面向图像分析领域的黑盒对抗攻击技术综述

武 阳 刘 靖
(内蒙古大学计算机学院 呼和浩特 010021)

摘 要 图像领域下的黑盒攻击(Black-box Attack)已成为当前深度神经网络对抗攻击领域的热点研究方向.黑盒攻击的特点在于仅利用模型输入与输出的映射关系,而无需模型内部参数信息及梯度信息,通过向图像数据加入人类难以察觉的微小扰动,进而造成深度神经网络(Deep Neural Network,DNN)推理与识别失准,导致图像分析任务的准确率下降,因此由黑盒攻击引起的鲁棒性问题成为当前 DNN 模型研究的关键问题.为提高黑盒攻击在图像分析任务下的攻击成效,现有相关研究以低查询次数、低扰动幅度、高攻击成功率作为优化目标,针对不同图像分析任务采用不同的攻击模式与评估方式.本文以主流的图像分析任务为出发点,阐述图像分类、目标检测与图像分割三类任务中黑盒攻击算法的核心思想和难点,总结黑盒对抗攻击领域中的关键概念与评估指标,分析不同图像分析任务中黑盒对抗攻击的实现策略与研究目标.阐明各个黑盒攻击算法间的关系与优势,从攻击成功率、查询次数以及相似性度量等多个方面对不同的黑盒攻击算法进行性能比较,以提出目前图像分析领域中黑盒对抗攻击仍存在的主要挑战与未来研究方向.

关键词 黑盒对抗攻击;深度神经网络;鲁棒性;图像分类;目标检测;图像分割
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.01138

A Survey on Black-Box Adversarial Attack in Image Analysis

WU Yang LIU Jing
(*College of Computer Science, Inner Mongolia University, Hohhot* 010021)

Abstract In the domain of image processing, black-box adversarial attacks have emerged as a prominent and hot area of research within the current landscape of adversarial attacks on deep neural networks (DNNs). Distinguished by their exclusive reliance on the input-output mapping of a model, black-box attacks forego internal model parameters and gradient information. By subtly introducing imperceptible perturbations into image data, these attacks induce misalignment in the inference and recognition capabilities of deep neural networks (DNNs), resulting in a deterioration of accuracy in image analysis tasks. Consequently, the robustness issues posed by black-box attacks have become a critical and focal concern in current DNN model research. To enhance the efficacy of black-box attacks in image analysis tasks, current research endeavors focus on optimizing objectives such as achieving low query counts, minimal perturbation amplitude, and high attack success rates. Different attack modes and evaluation methodologies are employed for distinct image analysis tasks. Beginning with mainstream image analysis tasks, including image classification, object detection, and image segmentation, this paper expounds on the core ideas and challenges presented by black-box attack algorithms within each category. The paper systematically summarizes key concepts and evaluation metrics in the domain of black-box adversarial attacks. The current evaluation metrics predominantly encompass three critical aspects. Firstly, the attack success rate

收稿日期:2023-04-06;在线发布日期:2024-01-16.本课题得到国家自然科学基金(61662051)、内蒙古科技计划项目(2020GG0187)、内蒙古自然科学基金重点项目(2023ZD18)资助.**武 阳**,博士研究生,主要研究方向为黑盒对抗攻击、深度学习模型鲁棒性分析等.E-mail:22109007@mail.imu.edu.cn.**刘 靖**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为可信人工智能、软件可靠性工程、云计算与大数据分析等.E-mail:liujing@imu.edu.cn.

is measured distinctively for various image analysis tasks. In image classification, the success of an attack implies a discrepancy between the model's output category and the original label category, often quantified through image misclassification rates. Object detection tasks frequently rely on the mean Average Precision (mAP) metric, where lower post-attack mAP values indicate heightened attack effectiveness. In image segmentation tasks, the success of a black-box attack is gauged by differences between generated pixel-wise segmentation images and labeled segmentation images, with certain pixels recognized as other categories. Presently, black-box attacks in segmentation tasks are frequently assessed using the mean Intersection over Union ($mIoU$) metric, where lower $mIoU$ values signify elevated attack performance. Secondly, considerations encompass query counts and attack time, instrumental in gauging the efficiency of black-box adversarial attacks. Reduced query counts or attack times denote enhanced efficiency in generating adversarial samples. Finally, similarity metrics center on the fundamental task of adversarial attacks which is ensuring model misalignment in inference and recognition while preserving perturbation imperceptibility. Consequently, generated adversarial samples need to closely resemble the original samples. This paper introduces current similarity metrics employed in black-box adversarial attacks. Based on the above content, the paper comprehensively analyzes the implementation strategies and research objectives of black-box adversarial attacks in various image analysis tasks. It elucidates the relationships and advantages among various black-box attack algorithms, categorizing them into four distinct types: meta-heuristic-based black-box adversarial attack techniques, proxy-model-based black-box adversarial attack techniques, direct-search-based black-box adversarial attack techniques, and zeroth-order optimization-based black-box adversarial attack techniques. Performance comparisons are systematically conducted across multiple facets, including attack success rates, query counts, and similarity metrics. The paper culminates by highlighting major challenges persisting in the realm of black-box adversarial attacks in image analysis and proposing comprehensive future research directions.

Keywords　black-box adversarial attack; deep neural network; robustness; image classification; object detection; image segmentation

1　引　言

近年来，深度神经网络(Deep Neural Network，DNN)被广泛应用于图像分类任务[1]、目标检测任务[2]以及图像分割任务[3]等各类复杂任务，这些任务通常作为自动驾驶、智慧医疗、人脸识别等实际任务中的基础任务或核心环节. 现有的 DNN 模型性能较优，但仍然存在一些问题，由对抗样本(Adversarial Samples)引起的鲁棒性问题是当前的研究热点之一.

部分研究人员聚焦于现实任务中深度学习系统出现的鲁棒性问题，2016 年 Sharif 等人[4]的研究团队通过佩戴打印带有扰动的太阳眼镜，可成功绕过先进的面部识别系统，甚至导致系统识别为其他人. 2019 年 Zhao 等人[5]通过将生成的扰动补丁覆盖于标志牌，可躲避自动驾驶系统的智能检测. 之后，Kumar 等人[6]进一步研究自动驾驶领域中的对抗攻击，算法产生的扰动可直接影响自动驾驶模型对于交通标志或交通信号灯的识别. 2021 年 Li 等人[7]针对自动驾驶模型提出一种黑盒攻击算法，以测试模型对于扰动街道标志图像的识别性能，实验表明即便添加微小扰动也可绕过模型检测，这极大影响之后自动驾驶模型的普及与应用. 2022 年 Wang 等人[8]提出基于偏见的补丁攻击，并将其应用于自动结账场景，成功攻击了淘宝与京东电子商务平台，粘贴的补丁导致平台对商品识别错误. 目前人脸活体检测已广泛用于许多安全敏感领域的身份验证，2022 年 Li 等人[9]聚焦于人脸活体检测系统的安全性，并提出一种对抗攻击框架可针对人脸活体检测系统进行定制化安全评估，实验表明利用生成的图像可成功绕过现实应用中人脸检测系统的识

别.目前对抗攻击的应用场景已经从自动驾驶领域等复杂工业场景逐步向电子商务平台、人脸活体检测系统等日常应用场景过渡,由此带来的信息安全问题极大影响 DNN 模型在实际任务场景和安全应用领域的运用.因此研究攻击算法如何产生扰动对于后续的防御工作,进而提升 DNN 模型鲁棒性至关重要.

从实际应用领域来看,虽然 DNN 模型被广泛部署在各类应用中,但普通用户仅可以通过查询来得到识别或预测结果,难以获取应用或第三方平台内部的模型详细信息,如目标模型的具体参数、结构和超参数等.因此仅利用模型的查询结果来扰动原始样本的黑盒攻击(Black-box Attack)更符合实际任务场景,越来越多的研究人员开始关注黑盒情况下如何生成有效的对抗样本,本文对现有的黑盒对抗攻击研究和技术进行分析与总结.

从攻击目标来看,黑盒攻击仅可利用模型输入与输出的映射关系生成对抗样本,因此需要大量查询目标模型以获取预测结果,攻击行为极易被第三方平台所监测.同时为欺骗人眼和平台检测算法,需要减小对抗样本的扰动幅度.由上分析可知,现有的黑盒攻击主要以低查询次数、低扰动幅度、高攻击成功率作为算法的优化目标.

从问题形式来看,黑盒攻击问题可表示为如式(1)形式的约束优化问题.其中,$f(\cdot)$表示当前任务中的目标模型,模型的具体形式、结构与各参数均未知.$h(\cdot)$表示当前黑盒攻击的目标函数,根据具体任务情况目标函数的形式不同[10-12].原始样本集表示为 X,样本集中的标注信息可表示为 y_i,针对不同的任务及不同的攻击类型,标注信息 y_i 的具体含义有所不同.针对不同任务,在图像分类任务中,标注信息 y_i 为图像的类别信息;在目标检测任务中,标注信息 y_i 包括目标类别与位置信息;在图像分割任务中,标注信息 y_i 包含图像像素级的标签信息.针对不同攻击类型,对于目标攻击,y_i 表示目标标签信息,而在非目标攻击中,y_i 表示除样本原始标签外的其他标签信息.$Dist(x_i, x_i^{adv})$ 表示原始样本与对抗样本间的距离,现有文献中主要采用 L_p 范数来衡量两者距离.阈值 δ 往往与具体任务可接受的最大扰动值有关.通过向样本 x_i 添加扰动 ε,即可生成对抗样本 x_i^{adv}.据此,目标函数 $h(f(x_i^{adv}), y_i)$ 用于衡量模型 $f(\cdot)$ 对于对抗样本 x_i^{adv} 的输出结果与标签 y_i 间的距离关系,在限制扰动距离情况下,目标函数值越小代表攻击成效越高.

$$\begin{aligned} &\min_{x_i^{adv}} h(f(x_i^{adv}), y_i) \\ &\text{s.t. } Dist(x_i, x_i^{adv}) \leq \delta \end{aligned} \tag{1}$$

由式(1)并结合黑盒攻击概念可知,黑盒攻击目标是满足样本间距离约束条件情况下,最小化 $h(f(x_i^{adv}), y_i)$ 函数,因此衡量黑盒攻击成功与否的关键在于评估指标如何正确表示 $h(\cdot)$ 函数的趋势以及约束条件.针对不同图像分析任务,评估指标间存在差异,具体指标情况列举在第 2 节.

由上分析可知,黑盒攻击问题可归结为目标函数 $f(\cdot)$ 形式未知的约束优化问题,因此利用黑盒优化问题(Black-Box Optimization,BBO)[13]来描述黑盒攻击,有助于分析不同黑盒攻击算法间的联系与特点.针对现有的黑盒优化方法,文献[14]以优化策略作为划分依据,将黑盒优化方法分为基于元启发式算法、基于代理模型算法、基于直接搜索算法、基于零阶优化算法,其重点在于无导数信息下如何获取原问题的解.面向图像分析领域的黑盒对抗攻击技术主要基于不同黑盒优化策略来构建对抗样本,各类黑盒攻击技术所解决的子问题也有所区别,本文以解决黑盒优化问题中的四类关键方法为切入点,分析各类黑盒对抗攻击技术优化策略和关键问题,将各类任务下的黑盒攻击算法分为以下四种类型.

(1)基于元启发式的黑盒对抗攻击技术

基于元启发式的黑盒对抗攻击技术首先构建样本种群,并利用适应度函数来筛选当前候选样本集,之后通过更新策略来获取对抗样本.整体优化过程中,样本种群如何初始化与后续对抗样本生成质量息息相关,如何定义适应度函数使其能够充分表示黑盒攻击的优化目标,以及如何制定更新策略来生成对抗样本是该黑盒对抗攻击技术需要解决的关键子问题.

(2)基于代理模型的黑盒对抗攻击技术

基于代理模型的黑盒对抗攻击技术将原约束优化问题转换为近似优化问题,因此,如何构建近似问题并保证获取的解与原问题下解的一致性,以及如何针对近似问题进行优化是该黑盒对抗攻击技术下需要解决的子问题.

(3)基于直接搜索的黑盒对抗攻击技术

基于直接搜索的黑盒对抗攻击技术在于利用当前解空间,迭代生成对抗样本,因此解空间如何构建以确保解空间与对抗样本空间对应,以及如何制定搜索策略来生成对抗样本是该黑盒对抗攻击技术下的核心子问题.

(4) 基于零阶优化的黑盒对抗攻击技术

相比于其他三类黑盒对抗攻击，基于零阶优化的黑盒对抗攻击技术仅依赖原始优化问题，通过估计目标函数梯度，并确定更新方向，之后利用更新策略优化候选解以生成对抗样本. 因此，如何提高原始目标函数梯度估计的准确性、如何确定更新方向以及制定何种更新策略是当前基于零阶优化的黑盒对抗攻击技术所需解决的关键问题.

本文重点对黑盒攻击算法的已有研究工作进行综述，基于上述三类主流图像分析任务和四种黑盒攻击技术类型，对各类方法下的攻击过程和关键问题进行分析和总结. 在第 2 节中主要介绍本文所涉及的关键概念与相关指标，包括对抗样本的概念、攻击类型以及文献中的实验评估指标等信息；在第 3 节中详细介绍本文的综述过程，重点从文献搜索、文献发表来源与等级统计情况来阐述文献收集工作. 在后续章节中详细论述了各类黑盒攻击算法的优化思想，通过分析各方法的改进策略与目的来梳理算法间的关系，总结了各类图像分析任务下的评价指标、数据集、目标网络模型等性能评估角度，详细汇总和分析各类任务下的各算法攻击性能，讨论了图像分析领域下黑盒攻击主要挑战及未来的研究方向.

2　关键概念与指标

本节对综述中出现的相关概念以及涉及的指标进行介绍，主要包括对抗样本的概念、对抗攻击的类型以及目前黑盒攻击中常见的相似度度量指标等信息.

(1) 对抗样本

对抗样本(Adversarial Samples)的相关概念最早可追溯到 Szegedy 等人[15]的研究. 通过向图像数据中加入人眼难以察觉的扰动，可以导致性能较优的 DNN 模型以较高置信度分类错误，使得 DNN 模型在现实应用中容易出现安全问题. 根据文献[15]中的定义，原始样本表示为 $x(x\in R^m)$，r 表示扰动，l 表示给定目标攻击标签，$f:R^m\rightarrow\{1\cdots k\}$ 代表分类器，$loss_f$ 表示模型输出与给定类别 l 之间的损失函数，生成对抗样本应满足式(2).

$$\begin{aligned}&\min\ \ c|r|+loss_f(x+r,l)\\&\text{s.t.}\ \ x+r\in[0,1]^m\end{aligned}\tag{2}$$

具体来说，需要找寻满足 $f(x+r)=l$ 的最小 r 值，以生成对抗样本 $x+r$，即保证错误分类情况下，找寻扰动最小的对抗样本.

(2) 攻击类型

① 黑盒攻击与白盒攻击. 现有的对抗样本生成方法可以依据攻击者是否掌握目标模型的先验信息分为白盒攻击(White-box Attack)与黑盒攻击(Black-box Attack)，这些先验信息主要包括目标模型的具体参数、网络结构和训练超参数等. 其中，白盒攻击正是在已知模型结构与参数信息情况下进行的攻击手段. 而黑盒攻击中攻击者仅仅只能访问目标模型的输出结果，利用有限的查询次数与结果以生成对抗样本.

② 目标攻击与非目标攻击. 对抗攻击按照是否针对具体目标标签攻击，分为目标攻击以及非目标攻击. 如式(1)中，若为目标攻击，则 y_i 表示目标标签信息，攻击成功代表 $f(x_i^{adv})=y_i$；若为非目标攻击，则 y_i 表示除原始标签 $\hat{y}_i$ 外的其他标签信息，攻击成功代表 $f(x_i^{adv})\neq\hat{y}_i$.

(3) 攻击成功率

攻击成功率泛指攻击成功的对抗样本占所有样本的比值，用于衡量当前生成的对抗样本是否具有干扰当前模型的预测和性能的能力，其值与式(1)中目标函数 $h(f(x_i^{adv}),y_i)$ 有关，多数情况下攻击成功率趋势与目标函数趋势相反. 根据不同的图像分析任务，攻击成功率的度量存在差异. 在图像分类任务中，攻击成功则意味着模型输出类别与原始标签类别不同，因此常用图像识别错误率表示攻击成功率. 在目标检测任务中，攻击成功代表三类情况发生，一是检测得到的目标类别不变，但识别的目标位置与标记位置不同；二是检测到的目标位置基本不变，但类别与原始样本类别不同；三是检测到的目标类别与位置均不同. 目前相关研究常用 mean Average Precision(mAP)指标来衡量攻击效果，攻击模型后 mAP 指标越低，代表攻击成效越高. 在图像分割任务中，黑盒攻击成功代表其生成的逐像素分割图像与标记的分割图像不同，部分像素被识别为其他类别. 目前分割任务下黑盒攻击常用均交并比 $mIoU$ (mean Intersection over Union)来评估攻击性能的优劣，$mIoU$ 指标越低，代表攻击性能越高.

(4) 查询次数

将单独输入到模型中的图像记录称为一次查询. 查询次数用于统计攻击所需的迭代查询访问次数. 最小化查询次数可以减少构建对抗样本所花费的时间，因此查询次数是衡量对抗样本生成效率的关键指标.

(5) 攻击时间

攻击时间也即生成对抗样本所需时间，观察生

成对抗样本的平均耗时.与攻击的查询次数、当前数据的复杂程度、目标模型的参数大小以及算法整体的复杂度相关.越低的攻击时间说明当前算法的生成效率更高.

(6)相似度度量指标

对抗攻击核心任务在于保证模型的推理与识别失准,并尽可能保证扰动幅度难以被人类察觉,因此生成的对抗样本需要与原始样本基本相同,对应于式(1)中的约束条件 $Dist(x_i, x_i^{adv})$.现有的黑盒攻击文献中采用不同的相似度指标来衡量扰动幅度,较为常见的指标包括 L_p 范数、均方误差 *MSE*(Mean Square Error)、结构相似性度量 *SSIM*(Structure Similarity Index Measure)以及峰值信噪比 *PSNR*(Peak Signal to Noise Ratio)等.

① 基于 L_p 范数的相似性指标

L_p 范数常用以度量向量间的距离,其定义如式(3)所示.

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \tag{3}$$

黑盒攻击中常见的范数主要包括 0、2、∞三种类型.L_0 范数(即 $p=0$)表示向量 $x_i - x_{adv}$ 中非 0 元素个数,由此定义可知满足 L_0 范数的对抗扰动较为稀疏,仅在图像的部分像素上加入对抗扰动.L_2 范数(即 $p=2$)将扰动幅度规定在可接受范围内,以接近原始样本的像素分布.L_∞ 范数(即 $p=\infty$)规定了图像中所添加扰动的最大幅度.

