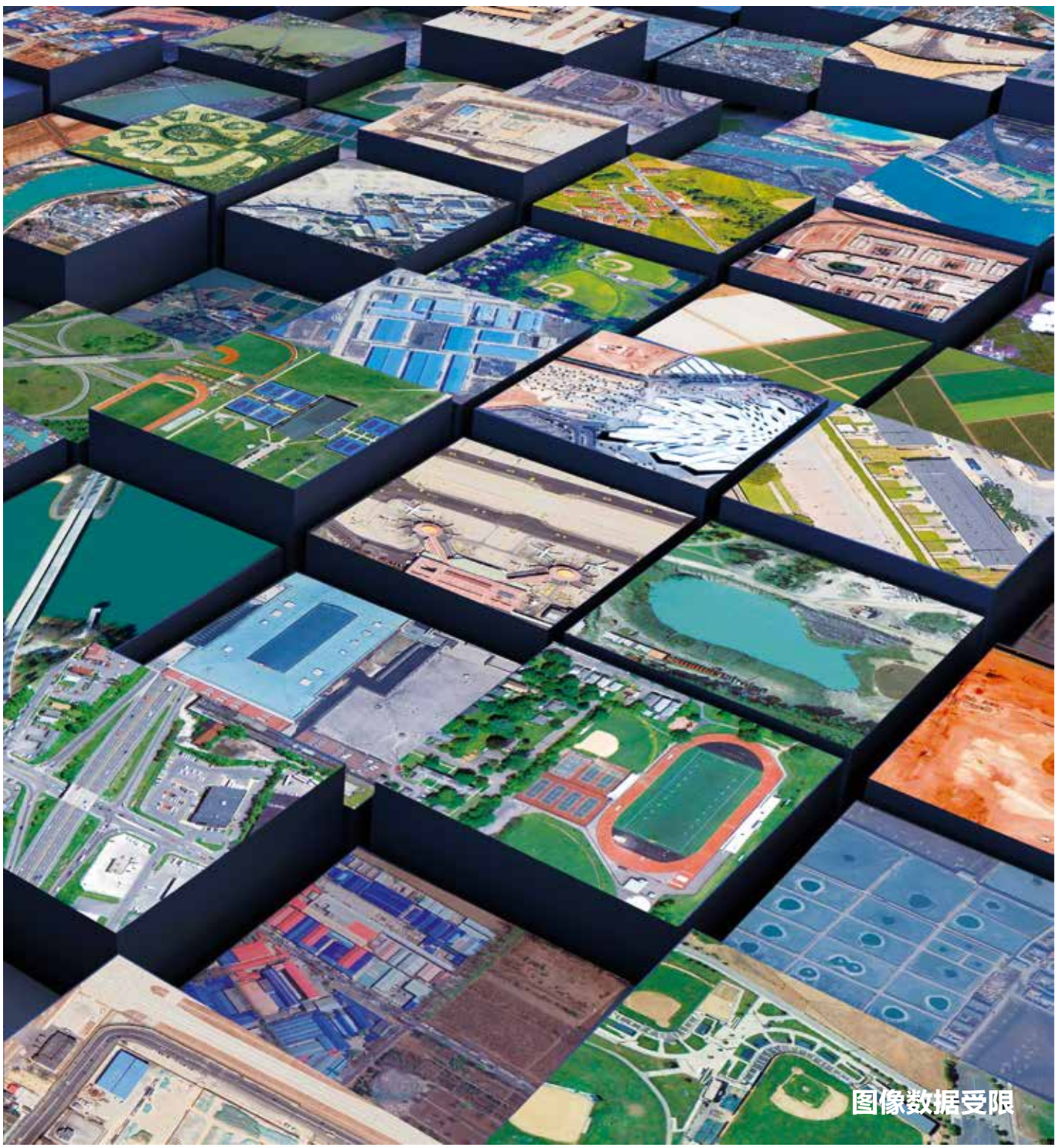
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
10
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第10期（总第318期）
2022年10月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

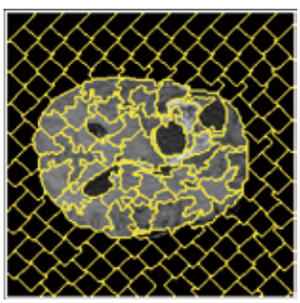
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

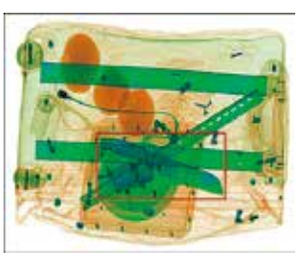
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年10月 第27卷 第10期（总第318期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

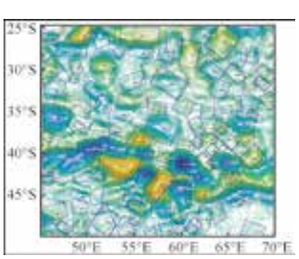
目次



MRI脑肿瘤图像的超像素/体素分割及发展现状(第2897页)



融合多尺度特征与全局上下文信息的X光违禁物品检测(第3043页)

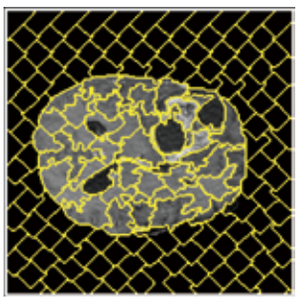


融合多尺度旋转锚机制的海洋中尺度涡自动检测(第3092页)

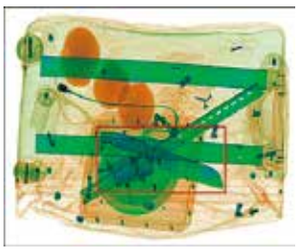
封面素材来自中国科学院空天信息创新研究院二部孙显研究员团队

Volume 27, Number 10
Published October 16, 2022

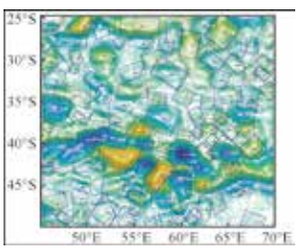
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



The review of superpixel/voxel segmentation of MRI brain tumor images(P2897)



Integrated multi-scale features and global context in x-ray detection for prohibited items(P3043)



Multi-scale rotating anchor mechanism based automatic detection of ocean mesoscale eddy(P3092)

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2022)10-2873-24

论文引用格式:Yuan L, Li X M, Pan Z X, Sun J M and Xiao L. 2022. Review of adversarial examples for object detection. Journal of Image and Graphics,27(10):2873-2896(袁珑,李秀梅,潘振雄,孙军梅,肖蕾. 2022. 面向目标检测的对抗样本综述. 中国图象图形学报,27(10):2873-2896)

[DOI:10.11834/jig.210209]

面向目标检测的对抗样本综述

袁珑[1],李秀梅[1],潘振雄[1],孙军梅[1*],肖蕾[2]

1. 杭州师范大学信息科学与技术学院,杭州　311121;2. 福建省软件评测工程技术研究中心,厦门　361024

摘　要:目标检测是一种广泛应用于工业控制和航空航天等安全攸关场景的重要技术。随着深度学习在目标检测领域的应用,检测精度得到较大提升,但由于深度学习固有的脆弱性,使得基于深度学习的目标检测技术的可靠性和安全性面临新的挑战。本文对面向目标检测的对抗样本生成及防御的研究分析和总结,致力于为增强目标检测模型的鲁棒性和提出更好的防御策略提供思路。首先,介绍对抗样本的概念、产生原因以及目标检测领域对抗样本生成常用的评价指标和数据集。然后,根据对抗样本生成的扰动范围将攻击分为全局扰动攻击和局部扰动攻击,并在此分类基础上,分别从攻击的目标检测器类型、损失函数设计等方面对目标检测的对抗样本生成方法进行分析和总结,通过实验对比了几种典型目标检测对抗攻击方法的性能,同时比较了这几种方法的跨模型迁移攻击能力。此外,本文对目前目标检测领域常用的对抗防御策略进行了分析和归纳。最后,总结了目标检测领域对抗样本的生成及防御面临的挑战,并对未来发展方向做出展望。

关键词:目标检测;对抗样本;深度学习;对抗防御;全局扰动;局部扰动

Review of adversarial examples for object detection

Yuan Long[1], Li Xiumei[1], Pan Zhenxiong[1], Sun Junmei[1*], Xiao Lei[2]

1. School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China;

2. Engineering Research Center for Software Testing and Evaluation of Fujian Province, Xiamen 361024, China

Abstract: Object detection is essential for various of applications like semantic segmentation and human facial recognition, and it has been widely employed in public security related scenarios, including automatic driving, industrial control, and aerospace applications. Traditional object detection technology requires manual-based feature extraction and machine learning methods for classification, which is costly and inaccuracy for detection. Recent deep learning based object detection technology has gradually replaced the traditional object detection technology due to its high detection efficiency and accuracy. However, it has been proved that convolutional neural network (CNN) can be easily fooled by some imperceptible perturbations. These images with the added imperceptible perturbations are called adversarial examples. Adversarial examples were first discovered in the field of image classification, and were gradually developed into other fields. To clarify the vulnerabilities of adversarial attack and deep object detection system it is of great significance to improve the robustness and security of the deep learning based object detection model by using a holistic approach. We aims to enhancing the robustness of object detection models and putting forward defense strategies better in terms of analyzing and summarizing the adversarial attack and

收稿日期:2021-03-19;**修回日期**:2021-07-29;**预印本日期**:2021-08-04

*通信作者:孙军梅　junmeisun@hznu.edu.cn

基金项目:国家自然科学基金项目(61801159);杭州市科技计划项目(20201203B124)

Supported by: National Natural Science Foundation of China (61801159); Science and Technology Plan Project of Hangzhou(20201203B124)

defense methods for object detection recently. First, our review is focused on the discussion of the development of object detection, and then introduces the origin, growth, motives of emergence and related terminologies of adversarial examples. The commonly evaluation metrics used and data sets in the generation of adversarial examples in object detection are also introduced. Then, 15 adversarial example generation algorithms for object detection, according to the generation of perturbation level classification, are classified as global perturbation attack and local perturbation attack. A secondary classification under the global perturbation attack is made in terms of the types of of attacks detector like attack on two-stage network, attack on one-stage network, and attack on both kinds of networks. Furthermore, these attack methods are classified and summarized from the following perspectives as mentioned below: 1) the attack methods can be divided into black box attack and white box attack based on the issue of whether the attacker knows the information of the model's internal structure and parameters or not. 2) The attack methods can be divided into target attack and non-target attack derived from the identification results of the generated adversarial examples. 3) The attack methods can be divided into three categories: L_0, L_2 and L_∞ via the perturbation norm used by the attack algorithm. 4) The attack methods can be divided into single loss function attack and combined loss function attack based on the loss function design of attack algorithm. These methods are summarized and analyzed on six aspects of the object detector type and the loss function design, and the following rules of the current adversarial example generation technology for object detection are obtained: 1) diversities of attack forms: a variety of adversarial loss functions are combined with the design of adversarial attack methods, such as background loss and context loss. In addition, the diversity of attack forms is also reflected in the context of diversity of attack methods. Global perturbations and local perturbations are represented by patch attacks both. 2) Diversities of attack objects: with the development of object detection technology, the types of detectors become more diverse, which makes the adversarial examples generation technology against detectors become more changeable, including one-stage attack, two-stage attack, as well as the attack against anchor-free detector. It is credible that future adversarial examples attacks against new techniques of object detection have its potentials. 3) Most of the existing adversarial attack methods are white box attack methods for specific detector, while few are black box attack methods. The reason is that object detection model is more complex and the training cycle is longer compared to image classification, so attacking object detection requires more model information to generate reliable adversarial examples. The issue of designing more and more effective black box attacks can be as a future research direction as well. Additionally, we select four classical methods of those are dense adversary generation (DAG), robust adversarial perturbation (RAP), unified and efficient adversary (UEA), and targeted adversarial objectness gradient attacks (TOG), and carry out comparative analysis through experiments. Then, the commonly attack and defense strategies are introduced from the perspectives of preprocessing and improving the robustness of the model, and these methods are summarized. The current methods of defense against examples are few, and the effect is not sufficient due to the specialty of object detection. Furthermore, the transferability of these models is compared to you only look once (YOLO)-Darknet and single shot multibox detector (SSD300) models, and the experimental results show that the UEA method has the best transferability among these methods. Finally, we summarize the challenges in the generation and defense of adversarial examples for object detection from the following three perspectives: 1) to enhance the transferability of adversarial examples for object detection. Transfer ability is one of the most important metrics to measure adversarial examples, especially in object detection technology. It is potential to enhance the transferability of adversarial examples to attack most object detection systems. 2) To facilitate adversarial defense for object detection. Current adversarial examples attack paths are lack of effective defenses. To enhance the robustness of object detection, it is developed for defense research against adversarial examples further. 3) Decrease the disturbance size and increase the generation speed of adversarial examples. Future development of it is possible to develop adversarial examples for object detection in related to shorter generation time and smaller generation disturbance in the future.

