Information

论文AdvRain：利用对抗性雨滴攻击基于摄像头的智能视觉系统

Amira Guesmi *、Muhammad Abdullah Hanif 和 Muhammad Shafique

eBrainLab，工程部，纽约大学阿布扎比分校（NYUAD），阿布扎比 129188，阿拉伯联合酋长国；
mh6117@nyu.edu (M.A.H.)；ms12713@nyu.edu (M.S.)* 通讯作者：ag9321@nyu.edu

arXiv:2303.01338v2 [cs.CV] 5 Oct 2023

摘要： 基于视觉的感知模块正越来越多地应用于各种场景，尤其是自动驾驶车辆和智能机器人。这些模块用于获取周围环境的信息并识别障碍物。因此，准确检测和分类对于始终做出恰当决策并采取适当、安全的行动至关重要。现有研究已表明，称为物理对抗攻击的“印刷式对抗攻击”能够成功误导目标检测器和图像分类器等感知模型。然而，这些物理攻击大多采用醒目、显眼的模式来生成扰动，因此人眼、实地测试或试驾都能识别或检测到这些扰动。本文提出一种基于摄像头的隐蔽对抗攻击（**AdvRain**），能够欺骗基于摄像头的感知系统，使其对同一类别的所有物体都产生误判。与需要访问底层计算硬件或图像存储器的基于掩膜的 FakeWeather 攻击不同，我们的攻击通过模拟自然天气条件（即雨滴）的影响来实现；这些雨滴可以印制在半透明贴纸上，并在对手准备发动攻击时贴在摄像头镜头外侧。需要指出的是，这类扰动在现实场景部署中仍不显眼，由于其与自然现象相关联，其存在不会引起注意。为此，我们设计了一种迭代过程，通过执行随机搜索来找出关键位置，确保所实施的变换能够对目标分类器构成对抗。我们的变换通过模糊捕获图像中预先设定的部分来实现，这些部分对应于雨滴覆盖的区域。仅使用20个雨滴，我们便使VGG19在ImageNet数据集上以及Resnet34在Caltech-101数据集上的平均模型准确率分别下降超过45%和40%。

关键词： 对抗机器学习；物理对抗攻击；安全；效率；扰动；物理攻击；深度神经网络；深度神经网络；分类；目标检测；摄像头；自主系统；机器人；自动驾驶车辆；Grad-CAM；随机搜索

check for updates

C引文：Guesmi, A.; Hanif, M.A.; SHafique, M. AdvRain：对抗性R雨滴攻击基于摄像头的S智能视觉系统。*Information*023, 2*1*, 0。https://doi.org/

A学术编辑：Alessandra Lumini

R 收稿：2023年7月31日
修订：2023年9月15日 接受：
2023年9月20日 出版：

版权： © 2023年 作者所有。许可方MDPI，瑞士巴塞尔。本文为开放获取文章，依据知识共享署名（CC BY）许可协议（https://creativecommons.org/licenses/by/ 4.0/）的条款和条件发布。

1. 引言

深度学习（DL）的革命性兴起对各个领域产生了深远影响，尤其是在自动驾驶领域[1]。汽车行业中的知名企业，如谷歌、奥迪、宝马和特斯拉，正积极将这项前沿技术与经济高效的摄像头相结合，以开发自动驾驶汽车（AVs）。这些自动驾驶汽车配备了先进的基于视觉的感知模块，使其能够在高压力情况下应对现实场景，做出明智决策，并执行安全且适当的操作。

因此，自动驾驶车辆的需求激增，推动自动驾驶汽车市场大幅增长。Strategic Market Research（SMR）预测，到2030年，自动驾驶汽车市场的估值将达到惊人的1969.7亿美元，复合年增长率（CAGR）高达25.7%（ACMS）。

深度学习驱动的视觉感知模块无疑加速了自动驾驶技术的发展，预示着汽车行业将迎来变革性时代。随着自动驾驶汽车日益普及，它们对道路安全、运输效率和整体用户体验的潜在影响，持续受到消费者、研究人员和投资者的广泛关注。

然而，尽管深度学习模型取得了显著进展，但它们并不能免受对抗攻击的影响，而这可能严重威胁其完整性和可靠性。对抗攻击是指通过引入精心设计的扰动来操纵深度学习分类器的输入，这些扰动由恶意行为者策略性地选择，旨在迫使分类器产生错误输出。攻击者可以利用此类漏洞，损害系统的安全性和完整性，并可能危及与系统交互人员的安全。例如，恶意行为者可以在停车标志上添加对抗噪声，导致自动驾驶汽车将其错误分类为限速标志 [2,3]。这种错误分类可能造成危险后果，包括事故和人员伤亡。值得注意的是，研究表明，对抗样本在现实条件下也有效 [4]。即使打印出来，专门设计成对抗样本的图像在不同光照条件和方向下仍能保持其对抗特性。

因此，了解并缓解这些对抗攻击，对于确保开发出安全、可信赖的智能系统至关重要。采取措施防御此类攻击，对于维护深度学习模型的可靠性和安全性不可或缺，尤其是在自动驾驶汽车、机器人技术以及其他与人互动的智能系统等关键应用中。

对抗攻击大致可分为两类：**数字攻击**和**物理攻击**，二者的区别在于各自独特的攻击形式 [3–5]。在**数字攻击**中，对手会在数字输入图像中引入难以察觉的扰动，专门用于欺骗特定的深度神经网络（DNN）模型。这些扰动经过精心优化，以确保人类观察者无法察觉。在生成过程中，攻击者会在预先设定的噪声预算内操作，确保扰动不超过一定幅度，从而保持其不可察觉性。相比之下，**物理攻击** 涉及设计能够转化到物理世界中的对抗扰动。随后，这些物理扰动会被部署到受害者深度神经网络模型所捕捉的场景中。与数字攻击不同，物理攻击不受噪声幅度限制，而主要受位置和可打印性因素约束，旨在生成能够有效打印并放置于现实环境中、且不会引起怀疑的扰动。

对抗攻击的首要目标，以及其在现实场景中的意义，在于不引人注意，看起来普通且合乎情理，而非明显带有敌意。以往许多针对图像分类开发对抗补丁的研究，主要着眼于最大化攻击效果并增强对抗噪声的强度。然而，这种方法往往会产生显眼的补丁，容易被人类观察者识别。另一类研究则试图提升所添加扰动的隐蔽性，使其自然融入在人类观察者看来合理可信的自然风格中。例如，将扰动伪装成彩色薄膜 [6],阴影 [7],或激光束 [8], 等。

在图1中，我们将 AdvRain 与现有的物理攻击进行了视觉对比。尽管图1 中的所有对抗样本都能成功攻击深度神经网络（DNN），但 AdvRain 能够生成带有自然模糊痕迹的对抗扰动（模拟与真实雨水相似的现象），这一特点使其脱颖而出；相比之下，AdvPatch 生成的模式较为醒目，而 FakeWeather 生成的模式则不切实际 [9]。这展示了 AdvRain 在生成能够与周围环境融为一体的对抗扰动方面的有效性，使其难以被人类观察者察觉。



图1。 AdvRain与现有物理世界攻击的对比：（**a**）AdvCF [7]，（**b**）Shadow [6]，（**c**）AdvPatch [10]，（**d**）AdvCam [8]，（**e**）FakeWeather [9]，和（**f**）AdvRain。

在本文中，我们提出了一种新颖的技术，旨在通过引入细微扰动来欺骗给定的深度神经网络模型，使属于特定类别的所有对象都被错误分类。我们的方法包括创建一种对抗性相机贴纸，并将其贴附在摄像头的镜头上。该贴纸采用精心设计的雨滴模式，当摄像头感知到这些模式时，会导致捕获图像被错误分类（见图2）。雨滴形成的模式在相机图像中呈现为水滴，使其不易被人类观察者察觉。