② 基于 *MSE* 的相似性指标

均方误差在深度神经网络训练中常用于构建损失函数,以衡量输出向量与标签向量间的误差.黑盒攻击实验中,部分文献采用 *MSE* 来逐像素计算差值并取平均以计算样本间的距离.

③ 基于 *SSIM* 的相似度指标

SSIM 主要考虑图像中的亮度、对比度以及结构信息,用于衡量两幅图像的相似程度.

④ 基于 *PSNR* 的相似度指标

PSNR 常用作图像压缩领域以衡量信号重建之后的信号质量,在黑盒攻击领域,*PSNR* 用于衡量生成的对抗样本质量,其值越大代表质量较高.

3 综述方法与过程

对于综述而言,文献采集工作的丰富性与综述内容的全面性直接相关,本节从多个方面介绍相关文献的收集工作,方便之后相关研究人员统计现有研究情况,整体文献收集工作如图 1 所示.

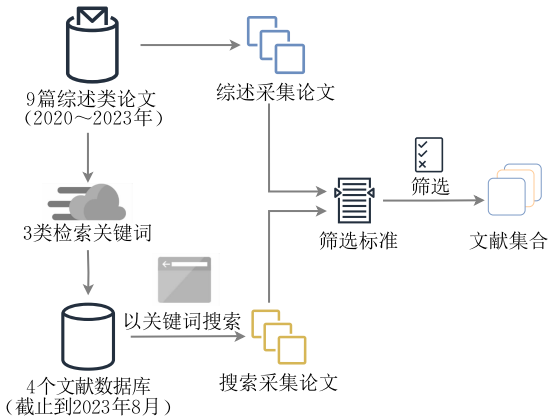

图 1 相关文献收集过程

首先是相关综述类论文,与对抗攻击相关的一部分综述论文中,包含有黑盒攻击相关算法研究,因此以这些综述类论文为导向,收集一部分黑盒攻击算法研究,本文共收集 9 篇相关综述类文献.

之后总结这些文献中的关键词,利用采集到的关键词或字符串来检索相关文献.本文将这些关键词或字符串按不同方向分为三类:一是与黑盒攻击概念相关的关键词,用以定位研究主题;二是与图像分析任务相关的关键词,用于检索对应图像任务下的黑盒攻击;三是与黑盒攻击特点相关的关键词或字符串,其来源于综述类论文中出现频率较高的关键词,用于进一步扩展搜索范围.本文采用的检索字符串如下所示:

(1) Black-box Adversarial (Attacks OR Examples)

(2) Image Classification OR Object Detection OR Image Segmentation

(3) Data-free OR Limited Queries OR Low-query OR Query-efficient OR Transferability OR Gradient Optimization OR Approximated Gradient OR Robustness OR Substitute (Training OR Model) OR Gradient-free OR Zeroth-Order Optimization

同时,本文收集文献主要来源于四个权威文献学术数据库:Science Direct、IEEE Xplore、ACM Digital Library 和 Scopus.以上述列举的不同角度的关键词为基础,通过不断组合来搜索相关文献.

最后将已收集文献中与图像分析任务关联度较小的文献,以及未在公开数据集评估的算法排除.

从发表时间来看,图 2 展示了本文中涉及文献的发表年份分布情况.对抗攻击相关研究主要开始于 2014 年,图中文献从 2014 年具体展示,2014 年之前主要包括黑盒优化相关研究与图像分析领域相关经典算法等内容.涉及文献数量的总趋势呈现逐年上升趋势,表明近年来,关于黑盒攻击的研究越来

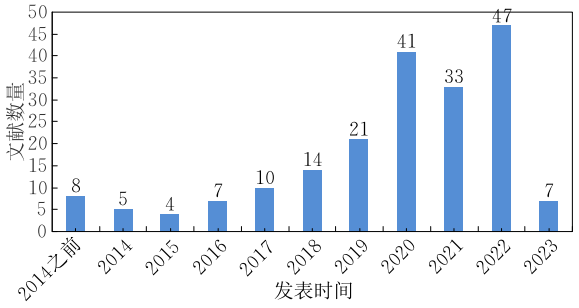

图 2　文献发表年份分布情况

越受到研究人员的关注.

从发表来源来看，图 3 中展示了涉及文献的来源情况，其中以会议(Conference)，期刊(Journal)，研讨会(Workshop)以及相关预印版(arXiv)为主要来源，总体发表渠道较为分散. 涉及文献中，会议与期刊论文为主要发表渠道，两者占所有论文的 84%. 会议论文占比最大，为整体文献数量的 49%. 图 4 统计会议发表情况，61%来自 CCF A 类推荐会议，18%来自 CCF B 类推荐会议，其中 81 篇论文发表在图像分析领域和人工智能领域相关的顶级会议，包括 AAAI(AAAI Conference on Artificial Intelligence)、ICLR(International Conference on Learning Representations)、CVPR(IEEE Conference on Computer Vision and Pattern Recognition)等.

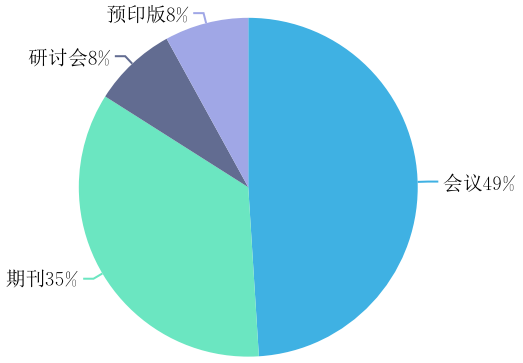

图 3　文献发表来源统计情况

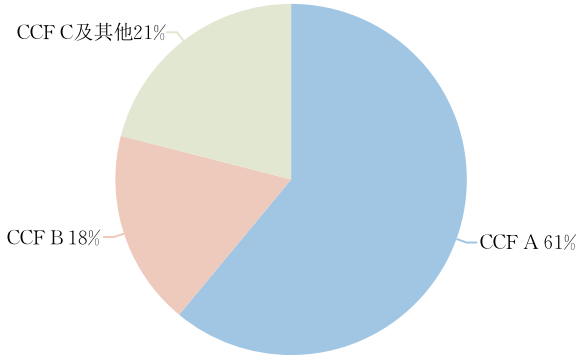

图 4　文献发表会议等级统计情况

从现有研究分析，自 2014 年起，图像分析领域下黑盒攻击相关研究呈现增长趋势. 而从发表来源来看，近年来相关文献多发表于相关顶级会议与期刊，表明黑盒攻击算法已逐步成为相关研究的热点问题之一.

在文献采集工作后，首先对综述类文献分析，以总结现有综述工作的角度与不足(第 4 节). 其次，以不同的图像分析任务为出发点对文献中的算法进行分类总结. 之后 4 节分别对应图像分类任务、目标检测任务、图像分割任务，和其他图像分析任务下的黑盒攻击. 在各任务中以四种黑盒优化算法为切入点，分别阐述基于元启发式算法、基于代理模型算法、基于直接搜索算法以及基于零阶优化算法的黑盒攻击. 重点汇总了现有算法的评估角度. 最后，在最后一节中总结本文的综述情况，并讨论了黑盒攻击的未来研究方向.

4　综述类论文分析

现有对抗攻击相关综述论文中主要是对整体对抗攻击与防御领域进行概述和总结，本文对近 3 年相关的 9 篇综述类论文进行概述.

文献[16]中对对抗攻击与对抗防御相关的文献进行了概述和汇总，包括对抗攻击、特定任务下的对抗攻击以及对抗防御相关内容. 文献[17]中概述了数据投毒攻击和对抗样本攻击，并从黑盒攻击以及白盒攻击角度对当前对抗样本攻击进行了概述. 文献[18]同样对攻击与防御进行介绍，并对目前对抗样本成因的几类假说进行总结. 在对抗防御相关内容中，文献[18]从不同角度对其进行分析，并对不同对抗防御算法的性能进行汇总. 文献[19]从白盒与黑盒两个角度对对抗攻击进行介绍，并总结了现有的防御方法. 文献[20]较为详细地总结了现有对抗攻击出现的时间轴，以及从不同角度对白盒与黑盒攻击进行分类. 文献[21]中将对抗样本生成算法分为两大类，即全像素扰动和部分像素扰动，并根据攻击类型与扰动可见性两个角度进行具体划分，在实验结果中采用 MNIST 数据集进行评估与分析. 文献[22]对当前黑盒攻击和防御相关文献进行了整理和分析，并将现有黑盒攻击技术分为梯度估计、基于迁移性的攻击、局部搜索以及组合优化方法，并对攻击方法的性能进行汇总. 文献[23]以攻击者所掌握的知识和对抗样本两个角度对对抗攻击研究进行描述，对抗样本主要关注扰动范围、扰动可见性以及扰动测量. 扰动范围主要包括两类，即攻击扰动面向个

体范围或者面向通用范围.扰动可见性主要考虑扰动是否被人类所发现.文献[23]采用 L_p 范数来进行扰动测量,其在攻击算法中主要用于限制插入图像的扰动大小以及数量.攻击者根据其掌握知识情况分为白盒、灰盒、与黑盒攻击.文献[24]将对抗攻击算法分为三个维度,即基于梯度的方法、基于分数的方法以及基于决策的算法,并汇报在 CIFAR-10 和 MNIST 数据集上的实验结果.具体文献整理情况如表 1 所示.

表 1 相关综述情况

文献	文献采集	实验结果整理	文献划分	任务调研情况
[16]	✓	×	×	✓
[22]	✓	✓	✓	×
[23]	✓	×	✓	×
[24]	✓	✓	✓	×
[17]	✓	×	✓	×
[18]	✓	×	✓	×
[19]	✓	×	✓	×
[20]	✓	×	✓	✓
[21]	✓	✓	✓	×

综上所述,目前多数文献主要关注 DNN 模型的对抗攻防算法,直接与黑盒攻击相关的文献较少、图像分析任务调查不充分、各算法间的联系与评估分析不全面.据此,本文从几个方面进行整理和汇总:(1)总结当前主流图像任务在黑盒攻击下的特点;(2)从不同图像分析任务与不同的黑盒优化算法两个维度对黑盒攻击算法进行划分和介绍;(3)对涉及的黑盒攻击算法实验结果按照不同数据集以及评估指标进行汇总与分析.

5 图像分类任务中的黑盒攻击

图像分类作为当前计算机视觉领域中的核心任务之一,同时也是人脸识别、自动驾驶等领域的基础任务.该任务目标在于利用图像中的信息对图像进行分类,即从已知图像类别标签库中为图像选择一个最为符合的类别标签,这些图像信息主要包括纹理信息、边缘信息等,因此图像分类的核心在于建立图像与标签的对应关系.现有研究中主要采用基于 *DNN* 模型结构进行建模,通过对图像信息的抽象和复杂组合实现分类,目前已在多个公开数据集中取得较优的性能,但由于图像本身复杂性较高,场景以及主体信息往往相互交叉覆盖,易受到干扰噪声影响,形成对抗样本,进而影响图像分类模型的分类精度.本文依据四类黑盒优化方法对图像分类中的黑盒攻击算法进行探讨.

在性能评估方面,本文总结了图像分类任务中现有黑盒攻击领域中较为常用的公开数据集,并从有目标攻击与无目标攻击两个维度对攻击算法进行汇总.在此基础上,选择多个评估指标以多角度更为全面地展示各个算法性能,图像分类任务中黑盒攻击评估所用数据集介绍如下:

(1) MNIST[25].该数据集由美国国家标准与技术研究院收集整理,其包括尺寸为 28×28 的灰度手写数字,并由 60 000 个训练示例和 10 000 个测试示例共同组成.

(2) CIFAR-10[26].该数据集用于识别常见物体的小型数据集,共有 10 个类,各个类中包含有 6000 个图像.主要包括 60 000 个尺寸为 32×32 的 RGB 图像,其中含有 50 000 张训练图像和 10 000 张测试图像.

(3) ImageNet[27].ImageNet 是由 14 197 122 张图像组成的大型数据集.主要使用类似于 WordNet 的语法集进行描述,以生成层级结构组织的图像数据集.共分为 21 841 个类别.

5.1 基于元启发式的黑盒攻击技术

元启发式算法(Meta-heuristic)是一类优化算法,其主要利用计算智能的机制对复杂优化问题进行求解,获得当前优化问题的最优解或满意解,通过模拟自然现象或动物行为过程,多次迭代以进行智能搜索.基于元启发式的对抗样本生成流程如图 5 所示.

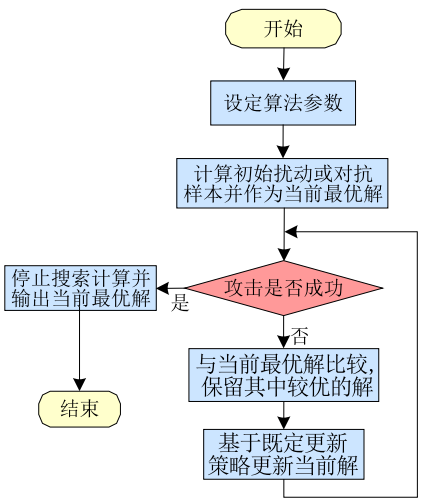

图 5 基于元启发式算法的黑盒对抗攻击流程

基于元启发式的对抗攻击首先需要设定初始解集,即初始的对抗样本或扰动,以这些初始对抗样本或扰动作为起始点,逐步搜索并与之前的最优解进行比较,更新当前的最优解,若攻击成功则表明对抗样本生成结束.其中,对抗样本生成的关键在于如何将元启发式算法与当前任务相结合,因此本节重点关注各类攻击算法利用何种元启发式算法,以及如何进行改进以生成对抗样本,并以这些问题对文献进行分析总结.相关研究情况如表 2 所示.

表 2　图像分类任务中基于元启发式的黑盒攻击技术研究情况

基础算法	研究目标	文献	改进策略	年份
粒子群优化算法(PSO)	加快收敛以提高扰动效率	[28]	与差分进化算法相结合	2018
	提高攻击成功率	[29]	在适应度函数中加入对原始样本与对抗样本之间距离的惩罚	2020
	提高对抗扰动的不可见性	[30]	基于自适应粒子群优化算法(APSO)，并通过平移与旋转来生成对抗样本	2022
多目标进化算法(MOEA)	提高对抗扰动的不可见性	[45]	问题转化为双目标优化问题且针对像素级和全局级提出四种类型攻击策略	2019
		[46]	引入注意力热图以获得突出区域，并基于多目标进化算法以生成对抗扰动	2022
		[47]	将 L_0 约束和原始优化问题作为两类优化目标采用进化算法生成对抗样本	2023
自然进化策略(NES)	减少查询次数	[32]	与高斯基上有限差分法相结合	2018
		[34]	以灰度图像作为初始图像并仅扰动部分区域	2018
	更符合少量信息场景	[35]	仅访问前 K 个标签信息	2018
差分进化算法(DE)	提高对抗扰动的不可见性	[40]	与协方差矩阵自适应进化策略相结合	2019
		[39]	仅修改单一像素	2019
		[42]	基于稀疏攻击思想并利用差分进化算法生成扰动	2022
		[43]	基于差分进化算法，针对图像属性构建对抗样本	2021
		[44]	设计一种多因素度量指标(MulFactorloss)并利用差分进化以搜索对抗扰动	2022
	搜索空间降维以提升查询效率	[41]	利用差分进化算法搜索梯度的符号而非扰动值	2022
协方差矩阵自适应进化策略(CMA-ES)	减少查询次数	[48]	以原始图像与目标图像的中间图像为起始图像	2019
		[49]	基于 CMA-ES 算法在子空间搜索扰动	2022
	提高攻击成功率	[50]	基于 CMA-ES 算法在图像像素域搜索对抗扰动	2021
	提高对抗扰动的不可见性	[51]	结合类间激活热力图信息以减少冗余扰动	2021
遗传算法(GA)	减少查询次数	[36]	采用退火策略以自适应调整超参数	2019
		[37]	基于遗传算法生成对抗样本，并利用代理模型信息以减少查询成本	2022
微生物遗传算法(MGA)	减少查询次数	[38]	利用基于替代模型生成的对抗样本作为初始种群	2020

部分文献采用粒子群优化算法(Particle Swarm Optimization，PSO)来生成对抗样本或扰动. 文献[28]中更新策略主要采用粒子群优化算法与差分进化相结合，通过在像素域进行搜索，以逐步更新当前最优的对抗样本，以获得迁移性较好的样本. 文献[29]中基于粒子群优化算法提出一种黑盒对抗攻击算法，对初始化种群、优化过程以及适应度函数做出相应的调整，在适应度函数中加入对原始样本以及对抗样本之间的距离进行惩罚，同时在优化过程中为避免对抗样本失真严重，采用修剪操作将其规定在图像范围内. 文献[30]基于自适应粒子群优化算法(Adaptive Particle Swarm Optimization，APSO)提出利用平移旋转操作进行有效对抗攻击的策略(Effective-AATR)，能够在给定有限的查询数内搜索最有效的对抗样本. 算法具体流程如图 6 所示.

图 6 中表示 Effective-AATR 的攻击策略，文献中将粒子表示为如式(4)所示，由位置 $\boldsymbol{p}$ 与速度 $\boldsymbol{v}$ 两个向量构成.

$$p_i=(p_{i1},p_{i2},\cdots,p_{iN}),v_i=(v_{i1},v_{i2},\cdots,v_{iN}) \quad (4)$$

其中，位置向量 $\boldsymbol{p}$ 中的各个元素由三个值构成，即表示原始样本的三维变换(包括旋转角度、水平平移以及垂直平移). 粒子群算法定义适应度函数值来评估各个粒子 $\boldsymbol{p}$ 位置的优劣，该文献[30]中将适应度函数定义为 $Pro(x,p_i)$，表示样本 x 经过 p_i 变换后

算法. Effective-AATR 搜索算法.

输入：加速度因子 c_1,c_2；随机权重 r_1,r_2；动量权重 w；迭代次数 T_2；群中粒子数 M

输出：对抗样本 x_{adv}

```
1. 初始化粒子群中各粒子 p_i的位置
2. For t=1 to T_2 do          /* 迭代搜索最优对抗变换 p_gb */
3.   For i=1 to M do          /* 对粒子群中各粒子进行更新 */
4.     v_i=wv_i+c_1r_1(p_i*-p_i)+c_2r_2(p_gb-p_i)
                              /* 计算各粒子更新方向 v_i */
5.     p_i=p_i+v_i            /* 更新粒子 p_i的位置 */
6.     If pro(x,p_i)<pro(x,p_i*) then
7.       p_i*=p_i             /* 经过 p_i变换后的样本正确分类概率更低则保存位置 p_i */
8.     End If
9.     If pro(x,p_i)<pro(x,p_gb) then
10.      p_gb=p_i             /* 保存当前最优对抗性变换 */
11.    End If
12.  End For
13. End For
14. 经过对抗性变换 p_gb后的样本 x 即为对抗样本 x_adv
15. Return x_adv              /* 得到对抗样本 x_adv */
```

图 6　Effective-AATR 算法的攻击流程[30]

被模型正确分类的概率，因此适应度函数值与正确分类概率相反，即正确分类概率越低，生成样本的适应度函数值越高，攻击成功的概率越大. 生成对抗样本的结果如图 7 所示，由图 7 可知各样本仅经过轻微平移或旋转则会导致模型分类错误.