Key words: object detection; adversarial examples; deep learning; adversarial defense; global perturbation; local perturbation

0 引 言

随着计算机软硬件技术的不断发展,基于卷积神经网络的深度学习技术广泛应用于社会各个领域(LeCun 等,2015),尤其在计算机视觉领域的图像分类(Deng 等,2009)、目标检测(Liu 等,2019a)、人脸识别(Liu 等,2017)和语义分割(Ding 和 Zhao,2018)等方面更是取得了巨大成功。但由于深度学习自身的脆弱性,在一些应用场景容易受到对抗样本(adversarial examples)对模型的攻击(Athalye 等,2018)。对抗样本最早在图像分类领域提出(Szegedy 等,2014),比较典型的研究有 FGSM(fast gradient sign method)(Goodfellow 等,2015;Dong 等,2018)、DeepFool(Moosavi-Dezfooli 等,2016)和 C&W(Carlini-Wagner)(Carlini 和 Wagner,2017)等。随着研究的不断深入,对抗样本不仅攻击图像分类,也开始攻击其他计算机视觉任务,如面部识别(Sharif 等,2016;Liu 等,2017)、视觉跟踪(Bertinetto 等,2016;Yan 等,2020)、语音识别(Du 等,2020)和自然语言处理(Ren 等,2019)等。

作为计算机视觉的核心任务,基于深度学习的目标检测(曹家乐 等,2022)在人工智能领域扮演着越来越重要的角色,许多其他计算机视觉任务诸如人脸识别、目标追踪和图像分割都是基于目标检测实现的。因此目标检测的安全对计算机视觉的发展至关重要。Xie 等人(2017)首次在目标检测和语义分割任务上证明,目标检测和图像分类都存在类似的安全性问题。对抗样本在目标检测领域的出现对目标检测器的鲁棒性提出了巨大考验。现有的对抗样本总结分析(Yuan 等,2019;Akhtar 和 Mian,2018;潘文雯等,2020)主要集中在图像分类领域,鲜有论文对目标检测领域的对抗样本生成方法及防御进行总结和分析。本文对目标检测领域的对抗样本生成和防御进行归纳总结,以期催生更多的防御策略,从而使未来的目标检测技术更加鲁棒,从容面对更复杂的环境。

为了梳理面向目标检测领域的对抗样本生成方法及防御策略,首先,根据对抗样本扰动生成的范围,将对抗攻击分为全局扰动攻击和局部扰动攻击。在全局扰动攻击的基础上,根据攻击的目标检测器类别分为针对两阶段网络的攻击、针对单阶段网络的攻击以及两种网络均针对的攻击,并对目标检测的对抗样本生成方法进行总结和分析。然后,通过实验对典型的目标检测对抗样本生成方法的性能进行分析对比。接着,从预处理方法和提高模型鲁棒性两个角度介绍了目标检测领域应对对抗攻击的防御策略。最后,对面向目标检测的对抗样本研究面临的挑战和发展趋势进行展望。

1 背景介绍

1.1 目标检测

目标检测作为计算机视觉领域众多任务的基础一直是研究热点,它的任务是从给定的图像中提取感兴趣区域并标记出类别和位置。目前,随着深度学习神经网络的快速发展,基于深度学习的目标检测技术(Liu 等,2020;Ding 和 Zhao,2018)凭借优越的检测性能已经取代了需要人工提取特征并分类的传统目标检测方法(Divvala 等,2012;Wang 等,2009;Viola 和 Jones,2004)。

基于深度学习的主流目标检测算法根据有无候选框生成阶段分为以 Faster R-CNN(region convolutional neural network)(Ren 等,2015)为代表的两阶段检测和以 YOLO(you only look once)(Redmon 等,2016)为代表的单阶段检测。两阶段检测网络将检测物体分为两个阶段,先检测物体的位置然后进行分类。Girshick 等人(2014)首次提出 R-CNN 算法,采用选择性搜索算法(Uijlings 等,2013)从图像中提取候选框进而分类。但是 R-CNN 每一个候选框都需要进行特征提取,比较耗费时间。为此,Girshick(2015)设计了 Fast R-CNN,对图像只提取一次特征,提高了检测速度。随后,Ren 等人(2015)提出 Faster R-CNN 算法,用区域建议网络(region proposal network,RPN)代替传统的选择搜索算法,加快提取候选框的过程。

与两阶段网络不同,单阶段网络不需要 RPN 而是直接将分类和定位一次完成。单阶段检测最具代表性的网络是 YOLO 系列网络(Redmon 等,2016;Redmon 和 Farhadi,2017,2018;Bochkovskiy 等,2020)和 SSD(single shot multibox detector)网络(Liu 等,2016)。2016 年提出的 YOLOv1 通过舍弃候选框生成阶段加快网络检测速度,但是降低了精度。同年提出的 SSD 网络通过引入多尺度信息,在保持速度的同时提高了精度。YOLOv2 和 YOLOv3 分别

添加了多尺度信息和设计了更强的骨干网络 DarkNet53 以提高提取特征的能力。YOLOv4 与 YOLOv3 相比,将骨干网络升级为学习能力更强的 CSPDarknet(cross stage partial Darknet),在 YOLOv3 的特征金字塔网络(feature pyramid networks,FPN)基础上加入路径聚合网络(path aggregation network,PAN)(Liu 等,2018)和空间金字塔池化模块(spatial pyramid pooling,SPP)(He 等,2015)。此外,近几年提出了一类检测框架 anchor-free。这类框架通过回归得到物体的关键点(例如左上角和右下角或者物体的中心点),进而得到边界框,这一类检测框架的代表网络有 CornerNet(Law 和 Deng,2018)、ExtremeNet(Zhou 等,2019b)、CenterNet(Duan 等,2019)和 Fcos(fully convolutional one-stage object detection)(Tian 等,2019)等。

1.2 对抗样本

1.2.1 对抗样本的概念

对抗样本由 Szegedy 等人(2014)首次提出。指在原本干净的数据集中,通过某种方式或遵循某种规律,向图像中加入一些细微的噪声(又称为扰动)形成的图像。在分类任务中,这类样本会使已经训练好的机器学习或者深度学习模型容易产生错误的分类结果。如图 1(Goodfellow 等,2015)所示,干净样本(左)通过人眼判断和模型输出的结果均为熊猫,但是添加对抗噪声后的图像送入模型则输出结果为长臂猿。对抗攻击前,模型输出熊猫的置信度为 57.7%,将对抗样本输入模型后,得到 99.3% 的长臂猿高置信度。在目标检测任务中,这类样本则会使模型输出错误的分类和定位结果。图 2(Xie 等,2017)为目标检测对抗样本示例。对原始样本,Faster R-CNN 能正确识别狗的类别和位置;而对对抗样本,在添加对抗扰动后,输出位置和分类都是错误的检测结果。检测器错误地检测出人和火车,却无法确定狗的存在,分类和定位都发生错误。

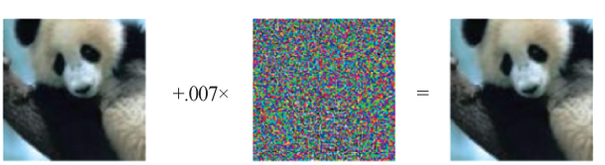
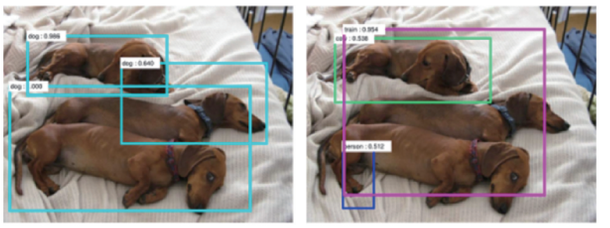


图 1 图像分类对抗样本示例(Goodfellow 等,2015)

Fig. 1 Instance of image classification adversarial examples(Goodfellow et al., 2015)



(a) 原始样本 (b) 对抗样本

图 2 目标检测对抗样本示例(Xie 等,2017)

Fig. 2 Instance of object detection adversarial examples(Xie et al., 2017)((a) original example;(b) adversarial example)

FGSM 是最早提出的图像分类攻击方法(Goodfellow 等,2015),该方法以简单的攻击思路和强大的攻击效果成为对抗攻击领域最为经典的方法之一。后续的许多方法都是在此方法的基础上做出改进,增强了攻击的稳定性。例如,将一步运算变成多步迭代的 I-FGSM(iterative-FGSM)(Kurakin 等,2017)和在迭代过程中加入动量的 MI-FGSM(momentum iterative-FGSM)(Dong 等,2018)等。除了 FGSM 系列,经典的对抗攻击方法还有 C&W 攻击(Carlini 和 Wagner,2017)、ATN(adversarial transformation networks)(Baluja 和 Fischer,2017)、单像素攻击(Su,2019)、通用对抗扰动(universal adversarial perturbations,UAP)(Moosavi-Dezfooli 等,2017)和 AdvGAN(adversarial GAN)(Xiao 等,2018)等。

1.2.2 对抗样本的产生原因

自 Szegedy 等人(2014)提出对抗样本以来,对其产生原因至今仍未有统一看法,以下是国内外学者比较认可的几个观点。Szegedy 等人(2014)认为对抗样本存在于数据流中的低概率(高维)区域,模型训练过程中只学习到了训练数据周围的局部空间,而对抗样本不处于模型训练这一局部空间,所以会使模型最后判断错误。Goodfellow 等人(2015)提出了与 Szegedy 完全相反的意见,认为正是因为神经网络模型的高维线性导致了对抗样本的产生,当在输入图像中加入少量噪声后,该细微噪声经过多层网络的传播,经过如 ReLU 或 Maxout 的线性激活函数后被无限放大,导致分类错误。Ilyas 等人(2019)指出对抗样本不是缺陷,它反映的更近似是一种特征。通常认为,模型训练会选择一些人类可以理解的特征进行分类,这些特征称为健壮性特征;但也会选择一些人类无法理解的其他特征用于区分目标,这类特征称为非健壮性特征。对抗样本归因

于非健壮性特征的存在,反映了数据的一种特征,具有高度可预测性,但这种特征是脆弱的且难以被人类理解。这类特征通常认为是模型训练的异常结果,对抗样本就是这类特征的代表。

1.2.3　对抗样本的相关术语

下面给出本文用到的对抗样本的相关术语。

1)对抗性扰动。指添加到干净样本使其成为对抗样本的噪声,一般对这种扰动有大小限制,使添加到图像上的扰动不被人眼察觉。

2)迁移性。指生成的对抗样本在不同模型、不同数据集上的攻击能力。

3)白盒攻击。指攻击时攻击者对攻击的目标模型内部的结构和参数都了解。

4)黑盒攻击。相对于白盒攻击而言,指攻击者对攻击模型的结构和参数等一切内部数据都未知。

5)欺骗率。指对抗样本进入模型以后,愚弄模型的对抗样本所占百分比。

6)目标攻击。指对抗攻击算法生成的对抗样本,能使模型将样本分类到攻击者想要的指定类别的攻击方式。

7)非目标攻击。不同于目标攻击,指攻击者生成的对抗样本使模型输出错误的分类结果,但不限制其错误类别。

1.3　目标检测的对抗样本

基于深度学习的目标检测在继承神经网络优点的同时,也容易遭受到对抗样本的攻击,这使得目标检测在实际使用时具有一定的安全隐患。由于目标检测不仅包含图像分类,还包含对目标定位,所以图像分类上的对抗攻击方法用在目标检测上效果较差,甚至绝大多数情况会攻击失败(Lu 等,2017)。目标检测是经典的多任务学习,对其进行攻击往往是根据目标检测所要达到的两个目标,即位置和类别来进行的。针对攻击方法的损失函数设计分为单损失攻击和组合损失攻击。单损失攻击在生成对抗样本时对物体进行分类损失函数攻击或者回归损失函数攻击,而组合损失攻击则综合考虑了两种损失函数来进行攻击。

基于对抗样本的目标检测攻击通过对输入图像 $\boldsymbol{x}$ 加入特定的扰动,得到扰动图像 $\boldsymbol{x}'$,并将其作为目标检测器的输入,旨在欺骗目标检测器生成随机或有目标的错误结果,其过程可以表示为

$$\min \|\boldsymbol{x}' - \boldsymbol{x}\|_p$$

$$\text{s.t.}\quad \hat{\theta}(\boldsymbol{x}') \neq \hat{\theta}(\boldsymbol{x}),\ \hat{\theta}(\boldsymbol{x}') = \theta^*(\boldsymbol{x}) \tag{1}$$

式中,p 表示两幅图像的差异距离度量,有 L_0 范数、L_2 范数和 L_∞ 范数 3 种。$\hat{\theta}(\boldsymbol{x})$ 表示网络的预测结果,$\theta^*(\boldsymbol{x})$ 为不正确的预测。

2　相关数据集和评价指标

2.1　相关数据集

目标检测中常用的数据集主要有:

1)PASCAL VOC(pattern analysis, statistical modeling and computational learning visual object classes)数据集。这是目标检测领域最常用的数据集之一,由于其轻量性,广泛应用于目标检测、图像分类和图像分割任务。数据集包含 20 个类别的物体,分为 4 大种类,每幅图像都有相应的 XML(extensible markup language)文件对应,文件包含图像物体的位置和类别。常用的 PASCAL VOC 数据集有 VOC2007(Everingham 等,2010)和 VOC2012(Shetty,2016)。其中,VOC2007 数据集包括 9 963 幅图像,由 train、val 和 test 组成。VOC2012 数据集包括 11 530 幅图像,由 train 和 test 两部分组成。现在常用的训练方法有两种。一种是使用 07_train + 12_train 作为训练集,用 07_test 作为测试集;另一种是使用 07_train + 07_test + 12_train 作为训练集,用 12_test 作为测试集。

2)MS COCO(Microsoft common objects in context)数据集(Lin 等,2014)。该数据集发布于 2014 年,是目标检测、语义分割和人体关键点检测任务较为权威的重要数据集,包括 91 个物体类别、328 000 幅图像和 250 万个标签,使用 JSON(JavaScript object notation)格式的标注文件给出每幅图像中目标像素级别的分割信息。数据集共包含 80 个对象类别的待检测目标,目标间的尺度变化大,具有较多的小目标物体。

3)ImageNet 数据集(Russakovsky 等,2015)是计算机视觉领域的一个大型数据库,广泛应用于图像分类和目标检测等任务,包括 1 400 多万幅图像,2 万多个类别。其中 103 万幅图像可以用于目标检测任务,包含 200 个物体类别,有明确的类别标注和物体的位置标注。

4)Open Image 数据集(Kuznetsova 等,2020)是谷歌团队发布的具有对象位置注释的现有最大的数

据集,包含 190 万幅图像,600 个种类,1 540 万个边界框标注。

2.2 评价标准

目标检测领域通常采用 mAP(mean average precision)(Shetty,2016)衡量对抗样本的攻击效果。mAP 为所有类别的平均精确率的均值,是衡量目标检测器检测效果最重要的一个指标。具体为

$$mAP = \frac{\sum_{i=1}^{m} AP_i}{m} \tag{2}$$

式中,m 为类别数目,AP_i 表示第 i 类物体的 AP(average precision)值。平均准确率 AP 为固定类别的精确率—召回率曲线下的面积和,表示检测器对该类别的检测能力,值越大代表检测器对该类物体的检测效果越好。

除了用 mAP 衡量模型的检测能力外,评价指标还包括精确率和召回率。

精确率(precision)表示分类正确的正样本个数与分类后判别为正样本个数的比值,衡量的是一个分类器分出来的正样本确实是正样本的概率。

召回率(recall)表示分类正确的正样本数与真正的正样本数的比值,衡量的是一个分类器能将所有的正样本都找出来的能力。在通常情况下,精确率越高,则召回率越低。

3 对抗攻击方法

Lu 等人(2017)提出在“停止”标志和人脸图像上添加扰动来误导相应的检测器,这是第一篇在目标检测领域提出对抗样本生成的文章。此后,出现了一系列针对目标检测的分类和定位两个任务进行对抗攻击的研究。根据对检测目标像素修改的数量,对抗攻击分为全局扰动攻击和局部扰动攻击。根据攻击的对象检测器类型,对抗攻击可以分为针对两阶段检测器的攻击、针对单阶段检测器的攻击以及针对两种检测器的攻击。

