面向目标检测的对抗攻击与防御综述

汪欣欣[1,2]，陈晶[1,2,3]，何琨[1,2]，张子君[1,2]，杜瑞颖[1,2,4]，李瞧[1,2]，佘计思[1,2]

（1. 武汉大学国家网络安全学院，湖北 武汉 430072；
2. 武汉大学空天信息安全与可信计算教育部重点实验室，湖北 武汉 430072；
3. 武汉大学日照信息技术研究院，山东 日照 276800；
4. 地球空间信息技术协同创新中心，湖北 武汉 430079）

摘 要：针对近年来目标检测对抗攻防领域的研究发展，首先介绍了目标检测及对抗学习的相关术语和概念。其次，按照方法的演进过程，全面回顾并梳理了目标检测中对抗攻击和防御方法的研究成果，特别地，根据攻击者知识及深度学习生命周期，对攻击方法和防御策略进行了分类，并对不同方法之间的特点和联系进行了深入分析和讨论。最后，鉴于现有研究的优势和不足，总结了目标检测中对抗攻防研究面临的挑战和有待进一步探索的方向。

关键词：目标检测；对抗攻击；对抗防御；鲁棒性；可转移性

中图分类号：TP181

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023223

Survey on adversarial attacks and defenses for object detection

WANG Xinxin[1,2], CHEN Jing[1,2,3], HE Kun[1,2], ZHANG Zijun[1,2], DU Ruiying[1,2,4], LI Qiao[1,2], SHE Jisi[1,2]

1. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
2. Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, Wuhan University, Wuhan 430072, China
3. Rizhao Institute of Information Technology, Wuhan University, Rizhao 276800, China
4. Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

Abstract: In response to recent developments in adversarial attacks and defenses for object detection, relevant terms and concepts associated with object detection and adversarial learning were first introduced. Subsequently, according to the evolution process of the methods, a comprehensive retrospective analysis was conducted on the research achievements in the realm of adversarial attacks and defense methods for object detection. Particularly, attack methods and defense strategies were categorized based on the attacker knowledge and the deep learning lifecycle. Furthermore, an in-depth analysis and discussion of the characteristics and relationships among different approaches were provided. Lastly, considering the strengths and limitations of existing research, the imminent challenges and directions were summarized for further exploration in adversarial attack and defense of object detection.

Keywords: object detection, adversarial attacks, adversarial defenses, robustness, transferability

---

收稿日期：2023−05−15；**修回日期：**2023−08−06

通信作者：陈晶，chenjing@whu.edu.cn

基金项目：国家重点研发计划基金资助项目（No.2022YFB3102100）；中央高校基本科研业务费专项资金资助项目（No.2042022kf1195）；国家自然科学基金资助项目（No.62076187, No.62172303）；湖北省重点研发计划基金资助项目（No.2022BAA039）；山东省重点研发计划基金资助项目（No.2022CXPT055）

Foundation Items: The National Key Research and Development Program of China (No.2022YFB3102100), The Fundamental Research Funds for the Central Universities (No.2042022kf1195), The National Natural Science Foundation of China (No.62076187, No.62172303), The Key Research and Development Program of Hubei Province (No.2022BAA039), The Key Research and Development Program of Shandong Province (No.2022CXPT055)

0　引言

目标检测的任务是对给定的视频或图像输出目标对象的位置及所属类别。随着深度学习技术的不断发展和取得突破性成就，基于深度学习的目标检测方法已经逐渐取代传统的基于手工特征的检测算法。目标检测作为计算机视觉的基础核心任务之一[1-2]，广泛应用于人脸识别、目标跟踪、自动驾驶、医疗诊断、智能安防等领域[3-8]，其安全性至关重要，对人类社会生活生产方式有着巨大而深刻的影响。

2013 年，Szegedy 等[9]发现面向图像分类的神经网络会遭受对抗攻击的影响，攻击者向目标图像分类器输入带有精心构造的对抗噪声的图像，即对抗样本，会使分类器预测错误。由此，关于对抗学习在深度学习中的研究被关注。

考虑到目标检测的关键性及其多任务多目标的复杂特性，Xie 等[10]首次将对抗攻击扩展到了目标检测任务上，提出了 DAG（dense adversary generation）攻击，使目标检测模型对对抗样本做出错误的类别和位置预测；Eykholt 等[11]采用 RP_2-based（robust physical perturbations based）策略增强了对抗噪声在物理世界中的鲁棒性，成功完成物理世界中对交通标志牌的误检测；Zolfi 等[12]提出 UAP（universal adversarial perturbation）攻击策略，成功干扰 Tesla 高级驾驶辅助系统，导致系统将红灯错误地识别为绿灯。由于目标检测作为基础任务广泛应用于其他重要领域，一旦遭受恶意攻击，将导致巨大损失。因此，目标检测的对抗攻防引起了大量研究者的关注，这些研究工作都具有不同侧重点，使用的机器学习算法也各有不同。

然而，鲜有文章对目标检测领域的对抗攻击和防御进行系统性的总结和分析，尽管文献[13]从局部扰动和全局扰动的角度出发，介绍了目标检测领域中关于对抗攻击的相关工作，并对防御策略进行了归纳和整理，但缺乏对众多研究成果之间联系的分析；同时，由于目标检测对抗工作的不断涌现，现有文献缺乏对最新成果的总结。本文从攻击者的根本意图以及攻击方法切入，主要从白盒攻击、黑盒攻击和数字世界、物理世界攻击的角度来总结和分析目标检测中最近的对抗攻击工作和联系，从模型训练和模型推理角度整理和归纳目标检测中的防御方法。目标检测任务的应用场景、存在的对抗攻击和防御方法如图 1 所示。

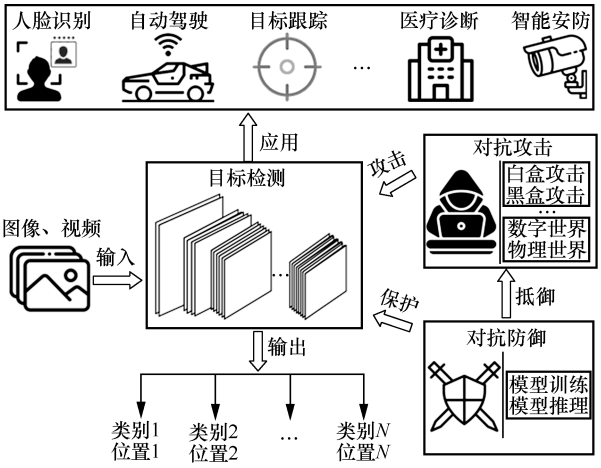


图 1　目标检测任务的应用场景、存在的对抗攻击和防御方法

1　背景知识

1.1　相关术语

1) 白盒攻击与黑盒攻击

根据攻击者对目标模型的知识掌握情况，对抗攻击可以分为白盒攻击和黑盒攻击[14]。在白盒攻击中，攻击者可以获取目标模型训练数据集、模型结构、参数、超参数等所有相关信息；在黑盒攻击中，攻击者无法获取或者缺失目标模型相关信息。通常情况下，实现黑盒攻击的难度远高于白盒攻击。

2) 定向攻击与非定向攻击

根据攻击者的攻击目标，对抗攻击可以分为定向攻击和非定向攻击。在定向攻击中，攻击者针对目标对象进行指定标签的攻击，旨在将目标对象分类成固定类别；在非定向攻击中，攻击者旨在将目标对象随机分类成其他错误标签。

3) 全局噪声与局部噪声

根据对抗噪声的分布范围，对抗噪声可以分为全局噪声和局部噪声。全局噪声 δ 的生成约束条件一般为 $\|\delta\|_p<\varepsilon$，即在 p 范式内噪声值小于 ε；局部噪声一般不限制噪声值，通常在图像上生成掩模矩阵来控制对抗补丁生成。相对于全局噪声，局部噪声更符合物理世界中的攻击场景。

4) 对象类别与目标类别

本文进行特殊定义，对象类别指图像中物体的分类类别，目标类别指攻击者进行定向攻击时将对象分类为某一指定的特殊类别，这一指定类别即目标类别。此外，目标图像指攻击者想要攻击的图像。

5) 目标模型与替代模型

目标模型指攻击者想要攻击的模型，即受害者模型。替代模型指攻击者用于生成对抗样本时的模型。白盒攻击中，目标模型一般与替代模型相同；黑盒攻击中，目标模型一般与替代模型不同。讨论可转移性时，目标模型与替代模型必然不同。

6) 交并比

交并比（IoU, intersection over union）[15]是用于目标检测中计算某一图像类别的预测区域与真实区域相互重叠比例的算法。IoU 值越大，意味着两张图像重叠越多。

7) 非极大值抑制

非极大值抑制（NMS, non-maximum suppression）算法[16]用于目标检测中去除冗余的检测框，NMS 值越大，保留的检测框越多。

8) 锚框

锚框[17-18]是目标检测中引入的固定的多尺度和多纵横比的先验参考框，用于辅助目标检测算法定位物体检测框。

9) 鲁棒性

鲁棒性[19]一词在各个领域应用广泛，本意指的是系统在被干扰或不确定的情况下仍能保持它们的特征行为的能力。神经网络的鲁棒性指神经网络模型防御对抗攻击的能力。鲁棒性越强，模型防御能力越高。鲁棒性也可用来指对抗噪声的能力。

10) 可转移性

对抗样本的可转移性[20]描述的是针对替代模型生成的对抗样本对其他目标模型也能攻击成功的一种能力。一般而言，非定向攻击生成的对抗噪声可转移性更强，具有相似结构的模型生成的噪声可转移性更强[21]。