图 7 由 Effective-AATR 算法生成的对抗样本[30]

此外，部分文献关注元启发式算法中的进化策略(Evolution Strategies)，其中自然进化策略(Natural Evolutionary Strategies，NES)[31]是一种常见的更新策略，其主要思想是基于搜索分布 $\pi(\theta|x)$ 思想的无导数优化算法，并非直接最大化目标函数，而是在搜索分布下最大化损失函数的期望值. 文献[32]中提出一种受启发于自然进化策略的算法变体，相比于基于有限差分的梯度估计方法[33]，生成速度快 2 到 3 倍. 文献[34]中同样采用自然进化策略以获得估计的梯度，但相比于之前的算法，该文献首先从灰度图像进行搜索以减少查询所需复杂度，同时生成扰动的过程并非遍布整张图像，而是通过启发式方法选取了图像的局部区域，并估计该区域中的梯度信息，之后将其平铺于整张图像，由于区域相对较小，且以灰度图像开始搜索，因此生成扰动的整体复杂度下降. 文献[35]中假设了三类攻击场景，即有限次的查询情况、部分信息情况以及仅有标签的攻击情况，针对有限次的查询情况则采用自然进化策略进行对抗扰动生成. 部分信息情况则仅可获取前 K 个类别概率和标签，且 K 个类别概率之和并非一定为 1，文献中针对该问题分为两类情况进行处理，在目标类别标签在前 K 个标签中，按照自然进化策略与投影梯度下降算法(Projected Gradient Descent，PGD)进行搜索，若不在前 K 个标签范围内，则取消此次更新，将学习率衰减后进行更新，以确保更新幅度较小，避免目标类别标签超出范围.

除上述两类元启发式算法外，在无梯度优化策略中，遗传算法由于搜索过程较为简单，且具有与问题领域无关的快速搜索能力，因此应用较为广泛. 文献[36]提出一种基于遗传算法的对抗样本生成算法，并且文献中考虑到评估各个候选解所需大量的查询和时间成本，因此候选解的群体规模设置较小，并且以小分辨率生成噪声掩模，之后采用双线性插值操作将噪声覆盖于整张图像，因此单一噪声点对应于图像中局部区域内的噪声情况. 同时，为了降低遗传算法对超参数的敏感性，如突变率、种群大小和突变范围等. 文献中采用一种退火策略，即当算法在多轮连续迭代下并未对搜索结果进行优化或优化程度较小，则算法参数(即突变率 ρ 和突变范围 α)逐渐减小. 文献[37]则同样基于遗传算法来生成对抗样本，不同的是在优化过程中，引入代理模型的信息以降低查询的成本. 文献[38]则基于一种遗传算法变体，即微生物遗传算法(Microbial Genetic Algorithm，MGA). 相比于其他元启发式算法，该文献将基于替代模型的迁移性攻击结合到算法中，利用迁移性攻击的特点生成一部分对抗样本，以这些对抗样本作为 MGA 算法的初始种群，减少了遗传算法在迭代更新过程中的查询次数.

此外，部分研究人员受启发于差分进化算法，其思想来源于遗传算法，将差分思想引入遗传算法的关键环节中. 文献[39]中提出一种基于差分进化算法的单像素黑盒攻击，相比于之前的黑盒攻击，该文献仅修改单一像素，因此在攻击隐匿性方面较强，修改距离较小. 文献[39]中将单一像素的攻击抽象为三维空间下可行解的搜索任务，三通道像素值采用差分进化进行更新，最终获取对抗样本，同时在实验中，尝试将该方法扩展到 3 个像素与 5 个像素以加快对抗样本搜索速度，该算法生成的对抗样本如图 8 所示，样本下方第一行表示原始图像类别及置信度，

图 8 单像素攻击所生成对抗样本[39]

第二行表示扰动后的图像类别及置信度，图像中圆圈标记的位置处表示扰动像素点的位置，仅通过单一像素导致模型识别有误.

文献[40]中提出一种划痕生成算法，用于在目标图像上叠加，以使神经网络做出误判，攻击流程如图 9 所示.

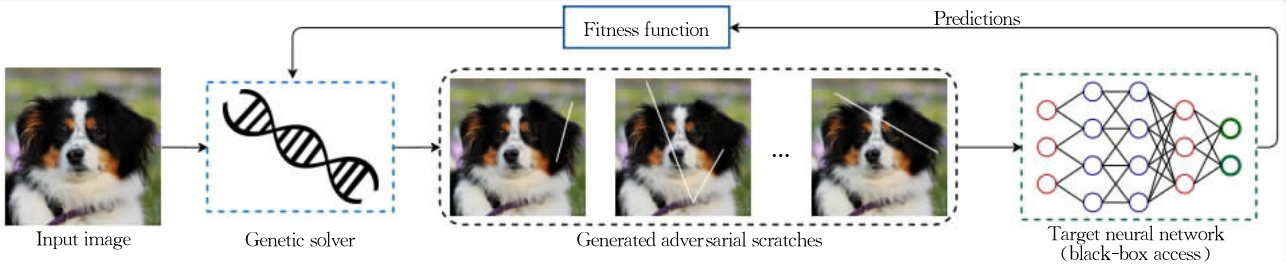

图 9　划痕黑盒攻击[40]

由图 9 可知，文献[40]利用模型查询结果，结合元启发式算法以生成不同数量、形状、长度，与角度的划痕. 该算法主要结合了差分进化算法(Differential Evolution，DE)与协方差矩阵自适应进化策略(Covariance-Matrix Adaptation Evolution Strategies，CMA-ES). 相比于之前的攻击算法，该文献主要对扰动形状进行约束，对扰动形状以及修改像素值大小进行建模，并采用进化算法进行优化. 文献[41]提出一种利用差分进化的近似梯度符号方法来解决黑盒攻击问题. 相比于之前算法，该文献并非通过差分进化来搜索扰动值，而是仅搜索梯度符号. 同时基于与原始图像相近的图像来降维，以提高进化算法搜索的效率. 文献[42]认为图像空间维度较高，难以直接搜索扰动信息，因此采用稀疏攻击策略. 为有效搜索扰动，该文献基于差分进化策略，并与均匀随机比例因子相结合，有效解决了高维优化问题. 文献[43]考虑图像属性对于对抗攻击的作用，包括图像亮度、对比度、锐度与色度等信息，并通过采用差分进化算法来获得对抗扰动. 文献[44]设计了一种多因素度量指标(*MulFactorLoss*)用于衡量原始样本与对抗样本间感知损失. 为减少查询次数，该文献基于差分进化思想与贪婪近似以搜索对抗扰动.

部分研究人员认为黑盒攻击属于多目标优化问题，文献[45]中将对抗攻击转为双目标优化问题，并提出多目标优化算法 MOEA-APGA，用于寻找最优扰动集，并选择满足滤波策略的扰动来实施最终攻击，并在像素级和全局级给出四类攻击扰动策略，其中，双目标中之一是针对分类概率，若是无目标攻击，则最小化当前类别的概率，若是目标攻击，则最大化目标类别概率；另一目标是针对像素修改的个数与距离，主要体现在条件设置中，若为像素级扰动，则采用 L_0 距离，保证修改像素点最少；若为全局级扰动，则采用 L_2 距离，以保证全局扰动时各个像素修改距离较小. 文献[46]提出一种基于像素跳跃与进化算法的注意力黑箱攻击算法. 利用注意力热图获取图像突出区域，通过跳跃像素以降低黑盒攻击的维度. 同时利用多目标进化算法以生成较优的对抗扰动. 文献[47]提出一种基于 L_0 约束的双目标优化对抗攻击算法，通过引入 L_0 约束将对抗扰动约束在少数像素点内，从而提升对抗扰动的不可感知性.

文献[48]中主要采用协方差矩阵自适应进化策略进行搜索，相比于传统的简单进化策略，CMA-ES 策略可基于之前一代的搜索结果，通过逐步增加搜索空间的标准差以探索更多可行解. 在迭代计算中，CMA-ES 为下一代采样提供多变量、正态分布参数. 由于 CMA-ES 可以利用最优解的信息同时调整其均值和方差，因此当最优解距离较远时，它可以扩大搜索范围，或者当最优解非常接近时，可以缩小搜索范围. 同时在初始解构建阶段，文献[48]并未直接搜索目标图像，而是同时输入原始图像与目标图像，在两类图像之间找寻中间图像，与直接从目标图像开始相比，从中间图像开始可以节省大量开销. 具体而言，采用投影法将目标图像投影在原始图像上，随着投影速率的增加，在中间图像周围找到有效样本的概率将更高. 文献[49]同样基于 CMA-ES 算法，提出一种子空间激活进化策略(Subspace Activation Evolution Strategy，SA-ES)，利用 CMA-ES 算法在子空间搜索扰动. 文献[50]中评估三类进化算法在生成对抗样本任务上的性能，其中基于 CMA-ES 算法的对抗攻击，在 ImageNet 数据集上的攻击成功率与平均查询次数方面较优. 文献[51]中将 CMA-ES 算法与代理模型的类间激活热力图信息相结合，以减少扰动像素个数，从而增强优化后对抗样本的不可感知性.

元启发式算法通过多次迭代来获得当前任务的可行解，目前已广泛应用于解决全局优化问题. 对于黑盒攻击而言，由于无法获得模型具体的参数与结构信息，导致梯度信息难以获取，因此一般的白盒攻

击方式难以适用于黑盒攻击场景.而元启发式算法在优化过程中无需梯度信息和模型的内部信息,因此较适合于黑盒场景下的对抗样本生成任务.通过分析可知,虽然元启发式算法对于黑盒优化问题具有良好的性能,但由于深度神经网络模型本身的复杂性,迭代过程中容易陷入局部最优,或者难以收敛,因此在设计初始种群、适应度函数以及更新策略时需要结合实际任务需求进行优化.

5.2 基于代理模型的黑盒攻击技术

代理模型(Surrogate models)或代理优化指在分析和优化设计过程中可替代复杂模型的近似学习模型.在实际优化问题中主要在以下两种情况下常用代理模型进行优化:其一是优化问题的解空间较为庞大且复杂,计算速度缓慢,难以采用直接优化的方式求得满意解.此时如果能够将问题或模型近似为相对较易优化的代理问题或模型,则可利用代理模型求解问题;其二是不能获取实际任务中模型的具体信息或显式表达式,因此无法参数化模型,且数据的获取成本较高,因此利用近似模型来获得相关数据并进行优化.以式(1)为例,$f(\cdot)$表示复杂模型,由于现实任务中难以获得 $f(\cdot)$的具体参数和信息,若能构建近似模型 $\tilde{f}(\cdot)$,则创建一个与原问题类似的近似问题进行优化.

在图像分析领域中,基于代理模型优化的方法具有相似的特点,即生成对抗样本的过程中,无法获取目标模型完整的内部信息包括梯度、参数与结构等.在获取数据方面,虽然实际任务中可以通过迭代查询获取数据,但目前相关网站已部署相应的流量监控软硬件,当监控到大量相似图像查询时会进行相关惩罚以阻止访问,因此需要尽可能减少查询次数.目前的基于代理模型的相关算法主要分为两类:一部分文献聚焦于寻找目标模型的代理模型,尝试在代理模型中生成迁移性较高的对抗样本进而干扰目标模型的预测.现有的白盒攻击算法由于可以获得模型全部信息,因此攻击性能较高.如果能够替换原有的黑盒模型,则可利用白盒攻击算法生成对抗样本;另一部分文献聚焦于改变整体优化问题,即转换为新目标函数并对其进行优化,此时新目标函数称为黑盒对抗攻击目标函数的代理模型.因此,对抗样本的生成质量取决于新目标函数与原问题的近似程度.

据此,本节从两类基于代理模型的黑盒攻击技术切入,并以不同算法改进的策略与研究目标为方向,对目前基于代理模型的黑盒攻击技术进行分析和总结.研究情况如表 3 所示.

表 3 图像分类任务中基于代理模型的黑盒攻击技术研究情况

代理模型思想	研究目标	文献	改进策略	年份
基于目标模型的代理模型	提高对抗样本的迁移性	[52]	利用对抗转换网络(ATN)并加入鲁棒性模块	2018
		[55]	采用一系列 FGSM 改进算法生成对抗样本	2020
		[56]	将图像与目标类别编码在一个相关空间中	2021
		[57]	基于元学习思想以学习多个模型梯度信息	2021
		[58]	关注 Top-K 攻击方面并提出一种归一化交叉熵损失函数	2022
		[59]	将注意力机制引入对抗攻击中以增强对抗样本的迁移性	2022
		[60]	提出一种循环优化算法用于对抗攻击	2022
		[61]	基于注意力热图思想提出一种敏感区域感知的黑盒对抗攻击	2023
	提高攻击成功率	[62]	基于生成式对抗网络,将攻击损失与可见性损失相结合	2018
		[67]	基于代理模型攻击与基于梯度攻击相结合	2021
	提高原始模型与代理模型的相似性	[62]	基于动态知识蒸馏	2018
		[63]	基于条件对抗分布	2022
		[55]	利用知识蒸馏来获取代理模型	2020
		[64]	基于低维空间探索一致性与敏感性引导相结合	2021
		[65]	基于动态神经网络思想	2022
	代理模型拟合目标模型的决策边界	[66]	基于梯度的线性数据集增强方法	2019
	实现补丁攻击	[70]	利用 VGG 特征激活格拉姆矩阵构建纹理字典	2020
		[71]	基于强化学习以优化补丁的位置以及超参数	2022
	在无法获取目标模型训练数据的情况下攻击	[68]	利用代理模型 Logit 层输出以最大化原始样本与对抗样本间散度	2020
		[69]	基于多任务生成模型,利用多个子网络生成目标样本	2022
	提高查询效率	[72]	利用代理模型训练元对抗扰动 并基于梯度估计方法进行攻击	2022
		[73]	基于元学习思想与代理模型信息以减少查询次数	2022
		[74]	构建模拟器模型以作为代理模型,并利用模拟器的特征层信息以优化对抗扰动	2022
	从数据与训练框架角度改进训练代理模型流程	[10]	引入特定于代理模型的损失函数,并提出一种三玩家框架以用于训练代理模型	2022
		[75]	提出一种替代元学习,利用元学习辅助代理模型的训练过程,以学习目标模型的输出	2022

（续　表）

代理模型思想	研究目标	文献	改进策略	年份
基于问题转化的代理模型	提高查询效率	[76]	基于贝叶斯优化算法并仅对局部区域生成扰动	2019
		[77]	基于贝叶斯优化算法，同时利用加性 GP 模型减小扰动生成的子空间复杂性	2020
		[78]	将连续约束替代为离散约束	2019
		[79]	基于插值方案的黑盒攻击，将对抗攻击视为 L_2 球体上的约束优化问题	2022
		[80]	基于贝叶斯优化算法，并利用生成器将图像映射到低维空间以减少查询次数	2022
	增强扰动隐蔽性	[81]	将扰动分为扰动大小与扰动方向，即转化为混合整数规划问题	2020
		[82]	基于深度强化学习并利用奖励反馈机制，以生成隐蔽性更高的对抗样本	2022

5.2.1　基于目标模型的代理模型

基于目标模型意味着，攻击者需要构造与原始目标模型类似的代理模型，以期望由攻击代理模型生成的对抗样本，可迁移至对目标模型的攻击中. 但在黑盒攻击任务中，代理模型与目标模型的性能往往差异较大，尤其在目标模型与代理模型的网络结构不同时差异更为明显，此时以代理模型生成的对抗样本往往性能较差，这种差异被称为代理偏差. 而不同的攻击技术也会导致对抗样本的迁移性不同，迁移性较高的对抗样本则更有可能误导目标模型的推理. 因此，提高对抗样本的迁移性，以及原始目标模型与代理模型的相似性，是研究基于代理模型黑盒攻击的重点内容.

对于提高对抗样本迁移性，各算法的策略不同. 文献[52]针对对抗转换网络(Adversarial Transformation Networks，ATN)[53]提出一种改进方式. 白盒对抗攻击与部分黑盒攻击专注于利用图像像素的梯度或直接求解图像像素的优化问题，原始的 ATN 算法则关注利用网络将原始样本转换为对抗样本. 而该文献在此基础上为了提升对抗样本的迁移性以针对黑盒攻击，在网络中加入了鲁棒性增强模块. 在实验中采用 InceptionV3[54] 模型作为代理模型，基于 InceptionV3 生成的对抗样本用以攻击其他模型结构. 文献[55]为提升对抗样本迁移性，采用 FGSM 的一系列改进算法进行白盒攻击. 通过集成不同攻击算法的特点与优势来提升样本迁移性. 文献[56]认为现有的黑盒攻击算法多数仅针对单一模型，如何同时针对多个模型进行攻击具有挑战性. 该文献通过构造编解码模型，将输入图像与目标类别编码在一个相关空间中，尝试让图像与类别的语义信息进行融合生成对应类别的对抗样本. 文献[57]受启发于元学习思想，提出一种元梯度对抗攻击(Meta Gradient Adversarial Attack，MGAA). 该文献集成多个模型的梯度信息，以提升对抗样本的迁移性从而满足黑盒攻击场景的需要. 文献[58]关注 Top-K 攻击方面，通过实验表明攻击强度更高的攻击算法在迁移性方面较好. 同时提出了一种新的标准化 CE 损失以提升攻击成功率. 文献[59]将注意力机制引入对抗攻击中，以破坏图像中注意力最为显著的特征. 由于注意力热图在不同模型之间可能存在显著差异，因此提出了一种转换不变聚集攻击策略来缓解对代理模型注意力的过拟合. 文献[60]中提出一种循环优化算法以生成对抗样本，通过结合之前累积的速度，以循环方式初始化对抗样本的动量以提高黑盒攻击的迁移性. 文献[61]提出一种针对敏感区域感知的黑盒对抗攻击，通过寻找图像中关键像素位置并进行扰动来生成对抗样本，为有效定位关键像素，该文献基于代理模型与注意力热图技术从而生成迁移性较高的对抗样本.