对于面向目标检测的对抗攻击方法,本文以全局像素攻击和局部像素攻击作为一级分类,以不同的目标检测器类型攻击作为二级分类。分类方法如图 3 所示。

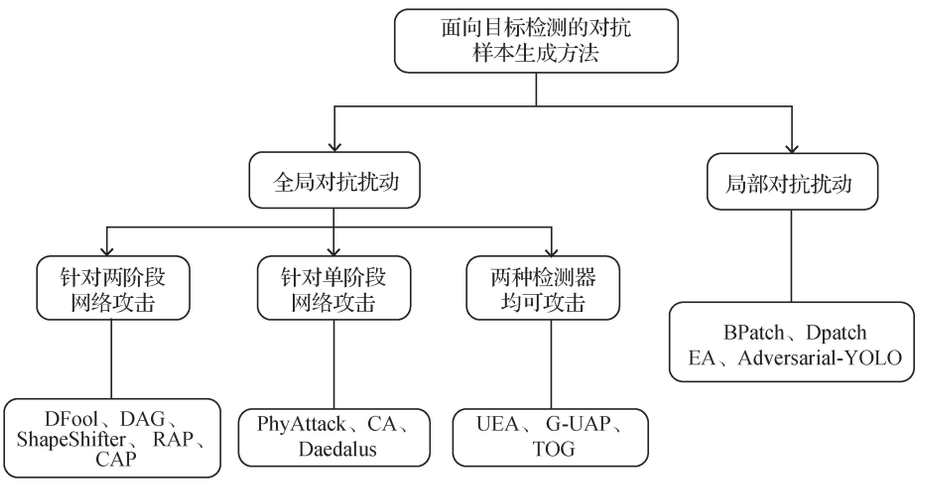


图 3 面向目标检测的对抗样本生成方法分类

Fig. 3 Classification of adversarial example generation methods for object detection

3.1 全局扰动攻击

全局扰动攻击是在对抗样本生成时对整幅图像进行像素修改,添加的噪声具有一个统一特点,即添加的噪声不是特定于一个地方而是遍布全图。

3.1.1 针对两阶段网络攻击

针对两阶段网络攻击的方法主要有 DFool(detectors fool)、DAG(dense adversary generation)、ShapeShifter、RAP(robust adversarial perturbation)和 CAP(contextual adversarial attack)方法。

1)DFool 方法。这是 Lu 等人(2017)提出的一种针对 Faster R-CNN 的白盒攻击方法,用 Faster R-CNN 对所有的停车标志进行得分测试。

$$\boldsymbol{\Phi}(T) = \frac{1}{N}\sum_{i=1}^{N} b \in \boldsymbol{B}_S(\boldsymbol{I}(M_i, \boldsymbol{T}))^{\varphi_s(b)} \tag{3}$$

式中,$\boldsymbol{T}$ 表示图像 $\boldsymbol{I}$ 在根坐标系的纹理特征,$\boldsymbol{I}(M_i, T)$ 表示对图像 $\boldsymbol{I}$ 使用映射 M_i 在 $\boldsymbol{T}$ 上进行叠加产生

的特征，$\boldsymbol{B}_S(\boldsymbol{I})$是图像 $\boldsymbol{I}$ 在 Faster R-CNN 产生的关于停车标志物体的候选框集合，$\varphi_s(b)$ 是 Faster R-CNN 对候选框的预测得分。通过最小化训练集所有图像的平均得分对式(3) 进行优化，提取梯度。在优化过程中，使用符号函数进行梯度的引导，并且通过多次迭代，当欺骗率达到 90% 时停止迭代。具体过程为

$$\begin{aligned} d^{(n)} &= \operatorname{sign}(\nabla_T \Phi(T^{(n)})) \\ T^{(n+1)} &= T^{(n)} + \varepsilon \times d^{(n)} \end{aligned} \tag{4}$$

式中，$\nabla_T \Phi(T^{(n)})$ 表示获取的梯度，ε 为较小的步长。DFool 首次证明了目标检测器上也存在对抗样本。同时，将对抗样本从数字维度转移到物理世界并成功欺骗了目标检测器，而且生成的对抗样本在一定程度上可以转移到 YOLO 9000 检测器。但是这个方法对扰动的大小、背景与前景的对比度有一定的要求。

2) DAG 方法。Xie 等人(2017) 将对抗样本从图像分类扩展到更加困难的语义分割和目标检测，针对两阶段检测器的分类损失函数提出 DAG 白盒攻击。考虑到两阶段检测器是通过 RPN 筛选含有物体的建议框，提出在一组像素点集或目标候选框集上优化目标任务损失函数进行攻击的方法。将攻击的目标放在感兴趣区域(region of interest，RoI)，其攻击表达式为

$$\begin{aligned} r_m &= \sum_{t_n \in T_m} [\nabla_{\boldsymbol{X}_m} f_{l'_n}(\boldsymbol{X}_m, t_n) - \nabla_{\boldsymbol{X}_m} f_{l_n}(\boldsymbol{X}_m, t_n)] \\ \boldsymbol{X}_{m+1} &= \boldsymbol{X}_m + \frac{\alpha_{\mathrm{DAG}}}{\|r_m\|_\infty} r_m \end{aligned} \tag{5}$$

式中，$\boldsymbol{X}_m$ 代表第 m 次迭代得到的图像，初始化为输入的图像，$f(\cdot)$ 代表目标检测器生成结果的函数，t_n 为输入图像的 T_m 个目标中的一个，l_n 代表该物体的正确类别，l'_n 代表指定的错误标签，r_m 是求得的对抗梯度，α_{DAG} 为迭代学习率。

DAG 的整个攻击思路为：首先对输入网络中的图像 $\boldsymbol{X}$，为其中的每个目标 t_n 随机指定一个标签作为要攻击的目标，该标签不同于目标的真实类别。接着进行迭代攻击，并在每次迭代中找出网络中仍然预测正确的 RoI 区域继续迭代。通过反向传播提高错误类别的得分，经过计算目标函数关于输入图像的梯度，通过 L_∞ 将梯度进行归一化后将梯度进行累计，直到达到最高迭代次数或者所有正样本均预测错误，停止迭代。

DAG 方法属于分类损失的单损失攻击，通过为目标设置一个非正确的标签，然后迭代朝着类别置信度低的方向进行移动，最终使检测器对输入图像的所有 RoI 都分类错误。从攻击的原理来看属于分类损失攻击，从攻击造成的结果来看属于目标分类错误攻击。

3) ShapeShifter 方法。这是 Chen 等人(2019) 提出的针对 Faster R-CNN 的第 1 个有目标的攻击方法。受图像分类对抗攻击方法 C&W(Carlini 和 Wagner，2017) 和期望转换(expectation over transformation，EOT)(Brown，2017) 的启发，对 Faster R-CNN 进行攻击。在图像分类中，C&W 通过优化的方式对图像进行有目标攻击，ShapeShifter 结合 C&W 的 L_2 攻击方式，具体攻击为

$$\arg\min L_{\mathrm{F}}(\tanh(\boldsymbol{x}'), \boldsymbol{y}') + c\|\tanh(\boldsymbol{x}') - \boldsymbol{x}_{\mathrm{o}}\|_2^2 \tag{6}$$

式中，L_{F} 为模型输出和目标标签 $\boldsymbol{y}'$ 的损失函数，通过 tanh 将像素约束在[- 1，1]之间，便于优化，用 c 控制修改后的图像 $\boldsymbol{x}'$ 和原图 $\boldsymbol{x}_{\mathrm{o}}$ 之间的差距。

在 C&W 基础上加上 EOT，期望变换就是在每次迭代过程中添加随机扰动，使加入的扰动更具有鲁棒性，操作方式为

$$M_t(\boldsymbol{x}_{\mathrm{b}}, \boldsymbol{x}_{\mathrm{o}}) = t(\boldsymbol{x}_{\mathrm{b}}) + \boldsymbol{x}_{\mathrm{o}} \tag{7}$$

式中，t 表示平移、旋转或者缩放操作，$\boldsymbol{x}_{\mathrm{o}}$ 为目标图像，$\boldsymbol{x}_{\mathrm{b}}$ 为背景图像。对目标图像 $\boldsymbol{x}_{\mathrm{o}}$ 进行 t 操作后加入到背景图像 $\boldsymbol{x}_{\mathrm{b}}$。

Chen(2019) 对 Faster R-CNN 第 1 阶段得到的多个区域建议提取其覆盖的子图像，然后对子图像进行分类，得到所有区域建议的分类损失，并且在迭代过程中利用 EOT 增强扰动的鲁棒性。

$$\begin{aligned} \arg\min E_{x\sim X, t\sim T}\Big[\frac{1}{m}\sum_{r_i \in rpn(M_t(\boldsymbol{x}'))} L_{F_i}(M_t(\boldsymbol{x}'), \boldsymbol{y}')\Big] + \\ c\|\tanh(\boldsymbol{x}') - \boldsymbol{x}_{\mathrm{o}}\|_2^2 \end{aligned} \tag{8}$$

式中，$\boldsymbol{x}'$ 为修改后的对抗样本，y' 为攻击指定的目标类别，$\boldsymbol{x}_{\mathrm{o}}$ 为干净图像。通过优化式(8)，同时攻击每个建议区域的所有分类。优化过程往往先通过 RPN 提取区域建议，并对区域建议进行修剪，在目标检测器第 2 阶段分类过程中进行式(8) 的优化。

4) RAP 方法。Li 等人(2018a) 针对两阶段目标检测模型提出的一种更为鲁棒的对抗性扰动生成算法，核心是通过破坏两阶段模型中特有的 RPN 网络对检测器进行攻击，设计了一种将分类损失与位置

损失结合在一起的损失函数,使用基于梯度的迭代算法对图像进行优化,具体为

$$\hat{\boldsymbol{p}}_t = \nabla_{I_t}\sum_{j=1}^{m} z_j(\log(s_j) + l_{\mathrm{SE}}(\boldsymbol{c}_j,\tau))$$
$$\boldsymbol{I}_{t+1} = Clip\left(\boldsymbol{I}_t - \frac{\lambda_{\mathrm{RAP}}}{\|\hat{\boldsymbol{p}}_t\|_2}\hat{\boldsymbol{p}}_t\right) \quad \text{s.t. } f_{\mathrm{PSNR}}(\boldsymbol{I}_t) \geqslant \varepsilon \tag{9}$$

式中,$\hat{\boldsymbol{p}}_t$ 表示计算的对抗梯度,$\boldsymbol{I}_t$ 表示第 t 次迭代的图像,Z_j 代表第 j 个建议的预测,$Z_j = 1$ 代表第 j 个建议包含物体,$Z_j = 0$ 代表第 j 个建议不包含物体,对这部分区域不予考虑。s_j 表示第 j 个建议区域预测的置信度得分,该置信度得分通过 RPN 网络的 sigmoid 函数计算得出。l_{SE} 为平方误差,$\boldsymbol{c}_j$ 为第 j 个建议区域的预测的具体信息,$\boldsymbol{c}_j = \{x_j, y_j, w_j, h_j\}$ 为边框的中心坐标、边框的宽和高;τ 为人为指定的具有较大偏移的新检测框,$Clip$ 为裁剪函数。PSNR(peak signal to noise ratio)为峰值信噪比,扰动越小,PSNR 越高。

计算网络预测的实际偏移量与人为指定的较大的偏移量之间的差值,将其作为损失函数,进行反向传播修改输入的图像,使最后网络预测的偏移量与真实偏移量之间产生很大的差距。

RAP 方法设计了一个同时包含位置损失函数和分类损失函数的组合损失函数,通过降低 RPN 网络得到的建议框置信度,使目标检测器将图像中的物体分类为背景,达到无法识别目标的目的。针对修改后仍然能识别出来的图像,又通过干扰其位置参数,使网络错误地定位该物体的正确位置,以此进行攻击。

5)CAP 方法。这是 Zhang 等人(2020)针对两阶段检测器提出的非目标攻击的方法。CAP 将分类损失和位置损失作为联合损失。在此基础上,又加入上下文损失。Zhang 等人(2020)考虑到图像中相邻像素的强相关性,发现候选区域的周边区域对目标检测器的定位和分类具有指导作用,将这块区域称为上下文区域。与单一候选区域相比,Zhang 等人(2020)提出的上下文区域具有更高的特征性,训练时能够极大地捕捉图像中物体的强特征,并对这些特征添加噪声,使攻击效果更好。具体为

$$L_c = \frac{1}{M}\sum_{j=1}^{M} z_j e_j^2 + \frac{1}{M}\sum_{j=1}^{M} - z_j\tilde{e}_j^2 \tag{10}$$

式中,上下文损失分为上下文区域的分类损失和背景损失,将两阶段检测器第 1 个阶段得到 RoI 中分类得分大于一定阈值记作正样本 RoI,数量为 M,e_j 和 $\tilde{e}_j$ 分别表示其正样本 RoI 上下文区域的最高类别得分和背景得分,通过对上下文损失函数的优化可以降低正样本 RoI 及候选区域的正确分类得分,同时提高背景得分,使攻击的成功率更高。

3.1.2 针对单阶段网络的攻击

针对单阶段网络的攻击方法主要有 PhyAttack、CA(category-wise attack)和 Daedalus 方法。

1)PhyAttack 方法。这是 Song 等人(2018b)提出的针对 YOLOv2 检测器的物理攻击方法。受图像分类领域的物理攻击 RP_2(robust physical perturbations)(Evtimov 等,2017)的启发,在 RP_2 基础上加入额外的对抗性损失函数,通过概率最小化式(11)降低图像中标志的得分,使检测器无法检测到停车标志。

$$J_d(\boldsymbol{x}, y) = \max_{s\in S^2,\ b\in B} P(s, b, y, f_\theta(\boldsymbol{x})) \tag{11}$$

式中,$f_\theta(\boldsymbol{x})$ 是检测器对图像 $\boldsymbol{x}$ 的输出,s 是 YOLO 网络单元格,b 是物体的检测框,y 是该物体对应的标签,$P(\cdot)$是用来从张量中提取对象类的概率。

同时,Song(2018b)设计出 creation attack,使目标检测器检测出不存在的物体。通过设计一个指定的位置,对该位置进行迭代优化,提高边框的置信度得分,使分类器将该位置分为前景区域,随后对位置之后的分类步骤提高类别概率,具体为

$$object = P_{\mathrm{box}}(s, b, f_\theta(\boldsymbol{x})) > \tau$$
$$J_c(\boldsymbol{x}, y) = object + (1 - object)\cdot P(s, b, y, f_\theta(\boldsymbol{x})) \tag{12}$$