11) 语义性

目前还缺乏对对抗噪声语义性的准确定义或评估，本文所指的语义性主要从人类主观角度出发，判断对抗噪声是否具有语义或者属于某类别。

1.2 目标检测模型

基于深度学习的目标检测算法分类如图 2 所示。

根据检测步骤可以分为单阶段目标检测算法和两阶段目标检测算法[22-23]。单阶段目标检测算法将目标检测问题抽象为回归问题，直接将特征提取、位置回归以及目标分类整合为一个阶段，对原始图像进行计算，单阶段目标检测代表算法包括 YOLO 系列[24-27]、SSD（single shot multibox detector）[28]、RetinaNet[29]、CornerNet[30]、CenterNet[31-32]、EfficientDet[33]等。两阶段目标检测算法先通过选择性搜索算法或者引入区域推荐网络（RPN, region proposal network）计算原始图像，生成候选区域；再针对第一阶段生成的候选区域进行二次修正得到分类和位置结果。两阶段目标检测代表算法包括基于区域的卷积神经网络（R-CNN）系列[34-36]、R-FCN（region-based fully convolutional network）[37]等。相较而言，两阶段算法需要生成大量候选框再进行定位和分类，一般速度慢、网络结构复杂但精度高；单阶段算法速度快但精度相对较低。

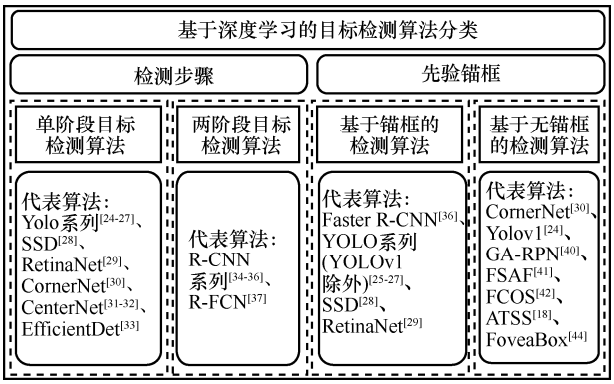


图 2 基于深度学习的目标检测算法分类

根据是否引入先验框来辅助定位检测，基于深度学习的目标检测算法可以大致分为两类[18]：基于锚框的检测算法和基于无锚框的检测算法。由于目标检测模型检测框定位不准确，Faster R-CNN[36]首次通过引入锚框解决目标多尺度和多纵横比的问题，提高了检测精度。基于锚框的目标检测代表算法包括 Faster R-CNN[36]、YOLO 系列（Yolov1 除外）[25-27]、SSD[28]、RetinaNet[29]等，但是由于引入锚框需要先验知识，计算过程中会产生大量负样本，因此存在锚框不通用、难以调优以及模型训练速度慢的问题。进而基于无锚框的目标检测算法大量涌现，一般基于 2 种原理：基于关键点的无锚框检测和基于中心点的无锚框检测。前者通过定位从特征图中学习到的关键点或者引入的先验关键点来预测边框，典型算法有 ATSS[18]、FreeAnchor[38]、CornerNet[30]及其优化 CornerNer-Lite[39]等。后者预测特征图的每个位置是目标中心点的概率，代表算法有 YOLOv1[24]、GA-RPN[40]、FSAF[41]、FCOS[42]、CSP[43]、FoveaBox[44]等。

1.3 数据集与评估指标

1.3.1 数据集

目前关于目标检测的对抗学习研究中干净数

据集主要有 Pascal VOC2007[45]（简称 VOC07）、Pascal VCO2012[46]（简称 VOC12）、MS COCO[47]（简称 COCO）、Inria[48]以及其他数据集[49-55]。其中，ImageNet 数据集[49]是指 ImageNet VID 视频目标检测数据集。相关对抗样本数据集主要有 APRICOT[50]，这是第一个开源的带物体对抗攻击的目标检测数据集，由 Braunegg 等在不同时间、地点、视角等拍摄收集，主要标记一个让目标检测器检测出某物体的对抗补丁，即假阳性攻击。尽管 APRICOT 为每个对抗补丁提供了边界框注释，但是缺乏像素级注释。Liu 等[51]为 APRICOT 数据集生成了 APRICOT 掩模（APRICOT-Mask）数据集，确定每个补丁像素级噪声位置，更适用于分割任务。目标检测数据集相关信息如表 1 所示。

1.3.2　评估指标

对抗攻击旨在构造特殊样本使模型输出错误，衡量对抗攻击与防御的评估指标主要包括与目标检测模型性能相关指标及攻击成功率指标。前者主要包括平均精度（AP, average precision）、平均精度均值（mAP, mean average precision）、PR 曲线以及 ROC 曲线等。其中，根据每个类别的 Precision 值和 Recall 值可以分别绘制出各个类别的 PR 曲线，Precision 为预测正确的正样本数占所有预测为正样本个数的比例，Recall 为预测正确的正样本数占所有真实值为正样本数的比例。ROC 曲线与 PR 曲线绘制方法相似，但文献中用来衡量对抗攻击的频率较低，此处不再赘述。AP 为 PR 曲线下面积之和，即在不同 IoU 阈值时的平均精度，mAP 为 AP 值在所有类别下的均值。关于攻击成功率（ASR, attack success rate）的定义，大多数文献的计算方式遵循 Eykholt 等[11]提出的方法，即 ASR 为攻击成功的图片占所有图片的比例。

2　面向目标检测的对抗攻击方法

根据攻击者能力或者掌握的知识程度，目标检测中的对抗攻击可以分为白盒攻击和黑盒攻击；根据攻击者的攻击目标，对抗攻击可以分为定向攻击和非定向攻击；根据攻击场景可以分为数字世界攻击和物理世界攻击；根据噪声类型或者分布范围，对抗攻击可以分为全局噪声攻击和局部噪声攻击。本文从攻击者的根本意图以及攻击方法的角度出发，主要从白盒攻击、黑盒攻击和数字世界、物理世界攻击的角度来进行总结和分析。表 2 展示了面向目标检测的对抗攻击方法[10-12,56-71]，其中，T 表示定向攻击，UT 表示非定向攻击；G 表示全局噪声，L 表示局部噪声。

2.1　白盒攻击

由于目标检测任务包括分类和定位，其本质上是一个多任务问题，并且目标检测对象具有多尺度、多纵横比、多数量的特性，其所处物理环境复杂。这些挑战导致关于目标检测的对抗攻击研究大多数集中于白盒攻击和数字世界。

在白盒攻击中，攻击者的思路一般是输入对抗样本后，使其特征图、模型的 logits 结果或模型输

表 1　　**目标检测数据集相关信息**

数据集	图像数量	实例数量	类别数量	年份	特点
VOC07[45]	9 963	24 640	20	2007 年	目标检测数据集，初步建立成一个完善的目标检测数据集
VOC12[46]	11 530	27 450	20	2012 年	目标检测数据集，与 VOC07 数据集互斥，测试数据集中只有图像，没有标签，可与 VOC07 结合使用
COCO[47]	330 000	1 500 000	80	2014 年	目标检测数据集，每一类图像多，检测难度大
Inria[48]	2 573	1 826	1	2005 年	行人检测数据集，标记的站立或行走的人的图像，部分标注不准确
ImageNet[49]	5 354	—	30	2015 年	视频目标检测数据集，每一类图像多，每个视频包括 56～458 帧图像
APRICOT[50]	1 011	—	60	2020 年	对抗补丁数据集，为每个补丁提供边界框注释，数据集包括室内和室外场景，不同时间、位置、比例、旋转和视角的补丁
APRICOT-Mask[51]	1 011	—	60	2022 年	对抗补丁掩模数据集，为 APRICOT 数据集中的每个补丁提供像素级注释
BDD[52]	100 000	1 841 435	10	2018 年	自动驾驶数据集，具有大规模、多样化、在街上采集的特点
MTSD[53]	100 000	325 172	313	2019 年	交通标志数据集，覆盖全球多个地区的街景
MPII[54]	25 000	40 000	1	2014 年	人体姿势数据集，图像涵盖了 410 个人类活动，从 YouTube 视频中提取
CCTV[55]	921	559	1	2022 年	此数据集来源于 CCTV 摄像头，只包含人类正例样本和负例样本
自制数据集	—	—	—	—	现实场景拍摄或虚拟场景截取图片等制作数据集，满足特殊场景要求

表 2　　**面向目标检测的对抗攻击方法**

攻击能力	具体分类	攻击方法	攻击目标	攻击场景	噪声类型	被攻击模型	数据集
白盒攻击	基于优化迭代的攻击方法	DAG[10]	T	数字世界	G	Faster R-CNN、R-FCN	VOC07、VOC12
		RP_2-based[11]	UT	物理世界	L	YOlOv2、Faster R-CNN	自制数据集
		UAP[12]	UT	物理世界	G	YOLOv5、YOLOv2、Faster R-CNN	LISA、MTSD、BDD
		AA-HA[56]	T、UT	物理世界	L	Faster R-CNN、YOLOv3、SSD、R-FCN、Mask R-CNN	自制数据集
		MeshAdv[57]	T、UT	数字世界	L	YOLOv3	COCO
		DTA[58]	UT	数字世界	L	EfficientDet、YOLOv4、SSD、Faster R-CNN、Mask R-CNN	自制数据集
		DAS[59]	UT	物理世界	L	YOLOv5、SSD、Faster R-CNN、Mask R-CNN	自制数据集
		UPC[60]	T、UT	物理世界	L	faster R-CNN、R-FCN、SSD、YOLOv2、YOLOv3、RetinaNet	自制数据集
		CAC[61]	T	物理世界	L	Faster R-CNN、YOLOv3、YOLOv5	自制数据集
		FCA[62]	UT	数字世界	L	YOLOv3、YOLOv5、SSD、Faster R-CNN、Mask R-CNN	自制数据集
		LAP[63]	UT	物理世界	L	YOLOv2	Inria
	基于生成器生成的攻击方法	UEA[64]	UT	数字世界	G	Faster R-CNN、SSD	VOC07、ImageNet
		NPA[65]	UT	物理世界	L	YOLOv2、YOLOv3、YOLOv4、Faster R-CNN	Inria、MPII、Mix
		TC-EGA[66]	UT	物理世界	L	YOLOv2、YOLOv3、Faster R-CNN、Mask R-CNN	Inria
黑盒攻击	基于可转移性的攻击	CAMOU[67]	UT	物理世界	L	Mask R-CNN、YOLOv3-SPP	自制数据集
		CAA[68]	T	数字世界	G	Faster R-CNN、YOLOv3、FoveaBox、DETR、Libra R-CNN、FreeAnchor、D-DETR、RetinaNet	VOC07、COCO
		T-SEA[69]	UT	物理世界	L	YOLOv2、YOLOv3、Faster R-CNN、YOLOv3tiny、YOLOv4、YOLOv4tiny、YOLOv5、SSD	Inria、COCO、CCTV
		ZQA[70]	T	数字世界	G	Faster R-CNN、RetinaNet、Libra R-CNN、FoveaBox	VOC07、COCO
	基于查询的攻击方法	PRFA[71]	UT	数字世界	G	Faster R-CNN、YOLOv3、FCOS、ATSS	COCO

