \lambda Le$	39.86	30.18	32.49	55.42

表 6。注意力分散部分的消融研究。我们将 λ 设为 105。

2 有关我们的实验详情，请访问 Our experimental details can be accessed at https://github.com/nlsde-safety-team/DualAttentionAttack。

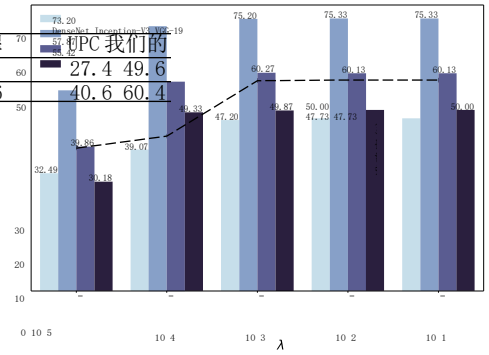


图 7。关于 λ 有效性的研究。虚线线代表精度变化的趋势，每个 λ 对应的值是四个模型的平均精度。

超参数 λ 的影响。关于超参数 λ，我们认为它控制着与情景语境的强语义关联的水平。我们使用精确度和 SSIM 在 ResNet-50 模型上评估了 λ 的有效性。具体来说，我们将 λ 分别设置为 105、104、103、102 和 101。如图所示 7 随着 λ 的增加，模型精度先增加，然后保持一个稳定值。我们计算每对干净的和相应的相反的例子之间的 SSIM 值，其显示出相似的趋势(即，0.7034、0.7551、0.8750、0.9982、0.9991 和 0.9998，越接近 1
SSIM 值越大，图像越相似)。根据结果，我们可以得出结论，λ 平衡了攻击能力和外观。λ 越大，精度和 SSIM 值越大，意味着攻击能力越低，外观越好。最后，SSIM 达到其上限，导致额外攻击能力的损失。

5. 结论

本文提出了双重注意抑制(DAS)攻击，通过抑制模型和人的注意来产生物理世界中的敌对伪装。为了提高对立标记的可迁移性，我们通过将模型共享的相似注意从目标区域转移到非目标区域来抑制模型注意。由于我们生成的伪装捕捉模型不可知的结构，它可以在不同的模型之间转移。为了生成视觉上更自然的伪装，我们通过回避人类特有的自下而上的注意来抑制人类的注意。通过保留与场景上下文具有强语义关联的种子内容补丁的形状，所生成的伪装可以与人类感知高度相关，这对于人类的注意力来说更加自然和无意识。我们在黑盒设置下的数字和物理世界中进行了广泛的分类和检测实验

DAS 的表现优于最先进的基准。

将来，我们有兴趣在真实世界的场景中使用真实的车辆来研究我们的对抗性伪装的攻击能力。使用投影或 3D 打印，我们可以简单地在真实世界的车辆上画出我们的伪装。此外，我们还想研究我们生成的伪装对提高模型对不同噪声的鲁棒性的有效性。

参考

[1] 马德雷·阿·马克洛夫·阿·施密特·齐普拉斯·阿·弗拉杜
A. 抗对抗性攻击的深度学习模型。在 arXiv 预印本 arXiv:1706.06083 中，2017 年 6 月。2

[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 e-prints，arXiv 页:1707.07397，2017 年 7 月。1,2

[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。2017 年 12 月 CoRR。3

[4] J. 精明。边缘检测的计算方法。
PAMI PAMI，1986 年 11 月 8 日。4

[5] A.Chattopadhay， A. Sarkar， P. Howlader 和 V. N. Balasub- ramanian。Grad-cam++:深度卷积网络的基于广义梯度的可视化解释。在 WACV 中，
2018 年 3 月。3

[6] 查尔斯·e·康纳、霍华德·e·伊格斯和史蒂文·扬蒂斯。视觉注意:自下而上与自上而下。当代生物学，14(19)，2004 年 10 月。2,3,4

[7] 董，廖，庞天宇，。推动对抗性攻击的势头。在 CVPR，
2018 年 6 月。1

[8] 董、、廖、庞天宇、、、、、。用动力推动对抗性攻击。2018 年 6 月，CVPR。2

[9] 、董、庞天宇、、。利用平移不变攻击规避对可转移对立范例的防御。2019 年 6 月，CVPR。2

[10] Alexey Dosovitskiy， German Ros， Felipe Codevilla，Antonio Lopez 和 Vladlen Koltun。卡拉:一个开放的城市驾驶模拟器。2017 年 11 月在科尔。5

[11] 段冉杰，马，王，，
A. 秦刚和杨云。用自然风格隐藏物理世界的攻击。2020 年 6 月在 CVPR。2,4

[12] 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。在 NeurIPS，2018 年 12 月。2

[13] 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。2018 年 6 月，CVPR。1,2,5,8

