中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

面向空间视觉目标检测的对抗攻击与防御算法

周栋, 孙光辉, 吴立刚

引用本文:
周栋,孙光辉,吴立刚. 面向空间视觉目标检测的对抗攻击与防御算法[J]. 控制与决策, 2024, 39(7): 2161−2168.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1669

面向空间视觉目标检测的对抗攻击与防御算法

周 栋, 孙光辉†, 吴立刚

(哈尔滨工业大学 航天学院，哈尔滨 150001)

摘 要: 随着航天航空技术的发展,空间目标视觉检测技术作为航天器智能在轨服务的重要技术支撑,获得了国内外研究学者的广泛关注. 考虑到太空中恶劣的光照条件以及未知的动态场景,空间目标视觉检测的鲁棒性问题亟待深入研究. 首先,提出一种黑盒迁移实例攻击方法,将图像识别领域的对抗样本攻击方法应用于空间目标检测任务,实现对 EfficientDet 目标检测模型的欺骗攻击; 同时,提出一种协同防御策略,将对抗训练和 SRMNet 去噪器相结合,有效增强目标检测模型的鲁棒性. 实验结果表明,所提出防御策略不仅能够成功抵御对抗样本攻击,还能取得高于原始空间目标检测模型的检测精度.

关键词: 空间目标视觉检测；黑盒迁移实例攻击；对抗训练；SRMNet

中图分类号: TP274+.5 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1669

引用格式: 周栋, 孙光辉, 吴立刚. 面向空间视觉目标检测的对抗攻击与防御算法 [J]. 控制与决策, 2024, 39(7): 2161-2168.

Adversarial attack and defense algorithms towards space visual object detection

ZHOU Dong, SUN Guang-hui†, WU Li-gang

(School of Astronautics，Harbin Institute of Technology，Harbin 150001，China)

Abstract: With the development of aerospace technology, space object visual detection as the key methodology of intelligent on-orbit service of spacecraft has garnered broad concerns. Considering the extreme illumination condition and unknown scenario dynamics, the robustness problem of space object visual detection is urged to be studied in depth. This paper proposes a black-box transferred instance attack, which applies adversarial attacks in image classification domain to the task of space object visual detection. It succeeds to fool the EfficientDet model. Meanwhile, we further put forward a cooperative defense strategy that combines adversarial training with the SRMNet denoiser, which effectively enhances the robustness of the object detector. Experimental results show that this defense strategy not only resists adversarial attacks successfully, but also makes a great improvement on the accuracy of the space object detection model.

Keywords: space object detection；black-box transferred instance attack；adversarial training；SRMNet

0 引 言

随着航空航天技术的快速发展,航天器智能在轨服务成为了近年来的研究热点[1],其应用范畴涵盖废弃卫星维护[2]、自主交会对接[3]和空间碎片清除[4]等诸多方面. 而空间在轨服务智能化水平的关键之处以及技术难点在于如何提升目标检测算法的准确性和鲁棒性.

多年以来,许多国内外学者针对空间目标检测算法的准确性问题进行了深入研究,涌现出各式各样的基于机器学习的视觉目标检测方法[5]. 最近,与深度学习相结合的空间目标检测算法,在准确性方面取得了令人印象深刻的结果[6-8]. 文献[9]提出了一个用于航天器检测、分割以及部件识别的数据集,并基于该数据集评估了 YOLOV 3、YOLOV 4 以及 EfficientDet 模型[10]在空间视觉检测任务中的实际性能; 文献[11]构造了一个空间非合作目标检测数据集(space non-cooperative object detection, SNCOD),并将计算机视觉领域最先进的 8 种目标检测模型在 SNCOD 数

收稿日期: 2022-09-20；**录用日期:** 2023-06-16.
基金项目: 国家自然科学基金面上项目 (62173107).
责任编委: 刘德荣.
†通讯作者. E-mail: guanghuisun@hit.edu.cn.

据集上进行了充分评估. 实验结果表明,EfficientDet能够取得高达0.977的平均检测精度(mean average precision, mAP). 然而,目前针对空间检测模型鲁棒性问题的研究尚处于空白状态. 文献[12]曾提出图像分类模型无论是基于传统机器学习,还是基于深度神经网络,均可通过在输入图像上添加不可察觉的扰动,对其进行欺骗攻击. 这类使得模型预测错误的输入图像称为对抗样本,而这个过程叫作对抗攻击.

随着关于对抗样本的研究逐渐深入,对抗攻击的概念不再局限于图像分类任务[13],如目标检测[14]、视频跟踪[15]、图像分割[16]等领域也出现了大量的对抗样本攻击方法,并在现实世界中完成了实验验证[17-18]. 此外,文献[19]验证了自然场景中的干净图像,由于数据分布不一致问题和场景类比相关性,也会欺骗深度神经网络,导致图像分类和目标检测任务失败. 考虑到太空中恶劣的光照条件以及未知的动态场景,这为日后空间目标检测模型提出了非常高的要求,因此,空间目标检测模型的鲁棒性问题亟待进一步研究.