出结果接近定向攻击的指定目标标签结果或远离图像的真实标签结果。其关键步骤就是攻击者构造损失函数，利用梯度求导优化，直接生成样本或者训练生成模型。进一步地，根据对抗噪声的生成方法，白盒攻击可以分为基于优化迭代的攻击方法和基于生成器生成的攻击方法。前者的优点在于生成方式简单、训练稳定，后者的优点在于速度快、效率高。

2.1.1　基于优化迭代的攻击方法

基于优化迭代的攻击方法框架如图 3 所示，其核心步骤是设计损失函数，利用梯度传播反向优化对抗样本，这个过程中不需要额外训练其他网络，生成对抗样本方式简单，但是速度较慢、效率较低。

2017 年，Xie 等[10]提出 DAG 攻击，首次将对抗攻击应用到目标检测任务中，其攻击思路是针对每个被预测正确的候选框进行攻击。但是在目标检测任务中，即使一个候选框被分类错误，其他候选框也可能被正确分类，因此其提高 NMS 阈值，保留更多的、密集的候选框集合进行优化。攻击者为图片中的每一个目标对象随机选择一个不同于目标的真实标签作为目标标签，在一组密集的候选框集合上不断进行迭代优化，提高目标类别的得分、降低真实类别的得分，直到所有的候选框被错误识别。DAG 攻击利用分类单损失优化全局噪声，最终使 RPN 生成的候选框全部分类错误。Li 等[72]尝试在背景区域生成局部噪声，利用分类损失和位置损失联合优化，重新排列区域候选框的预测得分，使正候选框分类为背景，负候选框分类为前景，同时对于仍能正确分类的预测框，干扰其位置参数，以此进行攻击。这 2 种方法均是面向目标检测的数字世界或二维图像的对抗攻击方法，旨在提高目标检测攻击成功率；由于目标检测在现实世界的广泛应用，关于目标检测的对抗学习也逐步向物理世界以及 3D 图像转移，对抗噪声的鲁棒性、语义性以及可转移性成为重要的研究关注点。

由于物理世界环境复杂，不仅存在相互遮挡等问题，而且环境因素如光照强度、光照亮度、角度、

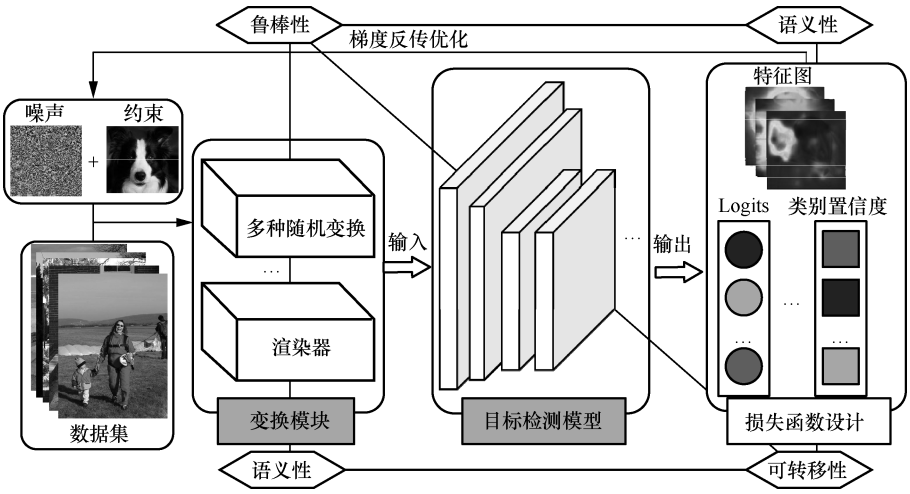


图 3　基于优化迭代的攻击方法框架

距离等都会影响对抗样本性能；另外，对抗样本还会被摄像头对对抗噪声以及图像的捕捉能力所影响，在物理世界和数字世界之间存在巨大的攻击性能差异。物理世界与数字世界的不同是鲁棒性对抗噪声需要考虑的问题。Eykholt 等[11]受图片分类领域中 RP_2-based 攻击[57]的启发，首次实现在物理世界中对交通标志牌的 RP_2-based 攻击。RP_2-based 攻击通过增强数据集，拍摄多角度和多距离的图片用于训练对抗样本，提高对抗样本在物理世界中的鲁棒性。同时，引入掩模方法，保证对抗噪声局限于目标对象的范围内，以此部署在物理世界中。RP_2-based 攻击通过最小化预测框的得分使检测器无法检测到交通标志牌，称为消失攻击；同时提出出现攻击，通过提高预测框的得分，使目标检测器检测出不存在的物体。与 RP_2-based 类似的还有 HA-AA（hiding attack-appearing attack）[56]，从数据集生成角度上，Zhao 等[56]提出 ERG（enhanced realistic constraints generation）方法，利用 Google 搜索交通标志牌和相应的公路环境合成更丰富的数据集；从优化方法的角度上，Zhao 等在特征层就开始进行对抗样本的干预，使其和目标图像的特征图不一样，以使目标图像消失。同时 Zhao 等利用生成小尺寸噪声使模型检测不存在的物体，这本身是利用 YOLOv3 等模型在小尺寸图片上检测效果好的原理。其核心思路是对于远距离物体而言，物体图像比较小，因此可以修改物体的全局区域；对于近距离物体而言，物体图像比较大，因此修改物体的局部噪声，如中心区域，以此制造一个远近距离皆有效的对抗样本。Zolfi 等[12]受相机对抗贴纸的启发[73]，提出 UAP 攻击，成功使 Tesla 高级驾驶辅助系统无法检测出交通标志牌。与普通的噪声打印贴纸[11,56,74]不同，UAP 攻击利用镜头的光学原理，直接利用贴在镜头上的透明噪声干扰目标检测器。UAP 攻击通过分类损失、位置损失以及其他损失构造损失函数，引入仿射变换解决噪声点中心位置离散不可微的问题，利用反向梯度传播生成类内通用噪声，部署在物理世界中。尽管这些鲁棒性的方法可以在物理世界中实现对抗攻击，但是考虑的攻击对象是平面、刚性物体，忽略了非平面、柔性物体由于不规则、形变等特性对对抗噪声的影响。

文献[57-62,75]探讨了针对非平面、柔性物体上的对抗攻击，由于对抗噪声部署在非平面物体如人体时，其数据分布随非平面物体本身形状以及运用发生改变，这些因素导致关于对抗样本的研究又延伸到 2 个方向，一是与图形学结合，利用可微分渲染解决 3D 场景中噪声与非平面物体完美融合的问题，同时提供多样性数据集；二是利用多种函数变换模拟形变，提高对抗噪声的鲁棒性。

文献[75]通过引入可微分渲染器，通过修改光照强度、3D 物体表面分布以及反射率（材质[76]）等因素，实现对分类器的对抗攻击。文献[57]利用未固定相机参数（位置、角度）的可微分渲染器生成针对目标检测器更实际的、合理的对抗噪声，同时探索了如何利用可微分渲染器去拟合逼近不可微分渲染器，解决渲染器黑盒问题。文献[58]提出针对车辆的 DTA（differentiable transformation attack），考虑到可微分渲染器不能完全表示真实物理世界中的变换，只能支持前景生成，背景仍然是采用传统的渲染器，前景与背景没有很好地融合，如阴影、光反射等，DTA 通过提出 DTN（differential

transformation network）学习 3D 场景中相机角度、光照、目标位置等变换，实现前景的完美融合与多变换，增加其鲁棒性，完成针对目标检测器的对抗攻击。文献[59]在引入可微分渲染器的基础上，提出模型不可知和人眼不可知的 DAS（dual attention suppression）对抗攻击，旨在生成可转移性更强以及更自然的对抗噪声。DAS 攻击的核心思想是利用一阶段模型和二阶段模型共有的特征提取网络来提高可转移性，并且认为大脑对物体的轮廓更敏感，尽可能保证初始噪声轮廓不变，优化对抗噪声，使其更自然。这些基于可微渲染器在 3D 世界中构造对抗噪声的攻击[57-62]主要依赖于虚拟现实模拟器来实现对抗攻击，如 CARLA（car learning to act）平台，其优点如下：1) 可以减少数据集收集成本，增强数据集多样性；2) 能实现一个全 3D 世界中的端到端的攻击，提高对抗噪声的鲁棒性，并且操作简单、效率高。但是这些研究也存在以下问题。1) 虚拟现实模拟器本质上和现实物理环境存在差异，很难模拟物理世界中的物体多样性、环境多样性，在真实度上存在差异，因此生成的对抗样本也很难在物理世界中真实的物体上使用。尽管文献[59,61]等也在物理世界中的汽车模型上评估对抗攻击效果，但是目标检测模型会受图像上下文、物体大小、环境复杂性的影响，抗噪声分辨率高低在贴合过程中也会受物体大小影响，因此在汽车模型上评估对抗攻击效果是不公正、不具有代表性的。2) 目前缺乏基于模拟器的公开数据集，大多数基于渲染器的目标检测对抗研究利用模拟器自制数据，所有的对抗攻击很难在一个公平标准下评估。

Huang 等[60]从可微分渲染器以外的角度探索物理世界中的针对非平面物体的目标检测对抗攻击，即在二维世界中模拟多种变换提高对抗噪声的鲁棒性，主要包括目标物体的内部变换和外部变换。物体的内部变换主要针对二维图像进行几何变换操作，如裁剪、调整大小、仿射变换等；外部变换主要利用数据集的多样性完成，以及控制明度、角度、位置来完成，类似于 EOT（expectation over transformation）操作[77]。Xu 等[78]在 Thys 等[79]提出的硬纸板对抗补丁基础上，考虑到人体行动过程中衣服上的褶皱会导致对抗噪声的性能变差，提出引入薄板样条插值法（TPS, thin plate spline）来模拟柔性变形。与文献[63,65,78-80]在人体胸部设置对抗噪声不同的是，Huang 等[60]在身体多个位置部署补丁，Hu 等[66]制作一件全覆盖对抗噪声的衣服，以此增强对抗攻击鲁棒性，但对抗噪声部署范围过大，易引起注意。