[14] 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。2017 年 3 月，ICCV。5

[15] 何、、、任、。用于图像识别的深度残差学习。2016 年 6 月，CVPR。5

[16] 迈克尔·亨特里希。方法学与冠心病治疗。可于 2015 年 8 月致电 SSRN 2645417 查询。3

[17] A. 霍尔和迪兹欧。图像质量指标:Psnr 与 ssim。
2010 年 8 月在 ICPR。7

[18] 黄高、刘庄和基利安·q·温伯格。密集连接的卷积网络。2016 年 8 月 CoRR。5

[19] 黄、高、周、谢慈航、阿兰 l .尤耶、邹长青和。对目标探测器的通用物理伪装攻击。在 CVPR，
2020 年 6 月。1,2,5

[20] 安德鲁·易勒雅斯、什巴尼·桑图尔卡、迪米特里斯·齐普拉斯、洛根·恩斯特罗姆、布兰登·特兰和亚历山大·马德瑞。反面例子不是错误，而是特征。在 NeurIPS，2019 年 5 月。1

[21] 内森·英卡威奇、魏文、海(海伦)李和陈。特征空间扰动产生更多可转移的对立例子。2019 年 6 月，CVPR。2

[22] 解释和利用对立的例子。arXiv 预印本 arXiv:1412.6572，2014 年，2014 年 12 月。1,2

[23] 贾，陆，塞内姆韦利帕萨拉尔，钟振宇和陶伟。利用离差减少增强对立范例的跨任务迁移性。2019 年 5 月 CoRR。2

[24] 亚历克斯·克里热夫斯基、伊利亚·苏茨基弗和杰弗里·E·辛顿。深度卷积神经网络的图像网分类。在 NeurIPS，2012 年 5 月。1

[25] 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。2016 年 7 月 CoRR。2

[26] 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 ICLR 车间，
2017 年 7 月。1

[27] 同伴在场对眼球运动和注意力表现的影响。在 Behav Neurosci 前面，2020 年 1 月。2

[28] 李天林，刘爱山，，，徐，张崇智，，谢。通过关键攻击路径理解对抗性竞争。信息科学，
2021 年 2 月。1

[29] 刘爱山，黄泰然，，，徐，，陈，斯蒂芬 j .梅班克，陶大成。对嵌入代理的时空攻击。在 ECCV，2020 年秋天。2,3,4

[30] 刘爱珊，，范佳欣，，，谢慧媛，陶大成。用于生成敌对补丁的感知敏感 GAN。2019 年 1 月，AAAI。2,3

[31] 刘爱山、王、、张崇智、曹博文和。自动检出的补丁攻击。2020 年 5 月在 ECCV。1,2,3

[32] 刘威、德拉戈米尔·安盖洛夫、杜米特鲁·尔汉、克里斯蒂安·塞格迪和斯科特·里德。Ssd:单次多盒探测器。2016 年 3 月，ECCV。5

[33] Sharif M，Bhagavatula S，Bauer L 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 年 10 月。2

[34] A.穆罕默德 G. E .达尔和 g .辛顿。使用深度信念网络的声学建模。IEEE-ACM T 音频专家会议，20(1)，2012 年 1 月。1

[35] 等。雷德蒙，约瑟夫。你只看一次:统一的，实时的物体检测。2016 年 9 月，CVPR。5

[36] 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。TPAMI，39 岁，2015 年 6 月。5

[37] Ramprasaath R. Selvaraju 、 Michael Cogswell 、 Abhishek Das 、 Ramakrishna Vedantam、Devi Parikh 和 Dhruv Ba- tra。Grad-cam:通过基于梯度的定位来自深度网络的视觉解释。2017 年 10 月，ICCV。3

[38] Zisserman A Simonyan K .用于大规模图像识别的非常深的卷积网络。arXiv 预印本 arXiv:1409.1556，2014 年 9 月。5

[39] 哦，温雅尔斯和 QV·勒。用神经网络进行序列间学习。NeurIPS，2014 年 12 月。1

[40] 克里斯蒂安·塞格迪、文森特·万霍克、谢尔盖·约菲、黄邦贤·史伦斯和兹比格涅夫·沃伊纳。重新思考计算机视觉的概念架构。2015 年 12 月，CVPR。5

[41] 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 年 12 月。1,2

[42] 季米特里斯·齐普拉斯、什巴尼·桑图尔卡、洛根·恩斯特罗姆、亚历山大·特纳和亚历山大·马德瑞。稳健性可能与准确性不一致。ICLR，2019 年 5 月。1

[43] 魏秀生、、雷洋、，还有刘凌桥。RPC:大规模零售产品结账数据集。2019 年 1 月 CoRR。2

[44] 肖、、、贾登和刘。Meshadv:用于视觉识别的对抗性网格。2019 年 6 月，CVPR。5

[45] 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 6 月。2

[46] 对空间和频率的听觉注意激活了相似的大脑系统。《神经影像》，1999 年 11 月。2,3

[47] 张崇智，刘爱山，，，徐，，李天林。用神经元敏感性解释和改进对抗鲁棒性。IEEE 图像处理汇刊，2020 年。1

[48] 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。2019 年 5 月，ICLR。1,2,5

[49] 周、Aditya Khosla、Agata Lapedriza、Aude Oliva 和 Antonio Torralba。学习区别性本地化的深层特征。2016 年 6 月，CVPR。3

[50] 纣王、博维克、谢赫和西蒙切利。图像质量评估:从误差可见性到结构相似。IEEE 图像处理汇刊，13(4)，2004 年 4 月。7