无人机视觉识别系统的自然物理对抗攻击方法

张 恒[1,2,3], 黄农森[1,3], 丁家松[1,3], 杭 芹[1,3]

[1] 重庆邮电大学计算机科学与技术学院 重庆 中国 400065
[2] 中国科学院合肥物质科学研究院 合肥 中国 230031
[3] 重庆邮电大学计算智能重庆市重点实验室 重庆 中国 400065

摘要 搭载目标检测算法的可见光视觉识别系统正逐渐成为无人机领域重要的感知系统。然而, 目标检测算法容易受到对抗攻击的威胁, 特别是物理攻击通常以对抗补丁的形式嵌入现实场景中, 其威胁程度远超过数字攻击。现有的物理攻击方法主要针对目标检测的地面近距离应用场景, 并且生成的物理对抗补丁容易被人类所察觉, 从而暴露攻击意图。为此, 提出了一种针对无人机视觉识别系统的自然物理对抗攻击方法(Natural Physical Patch Attack, NPAP)。首先, 设计了针对多尺度、多目标攻击的优化函数, 提升对抗补丁的攻击能力。接着, 为生成自然的对抗补丁, 引入相似性度量对补丁的外观进行约束。最后, 基于期望转换的原理, 设计了补丁物理增强变换模块, 采用多种物理增强变换, 提升对抗补丁对环境和尺度变化的鲁棒性。在数字攻击实验中, 该方法对 YOLOv3、YOLOv5、YOLOv7 三个主流目标检测器的攻击成功率分别为 72.6%、77.6%、75.0%。在物理攻击实验中, 将数字域中生成的对抗补丁打印到现实世界中进行测试, 该方法在 20～100m 高度范围内对三个目标检测器的平均攻击成功率分别为 63.6%、58.3%、56.8%。实验结果表明, 与 G/C、UPC、NAP 三种主流的攻击方法相比, 该方法在不增加复杂度的情况下能够生成与自然图像相似的对抗补丁, 并且生成的对抗补丁表现出优越的攻击性能和鲁棒性。

关键词 目标检测; 对抗样本生成; 无人机识别; 对抗攻击
中图法分类号 TP391 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2025.09.06

Natural Physical Adversarial Attack Method for UAV Visual Recognition System

ZHANG Heng[1,2,3], HUANG Nongsen[1,3], DING Jiasong[1,3], HANG Qin[1,3]

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
[3] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract The visible light visual recognition system equipped with object detection algorithms is progressively emerging as a vital perception system in the domain of unmanned aerial vehicle. However, object detection algorithms are susceptible to adversarial attack. In particular, physical adversarial attacks often manifest in the form of adversarial patches embedded within real-world scenes, posing a more significant threat compared to digital attacks. Existing physical attack methods primarily target close-range ground-based applications of object detection, and the generated physical adversarial patches are easily perceptible by humans, thereby exposing malicious intent. To address this issue, a natural physical adversarial attack method (Natural Physical Patch Attack, NPAP) targeting unmanned aerial vehicle visual recognition systems is proposed. Firstly, an optimization function tailored for multi-scale and multi-target attacks is designed to enhance the attack capability of adversarial patches. Subsequently, to generate natural adversarial patches, a similarity metric is introduced to constrain the appearance of the patches. Finally, based on the principle of expectation over transformation, a patches physical enhancement transformation module is designed. Multiple physical augmentation transformations are employed to enhanced the robustness of adversarial patches against environmental and scale variations. In the digital attack experiment, this method achieved success rates of 72.6%, 77.6%, and 75.0% against three mainstream object detectors: YOLOv3, YOLOv5, and YOLOv7, respectively. In the physical attack experiment, the adversarial patches generated in the digital domain were printed and tested in the real world. this method achieved average attack success rates of 63.6%, 58.3%, and 56.8% against the three object detectors within the altitude range of 20 meters to 100 meters. The experimental results demonstrate that, compared to three mainstream adversarial attack methods, namely G/C, UPC, and NAP, the proposed approach is capable of generating adversarial patches that resemble natural images without augmenting complexity. Concurrently, the generated adversarial patches exhibit superior adversarial performance and robustness.

通讯作者: 杭芹, 博士, 讲师, Email: hangqin@cqupt.edu.cn。
本课题得到国家自然科学基金项目(No. 12005030); 重庆市自然科学基金项目(No. cstc2021jcyj-bsh0252); 重庆邮电大学博士启动基金项目(No. A A2020-217 & A2020-216)资助。
收稿日期: 2023-12-15; 修改日期: 2024-03-18; 定稿日期: 2025-08-01

Key words object detection; adversarial sample generation; unmanned aerial vehicle recognition; adversarial attack

1 引言

深度神经网络(Deep Neural Network, DNN)的发展提升了智能视觉系统在各种实际应用中的性能, 其中, 目标检测算法提供了卓越的目标识别和定位能力。无人机作为灵活高效的航空平台, 为目标检测技术提供了独特视角。因此, 基于 DNN 的无人机目标检测技术广泛应用于无人巡检、应急救援和农业植保等领域。

然而, DNN 容易受到对抗攻击[1]的威胁。对抗攻击通过对输入样本添加微小扰动, 诱导 DNN 产生错误的预测结果。对抗攻击方法可以分为数字攻击和物理攻击两种类型。数字对抗攻击的目标是在输入样本中引入微小扰动, 并在数字域中追求强大的攻击性能[2]。但将数字攻击转移到物理场景时, 微小的扰动很容易被各种环境和设备因素过滤掉。与数字攻击不同, 物理攻击专注于攻击部署在真实世界中的 DNN 系统, 这类攻击通常以生成能够克服复杂物理环境的对抗补丁为目标。

目前对物理对抗攻击的研究主要集中在人脸识别、行人检测、自动驾驶等地面应用场景。在这些场景中, 视觉传感器通常从近距离获取图像, 环境对数据采集的干扰较小, 攻击者通常能够比较容易地生成具有高效攻击能力的对抗补丁。在空中场景, 无人机的视觉传感器通常从几十米至百米的距离采集图像, 图像中目标的数量和尺度变化较大, 并且运动拍摄过程中容易受到光照条件和设备引起的运动模糊、传感器噪声等影响, 这些因素导致了物理对抗攻击的难度增大。此外, 现有方法主要侧重于提高攻击性能, 忽略了对抗补丁的外观表现, 从而限制了其在无人机场景的实用性。

当前对无人机目标检测领域对抗攻击的研究未得到充分探索。文献[3-4]将基于补丁的攻击应用到无人机场景中, 但他们仅在数字世界中进行了攻击, 未扩展到物理世界中。文献[5]首次提出了针对无人机空中检测场景的物理对抗攻击方法, 但仅在固定高度进行了攻击, 并且该方法未考虑对补丁的外观进行优化。文献[6]设计了不同尺度的攻击方案, 但仍然未考虑物理补丁的外观表现和对环境影响的鲁棒性。

基于以上问题, 本文提出了一种适用于无人机空中目标检测场景的自然物理补丁攻击方法(Natural Physical Patch Attack, NPAP)。首先, 针对现有方法在攻击航拍图像中多尺度目标效果较差的问题, 本文将目标检测模型中不同尺度预测特征层的最大平均预测分数作为目标损失, 并结合 TOG[7]算法中目标消失攻击的损失函数对补丁进行联合优化。接着, 为了提升对抗补丁在远距离物理场景中的鲁棒性, 设计了物理增强变换模块: 除了采用期望转换(Expectation Over Transformation, EOT)[8]中定义的常见物理变换以外, 进一步引入运动模糊、透视变换、传感器噪声等物理变换方式。最后, 为了生成外观自然的对抗补丁, 提升其隐蔽性, 在优化函数中引入相似性度量, 确保对抗补丁的外观与目标图像相似。本文工作的主要亮点如下。

(1) 提出了一种针对无人机目标检测场景的 NPAP 方法。在优化函数设计上, 将目标检测器不同尺度的预测结果作为平均优化损失, 并将 TOG[7]算法中目标消失的损失函数作为置信度损失, 提升了对抗补丁对多尺度、多目标图像的攻击效果。引入相似性度量约束补丁的外观, 提升了对抗补丁的隐蔽性; 在补丁的优化过程中, 设计物理增强变换模块, 每次迭代对补丁施加 EOT 变换、运动模糊、透视变化以及传感器噪声等图像变换, 通过这些变换方式模拟对抗补丁在物理世界中可能遇到的各种干扰, 提升了对抗补丁在物理环境中的鲁棒性。