目前,目标检测的对抗样本攻击方法一般针对检测模型的重要构成模块进行对抗损失函数设计,如区域建议网络[20-21]、非极大值抑制[22]等,从而迭代生成对抗样本. 由于目标检测模型的复杂性,即需要同时定位目标位置并识别目标类别,大部分攻击方法效果不佳,同时无法实现快速攻击.

考虑到图像分类领域有大量对抗样本攻击的相关研究,若能够将这些工作有效迁移至目标检测领域,则将会有力地推动目标检测模型鲁棒性研究的发展,也能为对抗样本攻击在计算机视觉领域形成统一的理论框架提供支撑. 因此,本文遵循这一思路,根据DPatch算法[23]框架,提出一种黑盒迁移实例攻击算法(black-box transferred instance attack, BTIA),成功地将图像识别领域的诸多对抗攻击方法: FGSM[24]、PGD[25]、EOT[26]、Jitter[27]等直接迁移至目标检测任务中,实现了对EfficientDet目标检测模型的有效攻击.

针对上述所提出攻击方法,本文进一步地提出一种协同防御策略:将常用的对抗防御方法——对抗训练[24]和SRMNet去噪器[28]相结合,通过对抗训练提升模型的自身鲁棒性,利用SRMNet滤除对抗噪声,最终实现空间目标检测的协同防御. 实验结果表明,该防御算法不仅能够有效解决对抗样本攻击带来的性能下降问题,还能进一步提升相对于原始检测模型的准确性.

本文的主要内容如下.

1) 提出一种黑盒迁移实例攻击算法,将图像识别领域的诸多对抗攻击算法迁移至空间目标检测任务中,实现对EfficientDet的对抗攻击.

2) 提出一种针对目标检测模型的协同防御策略,通过结合对抗训练方法和SRMNet去噪器,不仅成功抵御了对抗样本攻击,还进一步提升了EfficientDet在干净数据集上的检测精度.

1 相关工作

1.1 目标检测的对抗攻击

根据对抗扰动作用在图像上的范围,主流的目标检测对抗攻击方法可分为全局攻击和局部攻击两大类. 全局对抗攻击需要对待检测图像的所有像素进行修改,经典全局对抗攻击算法包括稠密对抗生成法[29]、鲁棒对抗扰动法[30]等;而局部对抗攻击仅在待检测图像的某个区域内添加对抗噪声,进而改变目标检测模型的预测结果,主要方法有DPatch[23]以及蒸发攻击[31](evaporate attack, EA),所提出BTIA攻击方法也属于局部对抗攻击范畴.

大部分局部对抗攻击方法本质上还是针对目标检测模型的重要部件进行对抗损失函数设计,因此一般属于白盒攻击方法. 而文献[31]提出的蒸发攻击方法,是目标检测领域为数不多的黑盒局部对抗攻击方法. 该算法通过构造仅依赖于检测器输出的损失函数,利用改进的粒子群算法进行对抗扰动迭代生成,实现了对单阶段检测器和两阶段检测器的有效欺骗. 但是由于粒子群优化算法的局限性,EA算法需要设定较大的粒子数并经过上千次的迭代优化才能取得较好的攻击效果,这不可避免地导致该方法实时性很差,同时还易陷入局部优化问题. 而所提出BTIA攻击方法基于被攻击检测模型的推理结果,遵循模型替代以及任务简化的思路,直接利用简单的深度图像分类模型对检测实例图像进行对抗样本生成,有效避免了EA算法的问题,并最终取得了非常好的攻击成功率和算法实时性.

1.2 目标检测的对抗防御

在目标检测领域中,目前针对如何抵御对抗样本攻击的研究非常少. 现有的目标检测对抗防御方法主要分为两大类: 预处理防御方法以及模型鲁棒增强方法. 预处理防御方法将对抗扰动视为一种特殊的对抗噪声,利用滤波器[32]、去噪器[33]以及图像压缩[34]等方法,减轻对抗扰动对目标检测模型的影响,

该类方法具有即插即用的特点,但是其防御效果以及可迁移能力仍然存在明显的不足. 模型鲁棒增强方法从神经网络可解释性角度出发,旨在利用某种手段改变原有目标检测模型的参数值,使其具备更强的抗干扰能力. 典型的目标检测鲁棒增强方法主要有对抗训练方法[35]、正则化方法[36]等. 所提出协同对抗防御策略从上述两个方面同时着手,利用SRMNet网络大幅消除对抗扰动,同时结合对抗训练进行模型鲁棒增强,最终有效抵御高强度对抗样本攻击,并进一步提升原始模型的检测精度.

2　预备知识

2.1　EfficientDet目标检测模型

主流的深度目标检测器通常可分为两阶段目标检测器和单阶段目标检测器[37]. 两阶段目标检测模型首先生成大量区域建议框,再进行多层级特征提取,最后对各建议框对应的特征进行分类以及回归. 该方法一般具有较高的检测精度. 而单阶段目标检测模型直接从输入图像生成预测框,并同时进行分类和回归,因此,其具有较高的推理速度.