关于如何提高原始目标模型与代理模型间的相似性，文献[62]采用生成式模型，即生成式对抗网络(Generative Adversarial Networks，GANs)，基于 GANs 提出一种对抗样本生成模型(AdvGAN). 通过采用动态知识蒸馏的方式不断迭代优化，以提升模型间的相似性. 同时对于对抗扰动隐蔽性问题，该算法则利用生成式损失函数，以在优化过程中使对抗样本与原始样本间的距离较小. 文献[63]提出一种对抗可转移机制，其基本思想是利用条件对抗分布(Conditional Adversarial Distribution，CAD)，即以原始样本为条件的对抗扰动的分布. 整体信息的迁移往往由于代理偏差导致效果极差，因此该文献仅迁移 CAD 中的部分参数，其余参数则利用查询目标模型反馈进行学习. 文献[55]引入迁移学习思想，利用知识蒸馏生成黑盒模型的代理模型. 文献[64]认为现有的代理模型方法主要在高维空间中进行搜索和细化对抗扰动，因此需要大量的查询以确

保代理模型与目标模型近似. 该文献则提出一种低维空间探索一致性与敏感性引导相结合的集成攻击方法,在方法中集成多个多样化网络结构的代理模型,通过学习线性组合以近似目标模型. 文献[65]指出之前的方法多数将代理模型的结构固定于一种类型,如 ResNet 或者 Inception. 虽然利用白盒攻击可以获得具备一定迁移性的对抗样本,但由于代理模型与目标模型间的结构与参数存在差异,因此对抗样本迁移性并不高. 该文献受启发于动态神经网络(Dynamic Neural Networks)的思想,提出一种动态代理训练攻击算法以鼓励代理模型与目标模型近似. 根据不同的目标模型和任务,通过动态门自适应生成最优替代模型结构. 在此基础上,为提高数据质量以进一步提升代理模型与目标模型的相似性,该文献引入了基于任务驱动图的结构信息约束.

上述文献采用不同的损失函数以减小代理模型与目标模型间的差异,但由于无法获取模型内部信息,因此需要大量的查询样本供代理模型训练,部分文献认为模型的预测主要依赖于模型的决策边界,因此如果能够探明目标模型的决策边界,则可减少查询次数. 文献[66]针对该问题提出一种数据集扩充方法,即基于梯度的线性数据集增强方法(Gradient-Based Linear Dataset Augmentation Method). 通过改进数据集增强方式,使代理模型训练过程中更好地拟合目标模型的决策边界. 传统的基于雅克比矩阵的数据扩充方法往往随着 epoch 的增长,查询次数呈现指数级增长,而该算法则采用线性模式以尽可能减少查询次数. 此外,为了增强对抗样本的迁移性,该文献将集合攻击(Ensemble Attack)的思想集成在算法中以提升性能.

部分研究关注于优化代理模型的攻击方式以提高攻击成功率. 文献[62]黑盒攻击的目标函数的优化问题转换为不同模型结构的训练问题,通过引入带有代理模型信息的损失函数,以使生成式模型能够有效生成对抗扰动. 文献[67]中提出一种混合攻击算法,将基于梯度估计的对抗攻击算法与基于代理模型的对抗攻击算法相结合. 传统的基于代理模型的算法依赖于代理模型的近似程度,因此攻击效率较低,而基于零阶优化的算法由于采用梯度信息,因此生成对抗样本效率较高,但查询次数庞大. 因此该文献首先基于代理模型生成一部分对抗样本,将对抗样本混合之前的查询样本输入到梯度估计算法中,迭代生成下一批对抗样本.

上述文献中针对黑盒攻击场景提出不同的算法,但多数文献隐含假设可以获得目标模型的数据集,即在相同数据集下进行攻击. 文献[68]针对无法获取目标模型训练数据情况下进行黑盒攻击,利用预训练模型的 *Logit* 层输出,最大化原始样本与对抗样本间的散度. 文献[69]设计一个多任务生成模型来学习原始数据集的分布. 将分布信号通过反卷积操作转化为原始数据的共享特征,利用多个子网络以生成相应目标样本. 之后利用该数据集训练代理模型,采用白盒攻击算法生成对抗样本.

文献[70]受启发于白盒攻击中的补丁攻击,提出一种基于补丁的黑盒攻击. 传统补丁攻击主要采用白盒攻击方式生成补丁的面积大小、纹理与补丁位置. 该文献则利用对 VGG 主干网络特征激活格拉姆矩阵(Gram Matrices),并进行聚类来构建类相关的纹理字典,并以此字典参数化补丁的纹理. 文献[71]同样关注补丁攻击,该文献基于强化学习,同时优化补丁的位置以及补丁的超参数.

在提升查询效率方面,文献[72]利用了元学习思想. 首先在代理模型上训练元对抗扰动(Meta Adversarial perturbations, MAPs),并以此作为扰动初始值,之后通过对模型梯度进行估计来执行黑盒攻击. 由于算法加入了初始扰动,因此查询效率更高. 文献[73]同样基于元学习思想,通过训练元生成器以攻击原始样本. 同时,为减少查询成本,该文献利用代理模型来训练元生成器,之后利用对抗样本的迁移性进行攻击. 文献[74]则利用元学习思想构建模拟器模型,以作为黑盒攻击的代理模型,通过引入模拟器中的特征层信息以进一步优化对抗扰动.

部分文献从数据与训练框架角度改进代理模型的训练流程. 文献[10]主要通过引入特定损失函数以扩大类间相似性,同时为生成器引入与代理模型相关的损失函数,以增强类内相似性. 并提出一种三玩家框架以用于训练代理模型. 文献[75]则引入元学习思想,提出一种替代元学习(Substitude Meta-learning, SML). 将对抗扰动生成过程分为 SML 学习与 SML 训练两个阶段,利用元学习来辅助代理模型的训练.

5.2.2 基于问题转化的代理模型

上小节中主要叙述基于目标模型的代理模型,其关键在于如何构建与目标模型近似的代理模型,并通过查询目标模型的输入输出构建训练数据集,之后训练代理模型并利用白盒攻击算法来生成对抗

样本用以攻击目标模型. 而一部分文献则聚焦于目标函数的转化，现有的黑盒攻击往往定义目标函数，之后的训练过程则是迭代优化过程，因此黑盒攻击本质上属于优化问题. 部分文献尝试将现有复杂的目标函数优化问题转化为另一类型的目标函数优化问题，以期望可以提高查询效率，或者提高扰动对人眼视觉感知的不可感知性.

在提高查询效率方面，文献[76]利用贝叶斯优化来生成对抗样本. 传统基于查询的对抗攻击算法效率低下的部分原因在于查询过程随机性较高，而贝叶斯优化则可根据当前查询情况评估下一次查询位置，因此减少了查询次数，该算法也为之后的黑盒攻击算法提供了新的思想. 文献[77]利用高斯过程代理模型的贝叶斯优化来寻找有效的对抗样本. 同时利用统计代理以及查询数据信息，有效地优化了对抗扰动与潜在的搜索空间维度. 文献[78]将黑盒攻击算法目标函数优化为离散问题，即原有的黑盒攻击约束为式(5).

$$\|x_{adv}-x\|_{\infty}\leq\varepsilon \tag{5}$$

式(5)表示当前生成的对抗样本与原始样本间的 L_{∞} 范数小于扰动最大干扰值 ε，以保证当前生成的对抗扰动对于人眼是不可见的. 而文献[78]通过实验发现，现有的投影梯度下降攻击算法在 ε 为 8 时，攻击成功的攻击扰动值集中分布于 -8 和 8 两侧，因此文献将连续约束替换为离散约束即式(6).

$$x_{adv}-x\in\{\varepsilon,-\varepsilon\}^{p} \tag{6}$$

其中 p 代表图像的像素个数，由此攻击算法只需在各个像素点之间选择取离散的扰动值即可，之后利用不同的搜索算法即可进行迭代求解. 文献[79]提出一种基于插值方案的黑盒攻击策略. 将原先对抗攻击问题转化为 L_2 球体上的约束优化问题，并沿着球体上的测地曲线采样. 在生成扰动过程中，除查询目标模型的输出外，算法可利用当前状态周围信息对扰动损失进行二次近似，因此提高了查询效率. 文献[80]则利用生成器将图像映射到低维空间，之后基于贝叶斯优化思想以生成对抗样本.

在增强扰动隐蔽性方面，文献[81]关注稀疏对抗攻击，原有的黑盒攻击算法往往在迭代过程中不断优化扰动的方向和大小，而该文献则将扰动方向与大小进行分解，方向采用二进制选择因子(即 0 或 1)，即如果为 0 则表明当前像素不扰动，因子为 1 则进行扰动，并通过 Hadamard 乘积结合为最终扰动. 基于此分解，将稀疏攻击问题表示为混合整数规划问题(Mixed Integer Programming，MIP). 文献[82]则将生成对抗扰动问题转化为深度强化学习(Deep Reinforcement Learning，DRL)问题，根据目标代理的奖励反馈搜索最佳扰动. 在优化过程中利用语义信息，而非直接利用像素空间信息，在保证攻击成功率的同时提升对抗样本的隐蔽性.

基于代理模型的黑盒攻击算法依赖于代理模型的构建，根据代理模型的类型将该类型算法分为两类，即基于目标模型或基于问题转化. 前者方法通过构建与目标模型相似的代理模型进行攻击，以期望根据代理模型生成的对抗样本可成功迁移到目标模型中，因此模型的构建，数据的构建以及训练方式与最终代理模型的相似性息息相关. 之后采用白盒攻击策略攻击代理模型以生成迁移性较高的对抗样本；后者则将传统的黑盒攻击目标函数进行转化以从其他角度进行优化，包括贝叶斯优化、离散优化以及与混合整数优化等. 基于代理模型同样不需要梯度信息，因此较适合于黑盒攻击场景，但往往需要大量的查询次数，以及模型结构的精准选择和调试，因此如何在有限查询情况下提升模型的相似性与对抗样本的迁移性成为当前方法的重点问题.

5.3　基于直接搜索的黑盒攻击技术

直接搜索方法是现有黑盒优化问题中常用的方法之一，通过检查点的集合来确定候选点，之后不断对当前候选点更新迭代优化，组成新的试验点集合. 在更新优化过程中，需要不断计算候选点对应的函数值，以选择当前较优的点进行更新，在整体优化过程中不需要任何导数逼近. 在针对图像的黑盒攻击任务中，其搜索的候选解即图像的对抗样本，通过改变图像的各个像素值以获得对抗样本，因此搜索区域的选择对于候选解的生成和筛选至关重要. 依据现有算法的搜索域特点，本文将搜索域分为两类，即基于像素域的搜索策略与基于特征域的搜索策略. 其中，基于像素域的搜索策略主要在图像的像素值范围内修改，通过加减扰动值来生成对抗样本. 基于特征域的搜索策略认为图像中包含有大量的冗余信息以及噪声信息，这些信息对于图像任务并不是关键信息，因此部分文献尝试将图像映射到特征域或子空间，在映射域进行搜索以生成对抗样本.

据此，本节以不同搜索域以及各黑盒攻击算法的搜索策略为主要总结内容，表 4 中展示了现有基于直接搜索的黑盒攻击相关文献内容.

表 4 图像分类任务中基于直接搜索的黑盒攻击技术研究情况

搜索位置	研究目标	文献	搜索策略	年份
像素域	引入决策边界信息以提高样本攻击成功率	[94]	基于预定义的正交基以生成对抗样本	2019
		[95]	以决策边界超平面的法向量为指导来生成对抗样本	2020
		[96]	提出一种几何方法,以决策边界的几何特性为指导	2021
		[97]	通过估计与邻近决策边界的距离来确定更新扰动的步长	2023
		[98]	采用粗粒度以及细粒度随机搜索两种策略,以将样本移出决策边界外	2021
		[30]	采用贪婪策略对生成对抗样本涉及的旋转角度以及平移方向进行优化,以向决策边界移动	2022
		[99]	注意力热图与低频噪声信息结合,利用二分逼近至决策边界,通过邻域几何探测来缩短样本距离	2023
	以图像块为攻击单位来减少查询次数	[83]	以图像块为更新扰动的单位,并配合有限差分思想生成对抗样本	2020
		[84]	随机搜索与图像块攻击相结合,并在迭代中逐步缩减图像块边长	2020
		[85]	利用显著性目标检测获取突出区域并利用图像块来寻找最易扰动位置	2023
		[86]	基于随机搜索策略将图像块中的像素重新排列	2022
		[87]	基于差分进化思想进行补丁攻击	2023
	以图像像素点为扰动目标来提升扰动的不可见性	[88]	从最大化扰动值开始迭代以逐步减小扰动值,并随机改变扰动符号	2020
		[89]	提出一种随机搜索通用框架,以实现稀疏对抗攻击	2022
		[90]	采用基于搜索策略的攻击方式对图像离散整数域进行优化	2022
		[91]	黑盒攻击转化为强化学习任务,并预测成功率更高的扰动方向	2021
		[92]	在像素域内利用强化学习寻找对抗扰动的分布以生成对抗样本	2022
		[93]	取代理模型中间层信息作为先验信息,引入模型的雅克比矩阵来确定扰动方向	2022
特征域	将图像域做映射以减少查询次数	[100]	提出一种投影与概率驱动的黑盒攻击,并在图像低频区域生成对抗样本	2020
		[101]	合成对抗斑块并与原始图像低频部分融合构成对抗样本	2022
		[102]	利用神经过程对图像建模,并在小区域范围内生成对抗样本	2023
		[103]	在预定义的子空间中生成对抗扰动	2021
		[104]	定义两类子空间,并通过贝叶斯优化算法生成对抗样本	2021
		[105]	构建潜在空间并在潜在空间中优化对抗样本	2022
		[106]	构建颜色分布并利用该分布搜索图像扰动	2022

5.3.1 基于像素域的搜索策略

现有基于像素域的直接搜索算法中,部分研究人员关注于图像像素的搜索范围,即基于图像像素点进行搜索或者基于图像像素块进行搜索.

对于图像块攻击来说,文献[83]将图像进行分块,以图像块为单位进行搜索,并配合有限差分以及离散近似优化方法得到对抗样本.文献[84]采用基于图像块的随机搜索算法,在迭代优化中,方形区域的边长根据迭代次数的递增逐步缩减,类似于基于梯度优化中的步长缩减.文献[85]利用显著性目标检测模型以获得图像中目标区域.并通过以图像块为迭代单位,逐步缩小块的范围以获取最易扰动图像信息的图像块.文献[86]基于随机搜索策略,采用多种映射方法,对图像块区域内像素进行重新排列以生成对抗样本.文献[87]提出一种基于差分进化思想的补丁对抗攻击,该文献为简化解空间,利用配对算法匹配图像关键点与补丁位置,并使用目标图像作为补丁的初始化图像.

对于像素点攻击而言,文献[88]表明在对抗扰动中随机改变少量的符号可以显著提高攻击性能.首先初始化最大化的扰动值,之后通过逐步迭代减小扰动值,并在此过程中随机改变扰动的符号,以此提高攻击性能.文献[89]提出一种基于随机搜索的通用框架,即 Sparse-RS.用于黑盒环境下稀疏目标攻击与无目标攻击,扰动通过随机采样,并判断当前扰动是否有助于优化目标函数,若未变化则保持原有样本不变并采样下一次扰动值.文献[90]则认为现有攻击算法将像素值转为 0 到 1 的连续值,而图像像素是离散值.因此攻击产生的扰动由于保存为图像而失效.该文献则直接在离散整数域优化扰动,并基于搜索策略来生成对抗样本.文献[91]将黑盒攻击任务转化为强化学习任务,并探索一种基于模型的方法以预测攻击成功率更高的扰动搜索方向.文献[92]同样将强化学习思想引入黑盒攻击任务中,利用强化学习中的奖励机制引导扰动生成.文献[93]取代理模型的中间层作为先验信息,并求取其雅可比矩阵,利用矩阵右奇异向量来确定扰动的最佳方向.

另一部分研究将模型决策边界信息引入到攻击算法中.文献[94]采用较为简单的方式,即在搜索策略上采用正交方式,从预定义的正交基中随机抽取一个向量,之后将其作为扰动在图像像素域加入或减去,以此生成对抗样本.文献[95]认为深层网络的决策边界在数据样本附近通常具有较小的平均曲率如图 10 所示.

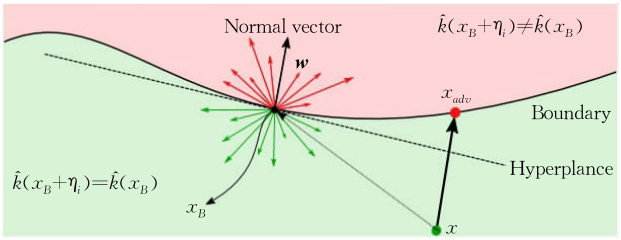

图 10　决策边界线性化[95]

图 10 中样本 x 向超平面靠近逐步迭代生成对抗样本，并以超平面的法向量 w 为指导，该算法实现对抗攻击的关键步骤在于超平面法向量的计算，文献[95]中采用多元高斯分布 $\eta_i \sim N(0,\Sigma)$，通过多次查询目标模型以估计超平面法向量. 文献[96]提出一种几何方法(SurFree)以大幅减少黑盒查询次数，在迭代中以分类器的决策边界的几何特性的精确指示为指导. 文献[97]提出一种查询较为高效的黑盒攻击算法，通过估计与邻近决策边界的距离来确定扰动更新的步长，同时利用双向轨迹来寻找中间对抗实例. 文献[98]基于直接随机搜索策略以生成对抗样本，首先通过粗粒度随机搜索以在决策边界附近定位样本，之后将样本移出决策边界外以生成对抗样本，最后通过细粒度随机搜索以优化当前对抗样本. 文献[30]中通过旋转与平移原始图像样本以生成对抗样本，采用贪婪策略以更新旋转的角度以及平移的方向，逐步将样本向决策边界移动. 文献[99]中将注意力热图作为掩膜，与通过随机采样得到的低频噪声信息相结合，通过二分逼近找寻对抗点. 最终在决策边界附近进行邻域几何探测来缩短样本距离.

5.3.2　基于特征域的搜索策略

搜索策略的性能与搜索域的空间大小息息相关，因此若图像分辨率较高，则基于像素域的攻击方法较为复杂. 部分研究基于特征域的搜索策略，重点关注如何将图像进行映射，以减少冗余信息，以提高攻击效率. 文献[100]提出一种投影与概率驱动的黑盒攻击(Projection & Probability-driven Black-box Attack). 该文献认为对抗扰动主要存在于图像的低频区域，因此从两方面进行优化，首先构造低频约束传感矩阵的简单方法用以降低维数，其次采用概率驱动策略执行随机游走减少查询次数. 文献[101]认为图像的高频部分是 DNN 能够正确识别与检测的关键信息. 利用几何图案为原型以合成对抗斑块，之后将其高频部分与原始图像的低频部分结合，通过投影得到对抗混合样本. 文献[102]认为多数黑盒攻击技术仅关注像素域的攻击，往往独立关注各个像素点的情况，而不考虑关于整幅图像的像素值结构以及其他信息，因此该文献首先利用神经过程对图像的信息进行建模，之后在原始样本的小区域范围内对对抗样本进行建模，以此分布提取的样本极有可能为对抗样本. 文献[103]认为像素域空间较为复杂且高维，因此尝试将任何随机攻击产生的对抗干扰限制在预定义的子空间中，并对保证存在最小对抗扰动的子空间进行了初步的理论分析，通过子空间的辅助信息加强黑盒攻击. 文献[104]定义了两类子空间，即 L_2 下的低维子空间以及 L_∞ 下的低分辨率子空间. 并在子空间中采用贝叶斯优化算法生成对抗样本. 文献[105]基于 StyleGAN 生成器来构建潜在空间，在潜在空间中优化对抗样本，以使对抗样本与原始样本相似性更高. 文献[106]提出一种自然颜色攻击方法，尝试在颜色空间搜索图像扰动. 该方法主要为图像公开数据集构建颜色分布，之后通过图像分割算法识别物体类别并与得到的颜色对应，最后通过映射颜色空间与优化对抗损失得到扰动样本.