(2) 为了训练针对无人机目标检测场景的对抗补丁, 使用无人机从不同高度采集并制作了高分辨率的航拍车辆数据集。为了验证方法的有效性, 使用所制作的数据集在 YOLOv3、YOLOv5 和 YOLOv7 三个目标检测模型上优化对抗补丁, 并分别进行了数字攻击实验和多高度、多方向的物理攻击实验。实验结果表明, NPAP 方法所生成的对抗补丁在数字世界和物理世界均能有效攻击无人机目标检测系统, 优于目前主流的攻击方法。

2 相关工作

2.1 面向目标检测的数字对抗攻击

目标检测是一种在图像中对目标进行定位和分类的技术, 广泛应用于视频监控、自动驾驶等领域。目前主流的目标检测器主要分为单阶段检测器和两阶段检测器。单阶段检测器直接输出检测结果, 典型的单阶段检测器包括 YOLO 系列[9-11]、SSD[12]等。两阶段检测器首先通过主干特征提取网络对输入数据进行特征提取, 然后进行分类和回归, 得到类别和边界框的预测信息。典型的两阶段检测器包括

Faster-RCNN [13]、Master RCNN [14]等。无论单阶段还是两阶段目标检测器, 其输出结果都包括边界框、置信度和类别信息。边界框用来准确定位目标在图像中的位置, 置信度则表明了该定位区域内包含目标的确信度, 类别信息则对目标进行分类。

根据目标检测技术的运行原理可知, 对目标检测进行对抗攻击至少需要修改其三个输出中的一个, 这相比攻击图像分类任务更具有挑战性。对目标检测进行数字对抗攻击主要通过在数字域中对输入图像添加全局或者局部的对抗性扰动来实现。2017 年, Lu 等人[15]通过在停车标志牌上添加对抗扰动, 成功地欺骗了 Faster-RCNN 和 YOLO 两个目标检测器。他们的工作为目标检测领域的对抗攻击提供了初步验证。随后, Xie 等人[16]提出了 DAG(Dense Adversary Generation)方法, 该方法通过在输入图像的多个区域施加微小的对抗扰动来优化对抗损失函数, 从而使目标检测模型产生错误的输出。Li 等人[17]提出了 R-AP 方法, 该方法通过攻击 Faster-RCNN 的 RPN(Region Proposal Network)网络, 结合分类损失和边界框的定位损失生成对抗样本。2019 年, Zhang 等人[18]从目标检测的多任务视角出发, 根据目标检测模型的分类和定位损失, 成功地将图像分类任务中的 PGD[19]攻击算法应用到目标检测任务当中。2020 年, Chow 等人[7]提出了 TOG 方法, 该方法通过精心设计三种不同的损失函数, 使得生成的对抗样本能够实现目标消失、目标冗余以及分类错误三种定向的攻击效果。2019 年, Liu 等人[20]首次提出了针对目标检测领域局部对抗攻击的 Dpatch 方法, 该方法通过在图像的随机区域添加一个特定大小的对抗补丁, 该补丁能够同时攻击目标检测器的边界框回归和分类任务。

2.2 面向目标检测的物理对抗攻击

与数字攻击相比, 物理攻击具有更高的威胁性, 其能够直接攻击部署在真实世界中的目标检测系统。2019 年, Thys 等人[21]提出了 Adversarial-YOLO 方法。该方法通过在数字世界中最小化置信度得分来优化对抗补丁, 并将补丁打印到现实世界以欺骗行人检测器。2021 年, 在 Adversarial-YOLO 方法的基础上, Wang 等人[22]验证了 YOLOv2 生成的对抗补丁在可迁移性方面较弱的问题, 并在 YOLOv3 网络上设计不同检测分数来优化对抗补丁, 他们采用打印和显示屏的方式在物理世界中致盲目标检测器。

对物理对抗攻击的研究大多集中在攻击性能上, 很少对对抗补丁的外观进行限制。Huang 等人[23]提出了 UPC 方法, 通过构建一个通用的目标隐蔽攻击模式来攻击目标检测器, 并引入优化约束来生成自然的对抗补丁。2021 年, Hu 等人[24]利用预训练的生成对抗网络来制作具有自然外观的物理对抗补丁。Tan 等人[25]提出了一个两阶段训练策略, 分别对补丁的外观和攻击能力进行优化。2022 年, Hu 等人[26]提出了对抗性纹理(Adversarial Texture)。该方法通过训练一个可扩展的纹理生成器, 并将其渲染到衣服上进行物理攻击, 由这些纹理渲染的衣服可以从不同角度欺骗行人检测器。Liu 等人[27]认为一些自然现象可以作为对抗样本, 他们使用生成对抗网络来模拟自然雨滴, 所生成的对抗样本看起来与现实世界的雨滴图像相似, 并且能够实现较好的攻击效果。Qin 等人[28]提出了基于生成对抗网络的两阶段对抗补丁生成方法, 在补丁生成过程中通过模拟不同遮挡情况, 生成的对抗补丁能够在不同遮挡情况下实现稳定的攻击效果。

2.3 面向无人机目标检测领域对抗攻击

现有的目标检测对抗攻击方法主要应用于地面场景, 而随着深度学习在无人机技术上的广泛应用, 相应地, 对无人机视觉检测系统的对抗攻击也引起了关注。2020 年, Den Hollander 等人[4]在航空图像数据集上训练了 YOLOv2 检测器, 并将 Adervsarial-YOLO 方法应用于无人机检测领域, 实现了对无人机空中目标检测的数字对抗攻击。2022 年, Du 等人[5]首次提出了针对空中图像检测的物理攻击方法, 他们通过在一组固定高度的数据集上优化对抗补丁, 并将其打印后放置于汽车车顶或者周围来实现目标隐藏攻击。随后, Zhang 等人[6]将图像高度标签作为比例因子, 根据拍摄高度对补丁进行缩放, 实现了不同高度下的物理攻击。Lian 等人[29]提出了 AP-PA 方法, 该方法生成的对抗补丁对地面飞机尺寸具有适应性。

然而, 现有对无人机目标检测领域的对抗攻击研究主要集中在数字攻击。尽管已经提出了几种物理攻击方法, 但这些方法没有考虑对抗补丁的外观表现, 导致对抗补丁在现实世界中容易暴露。此外, 现有的方法未充分考虑尺度变化和复杂环境对对抗补丁性能的影响, 导致对抗补丁的鲁棒性较差。因此, 本文的主要目标是生成具有自然外观且鲁棒性较强的对抗补丁, 用于攻击无人机航拍车辆检测器。

3 本文方法

本节介绍了 NPAP 方法的具体流程。首先, 本节阐述了 NPAP 方法所使用的威胁模型。其次, 介绍了方法的整体框架, 并对框架中所提出的补丁物理增强变换模块、优化函数分别进行了描述。

3.1 威胁模型

(1) **攻击者的知识**: 考虑到无人机目标检测的实时性, 本文以 YOLOv3、YOLOv5、YOLOv7 三个比较经典且应用广泛的检测器作为受害者模型, 并假设攻击者在白盒场景下对补丁进行优化。在白盒场景中, 攻击者能够获得目标模型的完整信息, 包括模型结构、模型参数、损失函数以及数据集等信息。攻击者可以利用这些信息制作对抗补丁, 但不能对受害者模型进行修改。

(2) **攻击者的目标**: 在本文中, 攻击者的目标是通过在数字域中优化和评估对抗补丁, 并将其打印后放置在现实场景中汽车车顶以执行物理攻击。物理攻击中, 使用预先训练好的无人机目标检测器从空中对地面车辆进行检测, 如果目标检测器无法识别到地面上被放置了对抗补丁的车辆, 则视为攻击成功。

(3) **攻击者的策略**: 在数字域中, 攻击者的策略是试图通过最小化目标函数来生成对抗补丁。在物理域中, 攻击者只能通过控制外部环境执行攻击, 而不能对无人机目标检测器的视觉传感器或数据传输管道进行访问。

3.2 方法整体框架

方法的整体架构如图 1 所示, 待优化的对抗补丁首先会输入补丁增强转换模块, 该模块的作用是对补丁进行一系列变换来模拟各种物理条件。接着, 变换后的对抗补丁在补丁应用模块中根据目标的标注信息构造掩码矩阵, 确定补丁附加的尺寸和位置, 然后附加到图像中的车辆上得到对抗样本。生成的对抗样本输入目标检测器中获得预测结果, 根据预测结果计算总体损失。最后, 结合 Adam 优化器进行梯度下降操作更新对抗补丁的像素值。