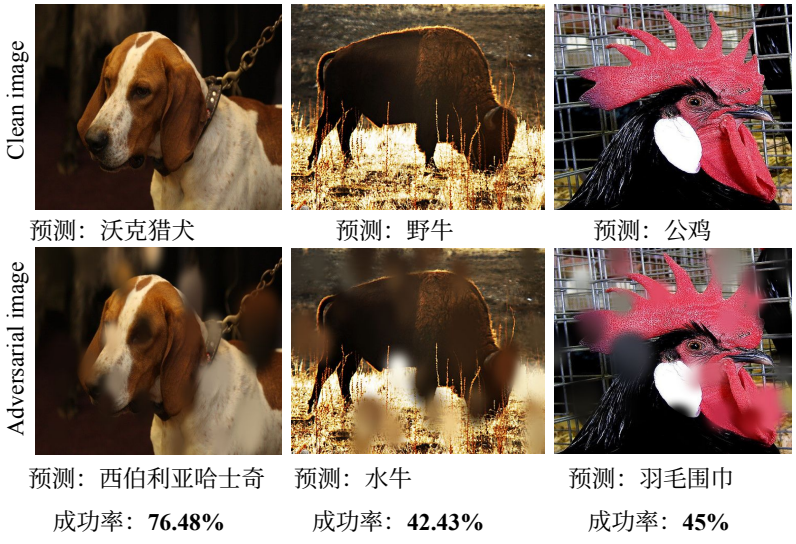


图 2. 仅使用 10 个雨滴时，**AdvRain** 生成的对抗样本及其对应的攻击成功率。

与通常作用于图像像素级粒度的传统对抗攻击不同，我们面临的主要挑战在于，使用这种摄像头贴纸模型所能引入的可行扰动空间有限。摄像头的物理光学特性对我们形成了约束，导致产生的主要扰动是模糊圆点。这些模糊圆点缺乏传统对抗攻击中常见的高频模式。我们的研究旨在应对物理对抗攻击带来的独特挑战，其目标是利用摄像头光学特性的局限性和特点，实现

隐蔽性。通过开发这项技术，我们有助于加深对物理对抗攻击的理解，并为在现实场景中开发智能系统的鲁棒防御机制提供见解。

本文提出了一种新颖的技术，利用Grad-CAM引导的随机搜索优化方法来构造对抗扰动 [11]。通过使用Grad-CAM，我们能够识别出可能带来更高攻击成功率的关键位置。这使我们能够在随机搜索过程中专注于探索较小的位置集合，最终确定能够确保AdvRain攻击最高有效性的最佳雨滴位置。我们的目标是创建可引入摄像头与物体之间视觉路径中的扰动，同时保持物体本身不变（如图3所示）。通过利用Grad-CAM，我们可以识别图像中显著影响深度学习模型决策过程的关键位置。该信息为我们的搜索优化方法提供指导，帮助我们在这些具有影响力的位置生成对抗扰动。因此，这些扰动能够有效欺骗目标分类器，而无需直接修改物体。我们方法的优势在于能够巧妙地操纵摄像头捕获的视觉信息。构造的对抗扰动能够无缝融入场景中，对人类观察者保持不易察觉。同时，它们会对模型的决策产生显著影响，导致物体被错误分类。

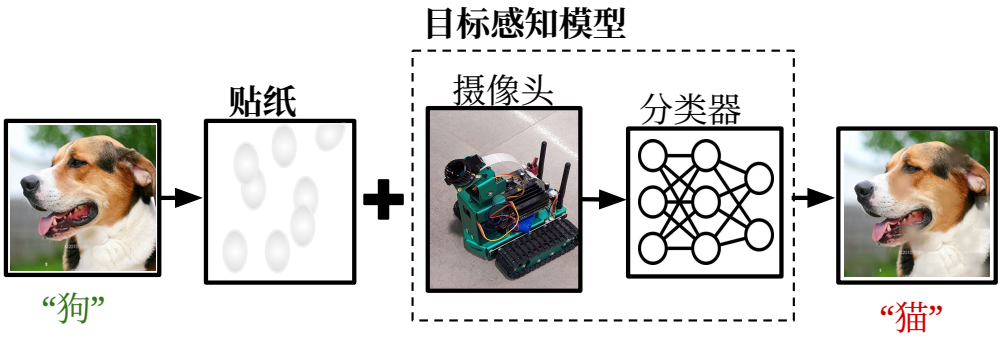


图 3. 攻击威胁模型：生成的模式被印在半透明贴纸上，并贴在摄像头的镜头上。因此，任何捕获图像都会包含对抗点，从而形成一种不显眼、外观自然的对抗图像，既能欺骗目标模型，也能骗过人眼。

我们新颖贡献的概述如图 4 所示。**总而言之，**本文的贡献**如下：**

- 我们提出了一种新颖技术，利用 Grad-CAM 引导的随机搜索优化方法来构造对抗扰动。这些扰动被引入摄像头与物体之间的视觉路径中，同时不改变物体本身的外观。

- 对抗扰动被设计成类似一种自然现象，具体而言就是雨滴，从而形成不显眼的模式。这些模式被印在半透明贴纸上，并贴附在摄像头镜头上，因此难以被检测到。

- 所提出的对抗贴纸会对属于同一类别的所有图像施加相同的扰动，从而构成针对目标类别的通用攻击。
- 我们的实验表明，AdvRain攻击具有强大的效果：与使用FakeWeather [9]时在相同结构相似性指数（结构相似性指数）下分别下降37%和40%相比，AdvRain攻击使ImageNet上的VGG-19准确率显著下降超过61%，并使Caltech-101上的Resnet34准确率下降57%。

- 我们研究了对图像特定部分进行模糊处理并引入低频模式对模型可解释性的影响。这为深入了解深度学习模型在所提出的对抗攻击下的行为提供了宝贵见解。

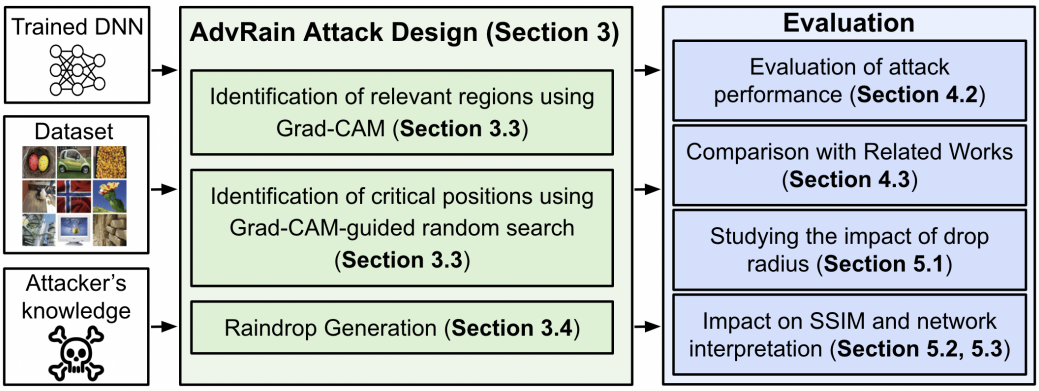


图4. 我们提出的新颖贡献概述。

论文结构：本文其余部分的结构如下。第 2节全面概述相关研究，为我们提出的方法奠定背景。第3节介绍生成对抗贴纸的方法，并概述 AdvRain攻击的关键组成部分。接着，第 4节对所提出的攻击进行详细评估。我们将 AdvRain 与 FakeWeather 和自然降雨进行比较，考察攻击的效力以及生成的对抗样本在视觉上的相似性。此外，我们还会考察攻击的效力以及生成的对抗样本在视觉上的相似性。第 5节深入讨论实验结果及其意义。此外，我们还探讨丢弃半径的影响，评估扰动对结构相似性指数（SSIM）的作用，并研究其对网络解释的影响。最后，第 6节简要总结本研究并给出结论，强调 AdvRain 作为一种隐蔽且有效的基于摄像头的视觉系统对抗攻击方法的重要性。

2. 背景与相关工作

2.1. 基于摄像头的视觉系统

环境感知已成为一项至关重要的应用，在工业界和研究界都引发了大量投入。开发强大的深度学习（DL）解决方案，尤其受到自主机器人和智能交通系统等应用领域的关注。设计可靠的识别系统，是实现稳健环境感知面临的主要挑战之一。在此背景下，汽车摄像头发挥着关键作用，并与目标检测器、图像分类器等相关感知模型共同构成基于视觉的感知模块的基础。这些模块在收集环境相关关键信息方面发挥着重要作用，帮助自动驾驶车辆（AVs）为安全驾驶做出关键决策。