文献[11]验证了EfficientDet单阶段目标检测器在SNCOD数据集上取得了非常好的检测精度. 因此,本文将其作为空间目标视觉检测鲁棒性研究的基础模型. 如图1所示,EfficientDet主要包括EfficientNet骨干网络、双向特征金字塔网络以及预测网络3部分. 其中: EfficientNet骨干网络用于提取多层级特征,即$\{P_1, P_2, \ldots, P_8\}$;双向特征金字塔网络负责融合多层级特征,以实现快速可靠的目标检测;而预测网络分为类别预测网络和包围框预测网络,分别用于各有效特征层级$\{P_3, P_4, \ldots, P_7\}$预设先验框的类别判断以及大小调整.

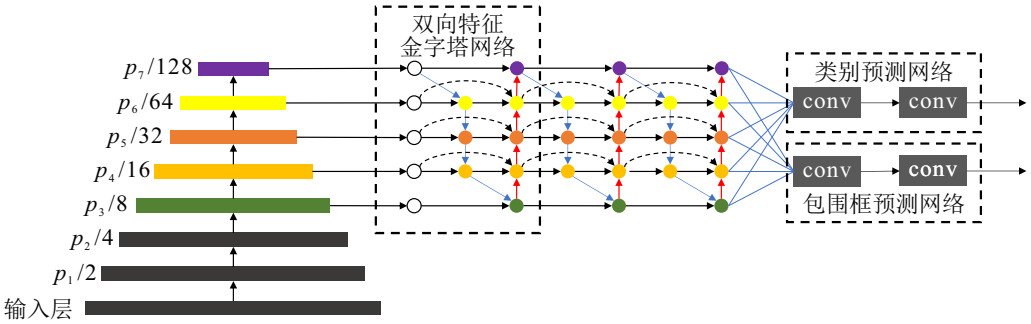


图 1　EfficientDet网络框架

2.2　对抗样本攻击

目前,已有大量针对图像分类模型的对抗样本攻击方法,若能够将这些方法迁移至目标检测领域,则将大幅推动目标检测模型鲁棒性问题的研究. 因此,本文考虑最典型的5种对抗样本攻击方法: 高斯噪声、快速梯度符号法[24]、抖动法[27](jitter)、投影梯度下降法[25]以及空间变换期望法[26](expectation over transformation, EOT).

2.2.1　高斯噪声

对输入图像加入服从正态分布的噪声,加入噪声后的像素值限制在$[0, 255]$间.

2.2.2　快速梯度符号法

FGSM是最先提出来的一种对抗样本攻击方法,它的核心思想是求取对抗损失函数对输入图像的梯度,使得模型预测值向某一指定类别移动,其算法可由下式表示:

$$\hat{x} = x + \epsilon \text{sign} \nabla_x L(x, y; \theta). \tag{1}$$

其中: $x \in R^{W \times H}$为输入图像,y为输入图像的类别标签,ϵ为噪声幅值项,$L(x, y; \theta)$为构造的对抗损失函数,θ为模型参数.

2.2.3　抖 动 法

Jitter攻击构造了如下目标函数L_J:

$$L_J(z, y) = \begin{cases} \dfrac{\|\hat{z} - y + N(0, \sigma)\|_2}{\|\gamma\|_p}, & f(\hat{x}) \neq f(x); \\ \|\hat{z} - y + N(0, \sigma)\|_2, & f(\hat{x}) = f(x). \end{cases} \tag{2}$$

其中: $\hat{z} = \text{softmax}\left(\alpha \dfrac{z}{\|z\|_\infty}\right)$,$z$为网络模型输出值,$\alpha$为输出缩放因子,旨在将网络模型输出值缩放至一定范围内,进而提高对抗样本攻击的成功率;y为原始输入图像x的类别标签;$N(0, \sigma)$为额外引入的高斯噪声,以实现从不同梯度方向探索对抗扰动;$\hat{x} = x + \gamma$为迭代对抗样本;γ为对抗扰动. 当对抗样本$\hat{x}$的预测值与输入图像x的预测值出现差异时,即$f(\hat{x}) \neq f(x)$,Jitter通过利用对抗扰动的p范数对损失函数进行归一化,进而生成人眼不可察觉的对抗噪声.

2.2.4　投影梯度下降法

PGD首先在原始输入图像允许范围内进行随机初始化搜索,然后多步迭代产生对抗样本,每步的对

抗扰动均被截断到规定范围内,其迭代表达式如下式所示:

$$x^{i+1}=\underset{x+\delta}{\text{Proj}}\{x^i+\alpha\text{sign}(\nabla_x L(x,y;\theta))\}. \tag{3}$$

其中:x^i 为第 i 次迭代的对抗样本,α 为对抗扰动幅度,$L(x,y;\theta)$ 为构造的对抗损失函数,θ 为模型参数.