进一步地，除了关注目标检测对抗噪声的鲁棒性，Tan 等[63]就对抗噪声的语义性提出 LAP（legitimate adversarial patch）攻击。文献[79]利用总变差损失[81]来保证对抗噪声的平滑性，与之类似的工作还有文献[78-80]，尽管提高了对抗噪声的平滑度，但生成的对抗噪声仍然缺乏语义性；Huang 等[60]在此基础上借助 PGD（project gradient descent）攻击约束噪声大小，使其从人类视觉角度出发有语义性，该方法实现简单，但是约束能力的大小会影响攻击性能，当噪声限制到一定大小时，难以保证攻击成功率。LAP 攻击从人类视觉角度重新构造对抗损失函数，包括颜色特征、边缘特征和纹理特征，LAP 攻击也指出，当语义性越强时，对抗攻击成功率越低。同时 LAP 攻击发现，相较于边缘特征和纹理特征，当提高语义性时，颜色特征在优化过程中会发生巨大变化。

2.1.2 基于生成器生成的攻击方法

基于优化迭代的攻击方法由于需要针对具体图片或者具体类别不断反向迭代优化噪声，消耗资源多，计算时间长，效率较低，但是该方法直观简单。基于生成器生成的攻击方法的核心思想是学习对抗噪声的分布，直接利用生成器生成对抗噪声。基于生成器生成的攻击方法框架如图 4 所示，一旦生成器训练稳定，只需要针对图像进行前向传播即可生成相应噪声，生成速度快、效率高，但是训练一个优良的生成器较困难，面临模型坍塌、训练不稳定、难以收敛等问题。

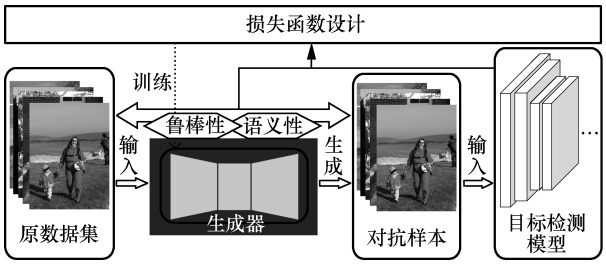


图 4 基于生成器生成的攻击方法框架

Wei 等[64]提出 UEA（unified and efficient adversary）攻击，这是一种针对目标检测统一且高效的数字世界对抗攻击，考虑到 DAG 攻击等只针对两阶段模型进行攻击，缺乏对单阶段模型的可转移性，因此提出对两阶段模型的 RPN 层、分类得分以及单模型的特征层进行攻击；同时，考虑到对图

片或者视频的单帧图像生成对抗噪声的效率问题，UEA 攻击提出利用生成对抗网络（GAN, generate adversarial network）[82-83]直接生成噪声，加快生成速度。

Hu 等[65]考虑到对抗样本缺乏语义性，由于生成对抗模型 BigGAN[84]或者 StyleGAN[85]能生成高分辨率、多样化的具有语义性的高质量图片，提出自然物理对抗补丁（NPA，naturalistic physical adversarial-patch）攻击，通过引入预训练模型 BigGAN 和 StyleGAN 来辅助生成噪声，利用渲染器渲染对抗噪声和原始图片，合成多场景、多样性（光照、位置、角度等）图片，最后将合成图片送入目标检测中，联合分类得分以及前景得分等构造损失函数，采用反向梯度求导优化隐空间变量。文献[65]在 NPA 攻击中引入渲染器以及多重变换，其目的是构造鲁棒性的物理世界对抗噪声，但是其发现引入更多的变换效果不佳，可能是因为生成模型的隐空间有限，无法找到适用于所有变换的鲁棒性好、语义性强的对抗噪声。

文献[66]考虑到当视角发生变化时，对抗噪声性能会下降，因此提出一种可扩展的、重复结构的对抗噪声，可以覆盖于任何形状、任何大小的衣服，只要摄像机捕获到衣物的某个局部图案，无论从何种视角，目标对象都将无法被检测模型识别，即 TC-EGA（toroidal-cropping-based expandable generative attack）。TC-EGA 包括以下 2 个步骤。第一步训练一个具有平移不变性的生成器，参考文献[86]方法，从能量函数以及互信息的角度构造损失函数训练全卷积网络（FCN, fully convolutional network）[87]，利用 FCN 拟合一个平移不变的对抗图案分布，实现可扩展原理。第二步引入循环切割（TC, toroidal cropping）方法随机采样一个固定的待优化的隐空间变量，并将采样后的隐空间表示输入生成器，利用其输出图案进一步优化图案的干扰能力。TC-EGA 生成的对抗噪声在不同的角度都具有较强的鲁棒性，可以实现对目标模型的攻击，但是生成的噪声缺乏语义性，易引起人眼注意。

2.2　黑盒攻击

在黑盒攻击中，由于攻击者不了解目标模型的结构、参数等信息，无法直接获取目标模型的梯度对样本进行优化，因此攻击者从另外 2 个角度出发优化对抗噪声，一是利用替代模型和对抗样本的可转移性构造对抗噪声，二是基于梯度估算或者随机搜索等方法，更新对抗样本。前者一般构造一个替代模型，利用反向梯度求导不断迭代对抗噪声，方法较简单，效率高，但是攻击成功率较低；后者一般通过查询等方式，基于零阶梯度优化估算样本梯度或者随机搜索样本空间，进行对抗样本更新，当达到查询次数上限或者攻击成功后停止查询，其攻击成功率更高，但是所需查询次数较多，一般需要成千上万次查询才可完成攻击。

2.2.1　基于可转移性的攻击方法

以往目标检测的对抗攻击[10-11]大多数是在简单的环境中实施，Zhang 等[67]提出针对较复杂的、现实的——物理世界中的非平面物体（如车辆等）进行伪装（CAMOU）攻击，考虑到时间和经济条件的约束，利用模拟器来模拟现实环境，如光线、角度、距离、位置等。针对模拟器的成像过程不可微分，CAMOU 攻击提出训练一个可微神经网络来近似成像过程，即输入为物理环境、对抗噪声和 3D 模型车辆，输出为合成图片，其挑战在于难以生成高分辨率合成图片。文献[67]意识到，与训练图像生成神经网络相比，将目标探测器和模拟器的成像过程视为一个完整的黑匣子会更容易学习一个函数来近似黑匣子行为。因此 CAMOU 攻击将图片生成和目标检测 2 个黑盒行为视为一个端到端的黑盒行为，训练一个可微的近似网络优化对抗噪声，利用对抗样本的可转移性，实现对目标模型的攻击。尽管 CAMOU 攻击仅在模拟器环境中测试对抗噪声针对 2 个目标检测模型的性能，但是 CAMOU 攻击首次利用 render 思想来解决物理世界对抗攻击问题，提出端到端方法解决了高分辨率数据集问题，利用对抗样本的可转移性完成黑盒攻击。Cai 等[68]提出一种上下文感知的对抗样本序列攻击，利用各类别之间的共现率、距离、大小等上下文关系，构建共现关系矩阵，通过干扰受害者类别以及其上下文的其他对象，提高攻击性能，与 CAMOU 攻击相同，Cai 等提出的 CCA（contextual camouflage attack）也是利用任务间的对抗样本可转移性实现黑盒攻击，由于相同的任务之间，模型可能具有相同的决策边界。尽管 CCA 需要对目标模型进行查询，但其本质上是利用一个或多个替代模型，通过反向梯度传播优化对抗噪声，实现黑盒攻击。其查询目的只是为了获取是否攻击成功，并且查询次数极少（为 6 次）。文献[78,80]同样利用集成多个替代模型

的方法实现对目标检测模型的攻击，以提高其可转移性。此类方法的缺点在于计算开销极大，成本高，并且选择替代模型需要先验条件或者专业知识。基于此挑战，Huang 等[69]提出利用随机深度[88]的思想对单模型进行训练，以此提高模型可转移性。与集成方法相比，此方法极大地减小了计算开销，缩短了对抗噪声优化时间，但需要修改模型结构。

除了利用任务内对抗样本的可转移性，零查询攻击（ZQA，zero-query attack）[70]利用任务间的可转移性完成针对目标检测黑盒攻击，在 CCA 上下文感知的基础上，对多个图像分类器进行攻击，计算扰乱单个物体分类的概率矩阵，根据共现关系矩阵和概率矩阵完成对目标检测器的攻击。ZQA 本质上是利用图像分类任务的对抗样本可转移性实现对目标检测分类任务的攻击，严重依赖于概率矩阵，因此计算开销大，效率低，分类器的选择会严重影响到 ZQA 的效果。

2.2.2 基于查询的攻击方法

Liang 等[71]针对目标检测提出基于查询的平行矩形反转黑盒攻击（PRFA），由于 NMS 机制，面临即使成功攻击一个最优预测框，其他次优边界框也可能出现，以及样本空间过大的挑战。受 Wei 等[64]提出的 UEA 的启发，Liang 等[71]等发现对抗噪声主要分布在目标对象的轮廓以及关键点上。因此其提出使用不同的先验知识来减少随机搜索空间，利用 Mask-R-CNN[89]模型的分割结果、RepPoints[90]模型的关键点检测结果以及目标模型的查询结果确定样本搜索空间，减少查询次数，提高效率。与 SquareAttack[91]每次增加一个方块噪声不同，Liang 等[71]提出并行查询策略，在查询的早期阶段，每次查询时随机更新多个方块的噪声，随着迭代次数的增加，更新方块数量逐渐减少，以此加快广度搜索；同时，其发现同一个预测框中出现多个方块噪声，提出利用符号翻转[92]进一步增加噪声多样性。PRFA 极大地提高了目标检测中黑盒攻击的查询效率问题，但是引入了较多先验知识，并且目前缺乏对鲁棒性的研究，只能部署于数字世界中。关于目标检测黑盒研究仍然是一个值得关注与探索的探究问题。面向目标检测的对抗攻击方法演进及联系如图 5 所示。