式中,$object$ 为筛选出来的建议框位置,τ 为边框置信度阈值,当超过该阈值时停止对位置进行优化,$P(s, b, y, f_\theta(\boldsymbol{x}))$ 代表网格单元 s 中含有的边框 b 属于类别 y 的概率,$P_{\mathrm{box}}(s, b, f_\theta(\boldsymbol{x}))$ 为网格单元 s 含有的边框 b 属于前景区域的置信度值。

2)CA 方法。这是 Liao 等人(2020)首次针对 anchor-free(Zhou,2019a)目标检测模型提出的非目标攻击的方法。anchor-free 模型将目标检测经典模型特有的 anchor 删去,从而减少网络参数,加快训练速度,并且能极大地减少背景误检率。CA 通过寻找重要的像素区域,利用其高级语义信息对检测器进行类别攻击。CA 有 L_0 和 L_∞ 两种优化方式,具体为

$$\min\|r\|_p \quad \text{s.t.} \quad \forall k, p \in \boldsymbol{P}_k, \boldsymbol{P}_k = \{p \mid p > t_{\mathrm{attack}}\}$$

$$\begin{aligned}&\operatorname{argmax}_n\{f_n(\boldsymbol{x}+r,p)\}\neq C_k\\&t_{\min}\leqslant \boldsymbol{x}+r\leqslant t_{\max}\end{aligned}\tag{13}$$

式中，r 是对抗扰动，k 代表检测到的物体类别，$\boldsymbol{P}_k$ 为 CenterNet 得到的包含类别 C_k 的热力图，p 为热力图中得分大于阈值 t_{attack} 的像素点，$\operatorname{argmax}_n\{f_n(\boldsymbol{x}+r,p)\}$ 表示第 n 次迭代对抗样本中像素 p 预测的类别。

对于 L_0 范数的攻击，Liao(2020)受图像分类的对抗攻击方法 DeepFool(Moosavi-Dezfooli 等，2016)和 SparseFool(Modas 等，2019)的启发，提出 Approx-Boundary 方法来近似目标检测器的决策边界，使原始图像朝着垂直于决策边界的方向一步步移动，生成较为稀疏的扰动来欺骗目标检测器，这种 L_0 范数攻击称为稀疏类别攻击(sparse category-wise attack，SCA)。考虑到目标检测与图像分类的区别，SCA 选取总概率最高的 $\boldsymbol{P}_k$ 来生成局部目标像素集，然后用改进后的 DeepFool 算法进行攻击，直到对原始图像所有对象攻击成功为止。

对于 L_∞ 范数的攻击，Liao(2020)学习图像分类的对抗攻击 PGD(project gradient descent)(Madry 等，2018)思想，沿着图像梯度的方向分步迭代生成攻击样本。这种 L_∞ 范数攻击称为密集类别攻击(dense category-wise attack，DCA)，首先计算图像中每个类别的像素的损失，并将损失求和。具体为

$$loss_{\text{sum}}=\sum_{p_n\in \boldsymbol{P}_j} CE(f(x_i,p_n),C_j)\tag{14}$$

式中，CE 为交叉熵损失，$\boldsymbol{P}_j$ 为图像中 j 个类别对应的像素集，C_j 为图像预测的第 j 个类别。对获得的损失求取梯度，然后用 L_∞ 范数对求得的梯度进行归一化处理后得到扰动。具体为

$$\begin{aligned}r_j&=\nabla_{\boldsymbol{x}_i} loss_{\text{sum}}\\ r'_j&=\frac{r_j}{\|r_j\|_\infty}\end{aligned}\tag{15}$$

式中，r_j 表示图像 $\boldsymbol{x}_i$ 上对所有物体像素计算的损失之和，r'_j 为用 L_∞ 范数对 r_j 进行标准化后的对抗梯度。

3)Daedalus 方法。这是 Wang 等人(2021)提出的一种破坏 YOLO 组件的攻击方法，通过破坏 YOLO 的非极大值抑制(non maximum suppression，NMS)机制，使检测器产生误报等错误的结果。NMS 是目标检测至关重要的组成部分，主要目的是消除冗余建议框，并确定物体的最佳位置。Daedalus 攻击的优化计算为

$$\begin{aligned}&\arg\min|\delta|_p+c\cdot f(\boldsymbol{x}+\delta)\\&\text{s.t.}\quad \boldsymbol{x}+\delta\in[0,1]^n\end{aligned}\tag{16}$$

式中，δ 是添加的扰动，要求扰动的 L_p 范数最小，以保证其肉眼的不可见性，f 是定义的对抗性损失函数，c 为平衡对抗损失和失真的超参数。

为了解决使生成的对抗样本像素值限制在[0，1]这样一个盒约束问题，采用 C&W 方法，将变量 δ 替换成 ω，具体为

$$\delta_i=\frac{1}{2}(\tanh(\omega_i)+1)-x_i\tag{17}$$

式中，$\boldsymbol{x}$ 为输入的原图，δ 为计算的加入扰动，对于整个公式，需要得到的就是导致扰动的变量 ω，通过加入 tanh 使值保持在[-1，1]，便于迭代优化。

对于对抗性损失函数 f，设计了 3 种不同的损失函数：1)最小化各建议框的 IoU(intersection over union)值；2)使所有建议框尺寸缩小，且最大化各建议框中心之间的欧几里得距离；3)为了节省成本，只缩小各建议框的尺寸。3 种损失函数的计算分别为

$$f_1=\frac{1}{\|\boldsymbol{\Lambda}\|}\sum_{\lambda\in\boldsymbol{\Lambda}}\mathop{E}_{i;\operatorname{argmax}(p_i)=\lambda}\{[b_i^0\times\max(p_i)-1]^2+\mathop{E}_{j;\operatorname{argmax}(p_j)=\lambda}IoU_{ij}\}\tag{18}$$

$$f_2=\frac{1}{\|\boldsymbol{\Lambda}\|}\sum_{\lambda\in\boldsymbol{\Lambda}}\mathop{E}_{i;\operatorname{argmax}(p_i)=\lambda}\{[b_i^0\times\max(p_i)-1]^2+\left(\frac{b_i^W\times b_i^H}{W\times H}\right)^2+\mathop{E}_{j;\operatorname{argmax}(p_j)=\lambda}\frac{1}{(b_i^x-b_j^x)^2+(b_i^y-b_j^y)^2}\}\tag{19}$$

$$f_3=\frac{1}{\|\boldsymbol{\Lambda}\|}\sum_{\lambda\in\boldsymbol{\Lambda}}\mathop{E}_{i;\operatorname{argmax}(p_i)=\lambda}\left\{[b_i^0\times\max(p_i)-1]^2+\left(\frac{b_i^W\times b_i^H}{W\times H}\right)^2\right\}\tag{20}$$

式中，$\boldsymbol{\Lambda}$ 表示要攻击的类集合，p_i 表示第 i 个建议框的类别概率，IoU_{ij} 表示第 i 个和第 j 个建议框的 IoU 值，b_i^0 表示第 i 个建议框的置信度得分，b_i^W 和 b_i^H 表示第 i 个建议框的宽和高，W 和 H 为原图的宽和高，b_i^x，b_i^y 表示第 i 个建议框的中心点坐标。

Daedalus 攻击的核心是使 NMS 失效，通过破坏 NMS 的筛选机制达到破坏的目的。设计的损失函数结合 C&W 算法使生成的对抗样本在攻击性和迁移性方面都有不错表现，但是也结合了 C&W 方法

的缺点,即生成对抗样本的时间成本很高,生成一个有效的对抗样本需要进行上千次迭代,这也是未来需要改进的地方。

3.1.3 针对两种检测器的攻击

针对两种检测器的攻击方法主要有 UEA(unified and efficient adversary)、G-UAP(generic universal adversarial perturbation)和 TOG(targeted adversarial objectness gradient attacks)方法。

1)UEA 方法。目标检测领域内对抗样本的生成,都需要将图像输入到网络,通过神经网络的前向传播获得需要的数据,然后根据设计的损失函数进行反向传播调整网络输入,这需要耗费一些时间来生成对抗样本。而且生成的对抗样本在 Faster R-CNN 表现很好,但是在 YOLO 网络上效果却很差,其攻击的迁移性较差。针对这两个问题,Wei 等人(2019)提出了 UEA 方法。通过引入对抗生成网络(generative adversarial networks, GAN)(Isola 等,2017),将 GAN 网络结合高级分类损失和底层特征损失来进行训练生成对抗样本。通过引入 GAN 网络的生成器和判别器,由生成器生成进入目标检测的对抗样本,然后由判别器区分输入的图像是对抗样本还是干净图像。GAN 的损失函数为

$$L_{cGAN}(G,D) = E_I[\log D(\boldsymbol{I})] + E_I[\log(1 - D(G(\boldsymbol{I})))] \tag{21}$$

式中,G 是生成器,D 是判别器,$\boldsymbol{I}$ 是输入的图像。为了衡量生成的对抗样本和原始样本的差别,$G(\boldsymbol{I})$ 表示对图像 $\boldsymbol{I}$ 生成的噪声,引入 L_2 loss 衡量它们的相似性,具体为

$$L_{L_2}(G) = E_I[\|\boldsymbol{I} - G(\boldsymbol{I})\|_2] \tag{22}$$

为了同时攻击两种类别的检测器,Wei(2019)提出在 GAN 网络中加入 DAG 方法提出的损失函数,用来攻击以 Faster R-CNN 为代表的基于建议的目标检测器,具体为

$$L_{DAG}(G) = E_I\left(\sum_{n=1}^{N}[f_{l_n}(\boldsymbol{X}_m,t_n) - f_{l'_n}(\boldsymbol{X}_m,t_n)]\right) \tag{23}$$

为了增加对抗样本的可迁移性,攻击基于回归的目标检测器,提出多尺度注意力特征损失,具体为

$$L_{Fea}(G) = E_I\left[\sum_{m=1}^{M}\|\boldsymbol{A}_m \circ (\boldsymbol{X}_m - \boldsymbol{R}_m)\|_2\right] \tag{24}$$

式中,“∘”表示两个矩阵之间的哈达玛(Hadmard)积。$\boldsymbol{X}_m$ 表示目标检测器的骨干网络第 m 层提取的特征子图,$\boldsymbol{R}_m$ 是一个随机预定义的特征图,在训练过程中固定。$\boldsymbol{A}_m$ 是根据 RPN 的建议区域计算出的注意力权重,是两个矩阵的 Hadmard 乘积。特征图的损失函数将注意力特征图强制为随机排列,从而更好地操纵前景区域的特征图。

最后的损失函数为以上 4 种损失函数的组合,即

$$L = L_{cGAN} + \alpha L_{L2} + \beta L_{DAG} + \varepsilon L_{Fea} \tag{25}$$

式中,α,β,ε 为每种损失函数所占的比重。

由于 UEA 是通过训练一个 GAN 来生成针对目标检测器的对抗样本,是通过生成机制代替传统的攻击算法的优化过程,是一个前向机制,省略了反向传播过程,所以生成样本的时间更少,损失函数中添加了多尺度的注意力特征损失,对骨干网络特征图进行多层提取,并对前景区域部分的特征区域进行重点关注,在训练过程中进行更好的学习,使最后的对抗样本具有更佳的迁移性。

2)G-UAP 方法。G-UAP(Wu 等,2019)是一种在 UAP(Moosavi-Dezfooli 等,2017)基础上改进、扩展到目标检测领域的黑盒攻击方法。G-UAP 方法通过攻击 RPN 网络,将攻击思路简化为一个二分类问题,即诱导 RPN 网络将前景物体误认为背景,通过优化式(26)寻找到通用扰动。

$$L(\delta) = -[l\log(\hat{l}(x_i + \delta)) + (1 - l)\log(1 - \hat{l}(x_i + \delta))] \tag{26}$$

式中,δ 为生成的扰动,l 代表前景的标签。式(26)等号右侧前半部分代表图像中前景的得分概率,后半部分为背景的得分概率。为了误导前景变为背景,使 l 为 0,式(26)变为

$$L(\delta) = -\log(1 - \hat{l}(x_i + \delta)) \tag{27}$$

通过最小化式(27),降低图像中所有物体的前景置信度得分,增加背景置信度得分。

G-UAP 选择一批图像,然后累计每个图像得到的扰动,将扰动作为网络的特征映射,用雅克比矩阵表示,这样可以从一批图像中学到通用的扰动以欺骗更多的样本。

3)TOG 方法。对两阶段检测器 Faster R-CNN 设计的对抗攻击方法往往通过攻击两阶段网络特有的组件 RPN 网络来欺骗检测器。Chow 等人(2020a)提出的基于迭代的 TOG 方法可以同时攻击两阶段和单阶段两种目标检测器,TOG 方法根据最

后的攻击效果分为目标消失攻击、伪造标签攻击和分类错误攻击 3 类。TOG 方法通过逆转训练过程，固定网络参数，每次反向传播时修改输入的图像，通过迭代生成对抗样本，迭代直到攻击成功或达到阈值停止，整体为

$$\boldsymbol{x}'_{t+1} = \prod_{\boldsymbol{x},\varepsilon}\left[\boldsymbol{x}'_t - \alpha_{\mathrm{TOG}} \Gamma\left(\frac{\partial L^*(x'_t;O^*;\omega)}{\partial x'_t}\right)\right] \tag{28}$$

式中，$\boldsymbol{x}'_t$ 代表第 t 次迭代的对抗图像，$\prod_{\boldsymbol{x},\varepsilon}[\cdot]$ 表示半径为 ε 的超球面在以 $\boldsymbol{x}$ 为中心的 L_∞ 上投影，每次图像更新后，将其限制在一定范围内，以便于修改。α_{TOG} 为攻击学习率，Γ 为符号函数，L^* 表示提出的攻击损失函数，O^* 表示认为攻击者设定的虚假标签，ω 为目标检测器的模型参数。

目标消失(vanish)攻击使目标检测器无法定位和识别任何物体。通过攻击目标检测器的 L_{obj} 损失函数，该函数是检测图像中是否存在物体的损失函数，在 Faster R-CNN 是 RPN 的置信度得分，在 YOLO 网络代表网格的物体置信度得分。设置 $O(\boldsymbol{x}) = \varnothing$，使目标检测器将目标划为背景区域，从而使检测器检测不到任何物体，具体为

$$\boldsymbol{x}'_{t+1} = \prod_{\boldsymbol{x},\varepsilon}[\boldsymbol{x}'_t - \alpha_{\mathrm{TOG}} \Gamma(\nabla_{\boldsymbol{x}'_t} L_{\mathrm{obj}}(\boldsymbol{x}'_t,\varnothing;\omega))] \tag{29}$$