基于直接搜索的黑盒攻击其优势在于整体优化过程中不需要任何导数逼近策略，且利用分布采样容易得到初始解集. 本文从搜索域的角度将算法分为两类，即基于像素域和基于特征域的攻击算法，总结如图 11 所示.

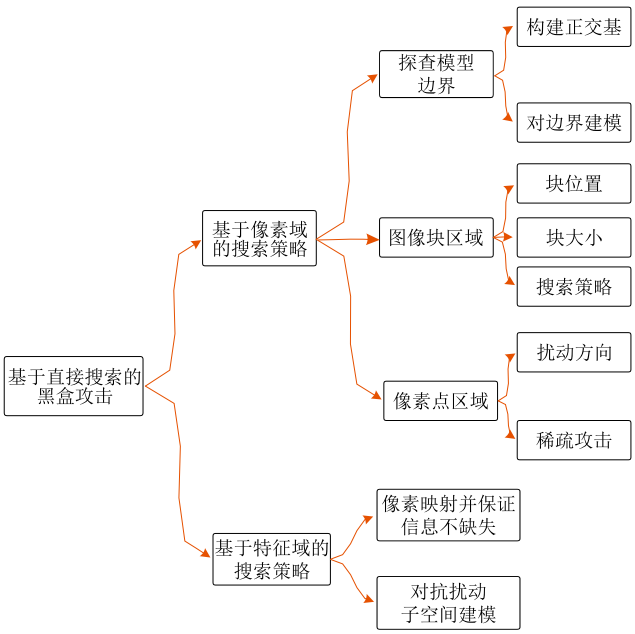

图 11　基于直接搜索的黑盒攻击

两种策略在搜索策略上各有优劣，其中，像素域的搜索范围固定在像素值内，因此可直接生成对抗样本，但由于图像分辨率差异，图像分辨率的增加会导致搜索区域激增，难度增大，因此如何提升算法的

攻击效率是重中之重. 而特征域的搜索由于将图像提前映射到子空间,因此搜索范围相对较小,搜索效率加快,但映射图像本质是对图像信息的压缩,因此如何在图像像素映射的同时保证关键信息的完整,是之后的对抗样本生成的关键问题.

5.4 基于零阶优化的黑盒攻击技术

梯度信息是现有优化问题中最为关键的信息之一,而对于深度神经网络模型优化,梯度下降算法与反向传播更新是模型能够学习的关键. 具体而言,通过给定损失函数形式,通过求取参数梯度,利用梯度下降算法进行反向传播,以迭代方式更新模型参数. 现有的白盒攻击算法多利用模型的一阶梯度信息以生成对抗样本. 以 FGSM[107] 算法为例,如图 12 所示,原始样本类别为"熊猫",在加入由梯度信息生成的微小扰动后,样本类别变为"长臂猿".

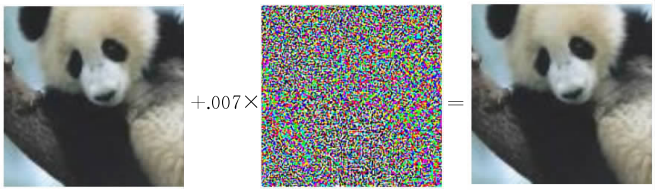

图 12 FGSM 算法生成对抗样本[107]

记 x 与 $\hat{x}$ 表示原始样本以及对抗样本,令 η 表示当前扰动,则 η 可表示为式(7).

$$\eta=\varepsilon\,\mathrm{sign}(\nabla_x J(\theta,x,y)) \tag{7}$$

因此对抗样本即为式(8)所示.

$$x=x+\eta \tag{8}$$

其中扰动由梯度与步长 ε 决定.

对于黑盒攻击而言,我们以 $f(x)$ 表示对抗攻击目标函数,可利用泰勒展开式近似为式(9).

$$f(x)\approx f(x_0)+\nabla^{\mathrm{T}} f(x_0)(x-x_0) \tag{9}$$

其中 x 表示当前对抗样本,通过迭代生成对抗样本 x_{t+1} 表示为式(10).

$$x_{t+1}=\arg\min_x f(x_t)+\nabla^{\mathrm{T}} f(x_t)(x-x_t)+\frac{1}{2a}L_\theta(x,x_t)\quad \theta\in\{0,1,2,\infty\} \tag{10}$$

式(10)中 x_t 与 x_{t+1} 分别表示当前生成样本以及下次迭代生成的样本,最后一项为惩罚项以控制不断迭代生成的样本间差异,主要通过不同的范数来衡量样本差异,以保证对抗样本与原始样本的相似性. 据此在整体优化过程中,模型的一阶梯度是迭代的关键信息,而由于黑盒攻击的定义,无法直接获得目标模型的结构与一阶梯度信息,因此难以使用与白盒攻击类似的梯度下降策略进行优化. 部分文献尝试采用零阶优化算法[14]以生成对抗样本,一阶优化与零阶优化算法如图 13 所示,其中,图 13 中 FO 表示一阶优化算法(First-order Optimization),ZO 表示零阶优化算法(Zero-order Optimization).

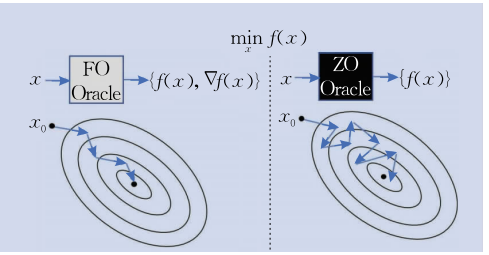

图 13 一阶优化与零阶优化算法[14]

由图 13 可知,对于统一的优化目标 $\min_x f(x)$,一阶优化算法在白盒目标函数 $f(x)$ 下利用一阶梯度信息求解优化问题,而零阶优化算法则针对于 $f(x)$ 是黑盒目标函数的优化场景,零阶优化算法围绕三个主要步骤,一是梯度估计策略,利用近似梯度替代白盒场景下的一阶梯度信息,常见方法包括单点估计以及多点零阶估计等;二是下降方向计算,得到近似梯度信息后,部分文献尝试利用白盒攻击原理进行优化,或者引入动量信息来计算下降方向;三是更新候选解,利用下降方向得到本次迭代的候选解,评估当前候选解是否满足目标函数与约束条件以更新当前最优解.

本节对现有文献中基于零阶优化的黑盒攻击技术进行总结,相关研究情况如表 5 中展示.

较为经典的采用零阶优化算法进行优化的黑盒攻击算法,可追溯到文献[33]的研究. 该文献发表前后的黑盒攻击领域主要基于代理模型进行优化,以期望生成的对抗样本可迁移到目标模型中,但往往迁移性不高. 该文献则从梯度信息方面考虑,基于有限差分思想提出一种零阶优化策略用以估计梯度信息. 在得到梯度信息后则采用相关的梯度下降策略进行优化,并不断迭代以生成对抗样本. 由于该研究中需要对各个像素进行梯度估计,单一像素需要两次查询才可用以估计梯度,因此查询数量庞大,难以应用于实际任务. 同时,沿梯度方向虽然可以较快地生成对抗样本,但也容易导致扰动隐蔽性差,从而易被识别. 因此,现有研究的主要内容在于:如何在保证攻击成功率的同时减少查询数量,以及如何提高对抗扰动的隐蔽性.

表 5　图像分类任务中基于零阶优化的黑盒攻击技术研究情况

研究目标	文献	梯度估计策略	年份
提高攻击成功率	[33]	基于有限差分思想提出一种零阶优化策略	2017
	[108]	提出一种黑盒动量迭代快速梯度符号法，利用差分进化以近似梯度	2020
	[109]	提出一种自适应近似梯度模拟算法用于黑盒攻击	2022
估计部分梯度以减少查询次数	[110]	采用基于局部搜索的技术来构造梯度的数值近似	2017
	[111]	采用估计特征组的梯度代替原有单一特征梯度	2018
优化估计策略以减少查询次数	[112]	提出一种基于自动编码器的零阶优化方法	2019
	[113]	引入基于运算符的拆分方法	2019
	[114]	将零阶梯度估计算法与二阶自然梯度下降算法相结合	2020
	[115]	仅估计一个不完美但信息较丰富的方向导数而无需准确估计梯度	2020
引入先验信息以减少查询次数	[116]	基于梯度估计与代理模型算法，并引入先验信息以减少查询次数	2022
	[117]	将元学习引入梯度估计算法中，以引入梯度先验信息	2020
	[67]	以代理模型的梯度信息作为先验信息并随机选取代理模型结构	2021
	[72]	利用代理模型训练元对抗扰动，并以此作为初始扰动	2022
	[118]	提出一种利用黑盒变分推理(BBVI)进行黑盒攻击的算法	2022
对抗扰动隐蔽性更高	[119]	提出一种基于 Frank-Wolfe 算法的对抗攻击算法	2020
	[120]	在局部图像区域生成扰动，并引入 *SSIM* 以求取局部梯度	2021

(1) 对于提高黑盒攻击的成功率方面

文献[108]受启发于白盒攻击中的快速梯度符号法(Fast Gradient Sign Method，FGSM)[107]，提出一种黑盒动量迭代快速梯度符号法(Black-box Momentum Iterative Fast Gradient Sign Method，BMI-FGSM)，通过利用差分进化近似梯度，并利用梯度信息构造对抗样本. 文献[109]基于差分进化思想提出一种自适应近似梯度黑盒攻击算法. 为提高攻击成功率，采用两步方式更新扰动图像，通过增加步长 β 对扰动图像进行微调以使其向对抗样本的边界移动.

(2) 在减少查询次数方面相关研究主要从三个方向切入：一是仅估计图像中少部分像素或特征的梯度. 文献[110]仅考虑图像的一部分像素，采用基于局部搜索的技术来构造网络梯度的数值近似，利用这些梯度信息对图像中的一组像素进行扰动. 而文献[111]同样关注查询次数过大的问题，采用估计特征组梯度的方式代替原有估计单一特征梯度的方式，而通过对特征进行分组来估计梯度，相当于通过沿着适当的方向投影来估计梯度的近似值.

二是优化梯度估计的策略. 文献[112]提出一种基于自动编码器的零阶优化方法(Autocoder based Zeroth Order Optimization Method)，采用两种方法减少迭代的查询次数，包括用于平衡查询计数和失真的自适应随机梯度估计策略，同时采用自编码器，该自编码器采用未标记数据进行离线训练或用于加速攻击的双线性调整操作. 文献[113]则引入了基于运算符的拆分方法，即交替方向乘法器(Alternating Direction Method of Multipliers，ADMM). 该方法主要利用了失真度量以及反馈设置，避免了高查询复杂度. 文献[114]针对梯度估计算法中的查询次数高，收敛速度慢等问题提出一种零阶自然梯度下降算法(Zeroth-Order Natural Gradient Descent，ZO-NGD). 该算法结合了黑盒场景下的零阶梯度估计算法以及二阶自然梯度下降算法来提高算法的查询效率. 文献[115]则认为基于梯度估计的对抗攻击算法其限制在于多次需要估计相对精确的梯度，但该文献认为不需要精确的方向导数值，获取一个不完美但信息较为丰富的方向导数即可使算法收敛，同时减少查询次数.

三是引入先验信息. 文献[116]采用一种混合算法即基于梯度估计以及代理模型算法，提出先验指导的随机无梯度算法(Prior-guided Random Gradient-free Method). 传统的梯度估计往往直接计算给定输入的梯度估计信息，因此查询效率低下. 该文献基于随机无梯度算法，将代理模型的梯度信息作为先验信息引入算法中，充分利用先验信息以提高算法的查询效率. 文献[67]同样采用代理模型的梯度信息以减少查询次数，但相比于之前采用单一模型结构，该文献在多步迭代时随机选择代理模型与代理模型参数，以提升对抗样本的迁移性. 文献[117]受启发于元学习在少样本学习问题中取得的成功，将元学习引入梯度估计算法中. 通过元攻击者可学习到各类模型梯度的先验信息，以推断出新目标模型的梯度. 文献[72]则提出一种混合攻击算法，首先利用代理模型训练元对抗扰动，并以此作为扰动初始值，之后通过梯度估计算法来执行黑盒攻击. 文献[118]基于黑盒变分推理(Black-box Variational Inference，BBVI)以估计对抗样本的分布. 为解决蒙特卡洛估计中方差大导致梯度不稳定的问题，通过重要性抽样并利用已知代理模型的梯度构造重要性分布，之后通过贝叶斯规则从估计分布中提取对抗样本.

(3) 对于提高对抗扰动隐蔽性

文献[119]认为现有的对抗攻击算法中，多数算

法尝试在扰动集边界或附近生成对抗样本，因此样本存在较大失真. 因此该文献提出一种基于 Frank-Wolfe 算法的对抗攻击算法，将对抗样本的失真约束集成在迭代算法中，以尽可能使样本像素在正常图像范围内. 文献[120]则关注对抗样本的生成质量问题. 因此在生成样本时考虑选择扰动对于图像质量影响较小但对于模型决策影响较大的区域，并引入结构相似性指数(Structural similarity index, SSIM)求取局部梯度用以判定图像区域.

现有对抗攻击相关研究中，相比于黑盒攻击，白盒攻击迭代次数较低且攻击成功率更高，这主要由于白盒攻击掌握有模型的结构信息以及具体的参数信息，直接利用参数反向求出模型梯度，在梯度方向上优化以生成对抗样本. 而黑盒攻击仅可通过查询模型预测结果来生成对抗样本，由于对于目标模型的结构与参数未知，因此难以获得梯度信息，采用零阶优化算法则可间接弥补该问题. 但零阶优化中最大的限制在于梯度估计查询数量庞大，以现有的 ImageNet 数据集为例，预处理后一张 $229\times229\times3$ 的图像利用梯度估计算法，获得各个像素的梯度信息需要百万次查询，难以满足实际任务的需要，因此在零阶优化相关研究中，如何减少查询次数以及如何高效利用查询结果尤为关键.

5.5 小 结

当前图像分类下黑盒攻击的评估工作主要从两方面开展. 一是采用公开数据集评估算法性能，包括攻击成功率与所需的时间成本. 目前应用较为广泛的数据集包括 MNIST、CIFAR-10 和 ImageNet，所涉图像类别基本覆盖常见应用场景，但实际任务中，相关图像分类模型需要在特定场景或数据集中微调，因此与公开数据集上训练的模型存在差异. 为了印证攻击算法的实用性，另一部分黑盒攻击算法的评估工作关注于实际任务中第三方视觉平台，其中包括谷歌云视觉 API(Google Cloud API)以及阿里云视觉智能开放平台(Vision Intelligent Application Programming Interface Platform)等，这些平台和服务供应商为全球图像相关应用提供识别与检测服务，因此平台上 DNN 模型的鲁棒性与相关图像分析模型的应用与推广息息相关. 本文汇总涉及文献在公开数据集上的性能结果，以此来比较各个黑盒攻击算法间的优势与不足.

表 6 给出了 MNIST 数据集中不同黑盒攻击算法的性能，表 7 给出了 CIFAR-10 数据集中不同黑盒攻击算法的性能，表 8 给出了 ImageNet 数据集

表 6 MNIST 数据集中黑盒攻击性能

文献	目标模型	非目标攻击				目标攻击			
		成功率/%	查询次数	所需时间	相似性度量/结果	成功率/%	查询次数	所需时间	相似性度量/结果
[45]	ACN[122]	100.0		0.465 min	*RMSD*/0.12	100.0		0.366 min	*RMSD*/0.17
[45]	AlexNet	100.0		0.835 min	*RMSD*/0.12	100.0		0.710 min	*RMSD*/0.15
[36]	C&W[121]					100.0	996		L_∞/0.3
[29]	C&W[121]	96.3	593	0.068 min	L_2/4.1	72.6	1882	0.238 min	L_2/4.8
[62]	C&W[121]	94.0			L_∞/0.3				
[66]	C&W[121]	96.9	4800		L_2/3.8				
[67]	⊖					90.0	<5000		L_2/3.7E-3
[65]	AlexNet	70.5				64.8			
[65]	VGG16	72.5				68.5			
[65]	ResNet18	63.7				65.8			
[10]	AlexNet	94.9				61.1			
[10]	VGG16	93.6				57.9			
[10]	ResNet18	85.7				57.1			
[77]	C&W[121]					98.0	75		$L_2/6.89\times10^{-3}$
[80]	⊖	100.0	4		L_∞/0.3	100.0	19		L_∞/0.3
[102]	⊖	100.0	1226		L_2/3.1	100.0	2605		L_2/3.9
[104]	C&W[121]	90.4	28		L_∞/0.3	26.2	130		L_∞/0.3
[33]	C&W[121]	100.0		1.38 min	L_2/1.5	98.9		1.62 min	L_2/2.0
[110]	NiN[123]	91.4		0.64 s	*PTBPIXELS*/2.24%				
[111]	Model A[124]	100.0	62720		L_∞/0.3	73.8	≤8000		L_∞/0.3
[111]	Model B[124]					73.7	≤8000		L_∞/0.3
[112]	C&W[121]					100.0	2428		$L_2/4.54\times10^{-3}$
[113]	⊖	100.0	7603		L_2/2.2				
[108]	C&W[121]	100.0		16.7 s	L_∞/0.2	98.2		23.7 s	L_∞/0.3
[119]	⊖					99.9	1133	0.1 s	L_∞/0.3
[114]	⊖	98.7	523		L_∞/0.4				
[117]	⊖	100.0	749		L_2/1.8	100.0	1299		L_2/2.7
[115]	C&W[121]	94.0	14000		L_2/1.1				