3 面向目标检测的对抗防御方法

本节研究针对目标检测对抗攻击的防御方法。现有的对抗防御方法根据防御目标可以分为两大类：1) 经验防御，指对现有攻击的一定理解基础上提出的防御方法，在实践中具有良好的性能，但其有效性缺乏理论保障；2) 认证防御，指有效性理论上在一定的假设条件下可以得到保证的防御方法，但在实践中其有效性普遍弱于经验防御，并且计算代价较高、效率较低。经验防御更注重实际应用，而认证防御则注重理论上的鲁棒性保证。根据防御策略的响应方式，又可以分为以下两种：1) 主动防御，在技术上预先采取措施，减少或消除潜在的对抗威胁，旨在使模型在遭受攻击时更鲁棒，涉及模型开发者自我模拟博弈的对抗过程；2) 被动防御，在对抗攻击发生后采取措施，以减轻或消除攻击的

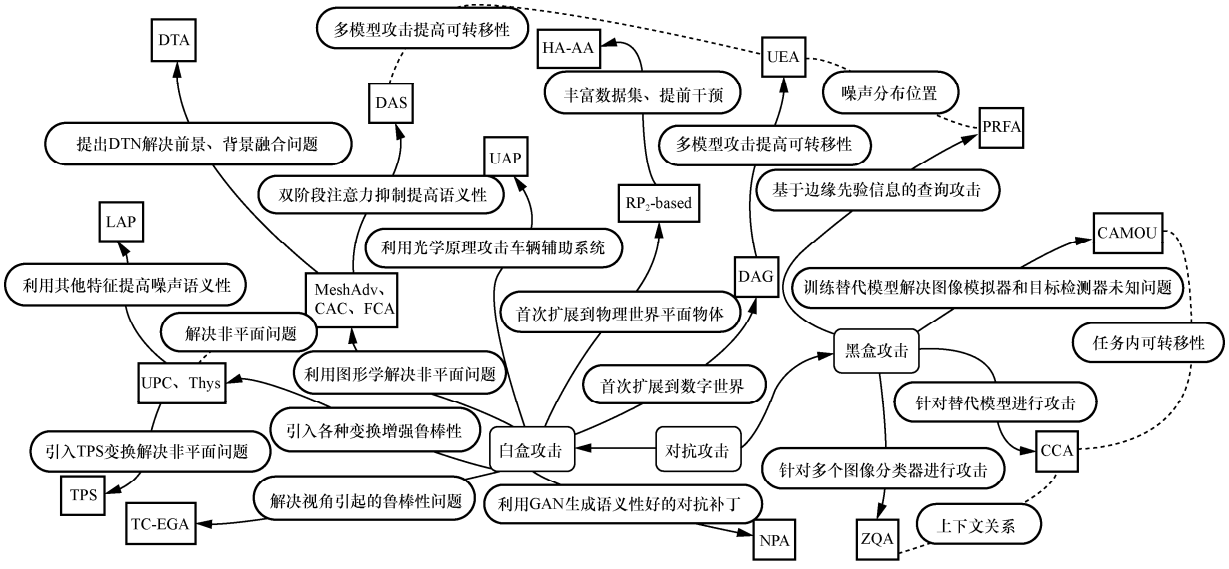


图 5 面向目标检测的对抗攻击方法演进及联系

影响，是模型攻击者与模型防御者之间的动态博弈进化。前者侧重于对对抗攻击的预防，而后者则侧重于对对抗攻击的响应。根据深度学习生命周期[93]的视角，可以从 2 个角度划分对抗防御方法：1) 模型训练角度，利用对抗训练[94]或者其他方法[95]提高模型鲁棒性，尽可能降低对抗噪声对模型的影响。其中，对抗训练是典型的经验防御和主动防御方法；2) 模型推理角度，利用降噪[51,96-97]或者对抗噪声检测[95]等手段消除噪声或者拒绝对抗样本输入。由于认证防御相关研究较少，本文主要从模型训练和模型推理角度对目标检测中的主要防御方法进行了归纳整理。面向目标检测的对抗防御方法如表 3 所示。

表 3　　**面向目标检测的对抗防御方法**

防御角度	具体分类	防御方法	防御目标	防御主动性	噪声类型	防御机制
基于模型训练的防御方法	对抗训练	MTD[99]	经验防御	主动防御	G	根据单任务生成的对抗样本集合筛选出使整体任务损失最大的对抗样本集合进行对抗训练，以此提高模型的鲁棒性
		CWAT[101]	经验防御	主动防御	G	利用相应类的对象数量对每个类损失进行归一化，保证无差别地攻击图像中的所有类别，提高模型对所有目标类的鲁棒性
	其他方法	文献[102]	经验防御	主动防御	L	限制上下文信息使用，如限制感受野范围或创建上下文无关的训练集，避免成为攻击者的攻击手段
		RobustDet[104]	经验防御	主动防御	G	利用多个不同的卷积核学习对抗样本和干净样本的鲁棒性特征，构建鲁棒性的目标检测器
基于模型推理的防御方法	降噪	文献[96]	认证防御	主动防御	G	利用随机中值平滑完成鲁棒性认证，保证目标检测器在大小范围内的认证鲁棒性
		ROSA[107]	经验防御	主动防御	G	利用超像素方法和随机排列消除全局噪声的影响，同时引入上下文感知尽可能恢复原图像分布
		文献[108]	经验防御	主动防御	L	根据对抗样本和干净样本的分布差异，利用基于熵和基于梯度的方法感知高频信息，去除对抗噪声
		APM[97]	经验防御	主动防御	L	学习一个数据预处理网络来确定对抗噪声位置，利用掩模去除对抗噪声
		文献[51]	经验防御	主动防御	L	利用分割模型确定对抗补丁位置，利用补丁完善方法微调补丁形状，利用掩模去除对抗噪声
	对抗噪声检测	Detector Guard[95]	认证防御	被动防御	L	利用鲁棒的小感受野图像分类器来判断目标是否存在
		文献[109]	经验防御	被动防御	L	以基础目标检测器和 BERT 语言模型计算上下文一致性来检测对抗噪声

3.1　基于模型训练的对抗防御

3.1.1　对抗训练

2014 年，GoodFellow 等[98]首次提出对抗训练方法，利用干净样本和对抗样本构成的训练数据集对模型进行训练，训练流程如图 6 所示，实验结果显示对抗训练可以有效降低模型对对抗样本的分类错误率，提高模型鲁棒性。

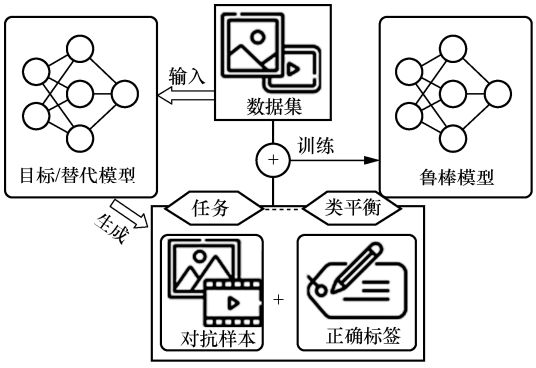


图 6　对抗训练流程

早期关于对抗训练的研究主要集中于图像分类领域，Zhang 等[99]将对抗训练方法扩展到目标检测领域中，从多任务学习角度，揭示了目标检测中对抗攻击[10,11,64,72,100]的共同底层机制，即遵循基于单任务损失或者组合损失的框架和设计原则。由于分类任务和位置任务共用相同的特征图，针对一个任务生成的对抗样本也会使另一个任务的性能下降，但 2 个任务又存在梯度错位，没有完全对齐，简单的基于任务未知、两者任务总损失的对抗训练会导致模型性能（mAP）和鲁棒性下降，因此 Zhang 等提出面向多任务监督源（MTD, multi task-oriented domain）的对抗训练方法来抵御目标检测中的对抗攻击，根据 2 个单任务生成的对抗样本集合筛选出使整体任务损失最大的对抗样本集合进行对抗训练，以此提高模型的鲁棒性。

尽管如此，MTD 方法在生成对抗样本时存在

类别不均衡的问题，可能存在特定类损失支配其他损失的情况，导致某些对象不会被攻击，最终引起对抗训练后模型性能以及鲁棒性在类别上的不平衡。Chen 等[101]提出基于类平衡的对抗训练（CWAT）方法，将总损失分解为类损失，利用相应类的对象数量对每个类损失进行归一化，以此减轻特定类损失支配其他类损失的情况，而不是按照对象数量对总损失进行归一化。这样可以保证无差别地攻击图像中的所有类别，有助于提高模型对所有目标类的鲁棒性。

MTD、CWAT 等基于对抗训练的防御方法能一定程度上提高模型的鲁棒性，但是实施的前提是能精确获取针对目标模型的对抗样本，需要了解攻击者能力，对模型进行重训练；其次，经过对抗训练的模型，其数据集分布由原来干净数据分布向对抗样本分布转移，因此模型精度大幅度降低；同时，基于对抗训练的防御方法总是存在新的对抗样本可以欺骗网络，无法抵御二次攻击。

3.1.2 *其他方法*

对于提高模型的鲁棒性，除了经典的对抗训练方法，文献[102]展开了对限制上下文信息方法的研究。文献[102-103]指出空间上下文信息可以有效提高目标检测模型性能，但是同时也成为攻击者实现对抗攻击[68,70]的手段。因此，文献[102]提出 2 种方法来限制上下文信息使用：1) 提出基于数据驱动的限制感受野方法，对边界框外的特征图进行非零值惩罚来限制感受野范围；2) 通过创建上下文无关的数据集来训练上下文无关的模型，如将一张图片中的目标对象剪裁出来粘贴到另一张图片的相同位置，进行边缘模糊、删除遮挡对象等，以此训练模型。但是此种防御方法制作数据集代价较大，并且仅针对上下文攻击，缺少泛化性。