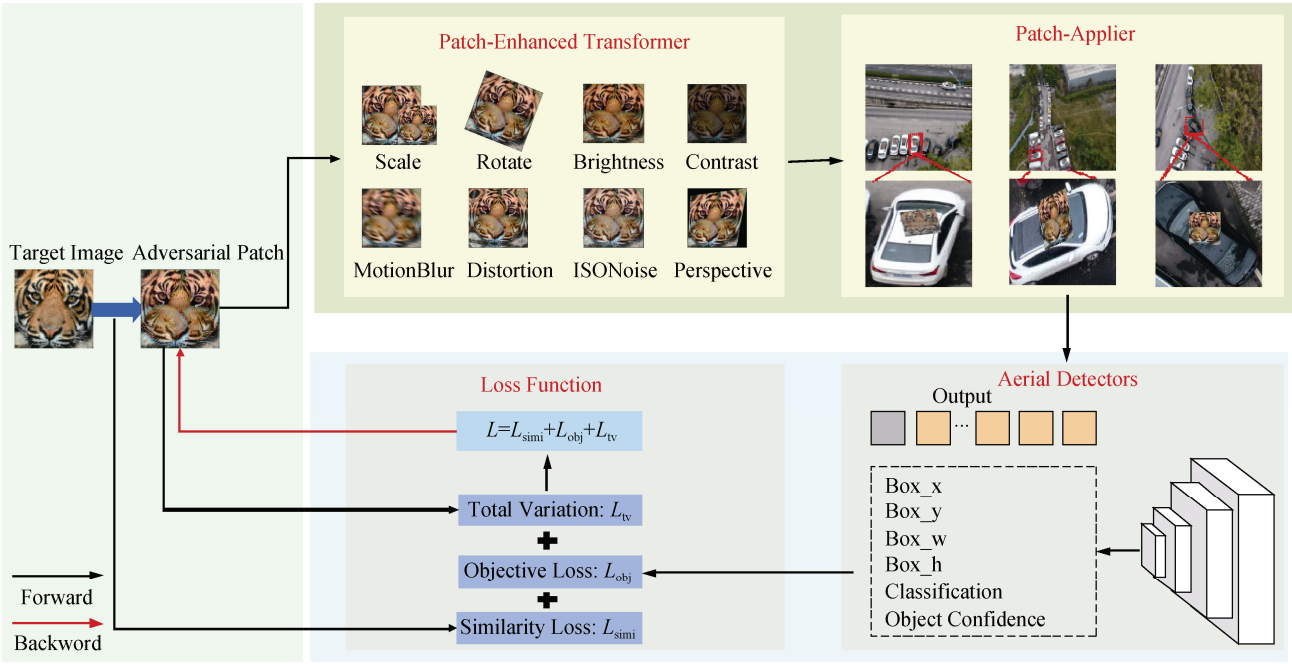


图 1 NPAP 方法整体流程

Figure 1 Overall process of the NPAP method

3.3 补丁物理增强变换

在现实场景中, 环境因素已被证实会对补丁造成负面影响[30], 特别是在远距离的无人机目标检测场景中。为了提升对抗补丁在物理场景中的鲁棒性, 本文引入一系列变换分布 T 以尽可能真实地模拟空中场景的各种物理因素。考虑到无人机视觉系统在运动状态下的检测, 在 EOT 变换的基础上, 本文进一步增加了下面几种图像变换方式。

(1) 运动模糊:无人机在飞行过程中, 传感器与对抗补丁之间可能产生相对位移, 导致图像可能出现模糊的情况, 本文通过对补丁添加运动模糊变换来模拟这种效应。

(2) 光学畸变: 由于无人机视觉传感器镜头和视场等原因, 拍摄的图像可能会产生畸变现象, 导致对抗补丁的形状发生改变。

(3) 传感器噪声: 无人机的传感器可能产生不同类型的噪声影响。通过对补丁施加传感器噪声, 模拟这些噪声影响, 以提高对抗补丁的鲁棒性。

(4) 透视变换: 无人机在不同高度和角度下进行拍摄, 物体可能会产生透视变换。通过引入透视变换,

模拟不同高度和角度下的产生透视变换情况。

本文使用 Albumentations[31]数据增强库对补丁进行可微变换, 表 1 为补丁物理变换的具体细节。

表 1 补丁变换细节

Table 1 Details of patch transformations

变换类型	表达式	参数说明	参数范围
Contrast	$P_a+(P-P_a)\cdot\beta$	β -常量 P_a -补丁平均像素	[0.8,1.2]
ISO Noise	$P+N(\alpha,\beta)$	α -色调方差 β -强度因子	[0.05,0.05] [0.1,0.5]
Gauss Noise	$P+N(\sigma)$	σ -方差	[0,0.3]
Rotate	$R_\theta P$	θ -旋转角度	[–20,20]
Motion Blur	$P\cdot K(\alpha)$	α -模糊核大小	[0,30]
Optical Distortion	$P\cdot O(\alpha,\beta)$	α -扭曲幅度 β -移动幅度	[–0.05.0.05] [–0.05,0.05]
Brightness	$\alpha\cdot P$	α -像素强度	[–0.1,0.1]
Perspective	$P\cdot A$	A -透视变换矩阵	/

补丁物理增强变换的表达式如式(1)所示。

$$P^*=t(P) \tag{1}$$

其中, P 为对抗补丁, P^* 表示变换后的对抗补丁, $t(\cdot)$ 为变换函数, 从总变换分布 T 中进行采样。

3.4 补丁应用模块

为了将对抗补丁成功放置到图像中不同尺度车辆的顶部, 本文设计了补丁应用模块。根据图像标注信息计算每一个补丁的缩放尺度, 接着构造掩码矩阵与输入图像进行叠加。表达式如式(2)所示。

$$\begin{aligned} S_{w,h}&=\varepsilon\cdot\sqrt{h^2+w^2}\\ X_a&=M(S_{w,h}\otimes P)\oplus X\end{aligned} \tag{2}$$

其中, $S_{w,h}$ 为补丁需要缩放的尺度, w 和 h 分别为标注框的宽高, ε 为缩放因子, 在本文中设置为 0.25 时, 补丁能够有效放置。 X_a 为对抗样本, X 为输入样本, P 为对抗补丁。$M(\cdot)$ 表示掩码矩阵, $\otimes$ 为补丁缩放操作, $\oplus$ 为叠加操作。

3.5 优化函数设计

3.5.1 相似度损失

衡量对抗补丁伪装性能的指标以损失函数的形式进行计算, 并纳入到补丁的优化过程中。本文引入相似度度量来约束对抗补丁的外观: 首先, 将自然图像裁剪到与对抗补丁同样的尺寸, 并将其作为目标图像。然后, 在每次迭代过程中, 计算对抗补丁与目标图像的相似度, 并将计算结果作为相似度损失进行反向迭代。本文选用均方误差(mean squared error, MSE)作为相似度损失。相似度损失的计算方式如式(3)所示。

$$L_{\text{simi}}=\frac{1}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(P_{i,j}-T_{i,j})^2 \tag{3}$$

其中, $P_{i,j}$ 和 $T_{i,j}$ 分别表示对抗补丁和目标图像对应位置的像素值。

3.5.2 目标损失

目标隐藏攻击的原理是通过降低目标检测器预测框的置信度分数, 使其低于目标检测网络的非极大值抑制(Non-Maximum Suppression, NMS)机制的置信度阈值。现有方法通常将最大置信度分数作为目标损失, 但在航拍图像中, 目标的数量通常比较多, 并且不同目标的尺度差异较大。对于 YOLO 系列目标检测器, 当输入一张图像时, 其特征提取网络分别以 8 倍、16 倍和 32 倍的下采样率得到三个不同尺度的预测特征图, 并在每个预测特征图上生成相应的预测结果。其中, 下采样率较小的预测特征层对小尺度目标的预测效果较好, 而下采样率较高的特征图适用于大尺度目标的检测。最大置信度分数的计算无法涉及图像中不同尺度的目标, 因此, 为了提升对多尺度目标的攻击效果, 本文将每个预测特征图的最大置信度的平均值作为平均优化损失, 具体表达如式(4)所示。

$$L_{\text{mean}}=\frac{1}{N}\sum_{i=1}^{N}\max\left(\text{bbox}_i^{\text{conf}}\right) \tag{4}$$