实现高性能环境感知的一大难题，是构建可靠的识别系统。这类系统的准确率和可靠性对于自动驾驶汽车的安全运行，以及确保其成功融入现实场景至关重要。然而，已有研究表明，这些识别系统容易受到对抗攻击，进而损害其完整性，并给自动驾驶汽车及其乘客带来潜在风险。鉴于这些挑战，当务之急是解决基于深度学习的环境感知系统在面对对抗攻击时存在的脆弱性。研究和开发工作必须着力构建鲁棒防御机制，增强这些系统抵御潜在威胁的能力，从而推动未来安全、可靠地部署自动驾驶汽车和智能交通系统。

对抗攻击可分为数字攻击和物理攻击。

2.2. 数字对抗攻击

在数字攻击场景中，攻击者可以灵活地操纵受害者深度神经网络（深度神经网络）的输入图像，精细到像素级。这类攻击假设攻击者能够访问深度神经网络的输入系统，例如摄像头或其他输入数据提供方式。对抗样本的概念最早由 Szegedy 等人提出：向输入中注入少量不可察觉噪声，使模型的预测偏向错误类别。 [12]。

随着时间的推移，人们开发了各种生成对抗样本的算法，推动了数字攻击的发展。一些值得关注的数字攻击包括 Carlini 和 Wagner 攻击（CW） [5], 快速梯度符号法（FGSM） [3], 基本迭代法（BIM） [13], 局部搜索攻击 [14], 以及 HopSkipJump 攻击（HSJ） [15],等。然而，在现实的威胁模型中，我们可以假设攻击者能够控制系统的外部环境或外部物体，而无法访问系统的内部传感器和数据管道。这种情形代表物理攻击，攻击者会构造对抗扰动，将其引入物理世界，从而欺骗深度神经网络。

在下一节中，我们将探讨一些针对图像分类的先进物理攻击。这些攻击旨在利用基于摄像头的视觉系统中的漏洞，并展示在现实场景中进行对抗操纵的潜在风险。了解并缓解此类物理攻击，对于构建稳健、安全的智能系统至关重要，尤其是在自动驾驶汽车和监控系统等安全关键型应用中。

2.3. 物理对抗攻击

物理攻击涉及在物理空间中添加扰动，以欺骗目标模型。构造物理扰动的过程通常包括两个主要步骤。首先，对手在数字空间中生成对抗扰动。然后，目标是在物理空间中复现该扰动，使其能够被摄像头和雷达等传感器感知，从而有效欺骗目标模型。

在不同位置添加对抗扰动的现有方法可分为四大类：直接修改目标物体的攻击：在这种方法中，攻击者直接修改目标物体以引入对抗扰动。例如，有研究提出了对抗性服装，其服装图案经过设计，可使目标检测器产生混淆 [16–18]。Hu 等人 [16] 利用预训练生成对抗网络模型生成逼真/自然的图像，这些图像可以印刷在T恤上，并能够隐藏穿着它们的人。Guesmi 等人 [17] 提出以基于语义约束的方法替代生成对抗网络：在损失函数中加入相似度项，从而直接操纵图像的像素。这样可以更加灵活地纳入多种变换。

通过修改背景进行攻击：对抗补丁是一类特殊的对抗扰动，旨在操控图像中的局部区域，从而误导分类模型。这类攻击利用模型对局部改动的敏感性，意图引入细微变化，对模型的预测结果产生重大影响。攻击者利用模型对特定图像特征或模式的依赖，制作对抗补丁，诱使模型错误分类图像，或将其感知为与实际内容不同的图像。适用于现实场景的实用攻击示例是 AdvPatch [10]。这种攻击会创建可应用于任意位置的通用补丁。此外，该攻击还采用变换期望（EOT） [19] ，以增强对抗补丁的鲁棒性。AdvCam 技术 [8]提出了一种创新的图像扰动方法，在风格空间中发挥作用。该方法结合神经风格迁移与对抗攻击的原理，生成能够与图像视觉风格自然融合的对抗扰动。例如，AdvCam 可以在停车标志上引入锈斑般的扰动，使其在周围环境中看起来自然、不显眼。

修改摄像头：这种方法通过修改摄像头本身来引入对抗扰动。一种方法是利用滚动快门效应，通过操控图像不同部分的拍摄时序来产生扰动 [20,21]。

改变摄像头与物体之间的介质：此类别包括通过改变摄像头与物体之间的介质实施的攻击。例如，基于光的攻击利用外部光源产生扰动，这些扰动被摄像头捕捉后会误导目标模型 [6,22]。物理对抗物体装置（OPAD） [22]采用结构化照明方法来改变目标物体的外观。该攻击系统由一台经济实惠的投影仪、一台摄像头和一台计算机构成，可在一次拍摄中操纵现实场景中的物体。Zhong 等人 [6] 利用阴影这一自然现象来创建对抗样本。此方法旨在兼顾数字环境和物理环境中的实用性。与传统的基于梯度的优化算法不同，它采用基于粒子群优化（PSO） [23]的优化策略。研究人员在模拟环境和现实场景中进行了广泛评估，结果揭示了阴影作为一种可行攻击途径所带来的潜在威胁。然而，需要注意的是，这些技术在光照条件变化时可能会降低效果。Zhang 等人 [7], 提出的对抗性色彩薄膜（AdvCF）方法，通过在相机镜头与关注对象之间放置彩色薄膜，实现有效的物理对抗攻击。AdvCF 旨在通过调整彩色薄膜的物理特性，在不改变目标物体外观的情况下，生成在各种光照条件下（包括白天和夜间）均能保持有效的对抗扰动。

FakeWeather [9] 攻击旨在模拟雨、雪和冰雹等各种天气条件对摄像头镜头造成的影响。该攻击通过在捕获图像中添加模拟恶劣天气所致失真的扰动，试图欺骗计算机视觉系统，尤其是自动驾驶汽车和其他基于图像的应用所使用的系统。在 FakeWeather 攻击中，对手设计特定的掩模或模式，以模拟不同天气条件产生的视觉伪影。然后将这些掩模应用于摄像头拍摄的图像中，引入可能误导图像识别模型的失真。其目标是让图像看起来像是在恶劣天气下拍摄的，从而可能导致模型作出错误的预测或分类。FakeWeather 攻击的一个局限是，生成的噪声或扰动可能呈现不自然的像素化模式，更稳健的图像识别系统可能会检测到这些模式。此外，该攻击的有效性可能仅限于特定场景和图像尺寸，因为最初的测试对象是 CIFAR-10 数据集中的小图像，每张图像含有 32×32个像素。

此外，研究人员还探索了利用自然风格生成对抗扰动的技术，以确保其对人类观察者而言隐蔽且可信。这些方法旨在使扰动看起来像场景中的自然现象。

3. 提出的方法

本节将概述我们设计对抗性相机贴纸的方法，这是一种为基于摄像头的视觉系统生成不显眼的对抗扰动的新方法。

3.1. 物理相机贴纸攻击的威胁模型

传统上，生成对抗样本的问题被表述为一个约束优化问题（公式（1）），给定原始输入图像 x 和目标分类模型 $f(.)$：

$$\min_{\delta} \|\delta\|_p s.t. f(x+\delta) \neq f(x) \tag{1}$$

目标是找到一个最小化的不显眼通用扰动， δ，使其添加到目标输入域 D中的任意输入后，能够导致底层基于深度神经网络的模型 $f(.)$ 发生误分类。请注意，由于基于深度神经网络的模型 $f(.)$ 是一个非凸机器学习模型，即深度神经网络，因此无法为该优化问题找到闭式解。因此，公式（1）被表述如下，以便使用经验近似技术对问题进行数值求解：

$$\arg\max_{\delta} \sum_{x \in \mathcal{D}} l(f(x+\delta), f(x)) \tag{2}$$

其中 l 是基于深度神经网络的模型损失函数， $\mathcal{D} \subset D$ 是攻击者的分类器训练数据集。要解决这个问题，可以使用现有的优化技术（例如 Adam [24]）。在每次训练迭代中，优化器都会更新对抗噪声 δ。