2.2.5 空间变换期望法

PGD 解决了约束优化问题,但是其不具有变换鲁棒性,即其生成的对抗样本在经过一定的空间变换后,不再具有对抗欺骗性. 为了解决这一问题,EOT 构造了如下优化问题:

$$\arg\max_{\hat{x}} \boldsymbol{E}_{t\sim T}[L(t(\hat{x}),y;\theta)];$$

$$\text{s.t. } \boldsymbol{E}_{t\sim T}[d(t(\hat{x}),t(x))]<\epsilon. \tag{4}$$

其中:$\hat{x}$ 为生成的对抗样本;x 为原始输入图像;y 为原始输入图像对应的真值;空间变换函数 t 隶属于空间变换集 T;此外,EOT 引入了距离度量函数 d,用于计算对抗样本与原始输入经空间变换 t 后的有效距离,而不再是两者间简单的差值;ϵ 用来控制期望有效距离.

3 黑盒迁移实例攻击

第 2 节对图像识别领域最典型的 5 种对抗样本攻击方法进行了介绍. 为了将上述方法迁移至目标检测领域,本文提出一种黑盒迁移实例攻击算法,成功实现了对 EfficientDet 网络的攻击,其原理见图 2.

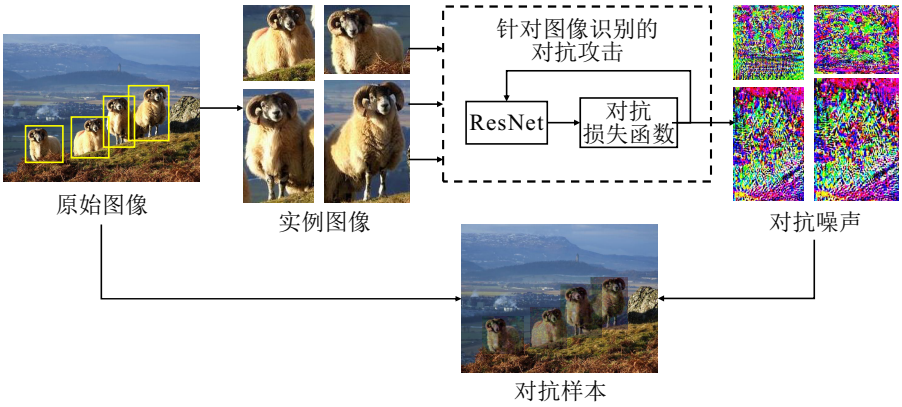


图 2 黑盒迁移实例攻击算法原理

首先,本文利用图像标注真值或目标检测模型获取原始图像中各目标实例的相关信息(类别、位置和大小),进而通过图像预处理操作生成固定尺寸的实例图像;然后,使用在 ImageNet 数据集上预训练的图像识别模型 ResNet-50[38],结合多种对抗攻击方法,同步生成多实例图像的对抗噪声;最后,将对抗干扰依据相关实例信息添加回原图像,得到适用于视觉目标检测任务的对抗样本.

所提出的黑盒迁移实例攻击方法不存在对目标检测模型架构和参数的依赖,仅利用常用的图像识别网络以及相应的对抗攻击方法,可实现对复杂检测模型的欺骗攻击.

4 协同免疫防御算法

本文提出一种结合对抗训练[24] 以及 SRMNet 去噪器[28] 的协同免疫防御算法,其算法流程如图 3 所示. 下面对 SRMNet 去噪器和对抗训练进行详细介绍.

4.1 对抗训练

对抗训练是应用最广泛、目前最有效的对抗防御机制之一. 该方法将受到攻击的样本加入训练集中对网络进行训练,从而使得网络具有一定的防御能力. 上述过程可描述为以下表达式:

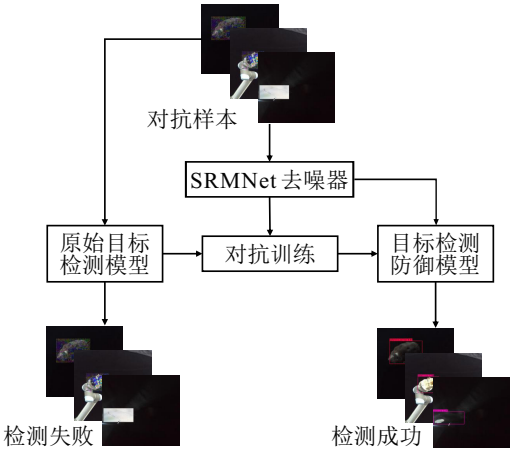


图 3 协同免疫防御算法

$$\min_{\theta} \boldsymbol{E}_{(x,y)\in D}\big[\max_{\|\delta\|_\infty\leqslant\epsilon} L(x+\delta,y;\theta)\big]. \tag{5}$$

其中:x 为输入数据;y 为输入样本的真实标签;δ 为对抗噪声;ϵ 为允许的扰动范围,为了保证生成的对抗扰动 δ 是人类难以察觉的. 式 (5) 内层表示对抗样本的生成过程,即通过对输入图像添加微小的扰动,使得对抗损失函数最大;外层表示利用对抗样本进行训

练, 使得模型的预测经验风险最小, 即保证模型在遇到对抗样本攻击时仍然具有良好的鲁棒性.