其中, N 表示模型预测特征层的数量, $\max(\text{bbox}_i^{\text{conf}})$ 代表预测特征层的预测框集合中的最大置信度分数。

为了提升对多目标图像的攻击效果, 本文引入了 TOG(Targeted Adversarial Objectness Gradient Attack)算法中目标消失攻击的损失函数。首先, 为训练数据分配虚假标签, 将标签内容设为空集。其次, 使用虚假标签和模型预测集合计算置信度损失, 目的是让模型输出的所有预测框的置信度分数向 0 进行迭代。在目标检测网络中, 使用二元交叉熵损失计算预测框的置信度损失, 由于目标的标签设置为空集, 因此置信度损失的表达式如式(5)所示。

$$L_{\text{conf}}=-\frac{1}{n}\sum_{i=1}^{n}\log(1-p_i) \tag{5}$$

其中, p_i 表示第 i 个预测框的置信度分数。

因此, NPAP 方法的目标损失由平均优化损失和置信度损失两部分构成, 表达式如式(6)所示。

$$L_{\text{obj}}=L_{\text{mean}}+L_{\text{conf}} \tag{6}$$

3.5.3 总变差损失

对抗补丁相邻像素之间的不连续性难以被视觉

传感器所捕捉, 从而造成对抗补丁因失真导致性能下降。为此, 引入总变差损失(Total Variation, TV)来平滑图像, 减少对抗补丁中的噪点。TV 损失的表达式如式(7)所示。

$$L_{\text{tv}}=\sum_{i,j}\sqrt{(p_{i,j}-p_{i+1,j})^2+(p_{i,j}-p_{i,j+1})^2} \tag{7}$$

其中, $p_{(i,j)}$ 表示补丁中每个位置对应的像素值。

综上所述, 本文方法的优化函数由相似度损失、目标损失和总变差损失三部分组成。其中, 相似度损失用于约束对抗补丁的外观, 使其表现出自然特征。目标损失用于降低模型预测框的置信度, 提升对抗补丁的攻击效果。总变差损失用来平衡图像, 减小对抗补丁在物理世界的失真。本文方法的优化问题如式(8)所示。

$$\begin{aligned}&\arg_p \min \text{Loss}\\&\text{s.t. Loss}=\alpha L_{\text{obj}}+\beta L_{tv}+\lambda L_{\text{simi}}\end{aligned} \tag{8}$$

其中, α、β、λ 为每一项损失的权重系数, 具体设置将在实验部分进行说明。

3.6 NPAP 方法生成对抗补丁过程

NPAP 方法每个迭代轮次会对数据集进行遍历, 在此过程中, 补丁首先会通过物理增强变换函数进行变换, 然后构造掩码矩阵与图像进行叠加制作对抗样本。对抗样本输入模型进行前向传播并分别计算目标损失、总变差损失和相似性损失, 将三部分损失进行累加求和后接着读取下一张图像。数据集完成一次遍历后, 将累计的总损失进行梯度下降更新对抗补丁。算法 1 为 NPAP 单次迭代更新对抗补丁的具体过程。

算法 1. NPAP 单次迭代更新对抗补丁过程

输入: 训练集 n、上一次迭代的对抗补丁 P_0;

输出: 更新后的对抗补丁 P;

1. Loss=0;

2. FOR n_i in n DO;

3. $P^*=t(P_0)$; //$t(\cdot)$ 为补丁物理增强变换函数.

4. $n_{p^*}=M(P^*)\oplus n_i$; //补丁自适应模块, $M(P^*)$ 为掩码矩阵, $\oplus$ 为叠加操作. n_{p^*} 为对抗样本.

5. $\text{Loss}=L_{\text{obj}}+L_{\text{tv}}+L_{\text{simi}}$; //前向传播计算总损失.

6. END FOR;

7. $P=\text{Adam}(\text{Loss})$; //梯度下降更新对抗补丁.

8. RETURN P;

从算法 1 可以看到, 算法总共有一个 For 循环, 更新一次对抗补丁所需要的时间复杂度与训练集样本大小呈线性关系, 因此单次迭代更新对抗补丁的时间复杂度可以表述为 $O(n_{\text{train}})$, n_{train} 为数据集大小。空间复杂度主要来源于模型正向传播过程中梯度信息所占的内存, 具体受到目标模型的参数量影响。

4 实验与结果分析

4.1 实验设置

1) 数据集制作

本文所讨论的场景需要使用到高分辨率并且包含多尺度特征的航拍车辆数据集, 以便于对补丁进行优化和评估, 但目前尚未有公开可访问的数据集。为此, 本文使用 DJ Mini2 型号的无人机从停车场和公路等场景采集数据。为了获得多尺度数据, 本文从 20～100m 的飞行高度范围进行采集, 所采集的图像分辨率为 3000 × 4000, 本文将其统一裁剪为 3000 × 3000。使用标注工具将采集的数据标注为 YOLO 格式, 用于在每次迭代过程确定对抗补丁施加的尺寸和位置。最终, 本文制作了一个多尺度的航拍车辆数据集, 其中包含 600 张训练图像, 4284 个目标和 300 张测试图像, 2140 个目标, 图 2 为了数据集的部分示意图。

图 2 本文制作的数据集示例

Figure 2 Data set examples created in this paper

2) 目标检测模型准备

本文使用 VisDrone2019[32]数据集分别训练了 YOLOv3、YOLOv5、YOLOv7 三个航拍车辆的目标检测器, 由于攻击的目标仅限汽车类别, 本文从数据集中剔除了其他类别数据, 仅保留包含汽车类别的样本数据。数据集共包含 6648 张车辆图片, 将其按照 9: 1 进行训练和测试, 在测试集上, 三个检测模型的平均精度(Mean Average Precision, MAP)分别为 82.3%、86.0%、87.2%。

3) 评价指标

衡量对抗攻击的性能常用 MAP 作为评价指标,

目标检测模型被攻击后的 MAP 数值越低，代表攻击算法的攻击性能越强。但对于目标隐藏攻击而言，需要统计成功误导目标检测器的对抗补丁数量，MAP 不是最佳的评价标准，目标检测器的其他漏检或者误检情况也会造成 MAP 的下降。为了更准确地衡量对抗补丁的攻击性能，本文使用攻击前后目标检测器输出的预测框的数量比作为攻击成功率(Attack Success Rate, ASR)，表达式如式(9)所示。

$$\mathrm{ASR}=\frac{\sum(x_{\mathrm{clean}}^{i}-x_{p}^{i})}{\sum x_{\mathrm{clean}}^{i}} \tag{9}$$

其中，x_{clean}^{i} 为攻击前的预测框数量，x_{p}^{i} 为攻击后的预测框数量。由公式可知, ASR 越高说明逃脱目标检测器的车辆越多，对抗补丁的攻击效果越好。

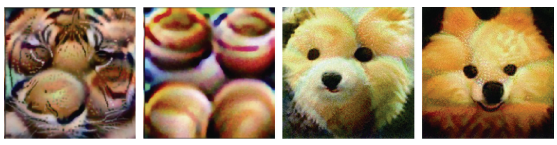
4) 实验细节

目标模型 YOLOv3、YOLOv5、YOLOv7 的输入图像尺寸分别为 416×416、640×640、640×640，三个目标检测器的置信度阈值均设定为 0.5。数字世界中，补丁的大小为 300×300。补丁优化过程中，使用 Adam 优化器进行梯度下降，并设置学习率为 0.01，总迭代轮次为 1000。本文的优化函数实际上涉及了一个联合优化问题，为了平衡补丁的外观和攻击能力，本文引入了动态调整机制: 在初始阶段，设置优化函数的权重分别为 $\alpha=1$，$\beta=1$，$\lambda=8$。当对抗补丁与目标图像的相似度损失下降到阈值(0.05)以后，在随后的迭代过程中，将损失函数的权重动态调整为 $\alpha=1$，$\beta=1$，$\lambda=4$。在此阶段，对抗补丁的外观已经具备自然特征，因此更加注重提升对抗补丁的攻击性能。