表 7　CIFAR-10 数据集中黑盒攻击性能

文献	目标模型	非目标攻击				目标攻击			
		成功率/%	查询次数	所需时间	相似性度量/结果	成功率/%	查询次数	所需时间	相似性度量/结果
[32]	C&W[121]					99.6	4910		L_∞/0.05
[40]	ResNet50					92.1	899		覆盖率[40]/7.48%
[39]	NiN[123]	71.6			L_0/1	23.2			L_0/1
[39]	VGG16	63.5			L_0/1	16.4			L_0/1
[36]	C&W[121]					96.5	804		L_∞/0.05
[38]	VGG19-BN	99.9	130		L_∞/8	95.9	1157		L_∞/8
[38]	ResNet50	95.3	551		L_∞/8	91.2	1879		L_∞/8
[38]	InceptionV3	99.7	98		L_∞/8	99.1	686		L_∞/8
[29]	C&W[121]	99.6	1224	0.139 min	L_2/1.4	71.9	6512	0.682 min	L_2/2.9
[41]	AllConv[125]	75.4	3200						
[41]	NiN[123]	59.4	5200						
[47]	AT1[47]	44.2	<1000		*SSIM*/0.93	84.4	<1000		*SSIM*/0.95
[43]	VGG16	70.6	154		L_2/—	44.9	122		L_2/—
[43]	ResNet50	83.6	117		L_2/—	48.1	111		L_2/—
[49]	C&W[121]					100.0	13933		L_2/0.7
[37]	ResNet50	93.4	29		L_2/2				
[37]	InceptionV3	94.7	37		L_2/2				
[62]	ResNet32	81.8			L_∞/8				
[63]	VGG16	99.9	56		L_∞/0.03125	98.8	861		L_∞/0.03125
[63]	DenseNet-BC-110	100.0	43		L_∞/0.03125	100.0	787		L_∞/0.03125
[55]	InceptionV1	85.6			L_∞/30				
[55]	VGG16-BN	78.1			L_∞/30				
[61]	VGG19	100.0	1087		L_2/1.38	100.0	1522		L_2/1.34
[68]	ResNet152	28.7	<10		L_∞/10				
[68]	DenseNet169	25.6	<10		L_∞/10				
[64]	ResNet-Preact-110	100.0	43		L_∞/(8/255)	97.2	860		L_∞/(8/255)
[64]	DenseNet-BC-110	100.0	11		L_∞/(8/255)	100.0	596		L_∞/(8/255)
[57]	PyramidNet-164	99.5			L_∞/8				
[67]	⊖					98.2	<5000		L_2/6.14E-4
[72]	VGG16					97.1	108		L_2/2.8
[72]	ResNet34					98.3	108		L_2/3.45
[72]	GoogleNet					99.0	148		L_2/3.52
[65]	VGG16	58.2			L_2/—	44.7			L_2/—
[10]	VGG16	77.5				41.2			
[77]	C&W[121]					87.0	154		$L_2/5.87\times10^{-4}$
[73]	DenseNet121	100.0	48		L_∞/0.031	100.0	133		L_∞/0.031
[73]	VGG19	100.0	58		L_∞/0.031	100.0	142		L_∞/0.031
[74]	PyramidNet-272	95.9	790		L_∞/—	73.8	1231		L_∞/—
[79]	InceptionV3	97.9	329		L_2/2.4	97.1	239		L_2/4
[80]	VGG19	99.4	10		L_∞/0.03125	95.5	17		L_∞/0.03125
[88]	ResNet18	95.4	409		L_∞/0.031	99.4	1807		L_∞/0.031
[102]	WideResNet	100.0	131		L_2/1.32	100.0	827		L_2/1.37
[104]	C&W[121]	70.4	76		L_∞/0.05	48.9	149		L_∞/0.1
[86]	ResNet18	100.0	119		L_0/26.8				
[93]	ResNet18	99.8	95		L_2/0.472	99.8	241		L_2/0.692
[33]	C&W[121]	100.0		3.43 min	L_2/0.1997	97.0		4.40 min	L_∞/0.5423
[110]	VGG	97.9		0.72 s	*PTBPIXELS*/2.99%				
[111]	ResNet32	100.0	61440		L_∞/8	97.0	<8000		L_∞/8
[112]	C&W[121]					100.0	1524		$L_2/1.20\times10^{-3}$
[113]	⊖	100.0	6213		L_2/0.415				
[108]	C&W[121]	100.0		20.6 s	L_∞/0.034	96.3		30.4 s	L_∞/0.047
[114]	⊖	99.2	131		L_∞/0.2				
[117]	ResNet18	94.0	1583		L_2/0.34	93.0	3667		L_2/0.77
[115]	C&W[121]	95.0	12000		L_2/0.13				
[120]	ResNet32	98.6	1593		*SSIM*/0.989				

表 8 ImageNet 数据集中黑盒攻击性能

文献	目标模型	非目标攻击				目标攻击			
		成功率/%	查询次数	所需时间	相似性度量/结果	成功率/%	查询次数	所需时间	相似性度量/结果
[34]	InceptionV3	100.0	1701		L_2/24.323				
[35]	InceptionV3					90.0	2.7×10^6		L_∞/0.001
[40]	ResNet50	73.0	21988		覆盖率[40]/1.3%	96.7	16838		覆盖率[40]/1.24%
[48]	⊖					100.0	74948		L_2/25.92
[32]	InceptionV3					99.2	24780		L_∞/0.05
[36]	InceptionV3					100.0	11081		$L_2/2.3\times10^{-4}$
[38]	InceptionV3	97.2	175		L_∞/12	96.7	7973		L_∞/12
[38]	ResNet50	99.9	100		L_∞/12	95.6	2857		L_∞/12
[41]	AlexNet	41.9	6300						
[42]	InceptionV3	98.8	620		L_0/50176				
[43]	InceptionV3	78.3	157		L_2/—				
[46]	InceptionV3	100.0	3000		L_2/3.82				
[63]	ResNet18	97.3	210		L_∞/0.05				
[63]	VGG16	99.4	77		L_∞/0.05				
[70]	ResNet50	99.7	983		*Avg_area*/3.10%	100.0	3747		*Avg_area*/15.36%
[70]	DenseNet121	99.7	1001		*Avg_area*/3.13%	100.0	3970		*Avg_area*/15.84%
[70]	ResNeXt50	99.5	1088		Avg_area/3.25%	100.0	3538		*Avg_area*/15.04%
[64]	InceptionV3	98.2	190		L_∞/(16/255)				
[57]	Adv InceptionV3	99.1		67.28 s	L_∞/16	37.3			L_∞/16
[67]	InceptionV3					98.5	<5000		L_2/7.18E-05
[65]*	ResNet50	32.3			L_2/—	22.9			L_2/—
[58]	ResNet152	99.0			L_∞/(16/255)	86.8			L_∞/(16/255)
[58]	InceptionV3	83.0			L_∞/(16/255)	30.7			L_∞/(16/255)
[10]*	ResNet50	62.8				60.9			
[77]	InceptionV3					60.0	1247		$L_2/1.74\times10^{-4}$
[78]	InceptionV3	98.5	722		L_∞/0.05	99.9	7485		L_∞/0.05
[73]	ResNet18	100.0	32		L_∞/0.05	100.0	798		L_∞/0.05
[73]	InceptionV3	100.0	124		L_∞/0.05	99.7	1455		L_∞/0.05
[79]	InceptionV3	87.9	1907		L_2/5	99.2	9429		L_2/12
[79]	ResNet50	98.0	1276		L_2/5	100.0	3670		L_2/12
[80]	ResNet50	100.0	4		L_∞/0.05	97.6	33		L_∞/0.05
[94]	ResNet50	98.6	1665		L_2/3.98	100.0	7899		L_2/9.53
[83]	VGG16	100.0	110		L_∞/0.05	98.1	2191		L_∞/0.05
[83]	DenseNet121	100.0	87		L_∞/0.05	99.4	2019		L_∞/0.05
[95]	ResNet50	91.2	10000		*Sparsity*/2.36%				
[88]	VGG16	98.7	1754		L_∞/0.031	99.2	16627		L_∞/0.031
[88]	InceptionV3	92.1	4501		L_∞/0.031	95.8	36681		L_∞/0.031
[84]	InceptionV3	92.9	1100		L_2/5				
[84]	ResNet50	99.3	616		L_2/5				
[100]	ResNet50	84.8	668		L_2/5				
[100]	InceptionV3	65.3	1051		L_2/5				
[102]	InceptionV3	100.0	273		L_2/14.01	98.0	5647		L_2/15.88
[103]	VGG16	97.5	263		$L_2/\sqrt{0.001\cdot D}$				
[103]	ResNet50	99.1	331		$L_2/\sqrt{0.001\cdot D}$				
[104]	ResNet50	67.5	46		L_∞/0.05				
[85]	ResNet50	97.2	676		L_2/4.16				
[87]	ResNet	100.0	1350		APA[87]/10.0	100.0	1262		APA[87]/23.8
[86]*	ResNet50	99.6	310		L_0/59.0				
[86]	ResNet50	98.0	341		L_0/155.7				
[93]	ResNet50	98.6	383		L_2/3.622	80.6	2730		L_2/7.926
[99]	ResNet152	76.0	500		L_2/20				
[106]	InceptionV3	50.0			L_∞/0.2				
[33]	InceptionV3	88.9			L_2/1.1992				
[110]	VGG	93.6		12.72 s	*PTBPIXELS*/0.43%				
[112]	C&W[121]					100.0	13525		$L_2/1.36\times10^{-4}$
[113]	InceptionV3	100.0	11742		L_2/—	94.0	1.52×10^6		L_2/—
[116]	InceptionV3	99.1	649		$L_2/\sqrt{0.001\cdot D}$				
[116]	VGG16	99.7	239		$L_2/\sqrt{0.001\cdot D}$				
[119]	InceptionV3					98.4	15099	50.6 s	L_∞/0.05
[114]	InceptionV3	97.0	582		L_∞/0.05				
[117]*	VGG19	99.0	3278		L_2/0.53	54.0	11498		L_2/1.24
[115]	ResNet50	90.0	160000		L_2/1.21				
[118]	VGG16	99.6	118		L_∞/0.025				
[118]	ResNet34	99.9	138		L_∞/0.035				

中不同黑盒攻击算法的性能. 各个表中以分割线表示不同的黑盒攻击方法类型，在本节中，各小节对应是各类方法的具体介绍，因此为便于比较各类方法的攻击成功率、查询次数以及相似性度量结果，本文按照前述方法的介绍顺序对不同算法的攻击性能进行归纳和分析，按表 6 至表 8 自上而下的顺序分别表示基于元启发式、基于代理模型、基于直接搜索以及基于零阶优化的黑盒攻击算法.

表 6、表 7 以及表 8 均汇总了不同算法所生成对抗样本与原始样本间的相似度指标结果，用于评估对抗样本的扰动可见性，多数文献采用 L_p 范数来衡量相似度，而部分文献则自定义相似度指标. 文献[40]中采用覆盖率来衡量扰动幅度和范围，该文献将覆盖率定义为扰动像素占图像总像素的比值. 文献[110]定义 *PTBPIXELS* 作为相似度衡量指标，其表示生成的对抗样本中受到扰动的像素百分比. 文献[120]采用结构相似性指标 *SSIM*(Structure Similarity Index Measure)来衡量样本生成的质量. 文献[70]将被斑块遮挡的平均面积百分比(Avg_area)作为衡量指标. 文献[95]中定义攻击的稀疏程度，即给定图像中扰动坐标的百分比($Sparsity$). 文献[103]与文献[116]采用 L_2 范数作为相似度指标，L_2 值取 $\sqrt{0.001 \cdot D}$，其中 D 在文献[103]中表示子空间大小，在文献[116]中表示图像空间大小.

表 6 至表 8 中文献标注“*”号则代表其数据集是微小版本(如 Tiny ImageNet)，“—”符号表示文献未提及相关数据. 目标模型标注为㊀表示文献采用自定义目标模型，其中文献[67]仅说明在 MNIST 与 CIFAR-10 数据集上的代理模型与目标模型的结构不同，但未提及具体结构信息. 文献[80]在 MNIST 数据集上采用四种自定义的模型结构进行攻击，各模型结构在层数、内部层结构、卷积核等方面存在差异. 文献[102]在 MNIST 数据集上训练了由三个全连接层构建的多层感知机模型(Multilayer Perceptron，MLP)进行对抗攻击. 文献[113]在 MNIST 和 CIFAR-10 数据集上均采用自定义的网络结构进行攻击，该网络结构由 4 层卷积层、2 层池化层、2 层全连接层以及 1 层 Softmax 层构成. 文献[119]在 MNIST 数据集上采用 4 层卷积层和 2 层全连接层构成，其中各层卷积层后采用最大池化和 Relu 激活进行处理. 文献[114]在 MNIST 和 CIFAR-10 数据集上训练相同的模型结构用于攻击，该模型结构由 4 层卷积层、2 层最大池化层、2 层全连接层以及 1 层 Softmax 层构成. 文献[117]在 MNIST 数据集上同样采用自定义的网络结构进行训练和攻击，该网络结构具体由 2 层卷积层，并采用 Tanh 激活函数，之后构建 2 层最大池化层、2 层全连接层，并采用 Relu 激活函数，最后利用 1 层 Softmax 层来输出结果.

(1) MNIST 数据集中结果分析

MNIST 数据集相对背景单一，图像主体目标较为清晰. 现有研究中目标模型复杂度较低，采用的经典模型结构包括 AlexNet、VGG、ResNet 等. 此外，文献[121]提出一种较为经典的对抗攻击方法(C&W)，之后的部分文献均采用与文献[121]相同的目标模型结构以便于对比实验结果. 其余文献采用与上述模型深度和复杂度近似的自定义模型结构. 从相关研究数量以及性能趋势分析，基于零阶优化和基于代理模型黑盒攻击相关文献较多，且基于零阶优化的黑盒攻击的成功率相对较高，但同时其所需查询次数与查询时间较高.

(2) CIFAR-10 数据集中结果分析

相比于 MNIST 数据集，CIFAR-10 数据集上的实验结果更丰富，涉及的分类模型结构更为全面，部分文献也采用与 C&W[121] 相同的模型结构. 类似于 MNIST 数据集上的结果，从攻击成功率分析，基于零阶优化的黑盒攻击性能较高，但查询次数较高. 相比于其他三类黑盒攻击方式，基于零阶优化黑盒攻击的查询次数多至少 10 倍左右，以 ResNet 分类模型为例，尽管不同文献所采用的网络层数有所差异，但其结构类似，因此各模型的性能相对近似. 而从非目标攻击的现有结果分析，文献[64]、文献[102]、文献[86]、文献[111]以及文献[61]的攻击算法在攻击成功率上均达到最优，但从攻击所需查询次数分析，相比于其他四种方法，文献[64]中的方法所需查询次数最少. 文献[37]与文献[93]中的算法在确保攻击成功率较高的同时(均大于 90%)，查询次数低于文献[111]的算法(至少 5×10^2 倍). 从目标攻击的现有结果分析，文献[93]与文献[102]的算法在攻击成功率方面较优，但文献[93]中的算法查询效率更高.

(3) ImageNet 数据集中结果分析

ImageNet 数据集所涉及类别数为 1000 类，因此实验结果更具代表性，表 8 中所列实验结果较多. 对比于 ImageNet 数据集上的性能，文献[46]、文献[34]、文献[73]、文献[80]、文献[83]、文献[102]以及文献[113]在非目标攻击上的成功率较高，其中文献[80]中算法的查询次数最少. InceptionV3 则是除文献[80]与文献[83]外共有的分类模型，虽然这五

种算法在攻击成功率上相同，但文献[73]的查询次数最低，因此在实际应用中，该方法更不易被第三方图像识别平台察觉. 对于目标攻击而言，攻击者需要使模型分类图像结果为指定类别，因此攻击难度更大. 在攻击成功率大于 90% 的现有攻击算法中，文献[80]的算法较优，其查询次数最低且攻击成功率仅损失 2.4%. 以 InceptionV3 作为分类模型，则文献[73]的查询次数最低，且攻击成功率大于 95%.

从这三个数据集的总体趋势分析，ImageNet 数据集上的查询次数高于其他数据集，这主要由于 ImageNet 的图像分辨率更高，攻击者为获得各张图像在分类过程中的关键信息和特点，则攻击查询所需的时间更长且次数更多. 而且 ImageNet 数据集的类别数量是 MNIST 以及 CIFAR-10 的 100 倍，因此图像中所含数据信息更多，图像中包含的物体及背景更复杂，这也导致相同算法在该数据集上需要更多的查询次数.

6 目标检测任务中的黑盒攻击

目标检测是计算机视觉任务中另一类重要任务，其被广泛应用于机器人导航、智能视频监控、工业领域以及无人驾驶领域. 同时，目标检测作为一种泛身份识别算法，是后续图像分割、人群密度估计、姿态估计等相关任务的前置任务. 大多数最先进的目标检测器利用 DNN 模型作为它们的主干和检测网络来提取图像特征，用于分类和定位. 并且由于图像中目标的尺度、位置和外观等信息特点不同，因此相比于图像分类任务更具挑战性.

现有的目标检测算法主要是两类，一类是双阶段目标检测算法，主要基于候选区域，包括检测与识别两个阶段. 该类算法主要根据图像中物体的纹理、颜色以及其他特征对可能出现物体的区域进行提取，以生成若干比例与尺寸不同的区域. 针对这些区域进一步采用检测器以识别物体位置，并对这些候选区域进行坐标修正. 该类算法一般具有较高的检测精度，但由于需要两个阶段进行检测，检测速度较差，较为典型的双阶段目标检测算法包括 R-CNN[126]、Fast R-CNN[127]、Faster R-CNN[128]、R-FCN[129]等算法.

另一类是单阶段目标检测算法，相比于双阶段目标检测算法，单阶段检测算法的结构较为简单，计算较为高效，在实时检测任务应用较为广泛. 该类算法将目标检测问题转化为对目标位置与类别信息的回归分析问题，在检测过程中，仅依赖于模型的单一检测过程，不存在候选区域的分类问题，但也正因如此，网络需要同时学习物体的类别信息和定位物体位置，因此网络参数训练更难. 此外，单阶段检测中为了有效匹配真实目标框，前期需要生成大量的 Anchor，因此存在正负样本数量不平衡以及多尺度检测的问题，导致检测性能低于双阶段目标检测. 在之后的文献中，尝试采用 Focal Loss[130]以及图像金字塔等方式提高其检测性能. 较为经典的单目标目标检测算法包括 SSD[131]、RetinaNet[130]、RefineDet[132]、RFB-Net[133]以及较为经典的 YOLO 系列，如 YOLO[134]、YOLO9000[135]、YOLOv3[136]和 YOLOX[137]. 上述单阶段目标检测主要属于 Anchor-based 目标检测，这类型检测算法依赖于 Anchor 的设计与初始化，模型通过学习 Anchor 的修正值来预测目标位置，但这种固定的 Anchor 设计损害了目标检测器的通用性，对于不同任务，Anchor 需要精细化设计. 基于此，部分研究尝试摒弃 Anchor，即研究 Anchor-free 目标检测，其研究方向可分为两类，即基于关键点与基于中心点的检测器，基于关键点则通过识别物体的左上右下点来框出物体位置，较为典型的算法包括 CornerNet[138]和 CornerNet-lite[139]. 而基于中心点的检测器，其核心思想主要通过预测物体中心点以及宽高来识别物体位置，典型算法包括 CenterNet[140]、ExtremeNet[141]、FCOS[142]、Fovea-Box[143].

6.1 基于元启发式的黑盒攻击技术

文献[144]提出一种蒸发攻击算法以实现黑盒情况下的对抗攻击，其将黑盒攻击问题表示为样本图像流形中的优化问题，之后利用混合元启发式优化算法，即遗传与粒子群混合优化算法(GA-PSO)来求解优化问题并生成对抗样本. 类比于图像分类算法，目标检测中的最后结果包含有物体的位置信息，而且现有的图像数据集分辨率较大，直接在图像中搜索干扰像素成本较大，复杂度较高，因此可利用预测的物体位置进行扰动以减小样本空间. 传统的简单进化策略中需要固定采样的标准差，不利于挖掘关键像素点，文献[145]则利用 CMA-ES 算法在预测的物体位置框中进行干扰，通过调控分布标准差以生成对抗样本.