对抗训练的基本流程如下: 1) 遍历 SNCOD 训练集, 以一定的概率利用 BTIA 方法生成 5 种不同的对抗样本, 并将其加入 SNCOD 训练集中; 2) 加载原始目标检测模型权重, 利用小批量训练方法, 先冻结骨干网络训练 50 轮, 再恢复骨干网络训练 100 轮, 获得最终更鲁棒的目标检测防御模型.

4.2　SRMNet 去噪器

考虑到对抗干扰实际上也是图像噪声, 因此, 自然而然地想到通过去噪器对对抗样本进行降噪处理, 使得对抗样本恢复为干净样本, 以抵御对抗攻击. 但是由于对抗噪声所具有的特殊形式, 传统去噪器的效果并不理想, 如高斯滤波、中值滤波等. 因此, 本文考虑利用编码-解码的图像复原框架, 引入 SRMNet 网络, 以无监督训练形式, 获得具有广谱对抗防御能力的深度去噪器.

SRMNet 的网络结构如图 4 所示. 它首先对输入图像进行双线性下采样, 得到 4 层图像金字塔; 然后, 利用 3×3 的共享卷积核对图像金字塔的不同层级进行初级特征提取; 接着, SRMNet 使用选择性残差块 (selective residual block, SRB) 提取高级语义信息; 同时, 底层特征张量通过像素解混操作 (pixel unshuffle) 进行下采样, 并进一步与上层语义信息融合; 最后, SRMNet 使用对称镜像结构, 实现自上而下的特征融合和恢复, 再利用选择性核特征融合模块 (selective kernel feature fusion, SKFF) 整合特征金字塔, 完成去噪图像的重建.

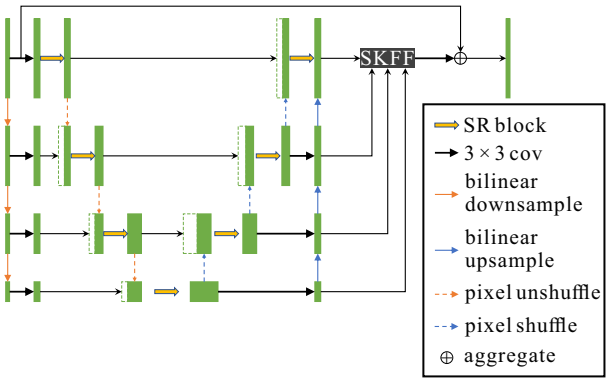


图 4　SRMNet 网络架构

本文通过如下目标损失函数 L_{deno} 对 SRMNet 去噪器进行训练:

$$L_{\text{deno}}=\sqrt{\|\tilde{x}-x\|^2+\epsilon^2}. \tag{6}$$

其中: $\tilde{x}$ 为去噪后的图像; x 为原始输入图像; ϵ 常量用于控制重建程度, 一般取 10^{-3}.

5　实验验证

本文依托于空间非合作目标检测数据集, 对前文所介绍的黑盒迁移实例攻击算法以及协同免疫防御方法进行了充分的实验, 有效地验证了所提出方法的有效性. SNCOD 数据集包含了 3 类非合作目标: C 类小行星、S 类小行星以及航天器, 共 2 337 张地面模拟图片. 其中: 训练集 1 501 张, 验证集 480 张, 测试集 356 张. 原始的 EfficientDet 目标检测模型是利用 SNCOD 训练集和 VOC 2007 数据集训练得到的, 而后续所有的对抗样本攻击和防御实验仅在 SNCOD 测试集上进行. 本文用于生成 BTIA 对抗扰动的图像分类器为 ResNet-50.

本文实验所采用的两个评估指标——攻击成功率和防御成功率的定义如下式所示:

$$\text{攻击成功率}=\frac{(\text{mAP}_v-\text{mAP}_o)}{\text{mAP}_o}, \tag{7}$$

$$\text{防御成功率}=\frac{(\text{mAP}_d-\text{mAP}_v)}{(\text{mAP}_o-\text{mAP}_v)}. \tag{8}$$

其中: mAP_o 为检测模型的原始平均检测精度, mAP_v 为检测器遭受攻击时的平均检测精度, mAP_d 为模型在引入防御方法的情况下所取得的平均检测精度.