4.2 数字域攻击实验

4.2.1 NPAP 方法与其他方法对比

本节在数字域中进行了实验。使用 NPAP 方法与 G/C[5]、UPC[23]、NAP[24]方法进行对比，四种方法均在相同实验设置下进行。图 3 展示了四种方法生成的部分对抗补丁，可以看到，G/C 方法所生成的对抗补丁不具有自然特征，其在视觉表现上抽象。相比之下，NPAP、NAP、UPC 方法所生成的补丁呈现出与自然图像相似的外观。



(a) NPAP (b) G/C (c) NAP (d) UPC

图 3 不同方法生成的对抗补丁示例

Figure 3 Examples of adversarial patches generated by different methods

在数字攻击实验中，本文通过添加不同的干扰模拟两种不同的物理场景来测试对抗补丁在数字域中的攻击性能。

(1) 典型物理场景: 在测试过程中，对补丁进行 EOT 变换，再将其覆盖到测试集的图像中。

(2) 增强物理场景: 在测试过程中，对补丁同时进行 EOT 变换和物理增强变换，再将其覆盖到测试集的图像中。

图 4～图 6 统计了在典型物理场景下，不同方法生成的对抗补丁在三种目标检测器上攻击前后的预测框数量。从图 4～图 6 可以看出，当对图像中的车辆添加随机补丁后，与攻击前相比，攻击后检测框的数量下降均不超过 200。

表 2 为不同对抗补丁在三种目标检测器上的攻击成功率。从表 2 可知，随机补丁在三种目标检测器上的攻击成功率分别为 5.1%、4.9%、6.2%，这说明未经优化的随机补丁，通常不具有对抗性，同时验证了局部遮挡在实验中对目标检测器的性能影响较小。相比之下，四种攻击方法生成的对抗补丁对三种目标检测器的性能产生了显著的影响，与攻击前相比，目标检测器被攻击后预测框数量显著减少。

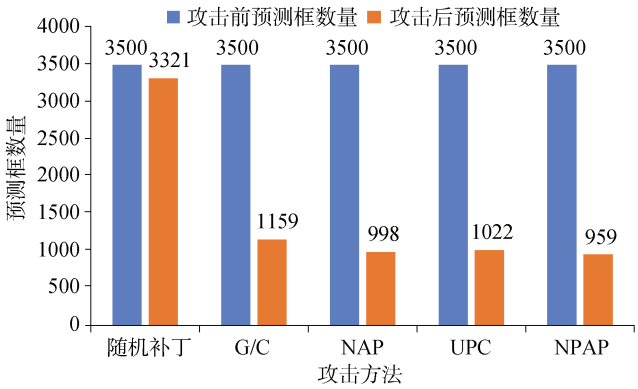


图 4 YOLOv3 上的攻击结果

Figure 4 The attack results on YOLOv3

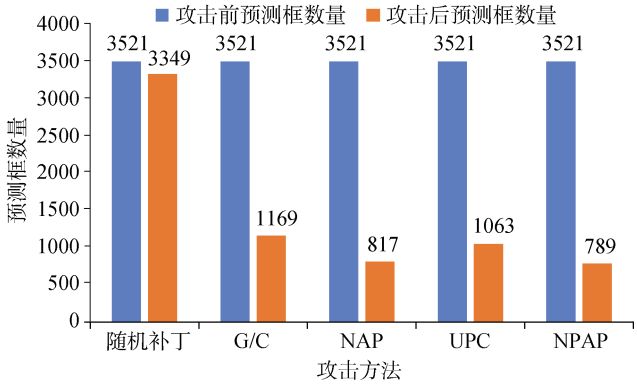


图 5 YOLOv5 上的攻击结果

Figure 5 The attack results on YOLOv5

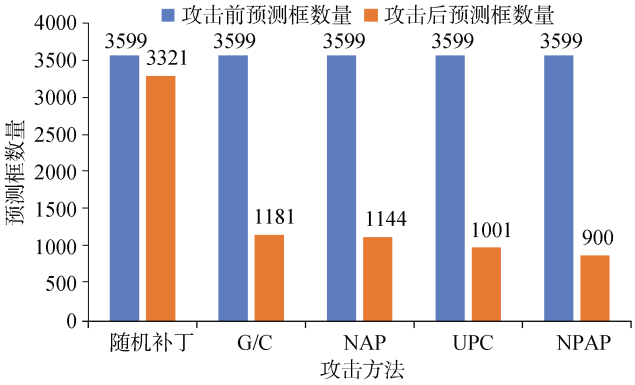


图 6　YOLOv7 上的攻击结果

Figure 6　The attack results on YOLOv7

表 2　不同方法的 ASR 对比结果

Table 2　Comparison of ASR with different attack methods

攻击方法	YOLOv3	YOLOv5	YOLOv7
随机补丁	5.1%	4.9%	6.2%
G/C	66.8%	68.8%	67.2%
NAP	71.5%	76.8%	68.2%
UPC	70.8%	69.8%	72.2%
NPAP	**72.6%**	**77.6%**	**75.0%**

从表 2 可知, NPAP 方法生成的对抗补丁在三种目标检测器上的攻击成功率分别为 72.6%、77.6%、75.0%, 攻击效果均优于 G/C、NAP、UPC 三种方法, 主要的原因在于 NPAP 方法中设计的目标损失更加合理, G/C、NAP、UPC 三种方法都采用了目标检测器预测集合中的最大置信度分数作为目标损失, 在攻击多目标图像时难以将所有目标的置信度降低到阈值以下, 从而导致了整体的攻击成功率下降。

图 7 对不同方法在 YOLOv5 上的攻击结果进行了定性分析。从图 7 可知, NPAP 方法对多目标和多尺度图像的攻击效果更好, 因为 NPAP 方法设计的优化函数中, 平均优化损失使对抗补丁能够同时影响目标检测器不同尺度的预测特征层。目标消失损失尽可能地降低了目标所在预测框的置信度分数。随着图像中的目标数量增多, NPAP 生成的对抗补丁使更多的车辆逃避检测。G/C、NAP、UPC 三种方法在应对多目标情况时攻击效果不足, 对抗补丁仅能使部分车辆逃避检测。以上结果说明了 NPAP 方法在对多目标和多尺度图像进行攻击时更有优势, 目标损失的设计使对抗补丁能够更好地抑制目标检测器检测框的置信度分数, 从而提高了攻击成功率。

(a) G/C　(b) NAP　(c) UPC　(d) NPAP

图 7　不同方法在数字域中的攻击样例

Figure 7　Attack examples of different methods in the digital domain

表 3 统计了不同对抗补丁在增强物理场景下的攻击结果。由表 3 可知, 对补丁进一步添加物理增强干扰以后, NPAP 方法生成的对抗补丁依然保持了较高的成功率, ASR 分别为 68.9%、75.3%、71.3%。而 G/C、NAP、UPC 三种方法生成的对抗补丁在三种目标检测器上的攻击成功率显著降低。这说明了 NPAP 方法在优化过程中添加的物理增强变换能够有效地应对这部分物理增强干扰, 对复杂的物理环境具有更好的鲁棒性。

表 3 物理增强场景下的攻击结果

Table 3 Attack results in enhanced physical environment

攻击方法	YOLOv3	YOLOv5	YOLOv7
G/C	55.2%	57.8%	58.9%
NAP	56.7%	66.8%	60.5%
UPC	65.8%	58.2%	57.4%
NPAP	**68.9%**	**75.3%**	**71.3%**

4.2.2 不同方法的复杂度对比

表 4 统计了不同方法在补丁优化过程中, 在训练集上迭代一轮的平均耗时和显存占用情况。从表 4 可以看到, NAP 方法迭代一次的平均耗时和平均显存占用最高, 这是由于 NAP 方法引入了生成对抗网络来生成对抗补丁, 在训练过程的前向传播和梯度下降过程中带来的额外的时间和空间消耗。而 G/C、UPC、NPAP 三种方法不需要引入新的网络, 计算复杂度主要与目标检测模型的参数量有关, 因此具有相近的时间和空间消耗。

表 4 不同方法复杂度对比

Table 4 Comparison of complexity of different methods

攻击方法	平均耗时/s	平均显存占用/Mb
G/C	1.55	2086.56
NAP	2.32	4096.20
UPC	1.62	2088.52
NPAP	1.56	2086.71