Dong 等[104]对 MTD、CWAT 等对抗训练方法进行分析研究，指出干净样本和对抗样本之间特征不一样，两者之间的冲突导致模型精度急剧下降，与其用一个模型学习两者之间的共同特征，不如正视特征不同的事实，提出一种新的基于对抗感知的鲁棒目标检测器（RobustDet），主要包括对抗图像判别器（AID, adversarial image discriminator）模块和基于重构图像的一致特征（CFR, consistent features with reconstruction）模块。受 Chen 等[105]的启发，RobustDet 利用多个不同的卷积核学习对抗样本和干净样本的鲁棒性特征，通过引入 AID 模块学习对抗样本和干净样本的不同分布，为多个动态卷积分发权重；同时，受 VAE 方法[106]启发，RobustDet 引入 CFR 模块，从对抗样本以及干净样本上学习到鲁棒性或一致性特征，再基于一致性特征重建出干净样本。RobustDet 的主要思路是先学习对抗样本和干净样本的不同分布，再利用动态卷积学习到两者的共同鲁棒性特征，进而利用鲁棒性特征完成目标检测。RobustDet 模型通过引入多个模块，一定程度上提高了模型鲁棒性，缓解了模型精度下降的问题，但对目标检测模型网络结构修改程度较大，同时需要提前掌握攻击者的对抗样本生成算法、模型结构等先验信息，并且 RobustDet 无法抵御二次对抗攻击。

3.2 基于模型推理的对抗防御

3.2.1 降噪

随着图像分类任务中认证鲁棒性工作[107-109]不断取得突破性进展，Chiang 等[96]提出基于中值平滑认证目标检测模型鲁棒性的方法，考虑到目标检测的复杂性，将目标检测认证鲁棒性问题构建为回归问题，利用预测框的 IoU 计算将鲁棒性认证从图像分类任务扩展到目标检测任务。由于传统的高斯平滑操作基于均值进行计算，易受基函数的极值影响而严重倾斜，当函数输出变化较大时，生成的边界相当松弛，因此均值平滑操作适用于分类任务的鲁棒性认证问题，却无法解决回归任务认证问题。为了得到更紧密的边界，Chiang 等[96]提出利用中值平滑代替均值平滑完成鲁棒性认证；同时，考虑到目标检测任务昂贵的训练代价，引入去噪模块，基于 DnCNN（denoising convolutional neural network）[110]训练去噪器，从而避免对目标检测器的重新训练，减小开销。由于目标检测器被视为一个黑盒来训练去噪器，因此其认证鲁棒性框架在其他未知任务上具有良好的扩展性。这是目标检测中为数不多的认证防御研究之一，将认证防御从图像分类问题扩展到目标检测问题上，虽然认证 AP 值比较低，但从理论上证明了随机中值平滑对目标检测对抗攻击防御的有效性。

文献[51,97,111-112]提出基于经验防御的对抗防御方法来去除对抗噪声的影响。Li 等[111]提出鲁棒的显著性目标检测（ROSA, robust salient object detection）方法来抵御对抗攻击，由于对抗噪声是攻击者根据目标模型精心构造的，相对

于普通噪声来说，目标检测模型对对抗噪声更敏感，对普通噪声更鲁棒。ROSA 通过分段屏蔽组件和上下文感知恢复组件完成鲁棒的显著性目标检测。首先，分段屏蔽组件通过引入随机噪声，利用超像素对载有随机噪声的图像进行区域划分；然后，对同一个区域内的像素进行随机排列，以此破坏对抗噪声。这不仅能限制新引入的噪声对图像造成影响，同时不会破坏图像的轮廓边界，还能进行数据增强，防止过拟合。然后，ROSA 方法将重新排列的图像送入全卷积网络中得到粗糙的显著图像，上下文感知恢复组件利用图模型来微调粗糙显著图像，从而得到最终的显著图像。ROSA 方法利用简单的超像素方法极大可能消除全局噪声对目标模型的影响，但是对于局部对抗补丁攻击可能会失效。文献[97,112]提出不同的降噪思想掩盖局部噪声消除对抗攻击，局部噪声的降噪流程如图 7 所示。Zhou 等[112]提出基于信息分布的局部补丁对抗防御方法，由于对抗样本和干净样本两者信息分布不一致，对抗样本一般携带高频噪声，因此提出先利用基于熵的建议组件确定对抗补丁位置；由于直接对建议框内的像素全部灰度化可能影响真正图片的信息，又提出利用基于梯度的过滤组件，筛选出不平滑的噪声，对其进行灰度化，去除对抗噪声的影响。这种方法对高频对抗样本防御效果较好，并且在干净样本上的任务精度影响较小，但是无法抵御低频对抗攻击。Chiang 等[97]基于此提出对抗像素掩模（APM, adversarial pixel masking）方法去除高频和低频噪声的影响。APM 方法认为对抗训练导致目标检测任务精度下降的原因是干净样本和对抗样本存在差异，提出利用一个数据预处理网络来学习对抗噪声位置，生成噪声掩模矩阵，只要将对抗样本的分布恢复为干净样本的分布即可弥补模型精度的差距，如利用 U-Net 结构。类似地，Liu 等[51]提出分割和完善对抗补丁形状的对抗噪声去除方法，利用二元交叉熵来训练分割模型，确定对抗补丁的位置，与 Zhou 等[112]方法不同的是，在训练分割模型的过程中引入对抗训练的思想，以此提高分割模型非鲁棒性，防止二次攻击。尽管这几种对抗噪声拥有类似的噪声去除思想，Li 等[111]提出的 ROSA 方法更适用于微小不可见的全局噪声，其他方法更适用于局部对抗补丁攻击的防御。

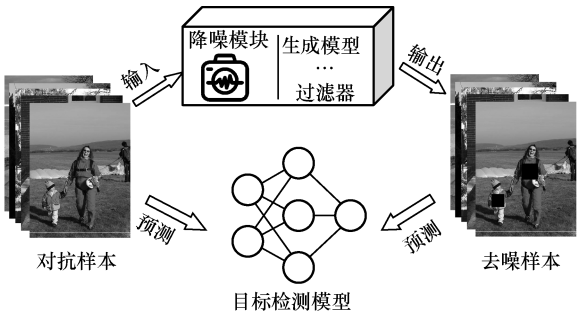


图 7　局部噪声的降噪流程

3.2.2　对抗噪声检测

Xiang 等[113]基于小感受野破坏局部对抗补丁的思想提出了针对图像分类对抗攻击的防御方法 PatchGuard，在此基础上提出了针对目标检测对抗防御算法 DetectorGuard[95]。DetectorGuard 认为鲁棒性的图像分类器[113]任务精度下降的主要原因如下。1) 目标对象类别标签错误；2) 把背景识别为目标；3) 把目标对象识别为背景。DetectorGuard 尝试将基于小感受野的鲁棒图像分类器迁移到目标检测中作为目标预测器，首先利用小感受野卷积核对各个区域进行分类，以此消除原因 1)引起的类别标签错误。然后，通过设定目标存在最低阈值，生成是否有目标的目标特征图，消除原因 2)引起的将背景识别为目标的错误。最后，通过目标解释器对比目标特征图和基础目标检测器结果，判断是否存在对抗攻击，此时可能出现 3 种情况：1) 如果两者相同位置均有目标，则不存在对抗样本；2) 如果预测器有目标，但是目标检测器相应位置没有，则认为对抗噪声攻击目标检测器并企图使其检测失败，此时发出攻击报警；3) 如果预测器没有目标，但是检测器相应位置存在目标，该文献声称是鲁棒的图像分类器的缺陷引起，则解释器认为不存在对抗样本，以此消除原因 3)引起的误判。至此，DetectorGuard 方法将鲁棒性的小感受野图像分类器迁移至目标检测对抗防御中，并且不会降低目标检测器的性能，文献[113]同时提出了鲁棒性认证方法，证明了其防御的有效性和可靠性，是一种认证防御方法的扩展。但是，DetectorGuard 针对情况 3)维持目标检测器的输出结果，这表明其无法抵御 Appearing Attack 这种出现型对抗攻击。

Yin 等[114]利用语言模型捕捉输入数据的内部依赖关系和上下文一致性来检测对抗样本，通过定义一种新语言 SCENE-Lang，将基础目标检测器的输出结果重新建模为语言模型的输入，输入

RoBERRTa 掩码语言模型[115]中，对所有目标对象逐一掩码遮蔽并预测遮蔽对象类别，最终与基础目标检测器输出结果进行对比，判断是否存在对抗样本。对于标签分类错误攻击和 Appearing Attack，通过对比语言模型和基础目标检测器结果即可发现，对于 Hiding Attack，Yin 等[114]提出根据语言模型预测结果计算上下文得分，由于某一个目标对象消失也会影响其他目标的预测，因此通过设定上下文得分阈值来判断某个目标对象是否存在。文献[114]也指出，如果少量个数对象（如一个）消失，检测方法的性能稍微有所下降。Yin 等[114]提出的对抗噪声检测方法利用上下文一致性来检测对抗攻击，仅捕捉目标对象的内部依赖关系，因此可能无法防御上下文攻击如 CCA[68]、ZQA[70]等。

面向目标检测对抗防御方法演进及联系如图 8 所示。

4 未来研究方向

目标检测对抗攻击的发展是其对抗防御研究进展的基础，有效的对抗攻击对目标检测模型鲁棒性的评估也至关重要。从目前已有的相关研究来看，本文认为今后关于目标检测对抗学习的研究，将主要围绕以下几个方向展开。