5.1　黑盒迁移实例攻击验证

首先, 将 5 种图像识别领域的对抗样本攻击方法: FGSM[24]、PGD[25]、Jitter[27]、EOT[26]、高斯噪声, 引入黑盒迁移实例攻击算法中. 然后, 将生成的 5 种对抗样本分别单独添加至 SNCOD 测试集中. 此外, 本文还以均匀概率随机抽取一种基于 BTIA 攻击框架的对抗样本生成方法, 对 SNCOD 测试集进行攻击, 称为 BTIA 混合攻击. 最后, 利用预先在干净数据集上训练的 EfficientDet 模型进行实验评估, 并以平均检测精度作为评估指标, 最终结果如表 1 所示.

表 1　黑盒迁移实例攻击评估结果

	干净样本	BTIA FGSM	BTIA PGD	BTIA Jitter	BTIA EOT	BTIA 高斯	BTIA 混合	EA
平均检测精度	0.707	0.333	0.330	0.313	0.380	0.557	0.402	0.396
攻击成功率/%	–	52.9	53.3	55.7	46.3	21.2	43.1	43.9

由表 1 可见, 所提出的黑盒迁移实例攻击方法, 无论基于何种对抗样本攻击算法, 均能够取得较好的攻击成功率, 尤其是基于 Jitter 的黑盒实例攻击方法能够将目标检测模型的平均精度降低至 0.313. 正如

文献[12]提到的,传统噪声对深度神经网络的性能影响非常微弱,如高斯噪声、椒盐噪声等. 因此,基于高斯噪声的黑盒实例攻击并未取得非常好的攻击效果.

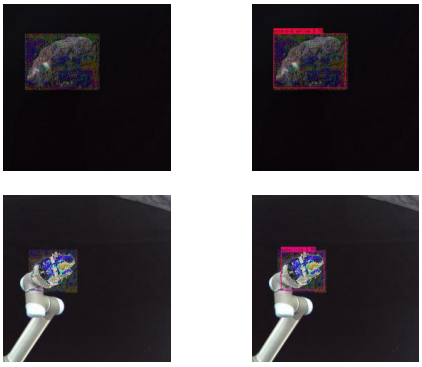
本文进一步地选用当前目标检测领域最先进的黑盒局部攻击算法——EA[31]作为对比. 实验结果表明,在SNCOD数据集上EA算法对EfficientDet目标检测模型的成功率仅为43.9%,远低于所提出BTIA Jitter算法所达到的55.7%攻击成功率. 此外,EA算法受限于粒子群优化算法,需要上千步的迭代过程才能取得较好的攻击效果,故其在算法实时性上较所提出方法仍然略逊一筹.

5.2 协同免疫防御验证

首先,单独评估对抗训练对于黑盒迁移实例攻击的防御效果,具体流程见第2.1节. 将不同的BTIA对抗样本添加至SNCOD测试集,评估经对抗训练后的EfficientDet模型对于不同攻击方法的防御效果,最终结果见表2的第3列. 可以发现,对抗训练在一定程度上能够抵挡BTIA攻击,尤其是基于Jitter方法的黑盒迁移实例攻击,其平均检测精度甚至能够达到0.737. 但是面对基于FGSM的BTIA攻击时,对抗样本训练的防御成功率仅有62.6%. 对抗训练部分防御效果如图5所示. 然后,本文再单独评估SRMNet去噪器对于黑盒迁移实例攻击的防御能力. 同样,利用BTIA方法对SNCOD训练集中的图像进行攻击,并将其与原始图像配对,作为SRMNet的训练样本. 本文中的SRMNet基于以上训练数据训练100个轮次,批训练大小设置为8,初始学习率为10^{-4}. 得到训练好的SRMNet网络后,将其作为EfficientDet的前置去噪模块,在SNCOD测试集上,对于不同的BTIA对抗攻击方法进行评估验证,最终结果见表2的第4列. 实验结果表明,SRMNet去噪器对于对抗样本攻击也具有一定的防御能力,但是相较于对抗训练而言,其防御效果一般. SRMNet去噪器部分防御效果如图6所示. 由图6可见,图像经SRMNet处理后,BTIA攻击所生成的对抗噪声基本被滤除,检测模型能够以较高的置信度识别出空间目标. 最后,本文对之前所提出的协同免疫防御方法进行了实验验证,其流程如图3所示. 先利用SRMNet进行去噪,再通过对抗训练获得鲁棒性更高的目标检测防御模型. 协同免疫防御方法的评估结果见表2的第5列,发现它具有非常优秀的对抗防御性能,即使同时面对5种不同的BTIA攻击,使用协同防御方法的EfficientDet网络也能取得高达0.75的平均检测精度,而原始检测模型在干净样本上的检测精度仅有0.707. 之所以协同免疫防御方法取得相较于原始模型更高的检测精度,是由于两方面原因: 1) SRMNet去噪器通过编码-解码架构改变了图像数据的分布,并在高维特征空间中减小了去噪图像与原始图像的差异; 2) 相较于原始模型,目标检测防御利用去噪后的图像进行了对抗训练,利用小步长迭代更新的优势,探索到了更优的局部极值.