4.2.3 消融实验

1) 损失函数对算法性能及复杂度的影响

本节定量分析了 NPAP 方法损失函数构成对算法攻击成功率以及复杂度的影响。表 5 统计了 NPAP 方法在不同损失函数组合下所生成的对抗补丁对 YOLOv5 的攻击成功率。从表 5 可知, 仅使用目标损失时, NPAP 方法的攻击效果最好, ASR 为 78.5%。将目标损失与相似性损失或者总变差损失进行组合时, NPAP 方法的 ASR 有所下降, 但仍在可以接受的范围内。这是因为相似性损失和总变差损失用于控制对抗补丁的外观表现, 会对补丁的像素进行约束, 与目标损失存在一定的互斥关系。仅使用目标损失无法生成外观自然的对抗补丁, 因此将目标损失与相似性损失、总变差损失进行组合是为了平衡 NPAP 方法的攻击性能和外观表现。

表 5 不同损失函数下的攻击结果

Table 5 Attack results in different loss functions

损失函数构成	ASR
L_{obj}	78.5%
$L_{\text{obj}} + L_{\text{tv}}$	78.2%
$L_{\text{obj}} + L_{\text{simi}}$	77.5%
$L_{\text{obj}} + L_{\text{tv}} + L_{\text{simi}}$	77.6%

进一步分析损失函数对方法复杂度的影响, 本文统计了 NPAP 不同损失函数在训练集上迭代一轮的平均耗时以及显存占用情况, 结果如表 6 所示。

表 6 不同损失函数复杂度分析

Table 6 Analysis of complexity for different loss functions

损失函数构成	平均耗时/s	平均显存占用/Mb
L_{obj}	1.57	2086.72
$L_{\text{obj}} + L_{\text{tv}}$	1.56	2086.71
$L_{\text{obj}} + L_{\text{simi}}$	1.58	2086.72
$L_{\text{obj}} + L_{\text{tv}} + L_{\text{simi}}$	1.56	2086.71

从表 6 可以看到, 在不同损失函数组合下, NPAP 方法迭代一次的平均耗时和平均显存占用几乎相同。这是因为方法的时间复杂度主要受到模型前向传播以及梯度下降过程的影响, 即目标损失的计算过程, 这与目标检测器的参数量成正相关。相似度损失与总变差损失与补丁单独计算, 对方法复杂度的影响远小于目标损失的计算过程。

2) 模型置信度阈值对攻击结果的影响

为了进一步探究目标检测器的置信度阈值对攻击结果的影响, 本文设计了不同的置信度阈值进行测试, 测试结果如图 8 所示。可以看到, 在置信度阈值大于 0.5 时, 随着置信度阈值的提升, NPAP 方法在三种目标检测器上的攻击成功率逐渐提高。当阈值达到 0.7 以后, 攻击成功率的提升幅度减小。这表明在测试集中存在部分困难样本, 尽管对抗补丁能降低模型对这部分样本的检测置信度, 但其预测分数仍然高于检测器的阈值。

当置信度阈值小于 0.5 时, 攻击成功率随着置信

度阈值的减小而降低, 当置信度阈值小于 0.3 时, 攻击成功率显著降低。这说明了目标检测器的阈值在 0.3 时, NPAP 方法生成的对抗补丁仍具有较好的攻击效果, 进一步说明了 NPAP 方法中设计的目标损失有助于尽可能地降低模型预测框的置信度。

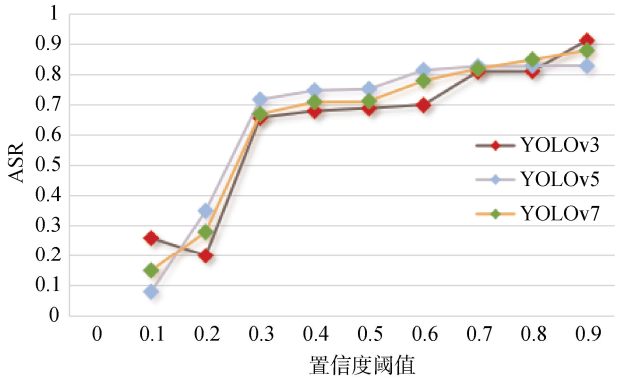


图 8 NPAP 方法不同置信度阈值下的 ASR

Figure 8 ASR with NPAP method at different confidence thresholds

4.3 物理域攻击实验

本节在现实世界中进行了实验, 将四种方法在数字域中生成的对抗补丁通过打印的方式转移到物理世界中。考虑到现实世界中汽车的实际尺寸和数字补丁的平衡, 对抗补丁打印的尺寸为 1m×1m, 图 9 为对抗补丁打印后的效果图。

图 9 对抗补丁打印后的效果图

Figure 9 The image after printing the adversarial patch

4.3.1 多高度、多方向攻击实验

1) 不同高度下的攻击结果

将对抗补丁依次放置在真实世界中的汽车顶部, 然后使用无人机从汽车正上方进行垂直拍摄。无人机从起始高度为 20m 的位置以 3m/s 的速度匀速上升至 100m 高度。本文将拍摄的视频根据高度进行切割, 每组视频进一步划分为(20～40m、40～60m、60～80m、80～100m)四组。将每组数据输入三个目标检测器进行检测, 并对检测结果进行逐帧处理以便于统计攻击成功率。表 7 统计了每组数据中覆盖了对抗补丁的视频帧数。

在物理攻击中, 仅对覆盖了对抗补丁的车辆进行统计, 将不同高度分组内攻击成功的视频帧数与总帧数之比作为平均攻击成功率(Average Attack Success Rate, AASR), 表达式如式(10)所示。

$$\mathrm{AASR}=\frac{N_s}{N_{\mathrm{all}}} \tag{10}$$

其中, N_s 表示当前分组中成功欺骗目标检测器的视频帧数, N_{all} 表示当前分组中视频的总帧数。

表 7 每组数据中覆盖了对抗补丁的视频帧数量

Table 7 The number of video frames with adversarial patch for each group

攻击方法	目标检测器	覆盖了对抗补丁的视频帧数				
		Group1 20-40m	Group2 40-60m	Group3 60-80m	Group4 80-100m	Total
G/C	YOLOv3	194	194	194	196	778
	YOLOv5	187	187	187	187	748
	YOLOv7	193	193	193	194	773
NAP	YOLOv3	185	185	185	185	740
	YOLOv5	192	192	192	192	768
	YOLOv7	190	190	190	191	761
UPC	YOLOv3	182	182	182	182	728
	YOLOv5	183	183	183	183	732
	YOLOv7	192	192	192	193	769
NPAP	YOLOv3	196	196	196	196	784
	YOLOv5	185	185	185	186	741
	YOLOv7	187	187	187	187	748

表 8 不同物理对抗补丁在不同高度范围内的攻击结果

Table 8 The attack results of different physical adversarial patches in different height ranges

攻击方法	目标检测器	平均攻击成功率(AASR)				
		Group1 20-40m	Group2 40-60m	Group3 60-80m	Group4 80-100m	Mean
G/C	YOLOv3	98.9%	91.0%	50.3%	0.6%	60.2%
	YOLOv5	95.8%	90.1%	43.2%	0.0%	57.3%
	YOLOv7	99.8%	70.2%	51.2%	0.8%	55.5%
NAP	YOLOv3	95.1%	56.0%	51.2%	2.5%	51.2%
	YOLOv5	93.9%	76.8%	40.0%	0.8%	52.9%
	YOLOv7	90.9%	70.1%	50.2%	1.0%	53.0%
UPC	YOLOv3	93.2%	64.7%	38.1%	1.1%	49.2%
	YOLOv5	95.8%	83.6%	38.5%	2.0%	55.0%
	YOLOv7	92.9%	82.1%	40.2%	1.5%	54.2%
NPAP	**YOLOv3**	**99.4%**	**90.2%**	**62.5%**	**2.5%**	**63.6 %**
	YOLOv5	**100.0%**	**83.5%**	**48.2%**	**1.6%**	**58.3 %**
	YOLOv7	**98.9%**	**80.6%**	**46.8%**	**1.0%**	**56.8 %**

表 8 为不同高度下的物理攻击结果。从表 8 可

知，对于三种目标检测器来说，四种方法生成的对抗补丁在分组 1 中的攻击效果远大于分组 2～4，AASR 达到 90%以上。而在分组 4 中，四种方法的攻击成功率显著降低。此外可以看到，NPAP 方法在 20～100m 范围的平均攻击成功率高于另外三种方法。

以上实验结果可以发现无人机的拍摄高度对物理对抗补丁的影响较大，随着无人机飞行高度的增加，物理对抗补丁的攻击成功率逐渐下降。主要的原因在于远距离拍摄会导致图像分辨率下降，对抗补丁在从物理空间转移到数字空间时，对抗补丁无法在图像中清晰地呈现。