与针对图像像素级粒度进行攻击的方法不同，我们提出的威胁模型——物理相机贴纸攻击——面临着一项重大挑战：可引入的可行扰动空间有限。相机的光学特性施加了约束，导致只能生成模糊圆点作为扰动，无法形成传统对抗攻击中常见的高频模式。为应对这一挑战，我们提出了一种创新方法，通过近似在贴纸上放置类似雨滴的小圆点所产生的效果，来设计对抗扰动。当在相机镜头上放置一小滴水时，捕获图像上会形成一个半透明斑块。这个斑块代表相机镜头光学特性所导致的低频扰动。对抗样本的构造如下：

$$x_{adv} = G(x, p, n, r) \tag{3}$$

其中，G 是雨滴生成器，p 表示雨滴的位置，n 表示雨滴数量，r 表示雨滴半径。具体而言，考虑一幅二维图像x ，其中$x(i,j)$ 表示位置(i,j) 处的像素。我们的目标是为n 个雨滴找到最佳候选位置$p(i,j)$ ，使模型错误地分类生成的对抗样本。

$$p(i,j)? \; s.t. \; f(G(x, p(i,j), n, r)) \neq f(x) \tag{4}$$

3.2. 所提方法概述

图5 概述了我们提出的攻击方法。其主要目标是生成具有特定模式的对抗扰动，使其呈现自然天气事件的影响，尤其是大气条件（即雨）导致摄像头镜头上的雨滴效果。为此，我们构造了能够模拟摄像头镜头上雨滴外观的模式。

AdvRain采用一种独特的方法来创建对抗扰动，以模拟自然天气条件（尤其是雨）对摄像头镜头的影响。该方法可分为几个关键步骤：首先生成一种模式，使其呈现雨滴落在摄像头镜头上的外观。这些雨滴属于低频扰动，会使捕获图像的部分区域变模糊。其目标是让这些扰动不易察觉，并与真实雨滴的视觉特征无缝融合。

为了优化雨滴模式以达到最佳效果，研究采用了由 Grad-CAM（梯度加权类激活映射）引导的随机搜索。Grad-CAM 有助于识别图像中影响模型决策的重要区域。从 Grad-CAM 标出的位置开始优化过程，可以缩小搜索范围，探索更少的潜在扰动模式。选定模式后，便将其作为模拟雨滴应用于输入图像。雨滴覆盖区域内的每个像素都会经过变换，具体来说就是进行高斯模糊。这种模糊效果模拟了雨滴在相机镜头上造成的失真。除了模糊处理之外，AdvRain 还加入了鱼眼效果，以模拟雨滴可能给相机视野带来的弯曲变形。雨滴的飞溅

则通过碰撞检测进行模拟，确保雨滴不会以不符合物理规律的方式相互重叠。

为了呈现自然的雨滴外观，AdvRain为每个雨滴组合使用圆形和椭圆形。最终形状模拟水滴表面，对该形状应用模糊效果则进一步增强了真实感。生成的扰动形似一组雨滴，旨在印制在半透明贴纸上，并将贴纸粘贴到摄像头镜头上。简而言之，AdvRain结合随机搜索优化、Grad-CAM引导以及精心设计的雨滴图案变换，生成视觉上不易察觉的对抗扰动。这些扰动施加到摄像头镜头上后，会引入低频失真，进而诱使深度神经网络对图像进行错误分类。

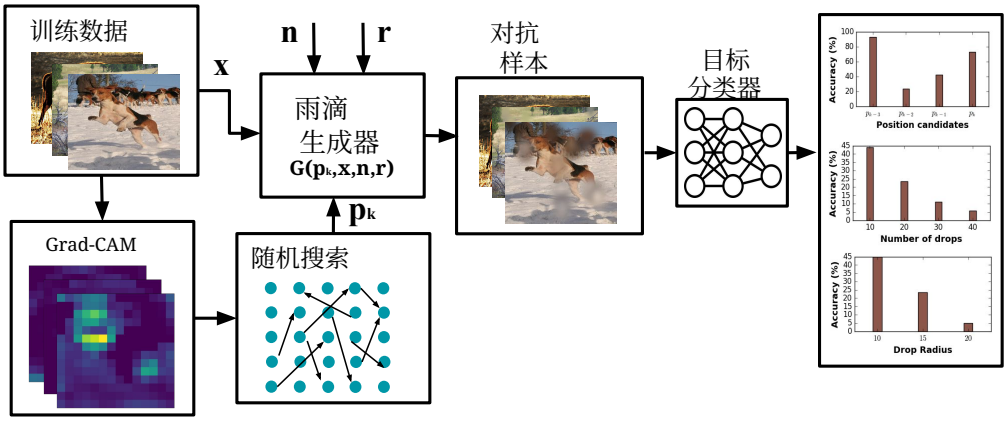


图5. 所提方法概述；**训练阶段：** 我们采用基于随机搜索的优化方法，并利用Grad-CAM提高优化过程的效率。借助Grad-CAM，我们能够识别可能带来更高攻击成功率的关键位置。这使我们能够在随机搜索期间专注于探索较小范围的位置，并最终确定最合适的雨滴位置，以确保AdvRain攻击具有最高效果。雨滴生成器的输入包括雨滴数量n、雨滴半径r、第k 次迭代中n 个雨滴的位置p_k，以及输入图像x。生成器的输出是添加到输入图像中的模拟雨滴。随后，将该图像输入分类器，以监测其分类准确率。

3.3. 识别关键位置

使用 Grad-CAM 引导的随机搜索方法来识别关键位置，如算法 1所示。该优化过程的主要目标是选择最佳候选位置，以引入能使分类准确率最低（即攻击成功率最高）的雨滴扰动。

我们提出的搜索方法如算法 1所示：首先初始化迭代次数 T 和候选位置数量 N。随后将扰动初始化为空集。对于从 1 到 t 的每次迭代 T，我们在图像中随机生成 N个候选位置。对于每个候选位置，我们应用高斯模糊和鱼眼效果来生成雨滴图案。我们使用候选位置模拟雨滴扰动对图像的影响，并使用 Grad-CAM 识别模型作出决策时最关键的区域。然后，选择能带来最高攻击成功率（即最低分类准确率）的候选位置。之后，我们使用所选候选位置对应的雨滴图案更新扰动。最后返回最终扰动，该扰动代表了用于引入对抗扰动、使分类准确率最低的最优雨滴图案。

这种由 Grad-CAM 引导、基于随机搜索的优化方法，能够高效地探索引入雨滴扰动的不同位置，并聚焦于显著影响模型决策的区域。通过选择使

攻击成功率最高的位置，我们确保对抗性相机贴纸能够有效欺骗目标深度学习模型，同时在捕获的图像中显得不易察觉。

算法1 Grad-CAM引导的随机搜索方法。1：初始化迭代次数 T 和候选位置数量 N2：将扰动初始化为空集。3：使用 Grad-CAM 识别模型做出决策时最关键的区域。4：**对于**$t = 1$：T **执行** 5：在已识别的关键区域内随机生成 N 个候选位置。6：**对于** $p = 1$： N **执行** 7：应用高斯模糊和鱼眼效果来生成雨滴图案。8：使用候选位置模拟雨滴扰动对图像的影响。9：选择能产生最高攻击成功率的候选位置。10：使用所选候选位置生成的雨滴图案更新扰动。11：**结束循环**12：**结束循环**

3.4. 雨滴生成器

雨滴生成器会对雨滴覆盖区域内的每个像素应用高斯模糊变换。该变换函数定义如下：设$I(i, j)$ 为输入图像，其中(i, j) 为像素坐标，而$I'(i,j)$ 为对雨滴应用高斯模糊变换后得到的图像。

高斯模糊变换表示为

$$I'(i,j) = \frac{1}{\sum_{a,b} W(n,m)} \sum_{a,b} I(i+a, j+b)\dot{W}(a,b) \tag{5}$$

其中$W(a,b)$ 是高斯核，定义如下：

$$W(a,b) = \frac{1}{2\pi\sigma^2} \exp - \frac{a^2+b^2}{2\sigma^2} \tag{6}$$

其中，a 和b 是雨滴区域内的像素偏移量，而 σ 是高斯核的标准差，用于控制模糊强度。

应用高斯模糊变换可确保雨滴图案在图像中呈现自然、无缝的效果，模拟大气条件导致雨滴附着在相机镜头上的效果。这种方法使对抗扰动不易察觉，令其难以被人类观察者发现，同时仍能有效欺骗目标深度学习模型。

为了创建逼真的雨滴表面效果，我们的方法结合了高斯模糊和鱼眼效果。通过碰撞检测来模拟雨滴及其飞溅过程，即检查一个雨滴的中心是否与另一个雨滴重叠。如果发生重叠，则将这些雨滴合并；否则不采取任何操作。为了构造雨滴的形状，我们结合使用一个圆形和一个椭圆形（图 6a）。通过调整这些形状的大小和方向，我们可以形成不同的间隙，并模拟水滴的外观（图 6b）。水滴表面的最终效果通过应用高斯模糊实现（图 6c）。生成的扰动代表雨滴图案，随后被印刷在半透明贴纸上。该贴纸被仔细地贴附在摄像头镜头上，使扰动能够引入视觉路径，同时不会妨碍图像捕获过程。