伪造标签(fabrication)攻击通过引入大量的伪造对象来增加目标检测的对象数量，达到攻击目标检测器的目的，具体为

$$\boldsymbol{x}'_{t+1} = \prod_{\boldsymbol{x},\varepsilon}[\boldsymbol{x}'_t + \alpha_{\mathrm{TOG}} \Gamma(\nabla_{\boldsymbol{x}'_t} L_{\mathrm{obj}}(\boldsymbol{x}'_t,\varnothing;\omega))] \tag{30}$$

式(30)与式(29)的区别在于式(29)是将前景换为背景区域，而目标伪造是使更多的背景区域变为前景，使所有的建议框或网格远离空标签。

目标分类错误(object-mislabeling)攻击使目标检测器对在输入图像上检测到的对象进行错误的分类。具体做法是对检测出来的对象进行分类时，用选定的目标类别代替原来的标签进行反向传播，得到对抗梯度，修改图像，使检测器虽然检测出来物体，但是分类错误，具体为

$$\boldsymbol{x}'_{t+1} = \prod_{x,\varepsilon}[\boldsymbol{x}'_t - \alpha_{\mathrm{TOG}} \Gamma(\nabla_{\boldsymbol{x}'_t} L(\boldsymbol{x}'_t,O^*(x);\omega))] \tag{31}$$

式中，O^* 表示错误的类别标签。

因为 TOG 不是针对目标检测器的特有结构(例如 RPN)设计的，而是从目标检测多任务角度进行攻击，所以可以为不同种类的目标检测器生成对抗样本。

3.2　局部扰动攻击

与全局扰动攻击需要针对全像素进行攻击的方法不同，局部扰动攻击只在原始图像的一个区域内添加扰动，使该区域的扰动能影响全图，达到欺骗目标检测器的目的。主要方法有 Bpatch、Dpatch、EA(evaporate attack)和 Adversarial-YOLO 方法。

1) Bpatch 方法。这是 Li 等人(2018a)率先提出的针对两阶段检测器的局部进行扰动攻击的方法，通过在图像目标之外的背景上添加扰动块来攻击目标检测器。BPatch 也是针对两阶段检测器中特有的部件 RPN(区域提议网络)进行攻击。由于 RPN 网络会生成大量包含候选框的候选区域，下一阶段的网络会针对 RPN 网络生成的候选框按照置信度进行排列，将高于置信度阈值的候选框挑选出来进行下一阶段的分类和位置回归。BPatch 针对 RPN 网络的筛选机制提出了攻击思路，通过降低 RPN 层得到的高置信度候选区的置信度，使得最后送入下一层网络的候选框少包含甚至不包含前景目标。

BPatch 补丁也是一种通过对损失函数优化来生成对抗扰动的方法，公式包含 3 种损失函数：1)真阳性置信度损失(true positive confidence loss，TPC)，该项损失的目的是降低包含图像目标区域候选框的置信度，从而无法提取正确的候选框进入下一层网络；2)真阳性形状损失(true positive shape loss，TPS)，该项损失是对目标的位置进行攻击，目的是使最后物体的位置定位是不精确甚至错误的；3)假阳性置信度损失(false positive confidence loss，FPC)，目的是提高背景区域的置信度，将背景补丁附近的区域选中送入 RPN 网络的下一层网络。通过这 3 个损失函数降低真实候选框的置信度，提高背景区域假候选区域的置信度，最后达到攻击目标检测的目的。

BPatch 损失函数具体计算为

$$\min_{I(Q)} L_{\mathrm{tpc}}(\boldsymbol{I}(Q);\boldsymbol{F}) + L_{\mathrm{shape}}(\boldsymbol{I}(Q);\boldsymbol{F}) + L_{\mathrm{fpc}}(\boldsymbol{I}(Q);\boldsymbol{F}) \quad \text{s.t. } f_{\mathrm{PSNR}}(\boldsymbol{I}(Q)) \geqslant \varepsilon \tag{32}$$

式中，L_{tpc}，L_{shape}，L_{fpc} 为上面提到的 3 个损失，$\boldsymbol{I}(Q)$

为加入补丁 Q 的图像 $\boldsymbol{I}$,$\boldsymbol{F}$ 为已经训练好的 RPN 网络,ε 为峰值信噪比的下限。

TPC 损失具体计算为

$$L_{\text{tpc}}(\boldsymbol{I}(Q);\boldsymbol{F}) = \sum_{j=1}^{m} z_j \log(s_j) \tag{33}$$

式中,S_j 表示第 j 个候选区域的置信度得分;Z_j 为权重,当第 j 个候选区域与任意标签比较,得到的 IoU 大于阈值(一般为 0.5),且该候选框的置信度大于阈值(一般为 0.1)时,令 $Z_j = 1$,否则 $Z_j = 0$。

TPS 损失具体计算为

$$L_{\text{shape}}(\boldsymbol{I}(Q);\boldsymbol{F}) = \exp\left(-\sum_{j=1}^{m} z_j((\Delta \boldsymbol{x}_j - \Delta\bar{\boldsymbol{x}})^2 + (\Delta y_j - \Delta\bar{y})^2 + (\Delta w_j - \Delta\bar{w})^2 + (\Delta h_j - \Delta\bar{h})^2)\right) \tag{34}$$

式中,Δx_j,Δy_j,Δw_j,Δh_j 表示检测器预测的中心坐标和矩形框宽高的偏移,$\Delta\bar{x}$,$\Delta\bar{y}$,$\Delta\bar{w}$,$\Delta\bar{h}$ 表示真实的偏移量,BPatch 的位置损失和 RAP 的位置损失是有区别的,它不像 RAP 通过指定特定的位置标签来执行目标攻击,而是对位置执行非目标攻击。通过优化式(34),使预测的偏移量逐渐远离真实偏移量。

FPC 损失函数定义为

$$L_{\text{fpc}}(\boldsymbol{I}(Q);\boldsymbol{F}) = \sum_{j=1}^{m} r_j \log(1 - s_j) \tag{35}$$

当第 j 个候选框与背景补丁 Q 的 IoU > 0.3 且与任意的真实矩形框的 IoU = 0 时,就选择该候选框进行优化,令 $r_j = 1$,否则,令 $r_j = 0$。通过优化式(35)提高背景补丁的置信度,使 RPN 网络给下一层网络输出更多的包含背景区域的候选框,导致检测失败。

2)Dpatch 方法。这是 Liu 等人(2019b)提出的针对目标检测器的目标攻击方法,核心是生成一个 patch,然后将该 patch 当做一个 GT(ground truth)检测框,通过反向传播使网络直接优化该 patch。因此,当分类损失和回归损失都收敛的情况下,只会产生一个检测框,即 patch 的坐标和类别。DPatch 可以针对 Faster R-CNN 和 YOLO 系列网络的特点同时进行攻击。

针对 Faster R-CNN 两阶段检测网络,攻击思路是使其 RPN 网络无法生成正确的候选区域,使 DPatch 所在的区域成为唯一有效的 RoI,而忽略其他可能的候选区域。

针对 YOLO 单阶段网络,核心要素是边界框预测和置信度分数。图像中的每个网络都可以预测边界框和这些边界的置信度分数。这些置信度得分反映了该边界框包含一个对象的概率以及该边界框的准确性。如果置信度得分相对较低,则由网格预测的边界框视为不包含真实对象。同样,攻击 YOLO 时,应将 DPatch 所在的网格视为对象,而其他网格则应忽略,即包含 DPatch 的网格比其他具有普通对象的网格具有更高的置信度得分。

DPatch 方法受谷歌对抗补丁(Brown 等,2017)的启发,通过类比图像分类的 patch,得

$$\hat{P} = \arg\max_{P} E_{\boldsymbol{x},t,l}[\log Pr(\hat{y} \mid \boldsymbol{A}(P,\boldsymbol{x},l,t))] \tag{36}$$

式中,$\boldsymbol{A}(P,\boldsymbol{x},l,t)$ 表示输入图像 $\boldsymbol{x}$,该图像通过变换 t,将补丁 P 加入到位置 l 处,这些变换包括缩放和旋转等操作,$Pr(\hat{y} \mid \boldsymbol{A})$ 表示输入 $\boldsymbol{A}$ 属于真实标签 $\hat{y}$ 的概率。DPatch 的原理是当输入一幅图像进入目标检测器时,最大化目标检测器对真实标签 $\hat{y}$ 和边界框标签 B 的损失,具体为

$$\hat{P}_u = \arg\max_{P} E_{\boldsymbol{x},s}[L(\boldsymbol{A}(P,\boldsymbol{x},l,t));\hat{y},\hat{B}] \tag{37}$$

在目标攻击中,DPatch 还可以提前指定想要攻击的目标类标签 y_t 和边界框标签 B_t,通过反向传播最小化损失函数。具体为

$$\hat{P}_t = \arg\max_{P} E_{\boldsymbol{x},s}[L(\boldsymbol{A}(P,\boldsymbol{x},l,t));y_t,B_t] \tag{38}$$

对于非目标攻击,DPatch 将图中的目标标签设置为 0,即将其训练为背景。DPatch 方法易受 patch 的大小影响,一般 patch 选的越大,攻击的成功率也就越大,相应的扰动的像素也就越多。

3)EA 方法。目标检测攻击的绝大部分算法,诸如 DAG、RAP,它们生成对抗样本的本质是通过损失函数进行优化得到,导致它们只能攻击白盒目标检测模型。对于未知的黑盒模型,其攻击效果不尽人意。针对这种情况,Wang 等人(2020)提出一种基于粒子群优化的黑盒攻击方法——EA 方法。这种方法仅利用模型预测的位置和标签信息来生成对抗样本。该算法将对抗样本的生成看做式(39)的优化。

$$\min L(\boldsymbol{x}') = d(\boldsymbol{x}',\boldsymbol{x}) - \delta(D(\boldsymbol{x}')) \tag{39}$$

式中,$d(\cdot,\cdot)$ 是距离度量,在此将距离度量通过 L_2 范数进行实例化;$\delta(\cdot)$ 是对抗标准,如果满足攻击

标准则取 0，否则取负无穷；D 代表目标检测模型。该方法首先向图像中添加随机噪声来生成初始图像粒子群，图像粒子的初始化计算为

$$\boldsymbol{x}'_i = \boldsymbol{x} + \varepsilon \times z$$
$$\text{s.t.} \quad z \,\%\, N(0,\delta^2 C), \delta(\boldsymbol{x}'_i) = 0 \tag{40}$$

式中，$\boldsymbol{x}$ 为原始图像，z 是随机生成的高斯噪声，ε 为限制噪声的超参数，C 表示正态分布中的样本种类，$\delta(\cdot)$ 表示对抗标准，如果扰动图像使检测器错误，则取 0，否则取 1。将添加了随机扰动的图像作为初始粒子 $\boldsymbol{x}'_i$，计算粒子群的适应度值。在满足对抗要求的情况下，图像粒子与原始图像的距离越小，适应度值就越大。

Wang(2020)修改了传统的 PSO(particle swarm optimization)算法(Kennedy 和 Eberhart，1995)的速度迭代公式，为了使生成的图像和原始图像尽可能相似，加入了最佳像素位置(U_{best})来引导粒子接近原始图像。同时，为了避免传统 PSO 易陷入局部最优问题，在速度迭代公式中添加高斯噪声。攻击分为两个阶段。第 1 阶段的粒子群移动方式为

$$pv = \mu_1 \times E + \mu_2 \times z \times (P_{\text{best}} - x) + \mu_3 \times z \times (G_{\text{best}} - x) + \mu_4 \times (U_{\text{best}} - x) + \mu_5 \times z \tag{41}$$

式中，μ_1 为初始权重因子，μ_2 和 μ_3 为初始化学习因子，μ_4 为原始图像投影的权重，μ_5 为高斯噪声的权重，P_{best} 为粒子个体的历史最优值，G_{best} 为粒子群全局最优值。第 2 阶段当粒子已经很接近目标则开始变慢速度，去除式(41)第 2 项，使粒子稳定地向前移动，直到达到迭代次数或者达到全局最优解，最终得到对抗样本。

4) Adversarial-YOLO 方法。这是 Thys 等人(2019)设计的一种基于 YOLO 网络的肉眼可见的对抗样本生成方法。该方法生成的对抗样本可以欺骗基于 YOLO 的行人检测，使其无法检测到人的存在。这种方法在数字世界和物理世界都有较强的攻击效果。在数字世界，为了使生成的样本具有攻击性，需要最小化检测器输出的对象的类损失 L_{obj}。为此，设计了 3 种 L_{obj}：1)将类标签为人的网格误导成其他种类；2)最小化物体的置信度得分；3)前两者的结合。实验证明，设计的第 2 种损失的效果最好(Thys 等，2019)。

同时，为了使生成的对抗样本可以转移到物理世界，加入了打印损失 L_{nps}，具体为

$$L_{\text{nps}} = \sum_{p_{\text{patch}} \in p} \min_{C_{\text{print}} \in \boldsymbol{C}} |p_{\text{patch}} - C_{\text{print}}| \tag{42}$$

式中，p_{patch} 为 patch 的像素，C_{print} 是一组可打印出来的颜色 $\boldsymbol{C}$ 中的一种颜色。为了使优化过程中 patch 的色彩过渡更为平滑以及防止噪声图像，提出了第 3 种损失 L_{tv}，具体为

$$L_{\text{tv}} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} \tag{43}$$

式中，$p_{i,j}$ 代表像素点。如果图像中相邻像素相似，则分数较低，反之，则分数较高。将以上 3 部分损失合并，得到最终的总损失，通过优化总损失得到对抗样本。最终的总损失为

$$L = \alpha L_{\text{nps}} + \beta L_{\text{tv}} + L_{\text{obj}} \tag{44}$$

3.3　对抗攻击方法总结

为了便于了解各种目标检测对抗样本生成方法的特点，表 1 对每种方法进行了简要总结。同时根据是否知道模型内部参数、是否属于定向攻击、主要攻击的检测器类型、损失函数的设计等 6 个方面对上述提到的对抗攻击方法进行总结分析，如表 2 所示。从表 1 和表 2 可以看出，自 Lu 等人(2017)提出 DFool 攻击以来，面向目标检测的对抗样本生成技术的发展具有以下几个规律：