图 10～图 13 为四种方法物理攻击结果的样例帧，同一行的图像为不同高度下的攻击结果。可以看到，在飞行高度分别为 20m、40m、60m 时，这些对抗补丁能够有效地隐藏目标。然而，在飞行高度上升至 80～100m 范围时，部分物理对抗补丁攻击失效。这主要归因于飞行高度的增加，对抗补丁在图像中的退化严重，后文将对高度带来的影响进行详细分析。

图 10　G/C 方法物理攻击结果的样例帧

Figure 10　Sample frames of physical attack results for G/C

图 11　NAP 方法物理攻击结果的样例帧

Figure 11　Sample frames of physical attack results for NAP

图 12　UPC 方法物理攻击结果的样例帧

Figure 12　Sample frames of physical attack results for UPC

图 13 NPAP 方法物理攻击结果的样例帧

Figure 13 Sample frames of physical attack results for NPAP

2) 不同拍摄方向下的攻击结果

本节进一步测试了四种方法所生成的物理对抗补丁在不同拍摄方向下的攻击结果。将对抗补丁放置在车顶后, 无人机分别从汽车正上方(摄像头与车身平行)、左侧(摄像头与车身垂直)、右侧(摄像头与车身垂直)三个不同的方向进行拍摄。表 9 统计了四种方法在 20m～100m 范围内的平均攻击成功率。

表 9 不同物理对抗补丁在不同方向下的攻击结果

Table 9 The attack results of different physical adversarial patches in different directions

攻击方法	目标检测器	拍摄方向		
		正上方	左侧	右侧
G/C	YOLOv3	60.2%	59.1%	60.0%
	YOLOv5	57.3%	57.2%	57.1%
	YOLOv7	55.5%	55.6%	55.0%
NAP	YOLOv3	51.2%	50.0%	51.5%
	YOLOv5	52.9%	52.1%	52.2%
	YOLOv7	53.0%	53.1%	52.0%
UPC	YOLOv3	49.2%	49.1%	48.5%
	YOLOv5	55.0%	55.5%	54.0%
	YOLOv7	54.2%	54.0%	54.1%
NPAP	**YOLOv3**	**63.6%**	**63.2%**	**63.2%**
	YOLOv5	**58.3%**	**58.1%**	**58.0%**
	YOLOv7	**56.8%**	**57.8%**	**56.2%**

从表 9 可知, 以不同方向进行拍摄时, 四种方法的攻击成功率变化较小。整体来看, 四种方法从正上方进行拍摄的攻击效果最好。进一步对结果进行定性分析, 图 14 为攻击结果的部分样例帧。

可以看到, 不同方向的拍摄主要使物理对抗补丁在图像中发生旋转, 对补丁的尺度和清晰度影响较小。由于每种方法在补丁优化过程中都添加了 EOT 变换中的旋转变换进行模拟, 补丁在不同方向拍摄的图像中仍具有较强的攻击效果。

图 14 不同方向下的样例帧

Figure 14 Sample frames of different directions

4.3.2 高度影响攻击结果的原因分析

从 4.3.1 节的实验结果可以发现, 无人机拍摄高度对补丁的攻击性能影响较大。本节对拍摄高度造成攻击性能下降的原因进行了分析。图 15 为其他场景下, NPAP 方法生成的对抗补丁在不同高度下的局部放大图, 可以看到, 随着无人机飞行高度的增加, 对抗补丁在视觉上的清晰度逐渐降低。

(a) 20 m (b) 40 m (c) 60 m (d) 80 m (e) 100 m

图 15 补丁在不同高度下的局部放大图

Figure 15 Local magnified images of the patch at different height

表 10 进一步统计了不同高度范围内, NPAP 方法生成的对抗补丁的像素量占图像总像素量的比例。从表 10 可以看到, 在飞行高度为 80～100 m 范围时, 对抗补丁像素量占比均不到 0.12%。从成像角度进行分析, 本文所采用的无人机影像传感器的尺寸为 1/2.3 英寸 (6.16 mm × 4.62 mm), 镜头的焦距为 24mm。图 16 为相机的成像关系, 公式(11)为视场的计算方式。

表 10 补丁在图像中的像素量占比

Table 10 Percentage of pixels occupied by the patch in the image

目标检测器	补丁的像素占比				
	20m	40m	60m	80m	100m
YOLOv3	1.27%	0.36%	0.16%	0.09%	0.07%
YOLOv5	1.32%	0.38%	0.15%	0.08%	0.05%
YOLOv7	0.95%	0.45%	0.23%	0.11%	0.08%

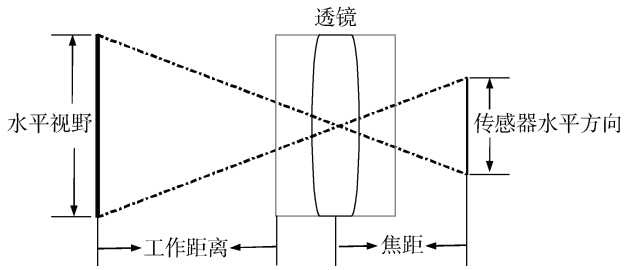


图 16 相机的成像关系

Figure 16 The imaging relationship of cameras

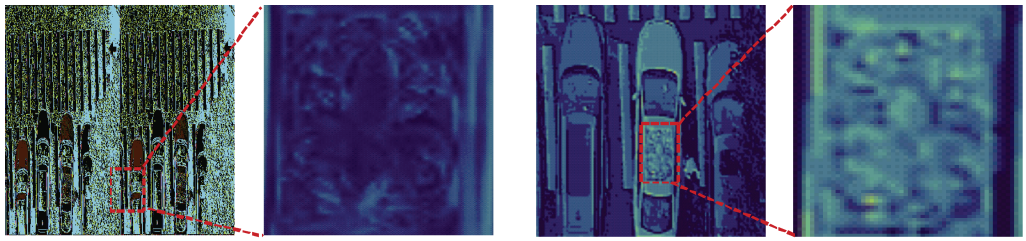
$$\text{视场(mm)} = \frac{\text{工作距离(mm)} \times \text{靶面尺寸(mm)}}{\text{镜头焦距(mm)}} \quad (11)$$

根据图 17 中的成像关系和公式(9)可以计算出, 在无人机飞行高度为 100m 时, 相机的视野范围约为 25.6m×19.3m。相机的分辨率为 3840×2160, 对抗补丁的实际尺寸为 1m×1m, 可以计算出此时对抗补丁在图像中所占的像素区域约为 150×112。图像在进入受害者模型之前, 会被缩放到一定的尺寸。以 YOLOv3 目标检测器为例, 图像在预处理阶段被缩放到 416×416, 经过预处理后, 对抗补丁在图像中所占的像素量约为 16×22, 此时单位像素在物理空间中所对应的区域已经达到了大约 28.4cm^2。从特征提取角度进行分析, YOLOv3 的 DarkNet53 特征提取网络采用多个大小为 3×3 的卷积核进行特征提取和下采样操作。图 17 为 DarkNet53 前两层输出的特征图, 这些特征图分别来自 20m 和 100m 高度下拍摄的图像, 图中红框表示对抗补丁所在的位置。

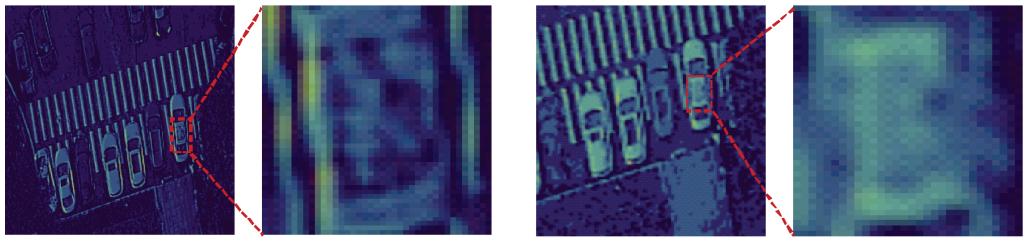
从图中可以看到, 20m 高度的图像经过卷积操作后, 神经网络的第一层和第二层特征图上能够捕捉到对抗补丁的特征。对于 100m 高度的图像, 由于对抗补丁在图像中的像素量较小, 3×3 的卷积核在补丁上的局部感受野相对较大, 神经网络第一层输出的特征图已经无法捕捉到对抗补丁中的细粒度特征, 因此攻击失效。以上分析说明了飞行高度的增加会使图像退化, 对抗补丁无法在低分辨率下清晰呈现, 导致对抗补丁具有诱导性的特征难以被神经网络提取到。