6.2 基于代理模型的黑盒攻击技术

文献[146]关注于语义分割任务与目标检测任务，提出一种密集对抗生成算法(Dense Adversary Generation，DAG)，首先找到图像中一组正确识别的目标，之后通过对代理模型求导，在指定区域利用梯度信息进行优化，以使模型最终对目标难以识别

和定位. 文献[147]则受启发于对抗补丁思想[148]，尝试在图像输入到网络前添加初始化补丁，并通过网络学习来更新补丁的位置和大小，以扰动模型使其难以定位目标. 文献[149]则尝试生成通用扰动以适用于多种黑盒攻击场景，同时为了解决通用扰动不能直接应用于目标检测器的问题，采用缩放与堆积两种方法添加在目标图像上. 文献[150]受启发于投影梯度下降算法[151]. 相比于图像分类任务，采用目标检测中的分类损失，边界框回归损失替代原有图像分类的损失函数. 文献[152]则首先利用检测器检测图像中的物体，之后在各物体上覆盖补丁，并且利用损失函数来优化补丁的位置和大小. 同时为了获得迁移性较高的对抗样本，基于集成学习思想综合考虑多个网络的反馈情况. 文献[153]提出一种通用密集对象抑制(Universal Dense Object Suppression，U-DOS)，图像分类任务往往只需要添加的扰动导致分类错误即可，但在目标检测中，攻击需要将检测器定位的所有区域进行修改. 该文献通过对损失函数优化，将所有定位区域的类别改为背景. 文献[154]基于 GAN 来生成对抗样本，同时为了应对过多噪声引起的失真问题，采用噪声抵消机制(即组优化与随机消除)来生成更难以察觉的对抗性扰动. 文献[155]同样利用 GAN 来生成对抗样本，但并不直接生成对抗样本，而是生成对抗补丁，同时对生成补丁进行变换后添加在数据上，利用得到的扰动数据与真实数据间的损失函数来训练 GAN 模型.

上述文献主要关注模型最后输出信息，最小化预测值与期望值之间的差异. 部分文献关注于模型中间特征，模型往往在学习过程中逐渐捕捉图像中的关键特征，如果尝试对中间特征修改则可影响最终模型的输出. 文献[5]尝试设计损失函数以扰动模型隐藏层的对象特征. 此外，该文献观察到目标检测器在学习目标特征的过程中，也在逐步学习目标所处的位置以及背景信息. 该文献尝试在图像中的停止标志牌区域内生成扰动以绕过检测，但最终目标检测器仍通过背景信息识别到标志牌的位置，导致攻击失败. 为提高攻击成功率，该文献[5]通过对背景进行扰动，并在无目标背景中人为添加扰动以生成对抗样本，从而成功绕过目标检测器的识别，具体攻击流程如图 14 所示.

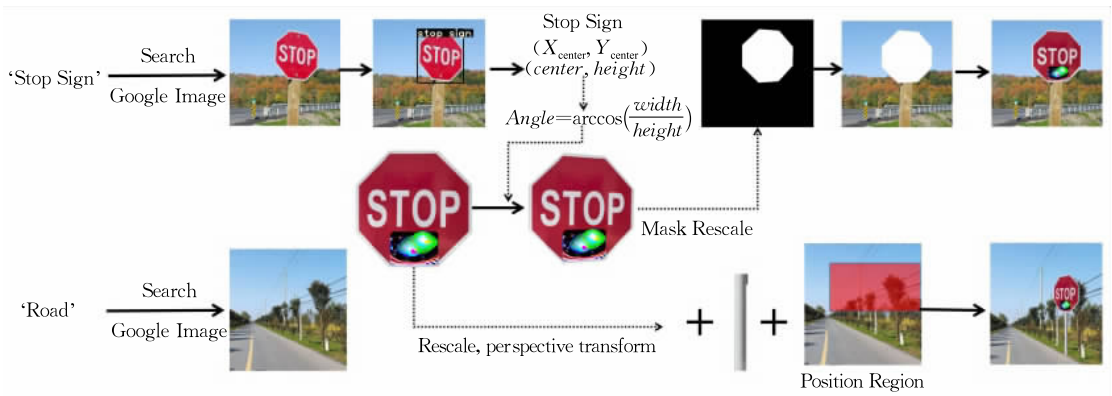

图 14　现实场景下对抗样本生成[5]

图 14 以停止标志牌数据为例来生成对抗样本. 算法主要针对两种情况，一是带有目标的图像，二是无目标图像. 针对带有目标的图像，首先通过目标检测模型识别标志牌位置，并结合识别到的目标长宽信息以及角度估计公式 $\arccos(width/height)$ 来计算原目标在图像中的角度. 之后根据边界框的大小对扰动目标进行缩放，利用估计的角度进行透视变换，以使扰动符合目标大小. 最后将原目标从图像中切除，将带有扰动的目标图像替换到原目标位置即生成对抗样本. 对于无目标图像，首先利用图像分割模型获取图像中最有可能出现目标的位置区域，位置区域设置为图像中最靠近或在语义相关对象区域内的矩形区域，图 14 中与停止标志牌相关的区域在街道旁，利用透视变换将扰动图像添加到目标图像中以生成对抗样本. 文献[156]受到图像对比度对于图像感知能力影响的启发，提出一种色散减少(Dispersion Reduction，DR)方法，通过降低模型内部特征图对比度，使图像目标难以识别从而绕过检测器.

6.3　基于直接搜索的黑盒攻击算法

基于搜索策略的黑盒攻击主要聚焦于搜索空间的选择，可将这类型算法分为在稀疏子空间搜索和在图像域搜索. 对于稀疏子空间，文献[157]尝试在稀疏域进行攻击，并提出一种稀疏对抗攻击(Sparse Adversarial Attack，SAA)，通过在图像中应用有界的 L_0 范数，优化的对抗扰动在空间上分布稀疏. 此外，文献[11]认为大型稀疏分量(Large Sparse

Components,LaS)在图像分类任务中起关键作用,因此将图像转换到稀疏域,通过设置阈值来选择对应 LaS 分量,并通过随机查询来获取最相关稀疏信号的方向以扰动 LaS 分量.对于图像域,文献[158]提出一种通过随机搜索的并行矩形翻转攻击(Parallel Rectangle Flip Attack,PRFA),在图像域中的各个矩形块内生成扰动,以避免在受攻击区域附近进行次优检测.同时,结合白盒攻击中扰动位于目标轮廓周围的特点,以减少被攻击矩形的搜索空间.

6.4 基于零阶优化的黑盒攻击算法

目标检测任务下基于零阶优化的黑盒攻击算法较少.文献[159]提出一种基于边界攻击加法机制的离散余弦变换改进攻击方法,并将其应用于离线和在线攻击目标检测器.

6.5 小 结

表 9 中对本文中涉及的目标检测任务黑盒攻击性能进行汇总,现有文献主要从数据集、目标模型、基础骨干网络、查询次数、评估指标以及相似性度量等内容对攻击算法进行评估,其中,相似性度量表示生成的对抗样本与原始样本间的差距.在本节中,将目标检测任务下的黑盒攻击技术按照四类方法进行分组,即基于元启发式的黑盒攻击技术、基于代理模型的黑盒攻击技术、基于直接搜索的黑盒攻击技术以及基于零阶优化的黑盒攻击技术,为便于对攻击方法的具体结果进行比较和分析,表 9 以分割线表示对各文献的攻击结果分组,按上述方法的介绍顺序对攻击结果进行汇总和分析.

表 9 目标检测任务下黑盒攻击性能

文献	数据集	目标模型	骨干网络	评估指标/结果	查询次数/所需时间	相似性度量/结果
[144]	MS COCO	YOLOv3	⊖	*ASR*/84.0%	29928/—	*MSE*/3.61E-03
[144]	Pascal VOC 2007	Faster-RCNN	⊖	*ASR*/48.0%	33662/—	*MSE*/8.65E-03
[145]	MS COCO	YOLOv3	⊖	*ASR*/74.0%	6154/—	L_2/<36
[146]	Pascal VOC 2007	Faster-RCNN	ZFNet	*mAP*/3.6%	47/—	
[146]	Pascal VOC 2007	Faster-RCNN	VGG	*mAP*/5.9%	41/—	
[147]	Pascal VOC 2007	Faster-RCNN	ResNet101	*mAP*/2.9%	180 k/—	
[147]	Pascal VOC 2007	YOLOv2	⊖	*mAP*/0.0%	180 k/—	
[149]	MS COCO	YOLOv3	DenseNet169	*mAP* 下降程度/30.9%	—/0.4 s	L_∞/20
[149]	Pascal VOC 2007	Faster-RCNN	DenseNet169	*mAP* 下降程度/45.4%	—/0.4 s	L_∞/20
[150]	Pascal VOC 2007	Faster-RCNN	VGG16	*mAP*/0.9%		
[150]	Pascal VOC 2007	Faster-RCNN	ResNet101	*mAP*/3.9%		
[152]	MS COCO	YOLOv2	⊖	*mAP*/10.7%		
[152]	MS COCO	YOLOv3	⊖	*mAP*/17.8%		
[152]	MS COCO	Faster-RCNN	ResNet50+FPN	*mAP*/23.5%		
[153]	Pascal VOC 2007	Faster-RCNN	VGG16	*mAP*/20.8%		L_∞/10
[153]	Pascal VOC 2007	Faster-RCNN	ResNet101	*mAP*/22.6%		L_∞/10
[154]	Pascal VOC 2007	Faster-RCNN	⊖	*mAP*/16.0%	—/1.8 s	*PSNR*/30.2
[154]	Pascal VOC 2007	SSD	⊖	*mAP*/6.0%	—/1.8 s	*PSNR*/30.2
[5]*	物理设备采集	YOLOv3	DarkNet53	成功率/51.8%		
[5]*	物理设备采集	Faster-RCNN	ResNet101	成功率/98.7%		
[156]	MS COCO2017	YOLOv3	DarkNet53	*mAP*/19.8%	100/—	L_∞/16
[156]	Pascal VOC 2012	YOLOv3	DarkNet53	*mAP*/38.2%	100/—	L_∞/16
[156]	MS COCO2017	SSD	MobileNetv2	*mAP*/3.9%	100/—	L_∞/16
[156]	Pascal VOC 2012	SSD	MobileNetv2	*mAP*/8.2%	100/—	L_∞/16
[156]	MS COCO2017	Faster-RCNN	ResNet50	*mAP*/2.5%	100/—	L_∞/16
[156]	Pascal VOC 2012	Faster-RCNN	ResNet50	*mAP*/2.8%	100/—	L_∞/16
[155]	MS COCO	Tiny-YOLOv3	DarkNet19	*mAP*/23.2%		
[157]*	MS COCO			*Evasion Score*/355.7		L_0/图像大小的 2%
[158]	MS COCO	Faster-RCNN	ResNet50	*mAP*/21.0%	3331/—	
[158]	MS COCO	YOLOv3	DarkNet53	*mAP*/24.0%	2949/—	
[158]	MS COCO	FCOS	ResNet50	*mAP*/23.0%	3395/—	
[158]	MS COCO	ATSS	⊖	*mAP*/20.0%	3500/—	
[159]*	MS COCO	YOLOv3	⊖	*ASR*/98.2%	100/—	L_2/<20

表 9 中标注为“—”表示文献中未提及相关具体数据,评估指标中 *ASR* 表示当前黑盒攻击的攻击成功率(Attack Success Rate).文献号后附带“*”号表示该文献中的实验部分与其他文献存在较大差异,其中,文献[5]主要以真实环境下的交通信号灯、行人、停车标志以及监控器获取的图像作为实验数据集,以攻击成功的帧数与监控获取总帧数的比值作为实验的评估指标.文献[157]的黑盒攻击实验部

分仅仅提及采用两种黑盒模型作为目标模型，但并未提及具体的模型类型以及结构，同时该文献采用 *Evasion Score* 作为评估标准以衡量攻击效果. 文献[159]采用主流的目标检测数据集 COCO 作为实验数据集，但在实验阶段仅选取了少量图像作为攻击目标. 现有的文献主要以 MS COCO 以及 Pascal VOC 作为黑盒攻击测试数据集，多数文献以 Mean Average Precision（*mAP*）作为衡量目标检测任务下黑盒攻击性能的指标，其值越低，则代表攻击成功率越高.

表 9 中基础骨干网络标注为㊀，表示文献中未提及是否更换目标检测模型的骨干网络，其中相关的目标检测模型主要包括 YOLOv2 与 v3、ATSS、SSD 以及 Faster-RCNN. 在对应的目标检测模型相关文献中，YOLOv2 与 YOLOv3 分别采用 DarkNet19 和 DarkNet53 作为骨干网络，Faster-RCNN 则分别采用 VGG16 以及 ResNet101 两种模型作为骨干网络，SSD 目标检测算法采用 VGG16 作为骨干网络，ATSS 则在多个骨干网络上进行训练，包括 ResNet101、ResNeXt 以及与 DCN 结合等. 此外，文献[155]中采用 Tiny-YOLOv3 作为目标检测模型，其骨干网络类似于 DarkNet19，但层数较少.

(1) MS COCO 数据集中结果分析

MS COCO（Microsoft COCO）数据集是现有目标检测领域应用较为广泛的数据集之一，其包含 91 个类别. 由于图像中部分类别间存在高度的依赖关系，因此为突出图像的上下文信息，数据集中各张图像平均包含 7 个类别标注，有助于模型学习图像语义信息以及测试多类别下的识别准确率. 表 9 中采用 MS COCO 作为数据集的文献主要以 YOLO 系列与 Faster-RCNN 系列为目标模型，在以 YOLO 作为目标模型的算法中，文献[152]、文献[156]与文献[158]均针对 YOLOv3 生成对抗样本，其中文献[152]所生成的对抗样本可导致目标检测模型 *mAP* 更低. 文献[152]中针对不同 YOLO 系列生成对抗样本，实验结果表明，相比于 YOLOv3，该算法所生成的对抗样本更易导致 YOLOv2 检测失误（*mAP* 相对较低）. 在以 Faster-RCNN 作为目标模型的算法中，文献[156]的攻击效果更优，文献[152]与文献[158]的实验结果中 *mAP* 远高于文献[156].

(2) Pascal VOC 数据集中结果分析

Pascal VOC（Pascal Visual Object Classes Challenge）数据集是较为经典的目标检测数据集，其包含 4 种类别，即人（Person）、动物（Animal）、交通工具（Vehicle）以及室内物品（Indoor）. 各个类别下进行细分，最后得到 20 个对象类别. 在以 Pascal VOC 作为目标数据集的文献中，多数文献采用 YOLO、SSD 以及 Faster-RCNN 作为目标模型. 由表 9 可知，以 YOLO 作为目标模型，文献[147]的算法所生成的对抗样本导致 YOLO 检测效果最差（*mAP* 下降至 0%）. 以 SSD 作为目标模型，文献[154]与文献[156]中的实验所得到的 *mAP* 近似，两者实验的 *mAP* 均低于 10%. 表 9 中现有文献多数以 Faster-RCNN 作为目标模型，其中文献[150]的攻击成功率较高（*mAP* 仅为 0.9%），此外，文献[146]、文献[147]以及文献[156]所报告的 *mAP* 均低于 6%，其中相比于文献[146]，文献[147]的攻击虽然导致模型的检测性能较低，但查询次数远高于文献[146]. 文献[153]与文献[156]均采用 L_∞ 作为对抗样本与原始样本间的相似性度量指标，相比于文献[153]，文献[156]的实验中 *mAP* 更低，但 L_∞ 的值比文献[153]高，因此文献[153]所生成的对抗样本更难以察觉.

为便于之后开展黑盒攻击研究与实验工作，本文从四个部分来汇总当前目标检测任务下，黑盒攻击实验所涉及的内容，包括评价指标、数据集、骨干网络以及目标检测模型等信息如表 10 所示，之后可根据表 10 的内容来组织黑盒攻击实验细节.

表 10　目标检测任务中黑盒攻击性能评估

评价指标	数据集	骨干网络	目标检测模型及模块
mAP	Pascal VOC 2007[160], 2012[161]	VGG	Faster-RCNN
AP	Microsoft COCO[162]	ZFNet	YOLO
PSNR	物理设备采集（如 Iphone 6s, HUWEI nova 3e）	ResNet	SSD
Success Rate	ISIC[163]	MobileNet	Mask-RCNN
MSE			FCOS
L_2 distance			
平均查询次数			
Evasion Score			

相比于图像分类任务，目标检测任务更为复杂，该任务包含对物体目标的类别预测，但同时需要判断物体的位置. 此外，图像中一般不止单一物体或同类物体，并且各物体间的尺寸、外观纹理、遮挡情况

等均不同,这些复杂情况导致轻微的扰动易使目标检测模型性能下降.一部分对抗样本生成算法是基于图像分类对抗攻击算法进行设计,包括对图像像素的扰动或潜在空间搜索,通过扰动图像像素或优化对抗空间,使其难以捕捉目标或者类别识别错误.此外,另一些文献利用目标检测任务自身的损失函数或设计扰动损失函数,利用反向传播以生成具有针对性的对抗样本.类比图像分类任务,本文将目标检测中对抗攻击按照四类黑盒优化方法进行分类,其中基于代理模型与基于直接搜索相关文献较多,而基于元启发式与基于零阶优化相关研究较少,可作为之后的研究方向.由上述文献可知,虽然各文献对算法性能均有评估,但采用的数据集、骨干网络与目标模型有所区别,因此难以直接比较攻击算法之间的性能差异.

7 图像分割任务中的黑盒攻击

图像分割是另一类应用较为广泛的视觉任务.相比于目标检测任务中主要对目标进行位置标记和分类,图像分割任务更为细粒度.该任务主要目标是对像素进行分类,即将图像分为互不相交的连通区域,因此图像分割有助于图像的全场景理解,对于自动驾驶、医学影像诊断以及人机交互等相关任务至关重要.按照不同的分类目标和结果,图像分割任务可分为语义分割任务、实例分割任务以及全景分割任务.

语义分割任务的主要目的是使用一组对象类别(例如人、车辆、树、天空等)对所有图像像素进行像素级标记.实例分割是在目标检测基础上的细化任务,对前景中的各个目标实现像素级别的分离.语义分割任务与实例分割任务仅侧重图像的一部分信息,而在现实应用中往往需要图像中出现的所有对象的信息,据此,全景分割综合考虑所有对象信息,背景与前景均进行分割,三类任务具体关系如图 15 所示.