1)攻击形式多样化。主要体现在 3 个方面。(1)攻击效果多样化。最开始的攻击方法如 DAG，攻击效果是使目标检测器对检测到的物体进行错误的分类。随着越来越多攻击方法的提出，造成的攻击效果不仅包含分类错误，还有使图像中的物体无法检测到、使检测到的物体错误分类、使检测到的物体的检测框错误、使图中出现许多未知标签等。可以看出，现在提出方法的攻击效果越来越多样化。(2)攻击损失函数更加多样化。由单一的分类损失变为分类损失结合回归损失的联合损失函数，有些方法在联合损失函数基础上，还加入背景损失、上下文损失函数，使生成的对抗样本更具有攻击性。(3)目标检测的对抗攻击不仅包含全局扰动攻击，也包含以 patch 为主的局部扰动攻击。全局扰动攻击和局部扰动攻击各有优缺点，分别适应不同场景。全局扰动攻击的扰动攻击全局，扰动分散不易察觉，而以补丁为主的局部攻击，常常延伸到物理世界，将对抗补丁做成贴纸或图案，以此实现物理世界的端到端攻击。

2)攻击对象更丰富。面向目标检测的对抗攻击对象既包括以 Faster R-CNN 为代表的两阶段检测器,也包括以 YOLO 为代表的单阶段检测器,此外也开始出现针对无锚框(anchor-free)(Huang 等,2015)的新型检测器的研究,例如上文对 CenterNet 攻击的 CA 算法。值得注意的是,目前的方法大多

表 1　对抗攻击方法描述

Table 1　Description of adversarial attacks

方法	描述	特点
DFool	第 1 个在停止标签上进行攻击,欺骗 Faster R-CNN,并且推广到数字世界,主要通过降低停止标签的得分来攻击。	首次尝试对停车标志检测和人脸检测模型进行攻击,能够一定程度上迁移到 YOLO;但成功率易受扰动大小和环境因素影响。
DAG	人工选取 RPN 生成的大量未修建过的建议框,为每个候选框随机分配 1 个错误标签,开始迭代优化,直到该建议框的标签变为分配的标签或不是原标签。	基于优化的白盒攻击方法,对 Faster R-CNN 具有较强攻击效果;但迭代生成对抗样本的时间较长,且迁移性差。
ShapeShifter	采用 C&W 方法攻击 RPN 生成的建议框,使其误分类,同时用 EOT 思想实现物理场景的攻击。	首个较为系统全面地面向目标检测的物理攻击方法,其 EOT 方法的运用有助于限制扰动的形状,提高攻击成功率;但黑盒攻击较差,生成的扰动易察觉。
RAP	是一种基于梯度迭代的攻击,主要攻击 RPN 组件,将目标检测的类别损失和位置损失结合起来同时攻击。	攻击形式更加多样化,不仅攻击对象的分类,也攻击对象的位置;但攻击强度一般,针对 RPN 的攻击导致其迁移性很差。
CAP	损失函数中加入了上下文损失,以破坏物体的上下文信息,同时增加上下文背景损失,抑制了前景得分,达到更强的攻击效果。	利用了上下文信息增强攻击效果,增加背景损失,使攻击不依赖于标签攻击;但上下文区域的大小影响攻击成本,迁移性较差。
PhyAttack	将 RP_2 方法延伸到目标检测领域,主要攻击 YOLO 网络,降低停车标志的得分。	对目标检测的物理攻击,可有效攻击 YOLO,对 Faster R-CNN 也有一定攻击性。
CA	首次对 Anchor-free 检测器进行攻击,利用热力图来捕获图像的高级语义信息,对语义信息进行类别攻击。	将图像分类的 DeepFool 和 PGD 思想作为优化方法,生成的对抗样本具有强攻击性,充分利用语义信息,生成的对抗样本的迁移能力较强。
UEA	将 GAN 和对抗样本结合,损失函数由 GAN 损失、DAG 损失、特征图损失和 L_2 损失组成,最后训练出一个可以实时生成对抗样本的网络。	生成对抗样本的速度快,且迁移性好;但由于训练的是一种偏向通用攻击的网络,生成的对抗样本的攻击性一般,且扰动相比其他方法较大。
Daedalus	通过破坏 YOLO 网络的 NMS 组件,使得 NMS 无法筛选出正确的建议框。	破坏 NMS 组件造成的攻击性很强,且难以被防御;但为了破坏所有的建议框,消耗的时间很长,扰动较大。
G-UAP	将 UAP 方法与目标检测器结合,通过误导 RPN 网络将前景误认为背景,将每幅图像的扰动累加,得到的通用扰动可以攻击其他未训练的图像。	证明了目标检测通用扰动的存在,具有较高的可迁移性;但不能保证通用扰动对每幅图像都起作用,有时更新过的通用扰动对更新前的数据会失效。
TOG	是一种基于梯度迭代的攻击,通过对物体的存在损失、位置损失以及类别损失进行梯度攻击。	基于梯度的迭代攻击具有很强的白盒攻击能力;但求取的梯度太依赖于模型内部参数,迁移性很差。
BPatch	通过向图像中加入补丁的方式攻击 Faster R-CNN,损失函数由 3 部分构成,分别可以实现降低真阳性数量,使真阳性的位置不准确,增加假阳性的数量。	在背景上添加补丁块的方式攻击形式较为新颖,对 Faster R-CNN 的攻击效果较好;但攻击效果与补丁的扰动有关,迁移性差。
DPatch	通过分类损失和位置损失联合训练 patch,迭代修改 patch 的值。	可迁移性强,对不同网络的攻击能力都表现很好;但训练成本较高,时间长。
EA	是一种基于粒子群优化的黑盒攻击方法,利用位置和标签信息优化扰动来攻击目标检测器。	可同时攻击基于回归和基于建议的检测器;但需要不断询问模型得到攻击信息,生成时间较长,计算成本大。
Adversarial-YOLO	攻击 YOLO 网络中对人的检测,设计 3 种不同的类损失函数来隐藏图像中的人,结合打印损失和平滑损失实现物理层面的攻击。	相比于之前的停车标志,对于人的物理攻击更加新颖,实际意义也更加大。

表 2　对抗攻击方法总结

Table 2　Summary of adversarial attacks

方法	攻击类型	是否定向	扰动范数	攻击的检测器类型	损失函数设计	全局攻击/局部攻击
DFool	白盒攻击	非定向	L_2	两阶段	单损失(分类损失)	全局
DAG	白盒攻击	定向;非定向	L_∞	两阶段	单损失(分类损失)	全局
ShapeShifter	白盒攻击	定向;非定向	L_2	两阶段	单损失(分类损失)	全局
RAP	白盒攻击	定向;非定向	L_2	两阶段	组合损失	全局
CAP	白盒攻击	非定向	L_2	两阶段	组合损失	全局
PhyAttack	白盒攻击	非定向	–	单阶段	单损失(分类损失)	全局
CA	白盒攻击	非定向	L_0, L_∞	单阶段	组合损失	全局
Daedalus	白盒攻击	非定向	L_0, L_2	单阶段	–	全局
UEA	白盒攻击	定向;非定向	L_2	均可	单损失	全局
G-UAP	黑盒攻击	非定向	L_∞	均可	–	全局
TOG	白盒攻击	定向;非定向	L_∞	均可	组合损失	全局
BPatch	白盒攻击	非定向	L_2	两阶段	组合损失	局部
DPatch	黑盒攻击	定向	L_∞	均可	组合损失	局部
EA	黑盒攻击	非定向	L_2	均可	–	局部
Adversarial-YOLO	白盒攻击	定向;非定向	–	单阶段	单损失(分类损失)	局部

注:“–”表示无法归属于当前类别。

数是攻击两阶段检测器。因为相比于单阶段检测器,两阶段检测器检测精度更高,更难攻击,因此大多研究更关注于攻击 Faster R-CNN。

3)白盒攻击普遍,黑盒攻击鲜有。本文提及的方法中,黑盒攻击方法仅有 3 种。这是因为目标检测相比于图像分类,网络更深,提取特征的能力更强。不仅可以进行图像分类,还可以用上下文信息对图像分类的结果进行纠正。因此目标检测对对抗样本的“免疫力”大幅高于图像分类,如果不了解模型内部的参数,很难构造出具有较强攻击性的对抗样本。这也造成了目标检测领域的黑盒攻击方法鲜有的现象。因此,如何设计出有效的黑盒攻击方法在未来是一个值得关注的方向,值得进一步探讨。

4　攻击方法对比

选取全局扰动攻击中有代表性的 DAG、RAP、UEA 和 TOG 方法,其中 TOG 包括 3 个子策略,共 6 种攻击方式进行对比实验。选择的数据集为 PASCAL VOC,其中训练集为 VOC 2007 + VOC 2012 的全部训练图像,测试集选择 VOC 2007 的全部测试图像。选择 Faster R-CNN 模型作为攻击的目标模型。首先采用干净样本训练 Faster R-CNN,用训练好的模型在测试集上进行测试。以 mAP 为评价指标,mAP 越大,说明检测器对数据集的检测效果越好。表 3 为 Faster R-CNN 的正常训练结果,经过 13 次迭代后,mAP 达到 70.1%,超过了 Faster R-CNN 论文中的效果,用此检测器测试不同的攻击方法。将测试集通过 6 种攻击方法生成对抗样本,然后用训练好的 Faster R-CNN 进行检测,实验结果如表 4 所示。

表 3　Faster R-CNN 正常训练结果

Table 3　Normal training results of Faster R-CNN

epoch	mAP/%
1	65.7
3	66.2
5	66.9
7	67.0
9	67.2
11	69.9
13	70.1

表 4　不同方法在 PASCAL VOC 2007 测试集上攻击的 AP 结果

Table 4　The AP results of different methods after attack in the PASCAL VOC 2007 test sets

/%

样本	攻击前	TOG-消失	TOG-假冒	TOG-误分类	DAG	RAP	UEA
plane	74.99	0.00	1.55	4.69	6.40	10.83	3.56
bicycle	79.21	0.00	2.19	3.89	5.66	2.49	11.96
bird	68.16	0.00	0.24	2.19	0.05	6.15	2.76
boat	55.88	0.00	3.65	4.08	2.20	7.42	1.91
bottle	53.80	0.00	0.33	1.54	2.42	0.14	2.25
bus	76.38	0.00	4.59	1.41	5.63	4.40	**25.20**
car	79.68	0.14	5.14	4.84	9.34	6.48	16.80
cat	84.54	0.00	0.95	4.17	7.03	1.10	14.71
chair	49.42	0.00	0.37	2.26	1.50	0.10	1.10
cow	76.31	0.00	0.97	4.62	6.36	4.23	2.20
table	69.37	0.00	0.02	2.24	0.04	5.14	11.24
dog	75.47	0.00	1.11	3.98	4.32	1.80	4.44
horse	79.76	0.00	1.43	3.07	8.89	2.28	6.76
motor	76.30	0.00	2.88	3.57	6.74	2.97	15.87
person	77.52	**0.18**	0.89	**7.03**	8.62	2.60	14.37
plant	44.23	0.00	2.53	1.72	3.33	0.52	4.58
sheep	72.79	0.00	4.86	4.24	5.64	4.11	3.38
sofa	62.39	0.00	0.04	2.62	2.03	5.77	12.86
train	75.06	0.00	1.31	5.20	**12.42**	11.59	10.23
tv	71.83	0.00	**5.62**	3.24	3.32	**20.11**	6.46
mAP	70.15	0.02	2.44	3.53	5.57	5.01	8.63

注：加粗字体表示对抗样本生成方法的最优结果。

表 4 给出了 6 种对抗样本生成方法对 PASCAL VOC 测试集中 20 个小类的检测情况。第 2 列是攻击前的检测精度 AP，第 3—8 列是 6 种对抗样本生成方法攻击后的检测精度。最后一行是 20 个小类 AP 的平均值，即 mAP 值。从表中可以看出，20 个小类的样本受不同方法攻击前后 AP 值的变化。这些方法均能够有效攻击 Faster R-CNN 检测器，但不同攻击方法的攻击强度不同。TOG 作为最新的一种攻击方法，攻击效果最好，TOG 的 3 种子策略都优于其他 3 种算法，尤其是 TOG-消失，几乎使整个测试集的 AP 下降到 0。

4.1　攻击效果分析

实验选择 Faster R-CNN 为测试模型，分别从攻击强度、攻击所需要的时间以及生成的扰动大小 3 个角度对上述 6 种攻击方法进行分析。

攻击强度用攻击成功率（attack success rate，ASR）评估，表示攻击前后 mAP 的变化情况，定义为

$$ASR = \frac{mAP_{\text{clean}} - mAP_{\text{attack}}}{mAP_{\text{clean}}} \tag{45}$$

式中，mAP_{clean} 代表攻击前的检测器的 mAP，mAP_{attack} 为攻击后的检测器的 mAP，ASR 越高代表攻击越强。

攻击算法的时间成本为整个测试集所有对抗样本生成时间除以图像数量求得的平均值。

生成的扰动大小是指生成的对抗样本图像与原

始图像的差距。实验使用对抗样本中较常用的 L_2 范数衡量扰动的程度。

表 5 是 6 种对抗样本生成方法在攻击强度、攻击所需要的时间以及生成的扰动大小 3 个方面的表现。从攻击强度看,3 种 TOG 方法的攻击成功率比其他 3 种方法好,UEA 的攻击成功率最差。从对抗样本生成时间看,UEA 的生成时间是最快的,这是因为 UEA 算法已经训练好了一个 GAN 网络,生成对抗样本的过程只涉及前向传播,而其他方法均需通过反向传播来生成扰动。同时,因为 UEA 缺少了反向传播来针对不同图像生成扰动,无法捕捉不同图像的细微差异性,所以为了达到攻击效果,只能增大扰动范围,表 5 显示的 6 种方法中,UEA 产生的对抗样本 L_2 范数指标为 0.124,是其他方法的数倍。RAP 与 DAG 相比,生成的时间较少,因为生成对抗样本时限制了生成扰动的大小,要求图像的 PSNR 大于固定的阈值,所以最后的迭代次数一般是少于 DAG 的,故时间代价比 DAG 要小。整体而言,综合考虑 ASR、时间代价和扰动大小,TOG 在各指标上的表现更加均衡,相比其他方法有更好的攻击效果。

表 5　几种攻击方法在 Faster R-CNN 上效果对比

Table 5　Comparison of several attack methods on Faster R-CNN

方法	mAP/%		攻击成功率/%	时间/s		L_2 范数
	攻击前	攻击后		攻击前	攻击后	
TOG-消失	70.1	**0.02**	**99.97**	0.11	1.33	0.056
TOG-假冒	70.1	2.44	96.51	0.11	1.43	0.055
TOG-误分类	70.1	3.53	94.96	0.11	1.52	0.052
DAG	70.1	5.57	92.05	0.11	9.83	**0.013**
RAP	70.1	5.01	92.83	0.11	3.77	0.039
UEA	70.1	8.63	87.68	0.11	**0.06**	0.124