5 结束语

本文针对无人机目标检测系统, 提出了一种生



(a) 20 m 高度的对抗补丁在神经网络上的特征图



(b) 100 m 高度的对抗补丁在神经网络上的特征图

图 17 不同高度下拍摄的对抗补丁在神经网络中的特征图

Figure 17 Feature maps of adversarial patches captured at different heights in neural network

成自然物理对抗补丁的方法。通过引入相似度度量进行约束, 以生成具有自然特征的对抗补丁。同时, 为了增强对抗补丁在物理世界中的攻击性能和鲁棒性, 设计了针对多尺度、多目标攻击的优化函数, 以提升物理补丁的攻击能力。此外, 采用多种物理变换方式以应对现实场景中环境和尺度变换导致补丁攻击能力下降的问题。在实验部分, 本文使用无人机采集并制作了多尺度的车辆数据集, 用于优化对抗补丁。对于已训练好的对抗补丁, 分别在数字世界和物理世界进行了测试。数字攻击实验中, NPAP 方法与主流方法相比, 在攻击效果和鲁棒性上均有着明显的优势。物理攻击实验中, 将对抗补丁打印后放置在车顶, 并在不同拍摄高度、拍摄方向下成功欺骗三种不同的目标检测器。

本文方法在数字攻击实验中表现出强大的攻击效果, 但在物理实验中, 现实世界中复杂的因素仍然可能对实验结果产生一定的影响, 特别是拍摄高度的增加对物理对抗补丁的攻击性能影响较大。本文方法生成的对抗补丁在 20～60m 范围内具有较好的攻击效果, 但随着高度增加, 对抗补丁的攻击效果逐渐下降, 本文从计算成像和神经网络特征提取的角度进行了分析。本文方法是在白盒假设前提下生成的对抗补丁, 而白盒假设通常是过于理想化的, 因此在今后的研究中, 本文将从以下两个角度进行深入探讨: (1) 尝试提升方法的黑盒攻击能力, 进一步提升方法在无人机目标检测物理对抗攻击领域的实用性; (2) 探索生成 3D 物理对抗补丁的方法, 全覆盖于汽车车身, 更全面地应对因拍摄高度所引起的分辨率下降的情况。

参考文献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[2] Wang W, Dong J, He Z W, et al. A Brief Introduction to Visual Adversarial Samples[J]. *Journal of Cyber Security*, 2020, 5(2): 39-48.
(王伟, 董晶, 何子文, 等. 视觉对抗样本生成技术概述[J]. *信息安全学报*, 2020, 5(2): 39-48.)

[3] Lu M M, Li Q, Chen L, et al. Scale-Adaptive Adversarial Patch Attack for Remote Sensing Image Aircraft Detection[J]. *Remote Sensing*, 2021, 13(20): 4078.

[4] den Hollander R, Adhikari A, Tolios I, et al. Adversarial Patch Camouflage Against Aerial Detection[C]. *Artificial Intelligence and Machine Learning in Defense Applications II*, 2020: 11.

[5] Du A, Chen B, Chin T J, et al. Physical Adversarial Attacks on an Aerial Imagery Object Detector[C]. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022: 3798-3808.

[6] Zhang Y C, Zhang Y, Qi J H, et al. Adversarial Patch Attack on Multi-Scale Object Detection for UAV Remote Sensing Images[J]. *Remote Sensing*, 2022, 14(21): 5298.

[7] Chow K H, Liu L, Loper M, et al. Adversarial Objectness Gradient Attacks in Real-Time Object Detection Systems[C]. *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications*, 2020: 263-272.

[8] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples[C]. *International conference on machine learning*, 2018: 284-293.

[9] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6517-6525.

[10] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779-788.

[11] Wang C Y, Bochkovskiy A, Liao H M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors[C]. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 7464-7475.

[12] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. *Computer Vision – ECCV 2016*, 2016: 21-37.

[13] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.

[14] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2980-2988.

[15] Lu J, Sibai H, Fabry E. Adversarial examples that fool detectors[J]. arXiv preprint arXiv:1712.02494, 2017.

[16] Xie C H, Wang J Y, Zhang Z S, et al. Adversarial Examples for Semantic Segmentation and Object Detection[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 1378-1387.

[17] Li Y, Tian D, Chang M C, et al. Robust adversarial perturbation on deep proposal-based models[J]. arXiv preprint arXiv:1809.05962, 2018.

[18] Zhang H C, Wang J Y. Towards Adversarially Robust Object Detection[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 421-430.

[19] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.

[20] Liu X, Yang H, Liu Z, et al. Dpatch: An adversarial patch attack on object detectors[J]. arXiv preprint arXiv:1806.02299, 2018.

[21] Thys S, Ranst W V, Goedemé T. Fooling Automated Surveillance

Cameras: Adversarial Patches to Attack Person Detection[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019: 49-55.

[22] Wang Y J, Lv H R, Kuang X H, et al. Towards a Physical-World Adversarial Patch for Blinding Object Detection Models[J]. *Information Sciences*, 2021, 556: 459-471.

[23] Huang L F, Gao C Y, Zhou Y Y, et al. Universal Physical Camouflage Attacks on Object Detectors[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 717-726.

[24] Hu Y C T, Chen J C, Kung B H, et al. Naturalistic Physical Adversarial Patch for Object Detectors[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 7828-7837.

[25] Tan J, Ji N, Xie H D, et al. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World[C]. *The 29th ACM International Conference on Multimedia*, 2021: 5307-5315.

[26] Hu Z H, Huang S Y, Zhu X P, et al. Adversarial Texture for Fooling Person Detectors in the Physical World[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 13297-13306.

[27] Liu J, Lu B, Xiong M, et al. Adversarial Attack with Raindrops[J]. arXiv preprint arXiv:2302.14267, 2023.

[28] Qin Y X, Zhang K J, Pan H W. Adversarial Attack for Object Detectors under Complex Conditions[J]. *Computers & Security*, 2023, 134: 103460.

[29] Lian J W, Mei S H, Zhang S, et al. Benchmarking Adversarial Patch Against Aerial Detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5634616.

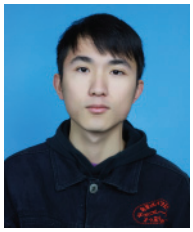
[30] Sava P A, Schulze J P, Sperl P, et al. Assessing the Impact of Transformations on Physical Adversarial Attacks[C]. *The 15th ACM Workshop on Artificial Intelligence and Security*, 2022: 79-90.

[31] Buslaev A, Iglovikov V I, Khvedchenya E, et al. Albumentations: Fast and Flexible Image Augmentations[J]. *Information*, 2020, 11(2): 125.

[32] Zhu P F, Wen L Y, Du D W, et al. Detection and Tracking Meet Drones Challenge[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7380-7399.

张恒 于 2019 年毕业于中国科学技术大学, 获计算机应用博士学位。现为重庆邮电大学计算机科学与技术学院/人工智能学院专任教师, 先进技术研究院副院长, 特种视觉联合实验室负责人, CCF 会员。研究领域为计算成像、人工智能安全。研究兴趣包括: 人工智能安全、特种机器视觉等。Email: zhangheng@cqupt.edu.cn

丁家松 于 2022 年毕业于成都理工大学, 获得网络空间安全学士学位。现就读于重庆邮电大学, 攻读计算机科学与技术硕士学位。研究领域为人工智能安全。研究兴趣包括: 对抗样本、目标检测、对抗防御等。Email: s2202011021@stu.cqupt.edu.cn



黄农森 于 2021 年毕业于西南科技大学, 获得过程装备与控制工程学士学位。现就读于重庆邮电大学, 攻读计算机技术硕士学位。研究领域为人工智能安全。研究兴趣包括: 计算机视觉、对抗样本等。Email: s210231076@stu.cqupt.edu.cn

杭芹 于 2019 年毕业于中国科学技术大学, 获核能科学与工程博士学位。现为重庆邮电大学计算机科学与技术学院/人工智能学院讲师, CCF 会员。主要研究方向为计算机视觉、图像处理。研究兴趣包括: 目标检测、对抗攻击等。Email: hangqin@cqupt.edu.cn