鱼眼效果、碰撞检测和雨滴形状设计与高斯模糊相结合，在

相机镜头上形成逼真且不易察觉的雨滴图案。这些图案在呈现出自然天气现象的同时，能够有效误导目标深度学习模型。这种方法展示了我们所提出的 AdvRain攻击的有效性，凸显了其对基于摄像头的视觉系统的潜在影响，以及在现实应用中采用鲁棒防御机制应对物理对抗攻击的重要性。图 6 展示了雨滴图案创建的概念和各个阶段。

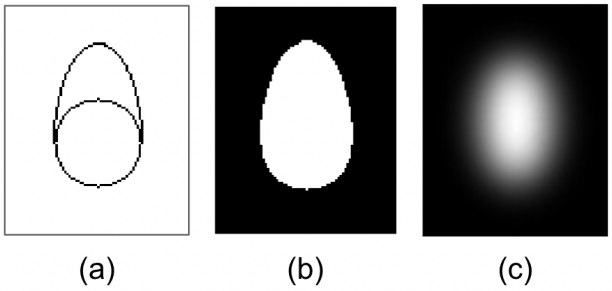


图6. 雨滴生成过程：**(a)** 使用一个圆形和一个椭圆形形成雨滴形状。**(b)** 最终形状。**(c)** 然后，我们通过添加模糊效果来模拟水滴表面的效果。

4. 实验结果

4.1. 实验设置

在本研究中，我们进行了实验，以评估不同大小的雨滴对 VGG-19 [25] 和 Resnet34 [26]，分类准确率的影响，这两个模型作为受害分类器。两个模型的输入图像分辨率均设置为 224×224 像素。

在评估中，我们使用了两个知名数据集中的图像：ImageNet 和 Caltech-101 [27]。ImageNet [28] 是一个大型视觉数据集，包含超过 1400 万张标注图像，分为 20,000 多个类别。相比之下，Caltech-101 包含描绘 101 个不同类别物体的图像，总计约 9000 张。每个类别包含数量不等的图像，从 40 张到 800 张不等；图像尺寸各异，典型边长为 200–300 像素。在整个实验过程中，我们通过评估模型处理未经修改的干净图像时的表现，确定了模型的基准分类准确率。通过模拟不同大小的雨滴并研究其对模型分类准确率的影响，我们深入了解了模型在对抗扰动下的鲁棒性。将结果与基准分类准确率进行比较，使我们能够定量评估 AdvRain攻击的有效性，以及它在对图像引入极小视觉变化的同时欺骗模型的能力。实验设置总结见图 7。

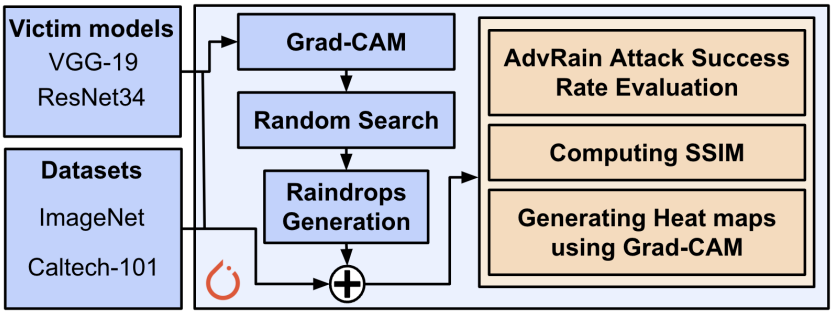


图7. 开展实验的实验设置和工具流程。

4.2. 攻击性能评估

在评估中，我们采用分类准确率作为主要指标来评估 s
我们 AdvRain攻击的有效性。为展示该攻击的影响，我们首先 t

为 10 个不同类别生成了对抗扰动，然后测量了不同雨滴数量下模型的准确率。

实验结果表明，我们的攻击成功降低了模型的分类准确率。具体而言，仅施加10个雨滴，ImageNet数据集上的准确率就显著下降了约30%，Caltech-101数据集上的准确率则下降了约20%。随着雨滴数量增加到20个，准确率下降得更加明显：ImageNet数据集上的降幅超过40%，Caltech-101数据集上的降幅超过30%（如表1所示）。

表1. 雨滴半径对模型准确率的影响：在ImageNet上训练的VGG-19和在Caltech-101上训练的Resnet34。雨滴半径越大，分类准确率下降越多。

雨滴数量	VGG-19	Resnet34
0	100%	100%
10	72%	81%
20	59%	69%
30	46%	50%
40	39%	43%

这些发现表明，AdvRain攻击能够将不同类别的物体误分类，同时保持不易察觉、类似自然天气事件的模式。仅需少量雨滴便能大幅降低模型准确率，这凸显了基于摄像头的视觉系统易受物理对抗攻击影响，也说明了在现实应用中制定鲁棒防御策略以应对此类威胁的必要性。

在表 2中，我们展示了八个不同类别的分类准确率：沃克猎犬、公鸡、蛇、蜘蛛、鱼、鹦鹉、美洲火烈鸟和野牛。在这些类别中，沃克猎犬类别在遭受我们的 AdvRain攻击后准确率下降幅度最大。这可以归因于 ImageNet数据集共包含 120 个犬种类别，其中许多犬种的特征和外观相似。因此，沃克猎犬图像中加入的雨滴扰动，由于其与其他犬种具有相似的视觉特征，导致分类准确率大幅下降。

表 2. ImageNet 上 VGG-19 对不同数量雨滴（丢弃半径 = 10）的各类别平均模型准确率。

数量 雨滴	Walker 猎犬	Cock	蛇	蜘蛛	Fish	鹦鹉	美国 火烈鸟	野牛
10	43%	70%	79%	68%	66%	86%	78%	68%
20	22%	56%	63%	59%	58%	82%	69%	57%
30	12%	48%	49%	51%	35%	68%	60%	33%
40	4%	40%	37%	43%	29%	60%	52%	31%

然而，对于具有更易区分且更独特特征的类别，例如“鹦鹉”类别和“美洲火烈鸟”，我们观察到准确率下降幅度相对较小。由于这些物体具有鲜明的视觉属性，欺骗它们更具挑战性，这使得雨滴扰动更难显著误导模型。

这些按类别划分的结果为我们深入了解 AdvRain攻击对不同物体类别的影响提供了宝贵信息。各类别准确率下降程度的差异凸显了这样一点：在评估对抗攻击的有效性时，必须考虑物体的视觉特征和复杂程度。这些知识对于理解该攻击的局限性和优势至关重要，并可指导定向防御策略的制定

，以提升基于摄像头的视觉系统面对物理对抗扰动时的鲁棒性。

4.3. 与最先进方法的比较：*AdvRain*与*FakeWeather*

在本节中，我们将AdvRain攻击与 [9]中提出的FakeWeather攻击进行比较。FakeWeather攻击通过创建三个模拟这些天气条件外观的掩模，旨在模拟雨、雪和冰雹对摄像头镜头的影响。然而，FakeWeather攻击的一个主要局限在于，生成噪声时使用了不真实且像素化的模式。此外，所添加的扰动覆盖了图像的很大一部分（图8），因此可能更加显眼，也更容易被人类观察者检测到。

此外，FakeWeather攻击的有效性仅在尺寸为 32×32 像素的CIFAR-10 数据集小图像上进行了评估。相比之下，我们提出的AdvRain攻击生成了更加逼真的扰动，能够密切模拟雨滴对相机镜头的影响。我们的攻击所构造的雨滴图案并不显眼，能够无缝融入图像中，因此在视觉上很难被检测出来。

此外，我们还将 AdvRain 的评估扩展到 ImageNet 和 Caltech-101 数据集中的更大图像，其尺寸为224×224 像素。这项范围更广的评估展现了我们的方法攻击高分辨率图像时的通用性和鲁棒性，使其更适用于现实场景和基于摄像头的视觉系统，例如自动驾驶汽车和智能机器人所使用的系统。