注:加粗字体表示各列最优结果。

4.2　迁移性分析

对抗样本的可迁移性是指生成的对抗样本对未知模型的攻击能力,是衡量对抗样本的重要指标。一种好的对抗样本生成方法,不仅具有较高的白盒攻击能力,同时也应具有较高的黑盒攻击能力。

实验选取上节提到的 6 种方法,以 Faster R-CNN 作为源模型生成对抗样本,被攻击的目标模型选择以 Darknet 为骨干网络的 YOLOv3(YOLOv3-D)和 SSD300 两种单阶段检测网络,实验结果如表 6 所示。从数据上看,6 种方法都使两种检测器的检测精度下降,但程度不同。除了 UEA,其他方法的攻击成功率都在 10% 之内。TOG-消失使 YOLOv3-D 从 81.6% 下降到 77.5%,攻击成功率为 5.0%,对 SSD300 的攻击成功率为 3.4%。RAP 对 YOLOv3-D 和 SSD300 的攻击成功率分别为 3.7% 和 2.8%,但对 Faster R-CNN 攻击成功率高达 92.83%(表 5)。结合表 5 和表 6 可以看出,TOG 系列、DAG 和 RAP 方法的可迁移性都较低。而 UEA 对 YOLOv3-D 有 43.2% 的攻击成功率,对 SSD300 的攻击成功率为 45.2%,相比其他方法,UEA 方法的可迁移性较好。从算法思想上看,DAG 方法生成对抗样本是根据 RPN 生成的建议框进行分类攻击,攻击成功率与 RPN 网络有关;RAP 算法的核心是破坏 RPN 网络;TOG 方法是通过 Faster R-CNN 的损失函数反向传播后求取的对抗梯度来生成对抗样本。这些方法都是基于 Faster R-CNN 的内部信息进行攻击,一旦检测器改变检测思路,例如 YOLO 直接得到建议框而不需要 RPN 网络,这些对抗样本就会失效,所以其可迁移性很低。从另一方面看,UEA 在训练过程中加入特征图损失这一方法能有效提高对抗样本的迁移能力。从以上数据也可以看出,目前面向目标检测的白盒攻击方法的可迁移性是有限的,无法对未知模型造成较大破坏。

表 6　迁移性比较结果

Table 6　The comparison results of transferability

/%

方法	YOLOv3-D			SSD300		
	攻击前 mAP	攻击后 mAP	攻击成功率	攻击前 mAP	攻击后 mAP	攻击成功率
TOG-消失	81.6	77.5	5.0	77.5	74.8	3.4
TOG-假冒	81.6	78.4	3.9	77.5	74.7	3.6
TOG-误分类	81.6	77.6	4.9	77.5	74.3	4.1
DAG	81.6	77.4	5.1	77.5	72.9	5.9
RAP	81.6	78.5	3.7	77.5	75.3	2.8
UEA	81.6	46.3	**43.2**	77.5	42.4	**45.2**

注：加粗字体表示各列最优结果。

5　防御策略

目标检测在许多计算机视觉任务中扮演着重要角色，所以如何有效防止目标检测器遭受对抗样本的侵害以提高模型鲁棒性显得越发重要。目前对于目标检测对抗样本的防御方法较少，本文按防御的时间段和作用，将防御策略分为预处理防御和提高模型鲁棒性防御两类。

5.1　预处理防御策略

预处理防御策略是指将图像在输入神经网络前先经过一系列操作，以减轻对抗样本的攻击。常见方法有去噪、滤波和图像压缩等。预处理防御策略在图像分类攻击领域是一项重要的防御措施，能够有效降低对抗样本的攻击性。一些学者尝试将图像分类的预处理操作运用到目标检测领域，发现也能起到一定的防御作用。

5.1.1　去噪、滤波和图像压缩

Saha 等人(2020)发现去噪(Akhtar 等，2018；Vincent 等，2008)和滤波器滤波图像(Wang，2016a，2016b)等经典防御方法在目标检测的对抗防御上能起到较好作用。Liao 等人(2020)将 DAG 产生的对抗样本经过 JPEG(joint photographic experts group)压缩(Dziugaite 等，2016)后，用 CenterNet 和 SSD 模型进行检测，发现这些对抗示例基本失去了攻击性。

5.1.2　随机中值平滑

Chiang 等人(2020)提出一种针对目标检测器的随机中值平滑方法来防御对抗样本。传统的高斯平滑操作由于计算的是平均值，易受基函数的影响而产生偏斜，而目标检测需要完成回归任务，这对于回归问题是重大缺陷，因此采用中值代替平均值的中值平滑方法。实验结果表明，DAG 攻击生成的对抗样本经过随机中值平滑操作，攻击成功率很低。

5.2　提高模型鲁棒性

5.2.1　对抗训练

对抗训练作为图像分类领域对抗攻击防御的常用方法能够有效提高模型鲁棒性。Goodfellow 等人(2015)提出通过对抗训练提高模型鲁棒性。Kurakin 等人(2017)提出在大型网络 Inception v3 模型和大型数据集 ImageNet 上用批量归一化的方法进行对抗训练。很多方法(Tramèr 等，2018；Li 等，2019；Song，2018a；Madry 等，2018)均将对抗训练用于对抗样本防御。

Zhang 和 Wang(2019)通过对目标检测器的一些经典攻击方法进行分析，从目标检测的多任务学习(Redmon 等，2016)角度指出，这些攻击是从单一的分类损失或分类损失与位置损失的组合实现对目标检测器的攻击，不同的任务损失对模型鲁棒性的作用不同，提出基于分类和位置损失来对抗训练目标检测器。结果表明，这种对抗训练在不同的攻击方法、数据集和检测器特征提取器上都能很好地提高鲁棒性。

5.2.2　限制上下文信息使用

Saha 等人(2020)认为现阶段的目标检测器之所以效果较好，是因为有效利用了上下文信息。这种上下文信息也因此被攻击者利用，Saha 提出在训练检测器时限制上下文信息的使用可以有效防御那

些通过上下文推理进行攻击的方法。

限制上下文信息的使用可以从两方面实现。1）Grad-defence。借助 Grad-CAM（Selvaraju 等，2017）的思想，将卷积层的中间进行可视化并进行裁剪，使其不超过被检测物体的边界框。如果超过了被检测物体的边界框，则对边界框周围的像素做非零惩罚，以限制最后一层的感受野范围，从而降低对抗样本在物体周围添加的噪声影响，获得更准确的预测。2）在训练数据中消除上下文的影响，通过人工在训练图像上粘贴一个脱离上下文的前景物体，用这种图像训练检测器，以限制上下文信息的使用。实验证明，这两种方法都能起到一定的防御作用。

5.2.3 正则化方法

Bouabid 和 Delaitre（2020）将图像分类领域的混合训练方法（Zhang 等，2018）扩展到目标检测领域来提高模型的鲁棒性。混合正则训练方法使神经网络由经验风险最小化变为领域风险最小化，可以使网络不再仅记忆训练的数据，而且更加关注泛化的数据。通过线性插值样本及标签，将图像从像素和锚点网络两方面进行混合，形成标签和样本的凸组合，然后在凸组合上训练神经网络，减少对错误标签的记忆，增加模型的鲁棒性。

5.2.4 特征对齐

Xu 等人（2021）提出使用中间层的特征对齐可以提高模型鲁棒性，降低对抗样本的攻击效果。通过知识蒸馏（knowledge distillation）和自监督学习（self-supervised learning）两种方法，将来自 siamese 网络和教师网络的先验特征进行特征对齐，这样的特征更加全面有效，通过指导中间特征层的输出来强化对抗训练，使得到的网络抗干扰能力更强。实验结果表明，新的特征对齐方法在防御上比 Zhang 和 Wang（2019）的对抗训练效果更好。

5.2.5 噪声混合

Li 等人（2020）观察到对抗样本的噪声是通过反向传播形成的，具有特定规律，如果破坏这种规律就会降低攻击效果。同时，神经网络对通用噪声的敏感度比对抗噪声低，更容易抵抗没有规律的通用噪声。因此设计了一种分段屏蔽框架，将图像分成小部分，随机对每部分的像素进行清洗，并添加通用噪声破坏对抗样本中的对抗模式，以达到防御目的。实验表明，加入分段屏蔽框架的目标检测模型在抵抗对抗样本的性能上优于未加入分段屏蔽框架的检测器。

5.2.6 检测器预警（DetectorGuard）

Xiang 和 Mittal（2021）针对对抗攻击中的局部攻击，设计了一种名为检测器预警（DetectorGuard）的通用框架报警机制。DetectorGuard 利用图像分类与目标检测的相互联系，将鲁棒性从分类器传递到检测器，设计了一种对象预测器的组件。对象预测器通过在整幅图像或特征图上使用比较鲁棒的分类器作为滑动窗口，对图像中的各位置分类，最后将分类结果进行聚合得到最后的预测结果。将预测结果与原始目标检测器的检测结果进行对比，如果两个检测结果不一致，则认定为对抗样本，触发检测器预警框架的攻击警报。

5.3 防御策略总结

表 7 列举了一些目标检测现有防御策略，这些防御策略各有优缺点且适用场合不同，往往只针对特定的攻击方法或数据集。目前，最有效的措施是进行对抗训练，但对抗训练需要优先生成对抗样本，对目标检测任务来说，开销大，速度慢。特别对于补丁类的攻击的防御代价更大，因为生成一个有效的补丁往往需要上千次乃至上万次迭代。

总体来说，目标检测领域的防御技术目前仍然十分匮乏，仅有的一些措施是对全局扰动攻击进行防御，对物理场景中检测的防御甚至没有。造成这种现象的原因是目标检测领域的对抗攻击出现时间短，研究远没有图像分类领域那么深入，且目标检测的网络结构更加复杂，不仅涉及分类，还涉及位置回归。所以许多分类器上的防御技术在目标检测领域遇到回归网络就会失效。如何有效防御目标检测领域的对抗攻击在将来是一个研究热点。

6 挑战与展望

目前，目标检测领域的对抗样本生成和防御技术处于探索阶段，还有很大发展空间。未来值得重点关注的研究方向如下：

1）目标检测对抗样本的可迁移性。对抗样本的可迁移性作为衡量对抗样本的重要属性之一，需要得到更多关注。而目标检测是计算机视觉中的一个热门领域，它不同于图像分类，由于其特有的方法和技术，使得目标检测对于一般的对抗样本有很好

表 7　防御方法总结

Table 7　Summary of defense methods

类别	方法	描述
预处理防御	去噪、滤波和图像压缩	通过滤波器去噪滤波以及图像压缩减少绝大多数噪声,降低攻击效果。
	随机中值平滑	通过改进高斯平滑在目标检测易受回归问题的影响来抑制噪声。
提高模型鲁棒性防御	对抗训练	从回归损失和分类损失两个角度提出目标检测器的特有对抗训练方式。
	限制上下文信息使用	在训练阶段减少检测器对上下文推理的过分依赖,以此增强检测器的鲁棒性。
	正则化方法	采用混合正则训练的方法使检测器泛化能力更强,增加对对抗样本的鲁棒性。
	特征对齐	通过知识蒸馏和自监督学习来对齐特征,使网络提取的特征更全面,令检测器的抗干扰能力更强。
	检测器预警	通过用更鲁棒的分类器在图像上进行滑动,然后将最后结果汇总与检测器结果比较,得到更可靠的输出。
	噪声混合	将图像分成若干份,随机进行像素清洗,同时加入新的噪声破坏对抗噪声的规律,使对抗样本失去攻击性。

的抗干扰性。目前该领域提出的大多数对抗样本白盒生成方法往往是针对特定的一类目标检测器进行攻击,例如针对两阶段检测器进行攻击,但是这样生成的对抗样本对单阶段检测器的攻击效果就不尽人意。说明现阶段方法生成的对抗样本的可迁移能力较差,无法在其他模型上取得较好效果。如何有效解决现有攻击方法过于依赖模型信息而导致生成的对抗样本缺乏泛化能力的问题,提高对抗样本的可迁移性和鲁棒性,使生成的对抗样本在各种目标检测模型上均具有较高的攻击性,是未来的一个研究热点。

2)目标检测的对抗防御。由于目标检测对抗样本生成是一个新兴领域,处于起步阶段,对于对抗样本产生的原因仍未达成共识。对抗防御方面的研究较少,已有的一些防御策略主要是借鉴图像分类中的对抗防御方法。如何增强目标检测器的鲁棒性以从容应对对抗攻击,以及如何从神经网络的根源抵抗这种攻击,仍然需要进一步研究。随着目标检测在安全攸关领域的普遍应用,设计更鲁棒、更安全的目标检测器越来越成为一项迫切的任务。

3)对抗样本的扰动大小和生成速度。由于目标检测模型是一个大型网络,而现有方法,如:DFool(Lu 等, 2017)、DAG(Xie 等, 2017)、RAP(Li, 2018b)、CAP(Zhang 等, 2020)、BPatch(Li 等, 2018a)、TOG(Chow 等,2020b)都是基于反向传播生成的,因此生成对抗样本需要较长时间。尽管 UEA 方法提出使用 GAN 网络提前训练好网络,这样生成对抗样本时仅通过前向传播就能生成,但是这种方法的攻击效果并不十分理想,而且生成的图像扰动较大。因此还需要在此基础上设计一种更好的网络,使生成对抗样本的速度比传统方法更快,在节省时间的同时,能保持较好的低扰动率和较高的攻击效果,这也是未来发展的一个重点方向。

7　结　语

本文从对抗样本扰动生成的范围、攻击的检测器类型以及使用的损失函数出发,归纳总结了面向目标检测的对抗样本生成方法。通过实验比较分析了几种典型对抗样本生成方法的性能。介绍了针对目标检测现有的对抗防御策略。

目前,目标检测领域的对抗样本无论生成还是防御都还存在很多问题,且由于对抗样本产生的原因还不是十分清楚,因而防御策略的研究较少。本文希望能够给研究者带来更多关于目标检测的对抗样本生成与防御的研究思路。随着对目标检测对抗样本生成与防御研究的深入开展,必然会推动目标检测技术进一步发展,为目标检测的广泛应用提供更安全的保障。

参考文献(References)

Akhtar N, Liu J and Mian A. 2018. Defense against universal adversarial perturbations//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3389-3398 [DOI: 10.1109/CVPR.2018.00357]

Akhtar N and Mian A. 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access, 6: 14410-14430 [DOI: 10.1109/ACCESS.2018.2807385]

Athalye A, Engstrom L, Ilyas A and Kwok K. 2018. Synthesizing robust adversarial examples//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 284-293

Baluja S and Fischer I. 2017. Adversarial transformation networks: learning to generate adversarial examples [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1703.09387.pdf

Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional Siamese networks for object tracking//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]

Bochkovskiy A, Wang C Y and Liao H Y M. 2020. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/2004.10934.pdf

Bouabid S and Delaitre V. 2020. Mixup regularization for region proposal based object detectors [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/2003.02065.pdf

Brown T B, Mané D, Roy A, Abadi M and Gilmer J. 2017. Adversarial patch [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1712.09665.pdf

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Cao J L, Li Y L, Sun H Q, Xie J, Huang K Q and Pang Y W. 2022. A survey on deep learning based visual object detection. Journal of Image and Graphics, 27(6): 1697-1722. (曹家乐, 李亚利, 孙汉卿, 谢今, 黄凯奇, 庞彦伟. 2022. 基于深度学习的视觉目标检测技术综述. 中国图象图形学报, 27(6): 1697-1722.) [DOI: 10.11834/jig.220069]

Chen S T, Cornelius C, Martin J and Chau D H. 2019. ShapeShifter: robust physical adversarial attack on faster R-CNN object detector//Proceedings of 2019 European Conference on Machine Learning and Knowledge Discovery in Databases. Dublin, Ireland: Springer: 52-68 [DOI: 10.1007/978-3-030-10925-7_4]

Chiang P Y, Curry M J, Abdelkader A, Kumar A, Dickerson J and Goldstein T. 2020. Detection as regression: certified object detection by median smoothing//Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Vancouver, Canada: [s.n.]