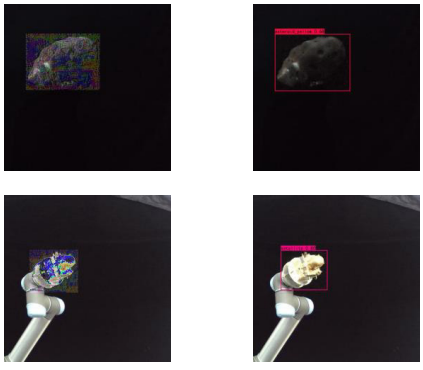
表 2 协同防御算法评估结果

	原始模型	对抗训练	SRMNet 去噪	协同免疫防御
干净样本	0.707	0.571	0.690	0.717
BTIA FGSM	0.333	0.567	0.573	0.737
BTIA PGD	0.330	0.687	0.604	0.754
BTIA Jitter	0.313	0.732	0.582	0.752
BTIA EOT	0.380	0.689	0.605	0.758
BTIA 高斯	0.557	0.699	0.705	0.721
BTIA 混合	0.402	0.670	0.660	0.750
EA	0.396	0.504	0.615	0.683



(a) 对抗防御前 (b) 对抗防御后

图 5 对抗训练防御效果



(a) SRMNet 去噪前 (b) SRMNet 去噪后

图 6 SRMNet去噪器防御效果

6 结 论

本文从对抗攻击和防御两方面着手,对空间目标检测模型的鲁棒性问题进行了深入研究. 首先,本文提出了一种黑盒迁移实例攻击算法,将图像识别领域的多种对抗攻击算法推广至目标检测任务,实现了对EfficientDet检测模型的攻击. 其中:基于Jitter的黑盒迁移实例攻击算法甚至能够达到55.7%的攻击成功率,这对空间在轨任务的稳定开展带来了

严重威胁. 因此,本文进一步提出了结合对抗训练和SRMNet去噪器的协同免疫防御方法. 实验结果表明,该防御方法不仅能够完全抵御不同的高强度对抗攻击,还会进一步提升原始检测模型的平均检测精度.

然而,所提出攻击与防御方法尚存在某些方面的不足,值得在之后的工作中进一步研究: 1) 黑盒迁移实例攻击方法仅引入了5种较为基础的对抗样本攻击方法. 目前还有其他更有效的攻击算法可以考虑与其结合,如C & W[39]、One-Pixel[40]、DeepFool[41]等. 2) 由于考虑到EfficientDet检测模型在空间非合作目标检测数据集上取得了最佳成绩,本文仅针对EfficientDet网络做了实验验证. 所提出对抗攻击和协同防御方法在其他主流的目标检测模型上的实际效果,如YOLO系列[42]、SSD[43]以及FCOS[44]等,将在后续工作中作进一步评估.

参考文献(References)

[1] 岳程斐, 孙英杰, 柳子然, 等. 基于一致性理论的多臂航天器协同控制方法[J]. 控制与决策, 2023, 38(5): 1430-1437.
(Yue C F, Sun Y J, Liu Z R, et al. Cooperative control method of multi-arm spacecraft based on consistency theory[J]. Control and Decision, 2023, 38(5): 1430-1437.)

[2] 聂媛媛, 方志耕, 刘思峰, 等. 基于节点修复的低轨卫星网络动态抗毁性模型[J]. 控制与决策, 2020, 35(5): 1247-1252.
(Nie Y Y, Fang Z G, Liu S F, et al. Dynamic invulnerability model of LEO satellite network based on node repair[J]. Control and Decision, 2020, 35(5): 1247-1252.)

[3] Zhou D, Sun G H, Lei W X, et al. Space noncooperative object active tracking with deep reinforcement learning[J]. IEEE Transactions on Aerospace and Electronic Systems, 2022, 58(6): 4902-4916.

[4] 艾海平, 陈力. 空间机器人捕获航天器操作的避撞柔顺复合自抗扰控制[J]. 控制与决策, 2021, 36(2): 355-362.
(Ai H P, Chen L. Collision avoidance and compliant composite active disturbance rejection control of space robot capture spacecraft[J]. Control and Decision, 2021, 36(2): 355-362.)

[5] Zhou D, Tian Y X, Li X, et al. ORB-based template matching through convolutional features map[C]. Chinese Automation Congress. Hangzhou, 2020: 4695-4699.

[6] Huan W X, Liu M M, Hu Q L. Pose estimation for non-cooperative spacecraft based on deep learning[C]. The 39th Chinese Control Conference. Shenyang, 2020: 3339-3343.

[7] Dumitrescu F, Ceachi B, Truică C O, et al. A novel deep learning-based relabeling architecture for space objects detection from partially annotated astronomical images[J]. Aerospace, 2022, 9(9): 520.

[8] Mahendrakar T, White R T, Wilde M, et al. SpaceYOLO: A human-inspired model for real-time, on-board spacecraft feature detection[J/OL]. 2023, arXiv: 2302.00824.