如表3所示，我们提出的攻击方法在攻击基于 ImageNet 和 Caltech-101 训练的模型时优于FakeWeather。例如，AdvRain 的攻击成功率达到 65%；然而，在结构相似性指数相同的情况下，FakeWeather 的攻击成功率仅为 37%。

总之，我们的 AdvRain攻击通过生成更加逼真且不易察觉的雨滴扰动，优于FakeWeather攻击。对雨效的模拟更加出色，并且已证明该方法对较大图像同样有效，这增强了我们所提出攻击的适用性和影响力，使其能够挑战基于摄像头的视觉系统中深度学习模型的完整性。这一比较凸显了开发不仅强大，而且在现实应用中逼真且不易察觉的攻击的重要性。

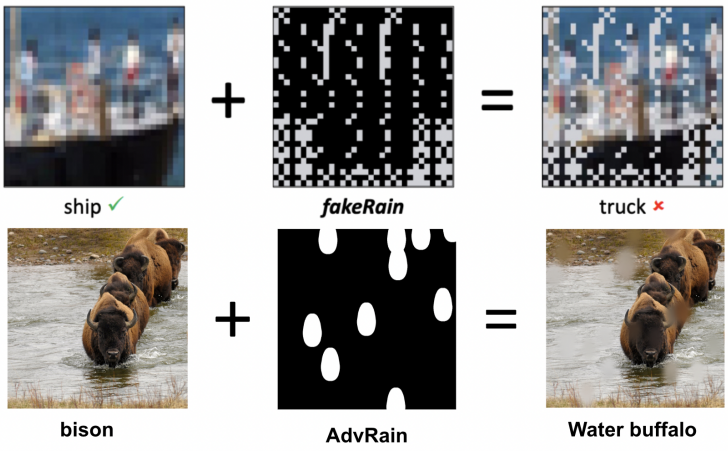


图8。 AdvRain 与 FakeWeather 攻击的对比。FakeWeather 攻击试图通过设计一种掩码来模拟雨的效果，该掩码通过改变像素值，伪造天气条件对摄像头镜头造成的影响。然而，生成的掩码呈现出不真实且像素化的模式。相比之下，AdvRain 通过生成更逼真的雨滴模拟效果，其形状更接近真实雨滴。

表3。 攻击成功率：AdvRain 与 FakeWeather [9]。

方法	ImageNet	Caltech-101
FakeWeather [9]	37%	40%
AdvRain（我们的方法）	65%	62%

4.4. AdvRain 与自然降雨的比较

在我们的评估中，我们比较了 AdvRain攻击与自然降雨。为模拟自然降雨，我们在图像中随机放置雨滴，这种方法会导致受害模型的准确率出现不同程度的下降。如图9所示，随机放置雨滴的位置不同，准确率下降幅度也不同。

这一对比凸显了我们的 AdvRain攻击 相较于自然降雨在显著降低分类准确率方面的有效性和优越性。对于左侧完全随机的雨滴组合，准确率仅下降了 3%，这表明随机放置雨滴对欺骗受害模型的影响有限。然而，使用我们的 AdvRain攻击 精心选择雨滴位置后，准确率下降了 60%，效果显著。

这些结果表明，与自然降雨的影响相比，我们的 AdvRain攻击 所构造的对抗扰动更能有效欺骗深度学习模型。AdvRain 对雨滴的策略性布局导致了更高程度的错误分类，这说明我们的攻击成功利用了雨滴图案的特性，从而构造出更有效的对抗扰动。

这一对比强调了 AdvRain 的独特优势：它能够生成不易察觉却极具效力的对抗扰动，有效误导目标模型，其效果超过了自然降雨相关的随机雨滴布局。通过精心选择雨滴位置实现准确率大幅下降的能力，凸显了 AdvRain 在现实世界中的潜在影响，尤其是在涉及基于摄像头的视觉系统的场景中，例如自动驾驶汽车和智能机器人所使用的系统。



图9. AdvRain 对分类准确率的影响，与自然降雨进行比较。*左*： 自然降雨（雨滴随机分布）导致准确率下降 3%。*中*： 非最优位置的组合导致准确率下降 23%。*右*： 精心挑选的位置组合（ AdvRain 的输出）导致准确率下降 60%。

5. 讨论

5.1. 在更多深度神经网络模型上的评估

AdvRain攻击已在其他卷积神经网络（CNN）模型上进行了测试，具体包括 Resnet-50 和 Inception v4。如表 4 所示，当将 AdvRain 应用于 Resnet-50 时，约 59% 的情况下会成功使目标模型发生错误分类。换言之，将经过 AdvRain 扰动的图像输入 Resnet-50 后，

与未受扰动的干净图像相比，模型的准确率下降了 59%。这证明了 AdvRain攻击对 Resnet-50 的有效性。此外，将 AdvRain 应用于 Inception v4 模型时，其成功率约为 57%。这表明该攻击应用于 Inception v4 时，会使其准确率下降 57%。这凸显了 AdvRain攻击在不同卷积神经网络模型（具体包括 Resnet-50 和 Inception v4）上的表现，也展示了该攻击的有效性，即 AdvRain 应用于图像时误导这些模型的能力。

表4. AdvRain 在更多卷积神经网络模型上的性能。

模型	**VGG-19**	**Resnet-50**	**Inception v4**
准确率	37%	41%	43%

5.2. 丢弃半径的影响

在我们的评估中，我们系统地研究了改变雨滴半径（即模糊区域）对 AdvRain攻击效果的影响。我们测试了三个在 ImageNet 上训练的不同模型（即 VGG-19、Resnet-50 和 Inception v4），以及另外三个在 Caltech-101 上训练的模型（即 ResNet-34、Resnet-50 和 Inception v4）。我们发现，增大半径会显著提高攻击成功率，进一步损害受害模型的准确率。例如，当我们将 ImageNet数据集的雨滴半径从 10 像素增大到 15 像素时，受害模型的准确率又下降了 25%（如表 5 和 6所示）。这一发现表明，雨滴大小对决定对抗扰动的强度至关重要。

表 5. 雨滴半径对 ImageNet 上模型准确率的影响：雨滴半径越大，分类准确率下降越多。

雨滴数量	**0**	**10**		**20**	
半径	-	10	15	10	15
VGG-19	100%	76%	57%	59%	37%

雨滴的半径越大，扰动覆盖的区域就越广，对欺骗深度学习模型的影响也越大。雨滴尺寸增大时攻击成功率随之提高，表明我们的AdvRain攻击在雨滴图案覆盖图像较大部分的情况下尤其有效，能够模拟现实世界中雨滴落在相机镜头上的效果。

表6. 雨滴半径对Caltech-101模型准确率的影响：雨滴半径越大，分类准确率下降得越多。

雨滴数量	**0**	**10**		**20**	
半径	-	10	15	10	15
ResNet-34	100%	79%	70%	69%	55%

5.3. 对结构相似性指数的影响

在本节中，我们使用结构相似性指数度量（结构相似性指数）评估对抗扰动对捕获图像的影响 [29]。鉴于我们的 AdvRain攻击旨在通过模糊图像的特定区域来引入雨滴扰动，与传统对抗攻击相比，我们观察到其对结构相似性指数的影响相对较小（表7）。例如，使用 10 个雨滴时，对抗样例与干净图像之间的结构相似性指数为 0.89。这表明

引入雨滴扰动对对抗图像与干净图像之间的结构相似性影响有限。

相对较高的结构相似性指数（SSIM）值表明，我们的 AdvRain攻击实现了其目标：创建不易察觉、且与自然天气现象极为相似的扰动。尽管雨滴扰动会显著降低分类准确率，但它们仍与原始图像保持高度的结构相似性，因此人类观察者很难通过视觉察觉。这一点使我们的 AdvRain攻击有别于传统对抗攻击；后者往往会在图像中引入明显且具有干扰性的模式，从而导致较低的结构相似性指数值。扰动不易察觉的特性，使 AdvRain 能够在不显著改变捕获图像视觉外观的情况下，隐蔽而有效地欺骗基于摄像头的视觉系统。

表7. 对抗样本与原始图像的结构相似性指数（SSIM）。

雨滴数量	**10**	**20**	**30**	**40**
SSIM	0.89	0.78	0.73	0.69

5.4. 对网络解释的影响

对抗补丁已被证明能够有效导致错误分类，但其使用也引发了对其可检测性的担忧。Grad-CAM 等标准网络解释方法可以凸显对抗补丁的存在，从而有效暴露对手的身份 [11]。Grad-CAM 是最知名的网络解释算法之一，已被证明在各种场景下优于其他最先进的解释方法。