1) 面向目标检测的黑盒攻击。一旦黑盒攻击达到一定的攻击成功率，对基于目标检测的其他重要领域，如自动驾驶领域等，将会造成不可磨灭的影响。目前关于目标检测的黑盒攻击研究尚处于起步阶段，但已从基于可转移性的方法和查询攻击 2 个维度进行了探索。一方面，目前基于可转移性的方法主要利用任务间以及任务内的可转移性来实现黑盒攻击，两者的核心思想是通过学习数据的多样性，即面向更多的模型，以增强对抗样本的可转移性，以此实现黑盒攻击。然而，如文献[78,80]所示，由于昂贵的计算开销以及成本限制，集成的模型数量有限，且缺乏模型选择的可解释性，不能确保生成的对抗噪声对其他所有模型有效；同时，由于目标检测模型的原理差异，针对不同模型生成的对抗噪声可能存在杠杆效应，增强某一模型的攻击成功率可能会削弱对其他模型的攻击效果，对不同模型的攻击成功率难以平衡，模型之间的可转移性难以提高。因此，如何选择集成模型或从其他领域（如域泛化、深度学习解释性等）提出更具有可转移性的黑盒攻击方法是一个重要的研究方向。另一方面，由于目标检测输入数据的样本空间过大，导致查询效率和攻击成功率难以同时提高。以查询攻击 PRFA 为例，一是需要利用先验知识提高查询效率，二是限制了对抗噪声的生成区域和攻击场景，尽管在数字世界中研究目标检测对抗攻击具有价值，但是目标检测基础任务的特性决定了对于物理世界中的攻击研究不可忽视。如何在弱假设下，在有限的查询次数下实现物理世界黑盒攻击仍然是一个挑战。深入研究目标检测的黑盒攻击不仅可以加深对目标检测模型内部逻辑结构和对抗样本的理解，更有助于探索和提高目标检测模型的鲁棒性和安全性。

2) 对抗攻击的语义性。在公众场所等环境中，具有语义性的自然对抗噪声显得尤为重要，能够避免引起注意力。虽然当前的研究已经尝试通过图像平滑度、颜色特征、边缘特征等生成具有语义性的对抗噪声，但这方面的研究仍然相对有限。一方面，现有研究在对抗噪声的语义性定义上缺乏统一标准，大多数研究依赖于人类的主观评估，效率低且缺乏公平性，如何量化对抗噪声的语义性并提出可靠的语义性指标是一个尚需解决的热点问题。另一

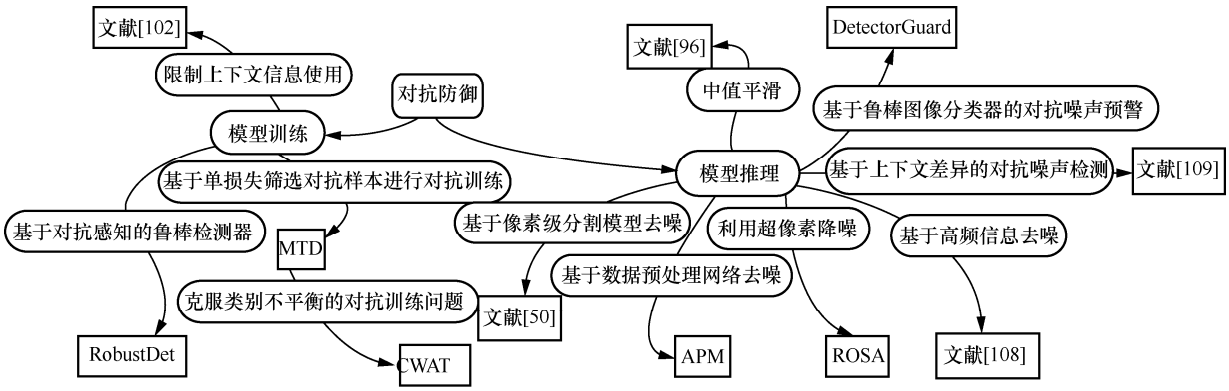


图 8 面向目标检测对抗防御方法演进及联系

方面，相关研究显示，对抗样本的语义性和攻击成功率之间存在负相关关系，为了提高对抗攻击成功率，往往需要以牺牲对抗样本的语义性为代价。此外，由于全局噪声语义性较高，但是难以部署在物理世界中。因此，如何在保持攻击成功率的前提下生成语义性强的局部对抗噪声也是未来研究的一个热点。

3) 面向目标检测的对抗防御。关于目标检测的对抗防御的研究，不仅能够提高模型的安全性，同时有助于揭示对抗样本产生的原因。目前关于对抗噪声的研究主要集中于图像分类领域，对抗训练已被公认为对抗攻击的一种可靠防御手段，然而，对抗训练的实施需要为特定的模型和攻击策略生成对抗样本，并进行模型的重训练，这不仅需要大量的额外计算资源，还可能对目标检测模型的任务准确率产生负面影响，同时对抗训练针对其他攻击方法生成的对抗样本的防御泛化性难以保证。因此，考虑到目标检测多任务的复杂性和重要性，如何设计出更鲁棒、高效且有效的目标检测防御策略是一个亟待解决的问题。此外，如何防止攻击者在已知防御措施的情况下进行二次对抗攻击的问题，只依赖于鲁棒的目标检测模型是无法解决的，为了确保生成的对抗噪声在实际应用中缺乏攻击力，可能需要结合语义性和多种防御策略，如何有效防御二次对抗攻击是一个长期且值得深入探索的问题。

5　结束语

目标检测作为计算机视觉的基础任务之一，是其他重要领域的奠基石，关于目标检测的对抗攻防吸引了众多研究者的关注。为厘清现有研究成果的优势与不足，明确未来的研究方向，本文从攻击者知识、攻击场景系统地研究和分析了目标检测中的对抗攻击方法和联系，从模型训练和模型推理角度梳理和归纳了对抗防御方法及其不同，回顾了大量极具影响力的目标检测对抗攻防研究成果。同时，本文指出了目标检测对抗学习研究中面临的挑战，探讨了未来可行的研究方向，旨在为推动目标检测性能和安全的进一步发展和应用提供指导和参考。

参考文献：

[1] CHEN D J, HSIEH H Y, LIU T L. Adaptive image transformer for one-shot object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 12242-12251.

[2] WANG Y, LV H, KUANG X, et al. Towards a physical-world adversarial patch for blinding object detection models[J]. Information Sciences, 2021, 556: 459-471.

[3] CHAUDHURI B, VESDAPUNT N, WANG B Y. Joint face detection and facial motion retargeting for multiple faces[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9711-9720.

[4] ZHENG L Y, TANG M, CHEN Y Y, et al. Improving multiple object tracking with single object tracking[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 2453-2462.

[5] SHEKAR A K, GOU L, REN L, et al. Label-free robustness estimation of object detection CNNs for autonomous driving applications[J]. International Journal of Computer Vision, 2021, 129(4): 1185-1201.

[6] LOEY M, MANOGARAN G, TAHA M H N, et al. Fighting against COVID-19: a novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection[J]. Sustainable Cities and Society, 2021, 65: 102600.

[7] ZENG Y L, ZHANG L H, ZHAO J H, et al. JRL-YOLO: a novel jump-join repetitious learning structure for real-time dangerous object detection[J]. Computational Intelligence and Neuroscience, 2021, 2021: 1-16.

[8] BI H B, ZHANG C, WANG K, et al. Rethinking camouflaged object detection: models and datasets[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 5708-5724.

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[10] XIE C H, WANG J Y, ZHANG Z S, et al. Adversarial examples for semantic segmentation and object detection[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 1378-1387.

[11] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Physical adversarial examples for object detectors[J]. arXiv Preprint, arXiv: 1807.07769, 2018.

[12] ZOLFI A, KRAVCHIK M, ELOVICI Y, et al. The translucent patch: a physical and universal attack on object detectors[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 15227-15236.

[13] 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(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(10): 2873-2896.

[14] ZHANG C N, BENZ P, KARJAUV A, et al. Investigating top-k white-box and transferable black-box attack[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15064-15073.

[15] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 658-666.

[16] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[C]//Proceedings of 18th International Conference on Pattern

Recognition. Piscataway: IEEE Press, 2006: 850-855.

[17] ZOU Z, CHEN K, SHI Z, et al. Object detection in 20 years: a survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[18] ZHANG S F, CHI C, YAO Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9756-9765.

[19] 陈梦轩, 张振永, 纪守领, 等. 图像对抗样本研究综述[J]. 计算机科学, 2022, 49(2): 92-106.
CHEN M X, ZHANG Z Y, JI S L, et al. Survey of research progress on adversarial examples in images[J]. Computer Science, 2022, 49(2): 92-106.

[20] LI Y W, BAI S, ZHOU Y Y, et al. Learning transferable adversarial examples via ghost networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11458-11465.

[21] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[J]. arXiv Preprint, arXiv: 1611.02770, 2016.

[22] ZAIDI S S A, ANSARI M S, ASLAM A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2022, 126: 103514.

[23] CARRANZA-GARCÍA M, TORRES-MATEO J, LARA-BENÍTEZ P, et al. On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data[J]. Remote Sensing, 2020, 13(1): 89.

[24] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 779-788.

[25] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 6517-6525.

[26] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv Preprint, arXiv: 1804.02767, 2018.

[27] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv Preprint, arXiv: 2004.10934, 2020.

[28] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.

[29] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[30] LAW H, DENG J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[31] DUAN K W, BAI S, XIE L X, et al. CenterNet: keypoint triplets for object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 6568-6577.

[32] ZHOU X, WANG D, KRÄHENBÜHL P. Objects as points[J]. arXiv Preprint, arXiv: 1904.07850, 2019.

[33] TAN M X, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10778-10787.

[34] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[35] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2016: 1440-1448.

[36] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2016: 1137-1149.

[37] DAI J, LI Y, HE K, et al. R-FCN: object detection via region-based fully convolutional networks[J]. arXiv Preprint, arXiv: 1605.06409, 2016.

[38] ZHANG X S, WAN F, LIU C, et al. Learning to match anchors for visual object detection[C]//Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2021: 3096-3109.

[39] LAW H, TENG Y, RUSSAKOVSKY O, et al. CornerNet-Lite: efficient keypoint based object detection[J]. arXiv Preprint, arXiv: 1904.08900, 2019.

[40] WANG J Q, CHEN K, YANG S, et al. Region proposal by guided anchoring[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 2960-2969.

[41] ZHU C C, HE Y H, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 840-849.

[42] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 9626-9635.

[43] LIU W, LIAO S C, REN W Q, et al. High-level semantic feature detection: a new perspective for pedestrian detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 5182-5191.