Chow K H, Liu L, Gursoy M E, Truex S, Wei W Q and Wu Y Z. 2020a. TOG: targeted adversarial objectness gradient attacks on real-time object detection system [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/2004.04320.pdf

Chow K H, Liu L, Gursoy M E, Truex S, Wei W Q and Wu Y Z. 2020b. Understanding object detection through an adversarial lens//Proceedings of the 25th European Symposium on Research in Computer Security. Guildford, UK: Springer: 460-481 [DOI: 10.1007/978-3-030-59013-0_23]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/cvpr.2009.5206848]

Ding S and Zhao K. 2018. Research on daily objects detection based on deep neural network. IOP Conference Series: Materials Science and Engineering, 322(6): #062024 [DOI: 10.1088/1757-899x/322/6/062024]

Divvala S K, Efros A A and Hebert M. 2012. How important are "deformable parts" in the deformable parts model?//Proceedings of 2012 European Conference on Computer Vision. Florence, Italy: Springer: 31-40 [DOI: 10.1007/978-3-642-33885-4_4]

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Du T Y, Ji S L, Li J F, Gu Q C, Wang T and Beyah R. 2020. SirenAttack: generating adversarial audio for end-to-end acoustic systems//Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. Taipei, China: ACM: 357-369 [DOI: 10.1145/3320269.3384733]

Duan K W, Bai S, Xie L X, Qi H G, Huang Q M and Tian Q. 2019. CenterNet: keypoint triplets for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6568-6577 [DOI: 10.1109/ICCV.2019.00667]

Dziugaite G K, Ghahramani Z and Roy D M. 2016. A study of the effect of JPG compression on adversarial images [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1608.00853.pdf

Everingham M, Van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 88(2): 303-338 [DOI: 10.1007/s11263-009-0275-4]

Evtimov I, Eykholt K, Fernandes E, Kohno T, Li B, Prakash A, Rahmati A and Song D. 2017. Robust physical-world attacks on machine learning models [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1707.08945v2.pdf

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s. n.]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

Huang L C, Yang Y, Deng Y F and Yu Y N. 2015. DenseBox: unifying landmark localization with end to end object detection [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1509.04874.pdf

Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B and Mądry A. 2019. Adversarial examples are not bugs, they are features//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #12

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/cvpr.2017.632]

Kennedy J and Eberhart R. 1995. Particle swarm optimization//Proceedings of the ICNN'95-International Conference on Neural Networks. Perth, Australia: IEEE: 1942-1948 [DOI: 10.1109/ICNN.1995.488968]

Kurakin A, Goodfellow I J and Bengio S. 2017. Adversarial machine learning at scale//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Kuznetsova A, Rom H, Alldrin N, Uijlings, J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Kolesnikov A, Duerig T and Ferrari V. 2020. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 128(7): 1956-1981 [DOI: 10.1007/s11263-020-01316-z]

Law H and Deng J. 2018. CornerNet: detecting objects as paired keypoints//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 765-781 [DOI: 10.1007/978-3-030-01264-9_45]

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444 [DOI: 10.1038/nature14539]

Li H F, Li G B and Yu Y Z. 2020. ROSA: robust salient object detection against adversarial attacks. IEEE Transactions on Cybernetics, 50(11): 4835-4847 [DOI: 10.1109/tcyb.2019.2914099]

Li P C, Yi J F, Zhou B W and Zhang L J. 2019. Improving the robustness of deep neural networks via adversarial training with triplet loss//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI: 2909-2915 [DOI: 10.24963/ijcai.2019/403]

Li Y Z, Bian X and Lyu S W. 2018a. Attacking object detectors via imperceptible patches on background [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1809.05966v1.pdf

Li Y Z, Tian D, Chang M C, Bian X and Lyu S W. 2018b. Robust adversarial perturbation on deep proposal-based models//Proceedings of 2018 British Machine Vision Conference 2018. Newcastle, UK: BMVA Press

Liao Q Y, Wang X, Kong B, Lyu S W, Yin Y B, Song Q and Wu X. 2020. Category-wise attack: transferable adversarial examples for anchor free object detection [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/2003.04367.pdf

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu L, Ouyang W L, Wang X G, Fieguth P, Chen J, Liu X W and Pietikäinen M. 2020. Deep learning for generic object detection: a survey. International Journal of Computer Vision, 128(2): 261-318 [DOI: 10.1007/s11263-019-01247-4]

Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Computer Society: 8759-8768 [DOI: 10.1109/cvpr.2018.00913]

Liu T, Zhao Y, Wei Y C, Zhao Y F and Wei S K. 2019a. Concealed object detection for activate millimeter wave image. IEEE Transactions on Industrial Electronics, 66(12): 9909-9917 [DOI: 10.1109/tie.2019.2893843]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Liu W Y, Wen Y D, Yu Z D, Li M, Raj B and Song L. 2017. SphereFace: deep hypersphere embedding for face recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6738-6746 [DOI: 10.1109/cvpr.2017.713]

Liu X, Yang H R, Liu Z W, Song L H, Li H and Chen Y R. 2019b. DPATCH: an adversarial patch attack on object detectors//Proceedings of Workshop on Artificial Intelligence Safety 2019 Co-located with the 33rd AAAI Conference on Artificial Intelligence 2019. Honolulu, USA: CEUR-WS.org

Lu J J, Sibai H and Fabry E. 2017. Adversarial examples that fool detec-

tors [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1712.02494.pdf

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Modas A, Moosavi-Dezfooli S M and Frossard P. 2019. SparseFool: a few pixels make a big difference//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9079-9088 [DOI: 10.1109/cvpr.2019.00930]

Moosavi-Dezfooli S M, Fawzi A, Fawzi O and Frossard P. 2017. Universal adversarial perturbations//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 86-94 [DOI: 10.1109/CVPR.2017.17]

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/cvpr.2016.282]

Pan W W, Wang X Y, Song M L and Chen C. 2020. Survey on generating adversarial examples. Journal of Software, 31(1): 67-81 (潘文雯, 王新宇, 宋明黎, 陈纯. 2020. 对抗样本生成技术综. 软件学报, 31(1): 67-81) [DOI: 10.13328/j.cnki.jos.005884]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2021-04-08]. https://arxiv.org/pdf/1804.02767.pdf

Ren S H, Deng Y H, He K and Che W X. 2019. Generating natural language adversarial examples through probability weighted word saliency//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL: 1085-1097 [DOI: 10.18653/v1/p19-1103]

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 91-99

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Saha A, Subramanya A, Patil K and Pirsiavash H. 2020. Role of spatial context in adversarial robustness for object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 3403-3412 [DOI: 10.1109/cvprw50498.2020.00400]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 618-626 [DOI: 10.1109/iccv.2017.74]

Sharif M, Bhagavatula S, Bauer L and Reiter M K. 2016. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM: 1528-1540 [DOI: 10.1145/2976749.2978392]

Shetty S. 2016. Application of convolutional neural network for image classification on Pascal VOC challenge 2012 dataset [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1607.03785.pdf

Song C, Cheng H P, Yang H R, Li S C, Wu C P, Wu Q, Chen Y R and Li H. 2018a. MAT: a multi-strength adversarial training method to mitigate adversarial attacks//Proceedings of 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). Hong Kong, China: IEEE: 476-481 [DOI: 10.1109/isvlsi.2018.00092]

Song D, Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Tramèr F, Prakash A and Kohno T. 2018b. Physical adversarial examples for object detectors//Proceedings of the 12th USENIX Workshop on Offensive Technologies. Baltimore, USA: USENIX Association

Su J, Vargas D V and Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5): 828-841 [DOI: 10.1109/TEVC.2019.2890858]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J and Fergus R. 2014. Intriguing properties of neural networks [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1312.6199v4.pdf

Thys S, Van Ranst W and Goedemé T. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 49-55 [DOI: 10.1109/cvprw.2019.00012]

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/iccv.2019.00972]

Tramèr F, Kurakin A, Papernot N, Goodfellow I J, Boneh D and McDaniel P D. 2018. Ensemble adversarial training: attacks and defenses//Proceedings of the 6th International Conference on Learning Representation. Vancouver, Canada: OpenReview.net: 1-20

Uijlings J R R, van de Sande K E A, Gevers T and Smeulders A W M. 2013. Selective search for object recognition. International Journal of Computer Vision, 104(2): 154-171 [DOI: 10.1007/s11263-013-0620-5]

Vincent P, Larochelle H, Bengio Y and Manzagol P A. 2008. Extracting

and composing robust features with denoising autoencoders//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM: 1096-1103 [DOI: 10.1145/1390156.1390294]

Viola P and Jones M J. 2004. Robust real-time face detection. International Journal of Computer Vision, 57(2): 137-154 [DOI: 10.1023/B:VISI.0000013087.49260.fb]

Wang D R, Li C R, Wen S, Han Q L, Nepal S, Zhang X Y and Xiang Y. 2021. Daedalus: breaking nonmaximum suppression in object detection via adversarial examples. IEEE Transactions on Cybernetics [DOI: 10.1109/tcyb.2020.3041481]

Wang Q L, Guo W B, Ororbia II A G, Xing X Y, Lin L, Giles C L, Liu X, Liu P and Xiong G. 2016a. Using non-invertible data transformations to build adversary-resistant deep neural networks [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1610.01934v4.pdf

Wang Q L, Guo W B, Zhang K X, Xing X Y, Giles C L and Liu X. 2016b. Random feature nullification for adversary resistant deep architecture [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1610.01239v3.pdf

Wang X Y, Han T X and Yan S C. 2009. An HOG-LBP human detector with partial occlusion handling//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE: 32-39 [DOI: 10.1109/ICCV.2009.5459207]

Wang Y J, Tan Y A, Zhang W J, Zhao Y H and Kuang X H. 2020. An adversarial attack on DNN-based black-box object detectors. Journal of Network and Computer Applications, 161: #102634 [DOI: 10.1016/j.jnca.2020.102634]

Wei X X, Liang S Y, Chen N and Cao X C. 2019. Transferable adversarial attacks for image and video object detection//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI: 954-960 [DOI: 10.24963/ijcai.2019/134]

Wu X, Huang L F and Gao C Y. 2019. G-UAP: generic universal adversarial perturbation that fools RPN-based detectors//Proceedings of the 11th Asian Conference on Machine Learning. Nagoya, Japan: PMLR: 1204-1217

Xiang C and Mittal P. 2021. DetectorGuard: provably securing object detectors against localized patch hiding attacks//Proceedings of 2021 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event: ACM: 3177-3196 [DOI: 10.1145/3460120.3484757]

Xiao C W, Li B, Zhu J Y, He W, Liu M Y and Song D. 2018. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI: 3905-3911 [DOI: 10.24963/ijcai.2018/543]

Xie C H, Wang J Y, Zhang Z S, Zhou Y Y, Xie L X and Yuille A. 2017. Adversarial examples for semantic segmentation and object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1378-1387 [DOI: 10.1109/iccv.2017.153]

Xu W P, Huang H C and Pan S Y. 2021. Using feature alignment can improve clean average precision and adversarial robustness in object detection//Proceedings of 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE: 2184-2188 [DOI: 10.1109/ICIP42928.2021.9506689]

Yan B, Wang D, Lu H C and Yang X Y. 2020. Cooling-shrinking attack: blinding the tracker with imperceptible noises//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 987-996 [DOI: 10.1109/cvpr42600.2020.00107]

Yuan X Y, He P, Zhu Q L and Li X L. 2019. Adversarial examples: attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9): 2805-2824 [DOI: 10.1109/tnnls.2018.2886017]

Zhang H C and Wang J Y. 2019. Towards adversarially robust object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 421-430 [DOI: 10.1109/iccv.2019.00051]

Zhang H T, Zhou W G and Li H Q. 2020. Contextual adversarial attacks for object detection//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE: 1-6 [DOI: 10.1109/icme46284.2020.9102805]

Zhang H Y, Cissé M, Dauphin Y N and Lopez-Paz D. 2018. mixup: beyond empirical risk minimization//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net: 1-13

Zhou X Y, Wang D Q and Krähenbühl P. 2019a. Objects as points [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/1904.07850.pdf

Zhou X Y, Zhuo J C and Krähenbühl P. 2019b. Bottom-up object detection by grouping extreme and center points//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 850-859 [DOI: 10.1109/cvpr.2019.00094]

作者简介

袁珑,男,硕士研究生,主要研究方向为深度学习、对抗样本攻击和防御、目标检测。E-mail: 835524032@qq.com

孙军梅,通信作者,女,副教授,主要研究方向为深度学习和智能软件系统。E-mail: junmeisun@hznu.edu.cn

李秀梅,女,教授,主要研究方向为视频分析及应用、压缩感知与机器学习。E-mail: lixiumei@hznu.edu.cn

潘振雄,男,硕士研究生,主要研究方向为深度学习、物理对抗样本攻击和防御、图像分类。

E-mail: 15167271541@163.com

肖蕾,女,副教授,主要研究方向为智能信息处理和软件测试。E-mail: lxiao@xmut.edu.cn