为了评估传统对抗补丁与我们的 AdvRain攻击的 Grad-CAM可视化结果，我们采用了一个经过 ImageNet 预训练的 VGG-19 分类器。评估过程比较了向模型中添加低频模式（传统对抗补丁）与高频模式（AdvRain）对模型解释结果的影响。

基于补丁的攻击会将模型的注意力从目标转移到补丁所在位置，因此可能容易被检测到；与此不同，我们的 AdvRain攻击表现出另一种行为。它会使模型忽略一些对模型做出准确判断至关重要的特征（如图 10所示）。通过策略性地引入雨滴扰动，我们将模型的注意力从关键特征上移开，使其更容易发生错误分类。

这种行为增强了我们的 AdvRain攻击的隐蔽性，因为它不像传统的对抗补丁那样会引起人们对某个特定补丁或区域的注意。雨滴扰动不易察觉，再加上它们对模型解读的影响，凸显了 AdvRain 在不引起人们怀疑存在对抗攻击的情况下欺骗基于摄像头的视觉系统的有效性。

Grad-CAM 可视化结果进一步证明了 AdvRain攻击的独特优势，凸显了它作为一种强大且隐蔽的技术，在现实应用中能够构造出逃避标准检测方法的对抗扰动，并削弱深度学习模型的可靠性。

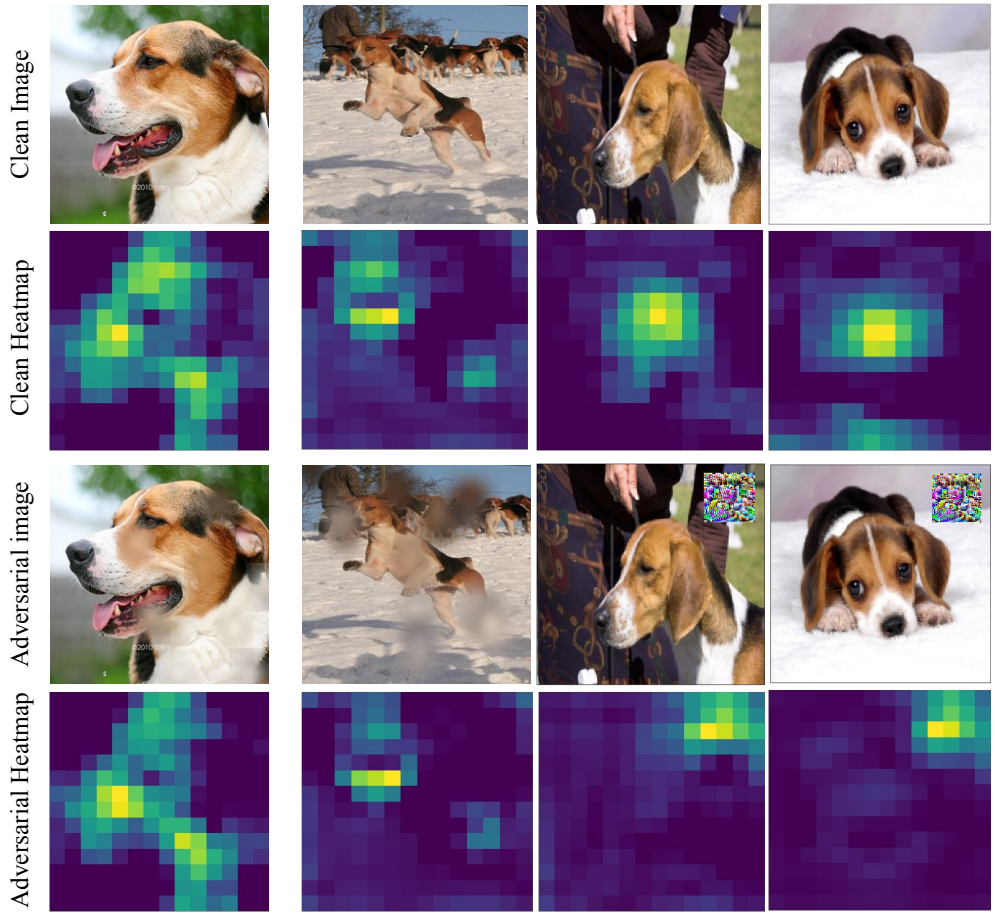


图10. 对比对抗补丁与 AdvRain 的 Grad-CAM 可视化结果。

5.5. 可能的防御措施

一种针对 AdvRain 的潜在防御策略是训练去噪/去雨模型。这种方法旨在增强深度神经网络抵御此类对抗攻击的能力。去噪器作为预处理阶段运行，在输入数据到达目标模型之前有效去除其中的对抗噪声。该去噪模型专门用于减轻 AdvRain 扰动的影响。

去噪/去雨模型可以设计为学习如何消除或抵消 AdvRain 扰动对模型输出的影响。可以通过定义一个引导去噪模型训练的损失函数来实现这一点。该过程可以这样进行：损失函数可以用于衡量受害模型在输入干净图像和去噪（去雨）图像时所作预测之间的差异。目标是尽量减小这种差异，从而有效地迫使去噪模型以确保受害模型的输出与干净图像时的输出保持一致的方式去除 AdvRain 扰动。训练数据由成对的图像组成：干净图像（即未遭受 AdvRain攻击的图像）和 AdvRain 感染图像（即遭受 AdvRain攻击并被引入扰动的图像）。

去噪模型在这些图像对上进行训练，以最小化定义的损失函数。在训练过程中，模型学习识别并理解与 AdvRain 扰动相关的模式。同时，它还学习如何修改 AdvRain 感染图像，以抵消这些扰动，从而有效生成与干净图像高度相似的去噪版本。训练完成后，当去噪模型接收到 AdvRain 感染图像作为输入时，它会应用学习到的变换来抵消 AdvRain 扰动。去除了 AdvRain 噪声的图像随后被传递

给受害模型进行分类或执行其他任务。由于去噪模型有效移除了 AdvRain 的影响，受害模型的输出仍会与输入干净图像时的输出保持一致。这种方法增强了系统抵御 AdvRain攻击的鲁棒性。去噪模型充当过滤器，在 AdvRain 感染图像到达受害模型之前对其进行“清理”，确保模型的预测结果不会受到这些扰动的不利影响。

另一种防御策略是考虑不同的雨滴模型生成对抗样本，并在针对给定深度神经网络的对抗训练过程中加入这些样本。通过使用涵盖多种由 AdvRain 生成的对抗样本的增强数据集训练深度神经网络，模型能够变得更加稳健，并能在更广泛的场景中抵御 AdvRain攻击。

5.6. AdvRain 在视频中的应用

AdvRain 的一个显著特点是能够实施类别范围攻击。这意味着，同一种对抗扰动（由贴在系统摄像头上的贴纸上的雨滴图案表示，例如汽车上的摄像头）可以通用于同一类别中不同物体的图像。这种通用方法简化了攻击流程，并能有效欺骗目标模型，使其在目标类别中面对各种物体时都受到影响。至于 AdvRain 在视频中的应用，其效果取决于多个因素，其中上下文稳定性是一个关键考量。我们认为，在上下文相对稳定的视频场景中，AdvRain 能够保持其有效性。在这种情况下，攻击可以在视频场景中持续有效地欺骗模型。这种现象在具有一致视觉元素的视频场景中十分常见。

5.7. 局限性

AdvRain 是一种有效且隐蔽的对抗攻击方法，但与任何技术一样，它也可能存在一些潜在局限性，具体如下：

- **设备特异性**：AdvRain 的成功可能取决于目标设备的具体摄像头和镜头特性。摄像头类型、镜头涂层和传感器分辨率的差异可能会影响攻击效果。这意味着，同一张 AdvRain 贴纸在所有设备上未必都能达到相同的效率。
- **贴纸放置位置**：攻击能否成功取决于 AdvRain 贴纸在相机镜头上的精确放置位置。如果贴纸未对准或部分受遮挡，就可能无法产生预期的扰动，从而降低攻击效果。
- **环境条件**：AdvRain 的效果可能受到光照、天气和能见度等环境条件的影响。在不同条件下，雨滴图案的逼真程度可能有所不同，从而可能限制攻击的可靠性。