[9] Dung H A, Chen B, Chin T J. A spacecraft dataset for detection, segmentation and parts recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, 2021: 2012-2019.

[10] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 10778-10787.

[11] Miao L N, Zhou D, Li X. Spatial non-cooperative object detection based on deep learning[C]. China Automation Congress. Beijing, 2022: 116-121.

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

[13] 孔锐, 蔡佳纯, 黄钢, 等. 基于复合生成对抗网络的对抗样本生成算法研究[J]. 控制与决策, 2023, 38(2): 528-536.
(Kong R, Cai J C, Huang G, et al. Research on generative adversarial example algorithm based on multiple GANs[J]. Control and Decision, 2023, 38(2): 528-536.)

[14] 袁珑, 李秀梅, 潘振雄, 等. 面向目标检测的对抗样本综述[J]. 中国图象图形学报, 2022, 27(10): 2873-2896.
(Yuan L, Li X M, Pan Z X, et al. Review of adversarial examples for object detection[J]. Journal of Image and Graphics, 2022, 27(10): 2873-2896.)

[15] Jia S, Ma C, Song Y, et al. Robust tracking against adversarial attacks[C]. European Conference on Computer Vision. Online, 2020: 69-84.

[16] 张宇. 不同目标先验下的视频目标分割及其对抗攻击算法研究[D]. 杭州: 杭州电子科技大学, 2022.
(Zhang Y. Video object segmentation with different object priors and its adversarial attack algorithm[D]. Hangzhou: Hangzhou Dianzi University, 2022.)

[17] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 1625-1634.

[18] Braunegg A, Chakraborty A, Krumdick M, et al. APRICOT: A dataset of physical adversarial attacks on object detection[C]. European Conference on Computer Vision. Online, 2020: 35-50.

[19] Hendrycks D, Zhao K, Basart S, et al. Natural adversarial examples[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville,

2021: 15257-15266.

[20] Xie C H, Wang J Y, Zhang Z S, et al. Adversarial examples for semantic segmentation and object detection[C]. IEEE International Conference on Computer Vision. Venice, 2017: 1378-1387.

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

[22] Wang D R, Li C R, Wen S, et al. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples[J]. IEEE Transactions on Cybernetics, DOI: 10.1109/TCYB.2020.3041481.

[23] Liu X, Yang H, Liu Z, et al. DPatch: An adversarial patch attack on object detectors[J/OL]. 2018, arXiv: 1806.02299.

[24] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J/OL]. 2014, arXiv: 1412.6572.

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

[26] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples[J/OL]. 2017, arXiv: 1707.07397.

[27] Schwinn L, Raab R, Nguyen A, et al. Exploring misclassifications of robust neural networks to enhance adversarial attacks[J/OL]. 2021, arXiv: 2105.10304.

[28] Fan C M, Liu T J, Liu K H. Selective residual M-net for real image denoising[J/OL]. 2022, arXiv: 2203.01645.

[29] Xie C H, Wang J Y, Zhang Z S, et al. Adversarial examples for semantic segmentation and object detection[C]. IEEE International Conference on Computer Vision. Venice, 2017: 1378-1387.

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

[31] Wang Y J, Tan Y A, Zhang W J, et al. An adversarial attack on DNN-based black-box object detectors[J]. Journal of Network and Computer Applications, 2020, 161: 102634.

[32] Chiang P Y, Curry M J, Abdelkader A, et al. Detection as regression: Certified object detection by median smoothing[J/OL]. 2020, arXiv: 2007.03730.

[33] Akhtar N, Liu J, Mian A. Defense against universal adversarial perturbations[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 3389-3398.

[34] Duan R J, Chen Y F, Niu D T, et al. AdvDrop: Adversarial attack to DNNs by dropping information[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 7486-7495.

[35] Zhang H C, Wang J Y. Towards adversarially robust object detection[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 421-430.

[36] Bouabid S, Delaitre V. Mixup regularization for region proposal based object detectors[J/OL]. 2020, arXiv: 2003.02065.

[37] Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: A review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[38] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[39] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]. IEEE Symposium on Security and Privacy. San Jose, 2017: 39-57.

[40] Su J W, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[41] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 2574-2582.

[42] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J/OL]. 2018, arXiv: 1804.02767.

[43] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[J/OL]. 2015, arXiv: 1512.02325.

[44] Tian Z, Shen C H, Chen H, et al. FCOS: Fully convolutional one-stage object detection[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 9626-9635.

作者简介

周栋(1996−), 男, 博士生, 从事对抗攻击与防御、空间目标视觉跟踪等研究, E-mail: dongzhou@hit.edu.cn;

孙光辉(1983−), 男, 教授, 博士生导师, 从事计算机视觉、空间机器人等研究, E-mail: guanghuisun@hit.edu.cn;

吴立刚(1976−), 男, 教授, 博士生导师, 从事自主无人系统、先进控制技术等研究, E-mail: ligangwu@hit.edu.cn.