[44] KONG T, SUN F C, LIU H P, et al. FoveaBox: beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[45] EVERINGHAM M. The pascal visual object classes challenge results[R]. 2007.

[46] EVERINGHAM M, GOOL L V, WILLIAMS C K, et al. The pascal visual object classes challenge 2012 results[R]. 2012.

[47] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.

[48] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2005: 886-893.

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

[50] BRAUNEGG A, CHAKRABORTY A, KRUMDICK M, et al. APRICOT: a dataset of physical adversarial attacks on object detection[J].

arXiv Preprint, arXiv: 1912.08166, 2019.

[51] LIU J, LEVINE A, LAU C P, et al. Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 14953-14962.

[52] YU F, CHEN H F, WANG X, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 2633-2642.

[53] ERTLER C, MISLEJ J, OLLMANN T, et al. The mapillary traffic sign dataset for detection and classification on a global scale[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 68-84.

[54] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 3686-3693.

[55] HOU C C. The application of human detection based on YOLOv5[J]. Highlights in Science, Engineering and Technology, 2023, 34: 203-208.

[56] ZHAO Y, ZHU H, LIANG R G, et al. Seeing isn't believing: towards more robust adversarial attack against real world object detectors[C]//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 1989-2004.

[57] XIAO C W, YANG D W, LI B, et al. MeshAdv: adversarial meshes for visual recognition[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 6891-6900.

[58] SURYANTO N, KIM Y, KANG H, et al. DTA: physical camouflage attacks using differentiable transformation network[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15284-15293.

[59] WANG J K, LIU A S, YIN Z X, et al. Dual attention suppression attack: generate adversarial camouflage in physical world[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 8561-8570.

[60] HUANG L F, GAO C Y, ZHOU Y Y, et al. Universal physical camouflage attacks on object detectors[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 717-726.

[61] DUAN Y X, CHEN J L, ZHOU X Y, et al. DPA: learning coated adversarial camouflages for object detectors[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022: 1-14.

[62] WANG D H, JIANG T S, SUN J L, et al. FCA: learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 2414-2422.

[63] TAN J, JI N, XIE H D, et al. Legitimate adversarial patches: evading human eyes and detection models in the physical world[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 5307-5315.

[64] WEI X X, LIANG S Y, CHEN N, et al. Transferable adversarial attacks for image and video object detection[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 954-960.

[65] HU Y C T, CHEN J C, KUNG B H, et al. Naturalistic physical adversarial patch for object detectors[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 7828-7837.

[66] HU Z H, HUANG S Y, ZHU X P, et al. Adversarial texture for fooling person detectors in the physical world[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 13297-13306.

[67] ZHANG Y, FOROOSH H, DAVID P, et al. CAMOU: learning physical vehicle camouflages to adversarially attack detectors in the wild[C]//Proceedings of International Conference on Learning Representations. Vancouver: ICLR, 2019: 1-20.

[68] CAI Z K, XIE X X, LI S S, et al. Context-aware transfer attacks for object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 149-157.

[69] HUANG H, CHEN Z Y, CHEN H R, et al. T-SEA: transfer-based self-ensemble attack on object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 20514-20523.

[70] CAI Z K, RANE S, BRITO A E, et al. Zero-query transfer attacks on context-aware object detectors[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15004-15014.

[71] LIANG S Y, WU B Y, FAN Y B, et al. Parallel rectangle flip attack: a query-based black-box attack against object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 7677-7687.

[72] LI Y, BIAN X, CHANG M C, et al. Exploring the vulnerability of single shot module in object detectors via imperceptible background patches[J]. arXiv Preprint, arXiv: 1809.0596612, 2018.

[73] LI J, SCHMIDT F R, KOLTER J Z. Adversarial camera stickers: a physical camera-based attack on deep learning systems[J]. arXiv Preprint, arXiv: 1904.00759. 2019.

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

[75] ZENG X H, LIU C X, WANG Y S, et al. Adversarial attacks beyond the image space[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 4297-4306.

[76] KO N. Directional statistics BRDF model[C]//Proceedings of IEEE 12th International Conference on Computer Vision. Piscataway: IEEE Press, 2010: 476-483.

[77] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[J]. arXiv Preprint, arXiv: 1707.07397, 2017.

[78] XU K D, ZHANG G Y, LIU S J, et al. Adversarial T-shirt! evading person detectors in a physical world[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 665-681.

[79] THYS S, RANST W V, GOEDEMÉ T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE

Press, 2020: 49-55.
[80] WU Z X, LIM S N, DAVIS L S, et al. Making an invisibility cloak: real world adversarial attacks on object detectors[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 1-17.
[81] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 1528-1540.
[82] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
[83] POURSAEED O, KATSMAN I, GAO B C, et al. Generative adversarial perturbations[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4422-4431.
[84] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[J]. arXiv Preprint, arXiv: 1809.11096, 2018.
[85] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4217-4228.
[86] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, et al. Learning deep representations by mutual information estimation and maximization[J]. arXiv Preprint, arXiv: 1808.06670, 2018.
[87] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2016: 640-651.
[88] HUANG G, SUN Y, LIU Z, et al. Deep networks with stochastic depth[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 646-661.
[89] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2980-2988.
[90] YANG Z, LIU S H, HU H, et al. RepPoints: point set representation for object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 9656-9665.
[91] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 484-501.
[92] CHEN W L, ZHANG Z X, HU X L, et al. Boosting decision-based black-box adversarial attacks with random sign flip[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 276-293.
[93] 李明慧, 江沛佩, 王骞, 等. 针对深度学习模型的对抗性攻击与防御[J]. 计算机研究与发展, 2021, 58(5): 909-926.
LI M H, JIANG P P, WANG Q, et al. Adversarial attacks and defenses for deep learning models[J]. Journal of Computer Research and Development. 2021, 58(5): 909-926.
[94] JIN G, YI X, HUANG W, et al. Enhancing adversarial training with second-order statistics of weights[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 1-15.
[95] XIANG C, MITTAL P. DetectorGuard: provably securing object detectors against localized patch hiding attacks[J]. arXiv Preprint, arXiv: 2012.02956, 2021.
[96] CHIANG P Y, CURRY M J, ABDELKADER A, et al. Detection as regression: certified object detection by Median smoothing[J]. arXiv Preprint, arXiv: 2007.03730, 2020.
[97] CHIANG P H, CHAN C S, WU S H. Adversarial pixel masking: a defense against physical attacks for pre-trained object detectors[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 1856-1865.
[98] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.
[99] ZHANG H C, WANG J Y. Towards adversarially robust object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 421-430.
[100] CHEN S T, CORNELIUS C, MARTIN J, et al. ShapeShifter: robust physical adversarial attack on faster R-CNN object detector[C]//Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2019: 52-68.
[101] CHEN P C, KUNG B H, CHEN J C. Class-aware robust adversarial training for object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 10415-10424.
[102] SAHA A, SUBRAMANYA A, PATIL K, et al. Role of spatial context in adversarial robustness for object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 3403-3412.
[103] BARNEA E, BEN-SHAHAR O. Exploring the bounds of the utility of context for object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 7404-7412.
[104] DONG Z Y, WEI P X, LIN L. Adversarially-aware robust object detector[M]. Berlin: Springer, 2022.
[105] CHEN Y P, DAI X Y, LIU M C, et al. Dynamic convolution: attention over convolution kernels[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 11027-11036.
[106] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. arXiv Preprint, arXiv: 1312.6114, 2013.
[107] COHEN J M, ROSENFELD E, KOLTER J Z. Certified adversarial robustness via randomized smoothing[J]. arXiv Preprint, arXiv: 1902.02918, 2019.
[108] LECUYER M, ATLIDAKIS V, GEAMBASU R, et al. Certified robustness to adversarial examples with differential privacy[C]// Proceedings of IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 656-672.
[109] SALMAN H, YANG G, LI J, et al. Provably robust deep learning via adversarially trained smoothed classifiers[J]. arXiv Preprint, arXiv: 1906.04584, 2019.
[110] ZUO W M, ZHANG K, ZHANG L. Convolutional neural networks for image denoising and restoration[M]. Berlin: Springer, 2018.
[111] LI H F, LI G B, YU Y Z. ROSA: robust salient object detection against adversarial attacks[J]. IEEE Transactions on Cybernetics, 2020, 50(11): 4835-4847.
[112] ZHOU G Z, GAO H C, CHEN P, et al. Information distribution based defense against physical attacks on object detection[C]//Proceedings of IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway: IEEE Press, 2020: 1-6.
[113] XIANG C, BHAGOJI A N, SEHWAG V, et al. PatchGuard: a provably robust defense against adversarial patches via small receptive fields

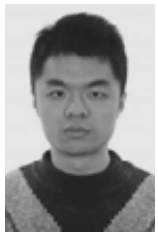
and masking[J]. arXiv Preprint, arXiv: 2005.10884, 2020.
[114] YIN M J, LI S S, CAI Z K, et al. Exploiting multi-object relationships for detecting adversarial attacks in complex scenes[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 7838-7847.
[115] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized bert pretraining approach[J]. arXiv Preprint, arXiv: 1907.11692, 2019.

[作者简介]

汪欣欣（1995− ），女，湖北随州人，武汉大学博士生，主要研究方向为目标检测、对抗学习和后门学习等。

陈晶（1981− ），男，湖北武汉人，博士，武汉大学教授、博士生导师，主要研究方向为网络安全、人工智能安全、分布式系统安全和区块链等。

何琨（1986− ），男，湖北武汉人，博士，武汉大学副教授、博士生导师，主要研究方向为应用密码学、网络安全、云计算安全、人工智能安全和区块链安全等。



张子君（1989− ），男，湖北武汉人，博士，武汉大学副教授，主要研究方向为神经网络优化算法、正则化、网络架构、表示学习和强化学习等。

杜瑞颖（1964− ），女，河南新乡人，博士，武汉大学教授、博士生导师，主要研究方向为网络安全、隐私保护、云安全和移动安全等。

李瞧（1995− ），女，辽宁辽阳人，武汉大学博士生，主要研究方向为人工智能安全、对抗学习和后门学习等。

佘计思（1999− ），女，湖北随州人，武汉大学硕士生，主要研究方向为人工智能安全、目标检测。