6. 结论

本文提出了一种新颖的方法，用于制作逼真的物理对抗相机贴纸，以欺骗图像分类器。我们提出的攻击称为 AdvRain，通过将雨滴置于相机镜头上，模拟自然天气条件的影响。我们的方法围绕采用基于随机搜索的优化方法来确定最佳雨滴位置展开。通过大量评估，我们证明了 AdvRain 的有效性。仅使用 20 个雨滴，我们就显著降低了平均分类准确率：在 ImageNet 上使用 VGG19 时，准确率下降超过 45%；在 Caltech-101 上使用 Resnet34 时，准确率下降超过 40%。我们的结果展现了 AdvRain 作为一种隐蔽而强大的基于摄像头的视觉系统对抗攻击方法的威力。

作者贡献： 概念化，A.G.、M.A.H. 和 M.S.；方法，A.G.、M.A.H. 和 M.S.；软件，A.G.；验证，A.G.；正式分析，A.G.；调查，A.G.；资源，A.G.；数据整理，A.G.；初稿撰写，A.G.、M.A.H. 和 M.S.；审阅与编辑，M.S.；可视化，A.G.；监督，M.S.；项目管理，M.S.；经费获取，M.S.。所有作者均已阅读并同意稿件的已发表版本。

资助：本研究部分由NYUAD网络安全中心（CCS）提供支持，该中心由Tamkeen通过NYUAD研究所奖项G1104资助；交互式城市网络中心（CITIES）由Tamkeen通过NYUAD研究所奖项CG001资助；人工智能与机器人中心（CAIR）由Tamkeen通过NYUAD研究所奖项CG010资助。

数据可用性声明：未创建新数据。

利益冲突：作者声明不存在利益冲突。

参考文献

1. Al-Qizwini, M.；Barjasteh, I.；Al-Qassab, H.；Radha, H. 使用 GoogLeNet 实现自动驾驶的深度学习算法。IEEE智能车辆研讨会（IV）2017，89–96。2. Tramèr, F.；Zhang, F.；Juels, A.；Reiter, M.K.；Ristenpart, T. 通过预测API窃取机器学习模型。载于第25届 USENIX安全研讨会（USENIX Security 16）论文集，美国得克萨斯州奥斯汀，2016年8月10日至12日；第601–618页。3. Goodfellow, I.J.；Shlens, J.；Szegedy, C. 解释并利用对抗样本。arXiv 2015，arXiv:1412.6572。4. Kurakin, A.；Goodfellow, I.J.；Bengio, S. 物理世界中的对抗样本。arXiv 2016，arXiv:1607.02533。5. Carlini, N.；Wagner, D.A. 迈向神经网络鲁棒性的评估。arXiv 2016，arXiv:1608.04644。6. Zhong, Y.；Liu, X.；Zhai, D.；Jiang, J.；Ji, X. 阴影可能很危险：利用自然现象实施隐蔽且有效的物理世界对抗攻击。arXiv 2022，arXiv:2209.02430。https://doi.org/10.48550/ARXIV.2203.03818。7. Hu, C.；Shi, W. 对抗性色彩薄膜：针对深度神经网络的有效物理世界攻击。arXiv 2022，arXiv:2209.02430。https: // doi.org/10.48550/ARXIV.2209.02430。8. Duan, R.；Ma, X.；Wang, Y.；Bailey, J.；Qin, A.K.；Yang, Y. 对抗性伪装：以自然风格隐藏物理世界攻击。arXiv 2020，arXiv:2003.08757。https://doi.org/10.48550/ARXIV.2003.08757。9. Marchisio, A.；Caramia, G.；Martina, M.；Shafique, M. FakeWeather：在自主系统摄像头镜头上模拟天气条件，对深度神经网络发起对抗攻击。arXiv 2022，arXiv:2205.13807。https://doi.org/10.48550 /ARXIV.2205.13807。10. Brown, T.；Mane, D.；Roy, A.；Abadi, M.；Gilmer, J. 对抗补丁。arXiv 2017，arXiv:1712.09665。11. Subramanya, A.；Pillai, V.；Pirsiavash, H. 迈向将对抗样本隐藏于网络解释之外。arXiv 2018，arXiv:1812.02843。12. Szegedy, C.；Zaremba, W.；Sutskever, I.；Bruna, J.；Erhan, D.；Goodfellow, I.J.；Fergus, R. 神经网络的有趣特性。载于第二届国际学习表征会议论文集，ICLR 2014，加拿大艾伯塔省班夫，2014年4月14日至16日；会议专题论文集；Bengio, Y.、LeCun, Y. 编；2014。13. Brendel, W.；Rauber, J.；Bethge, M. 基于决策的对抗攻击：针对黑盒机器学习模型的可靠攻击。arXiv 2017，arXiv:1712.04248。14. Narodytska, N.；Kasiviswanathan, S.P. 针对深度网络的简单黑盒对抗扰动。arXiv 2016，arXiv:1612.06299。15. Chen, J.；Jordan, M.I. 边界攻击++：查询高效的基于决策的对抗攻击。arXiv 2019，arXiv:1904.02144。16. Hu, Y.C.T.；Chen, J.C.；Kung, B.H.；Hua, K.L.；Tan, D.S. 面向目标检测器的自然风格物理对抗补丁。载于2021年IEEE/CVF国际计算机视觉会议（ICCV）论文集，加拿大魁北克省蒙特利尔，2021年10月10日至17日；第7828–7837页。https://doi.org/10.1109/ICCV48922.2021.00775。17. Guesmi, A.；Bilasco, I.M.；Shafique, M.；Alouani, I. AdvART：用于伪装目标检测攻击的对抗艺术。arXiv 2023，arXiv:2303.01734。18. Guesmi, A.；Ding, R.；Hanif, M.A.；Alouani, I.；Shafique, M. DAP：用于规避行人检测器的动态对抗补丁。arXiv 2023，arXiv:2305.11618。19. Athalye, A.；Engstrom, L.；Ilyas, A.；Kwok, K. 合成鲁棒对抗样本。PMLR 2018，80，284–293。20. Sayles, A.；Hooda, A.；Gupta, M.；Chatterjee, R.；Fernandes, E. 不可见扰动：利用滚动快门效应的物理对抗样本。arXiv 2020，arXiv:2011.13375。21. Kim, K.；Kim, J.；Song, S.；Choi, J.H.；Joo, C.；Lee, J.S. Light Lies：光学对抗攻击。arXiv 2021，arXiv:2106.09908。22. Gnanasambandam, A.；Sherman, A.M.；Chan, S.H. 光学对抗攻击。arXiv 2021，arXiv:2108.06247。23. Kennedy, J.；Eberhart, R. 粒子群优化。载于ICNN’95——国际神经网络会议论文集，澳大利亚珀斯，1995年11月27日至12月1日；第4卷，第1942–1948页。https://doi.org/10.1109/ ICNN.1995.488968。

24. Kingma, D.P.；Ba, J. Adam：一种随机优化方法。*arXiv* **2014年**，arXiv:1412.6980。https://doi.org/10.48550/ARXIV.1412.6980。25. Simonyan, K.；Zisserman, A. 用于大规模图像识别的超深卷积网络。*arXiv***2014年**，arXiv:1409.1556。26. He, K.；Zhang, X.；Ren, S.；Sun, J. 用于图像识别的深度残差学习。*arXiv***215**，arXiv:1512.03385。27. Li, F.F.；Andreeto, M.；Ranzato, M.；Perona, P.*Caltech 101*。CaltechDATA 2022年。https://doi.org/10.22002/D1.20086。28. Deng, J.；Dong, W.；Socher, R.；Li, L.J.；Li, K.；Fei-Fei, L. ImageNet：大型分层图像数据库。载于2009年IEEE计算机视觉与模式识别会议论文集，美国佛罗里达州迈阿密，2009年6月20日至25日；第248–255页。29. Wang, Z.；Bovik, A.；Sheikh, H.；Simoncelli, E. 图像质量评估：从误差可见性到结构相似性。*IEEE*图像处理汇刊 **2004年**，*13*，600–612。https://doi.org/10.1109/TIP.2003.819861。

免责声明/出版者说明：所有出版物中包含的陈述、观点和数据仅代表各作者及贡献者个人，并不代表 MDPI 和/或编辑的立场。对于因内容中提及的任何观点、方法、说明或产品而对人员或财产造成的任何损害，MDPI 和/或编辑概不负责。