由图 15 可知,原始图像中包含有多种对象信息,包括车辆、街灯、行人、楼房以及天空等信息,图像分割的任务在于像素级分类.其中,语义分割任务实现对各类对象像素的标记,同类对象标记为同种颜色;实例分割对前景对象进行标记,同时为区分不同的对象而采用不同颜色标记;全景分割将上述两种任务进行结合,需要对前景与背景中的对象进行像素标记,同时同一类别不同对象标记为不同

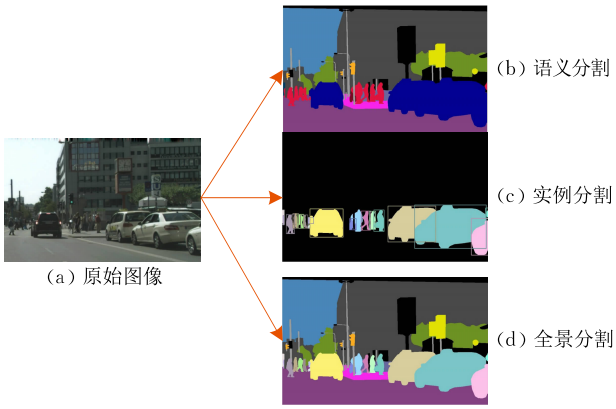

图 15 三类图像分割任务

类别.图像分割任务更为细致,因此像素的轻微扰动易造成分割有误,本文以黑盒优化问题的四类关键方法为出发点,对现有对抗攻击算法进行介绍和总结.

7.1 基于代理模型的黑盒攻击技术

文献[164]基于迭代最小似然类方法(Itertive Least-Likely Class Method)[165],将原有损失中对于图像整体的优化改为对图像像素类别的优化,以获取图像分割任务下的对抗样本.文献[146]在上一节中已介绍其攻击思想,由于其关注于图像本身特征,因此攻击可应用于语义分割以及目标检测两类任务.文献[166]则将现有的对抗攻击应用于语义分割任务,并进行评估以总结对抗攻击对于不同分割网络的性能.文献[167]采用条件生成对抗网络(Conditional Generative Adversarial Networks,CGAN)结构作为生成对抗样本的整体架构,同时为了解决生成图像质量差且攻击率低的问题,利用空间自适应非规范化(Spatially-Adaptive Denormalization,SPADE)结构以及添加额外的损失项来生成有效的对抗样本,文献[167]中 AdvSPADE 算法流程如图 16 所示.

图 16 中攻击算法由图像编码器、SPADE 生成器、目标图像分割模型以及 SPADE 鉴别器构成.其中,图像编码器以原始图像 x 为输入,计算均值和方差向量$(\mu(x),\sigma(x))$,并利用重参数化结合随机噪声 ε 生成随机扰动 z.SPADE 生成器 G 利用随机扰动 z 与原始语义标签以生成合成图像 x',之后利用设计的损失函数以欺骗 SPADE 鉴别器,同时最大化分割模型 S 的预测结果与语义标签间的距离.由于生成对抗样本的过程中引入随机噪声 ε,使得生成器 G 可以生成不同类型的对抗样本从而提高攻击迁移性.文献[156]利用图像对比度进行攻击,

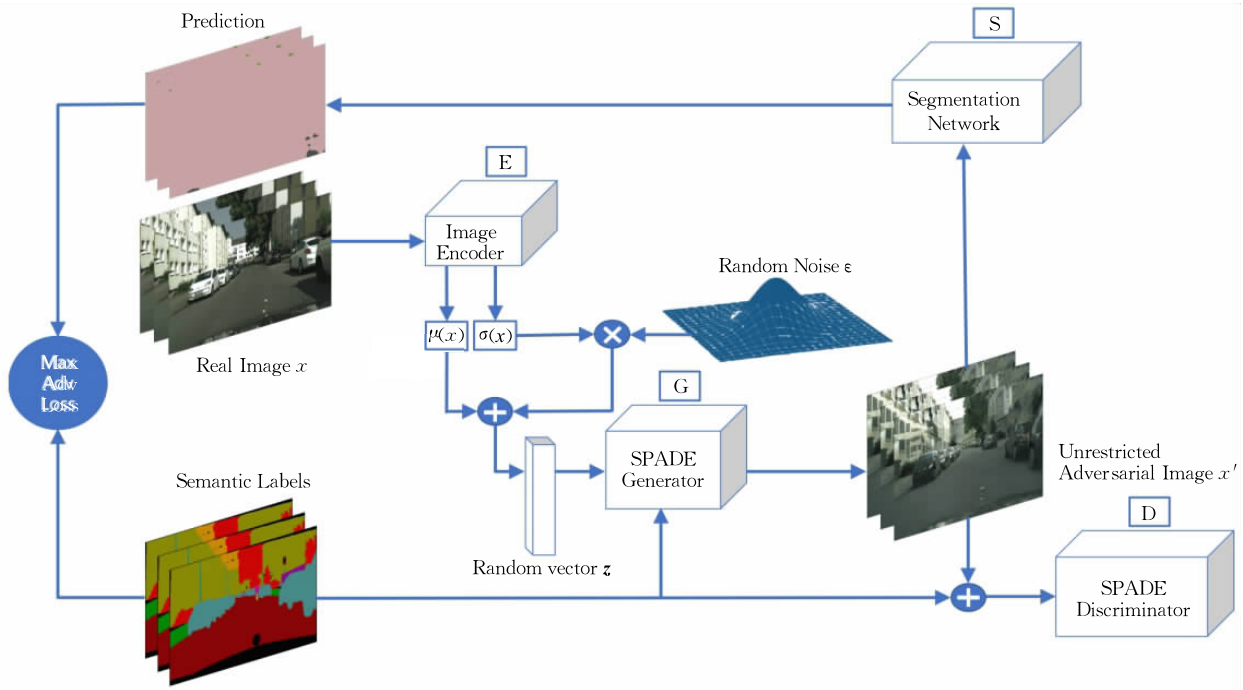

图 16　AdvSPADE 算法流程[167]

在上一节中主要讲述其目标检测下的攻击过程，由于其跨任务的攻击能力，该文献在语义分割任务中也进行了评估. 文献[168]引入一种间接攻击策略，即自适应局部攻击(Adaptive Local Attacks). 该方法旨在寻找干扰的最佳图像位置，并逐步通过学习来优化补丁的位置. 文献[169]认为图像分割任务主要针对单像素进行分类，现有针对该任务的攻击策略同样仅关注单像素的类别改变. 该文献认为单像素的信息相对较少，尝试聚合补丁以及完整图像的信息以实现更强的攻击. 文献[170]则受启发于图像分类任务中的白盒攻击思想，将其中的损失函数更改为图像分割任务下的损失函数，以对各个像素更新来干扰像素类别. 文献[171]基于 DeepFool 算法[172]提出一种通用性扰动策略，同时为了减少对原始样本的依赖，引入了自监督余弦相似性损失来优化通用扰动.

7.2　基于零阶优化的黑盒攻击技术

文献[173]认为扰动在网络前向传播中易逐步削弱，因此需要寻找在传播过程中保持不变或被逐步放大的扰动. 该文献首先给出添加扰动后 DNN 模型的泰勒展开式如式(11).

$$f(\bar{x}+\varepsilon\tilde{x})=f(\bar{x})+\varepsilon\Im\tilde{x}+O(\varepsilon^2) \tag{11}$$

该文献中指出可通过求得最大特征值以及对应的特征向量来构建对抗扰动，此外对于涉及到的雅克比矩阵计算，采用类似梯度估计的方式进行求解. 文献[174]提出一种基于查询的黑盒攻击算法，可在有限查询预算内改变前景像素的类别. 该算法采用更精确的损失梯度估计，以及采用可学习的方差替换自适应分布的固定方差，从而改进了自适应平方攻击，并在医学图像分割任务中进行评估.

7.3　基于元启发式的黑盒攻击技术

文献[175]中基于差分进化算法提出一种多次迭代的攻击算法(Differential Evolution Attack，DEAttack)，算法可自动识别输入图像的最敏感区域. 与基于梯度的算法相比，DEAttack 在整个优化过程中保持了解空间的多样性.

7.4　基于直接搜索的黑盒攻击技术

文献[176]尝试对现有自动驾驶语义分割任务中存在的对抗攻击进行评估，其主要从两个方面研究对抗扰动的影响，即模型结构的内部因素以及数据级环境扰动的外部因素. 图像的扰动主要通过 19 种不同的对抗噪声实现，该文献将其归为 4 类，且各类噪声均有 5 个等级，通过添加不同种类的扰动来生成对抗样本. 文献[177]认为部分对抗攻击算法中常利用截断函数以纠正扰动幅度，若生成的对抗样本超出图像值范围则进行截取，这种方式易导致图像出现轻微失真. 该文献尝试在 tanh 函数中搜索对抗样本，由于 tanh 函数值域范围在[−1,1]，因此生成的对抗样本满足图像值要求从而失真度更低.

7.5　小　结

表 11 中汇总了本文所涉及的图像分割任务下黑盒攻击的性能情况，并以分割线表示不同的黑盒

攻击算法类型.本节中的黑盒攻击技术按照基于代理模型、基于零阶优化、基于元启发式以及基于直接搜索的黑盒攻击技术进行汇总和分析,为便于方法与攻击结果对应,表 11 中自上而下按照本节攻击技术的介绍顺序对算法结果进行分析和总结.

表 11 中标注为"—"表示文献中未提及相关具体数据,㊀符号表示该文献实验中未提及是否更换图像分割模型的骨干网络,从涉及的图像分割模型来分析,主要包括 Deeplabv3、Upernet、PSANet、PSPNet、UNet、COPLENet、ESPNet 以及 FastSCNN 模型.在对应的图像分割模型文献中,Deeplabv3 与 PSPNet 模型均采用 ResNet 作为骨干网络进行特征提取,Upernet 则主要基于 FPN 网络进行训练,PSANet 采用 ResNet-FCN 网络作为骨干网络.UNet 模型则通过两条路径进行编码和分割,即收缩路径与扩展路径.COPLENet 类似于 UNet 结构,采用双路径实现图像分割.FastSCNN 则基于 MobileNetv2 设计骨干网络.

表 11 图像分割任务下黑盒攻击性能

文献	数据集	目标模型	骨干网络	评估指标/结果	查询次数/所需时间	相似性度量/结果
[146]	Pascal VOC	FCN	AlexNet	*mIoU*/4.0%	32/—	
[146]	Pascal VOC	FCN	VGG	*mIoU*/4.1%	54/—	
[166]	Pascal VOC	FCN8	VGG	*IoU Ratio*/65.4%		L_∞/8
[167]	Cityscapes	DRN38		*mIoU*/40.7%		*FID*/67.3
[167]	Cityscapes	Deeplabv3	㊀	*mIoU*/42.5%		*FID*/67.3
[167]	ADE20K	Upernet50	㊀	*mIoU*/9.6%		*FID*/53.5
[156]	MS COCO2017	Deeplabv3	ResNet101	*mIoU*/17.2%	100/—	L_∞/16
[156]	Pascal VOC 2012	Deeplabv3	ResNet101	*mIoU*/21.8%	100/—	L_∞/16
[156]	MS COCO2017	FCN	ResNet101	*mIoU*/12.9%	100/—	L_∞/16
[156]	Pascal VOC 2012	FCN	ResNet101	*mIoU*/14.4%	100/—	L_∞/16
[168]	Cityscapes	FCN	ResNet	*mIoU*/52.0%		
[168]	Cityscapes	PSPNet	ResNet	*mIoU*/19.0%		
[168]	Pascal VOC 2007	FCN	ResNet	*mIoU*/50.0%		
[168]	Pascal VOC 2007	PSANet	㊀	*mIoU*/28.0%		
[169]	Cityscapes	PSPNet	㊀	*F* 值/95.9		
[169]	Cityscapes	PSPNet	㊀	*AUC* 值/91.1		
[169]	Cityscapes	Upernet	㊀	*F* 值/96.2		
[169]	Cityscapes	Upernet	㊀	*AUC* 值/94.9		
[170]	Pascal VOC	PSPNet	ResNet101	*mIoU*/12.3%		L_∞/0.03
[170]	Pascal VOC	FCN8s	VGG	*mIoU*/19.6%		L_∞/0.03
[170]	Cityscapes	PSPNet	ResNet50	*mIoU*/3.7%		L_∞/0.03
[170]	Cityscapes	Deeplabv3	ResNet50	*mIoU*/3.8%		L_∞/0.03
[170]	Cityscapes	FCN8s	VGG16	*mIoU*/10.6%		L_∞/0.03
[171]	Pascal VOC	FCN	ResNet101	*mIoU*/26.9%		
[171]	Pascal VOC	Deeplabv3	ResNet101	*mIoU*/20.3%		
[173]	CamVid	UNet	EfficientNet-b3	*Dice*/0.25		*SSIM*/0.86
[174]	Chest X-ray	UNet	㊀	*ASR*/81.5%	1000/—	*SSIM*/0.81
[174]	Chest X-ray	UNet	㊀	*ASR*/81.5%	1000/—	L_∞/5
[174]	Chest X-ray	COPLENet	㊀	*ASR*/89.2%	1000/—	*SSIM*/0.84
[174]	Chest X-ray	COPLENet	㊀	*ASR*/89.2%	1000/—	L_∞/5
[175]	Glaucoma	UNet	㊀	*IoU*/0.0%		*APOPC*/0.2%
[175]	Lung	UNet	㊀	*IoU*/30.0%		*APOPC*/2.2%
[175]	ISIC	ResNet50		*IoU*/60.0%		*APOPC*/3.7%
[176]	Cityscapes	Deeplabv3+	ResNet50	*mIoU*/9.5%		
[176]	Cityscapes	FCN	ResNet50	*mIoU*/1.5%		
[176]	Cityscapes	SegNet	VGG16	*mIoU*/13.0%		
[176]	Cityscapes	PSPNet	ResNet50	*mIoU*/4.6%		
[177]	Cityscapes	ESPNet		成功率/77.0%		扰动比值/0.72
[177]	Cityscapes	FastSCNN	㊀	成功率/69.3%		扰动比值/0.52

表 11 中多数文献采用 *mIoU* 作为衡量攻击成功的指标,但少数文献采用其他指标或自定义指标来评估结果,文献[166]中采用 *IoU Ratio* 作为衡量黑盒攻击的指标,该文献将其定义为对抗样本的 *IoU* 与原始样本 *IoU* 的比率.文献[173]中采用 *Dice* 系数作为衡量攻击是否有效的指标,文献[174]则将被破坏的前景像素比例作为对抗攻击成功率,模型识别前景的准确性越差,对抗攻击成功率越高.对于

相似性度量指标，多数文献仍然采用扰动范数来评估对抗样本与原始样本间的差距，文献[167]采用 Frechet Inception Distance(*FID*)作为度量标准，文献[173]与文献[174]采用 SSIM 指标以衡量两幅图像的相似度. 文献[175]则是定义了扰动点变化的平均百分比 *APOPC*(Average Percent of Perturbation Points Changed)作为衡量相似性的指标. 文献[177]统计图像中被误分类为其他类别的像素个数，并计算与总像素个数的比值作为攻击成功率. 现有图像分割黑盒攻击主要集中在 Pascal VOC 与 Cityscapes 两个公开数据集进行测试和评估，本文对这两类数据集上的结果进行汇总与分析.

(1) Pascal VOC 数据集中结果分析

除目标检测任务外，Pascal VOC 数据集也包含有图像分割的实例标注信息，以便于相关分割任务进行评估. 图像分割任务下该数据集的类别数量与目标检测任务相似，除原有 20 个类别外，不属于这些类别的像素则标记为背景类. 图像分割模型需要对图像中各个像素按照这些类别进行分类，并计算最后分割准确率. 表 11 中采用 Pascal VOC 作为数据集的文献主要集中于基于代理模型的黑盒攻击，并且主要以 *mIoU* 作为衡量攻击是否有效的指标. 以 FCN 作为目标模型则文献[146]的算法攻击更为有效，但图像分割任务下的黑盒攻击除受攻击算法的影响外，骨干网络的结构以及训练后模型的性能均是影响攻击是否有效的因素. 文献[146]中以 AlexNet 作为骨干网络时其模型自身的性能低于 50%(*mIoU*)，而文献[168]与文献[171]以 ResNet 为骨干网络的图像分割模型 *mIoU* 则高于 60%，因此初始的模型性能差异较大. 此外，目标模型结构上的差异对攻击后算法的性能影响较大，文献[156]与文献[171]均采用 Deeplabv3 作为目标模型，且骨干网络结构相同，其攻击性能差异较小，但文献[156]的对比实验中，采用相同的骨干网络，目标模型分别采用 FCN 以及 Deeplabv3，攻击后模型的 *mIoU* 相差 7%左右，类似于文献[156]，文献[171]的 *mIoU* 差异约 6%.

(2) Cityscapes 数据集中结果分析

Pascal VOC 数据集主要针对自然图像中所存在的物体进行标注，Cityscapes 数据集则主要针对复杂城市街道场景下的图像分割任务. 该数据集通过拍摄 50 个不同城市的街道来构建复杂街道场景下的分割数据集，其中 5000 张图像采用像素级标注，以便模型进行监督学习，20000 张图像采用粗略的多边形标注，以支持弱监督信息下的图像分割模型训练. Cityscapes 数据集上的目标模型主要集中于三种结构，包括 Deeplabv3、FCN、PSPNet. 由表 11 可知，以这三类模型作为目标模型，文献[170]与文献[176]攻击算法的 *mIoU* 更低. 具体而言，对于 Deeplabv3 以及 PSPNet 模型，文献[170]的攻击性能较优(*mIoU* 值均低于其他算法)，对于 FCN 模型则文献[176]的攻击性能较优，其中文献[176]主要尝试采用不同种类噪声来攻击目标图像，在文献[176]报告的实验结果中，由脉冲噪声引起的对抗扰动更有效.

类似于表 10，本文从四个部分来汇总当前图像分割任务中黑盒攻击实验所涉及的内容，包括评价指标、数据集、骨干网络以及图像分割模型等信息如表 12 所示，之后可根据表 12 的内容来组织图像分割任务中黑盒攻击的评估实验.

表 12　图像分割任务中黑盒攻击性能评估

评价指标	数据集	骨干网络	图像分割模型及模块
均值、标准差	Cityscapes[178]	VGG 16, 8	FCN
mIOU	Pascal VOC 2007[160], 2012[161]	AlexNet	CRF-RNN
IOU-Rate	ADE20K[179]	ResNet 50, 101	DilatedNet
FID(Frechet Inception Distance)	BDD100K[180]	MobileNet	PSPNet
Accuracy	Mapillary Vistas[181]	Xception	ICNet
ASR(Attack Success Rate)	Semantic KITTI[182]	EfficientNet	SegNet
F 值	CamVid[183]		DeepLab v2, v3
AUC	SCR[184]		UperNet
SSIM	Glaucoma[185]		DRNet
Dice 系数	Lung[186]		PPM
AOPC	ISIC 2018[163]		PSANet
APOPC			DANet
			PointNet
			SqueezeSeg
			Cylinder 3D
			PointASNL
			U-Net
			COPLENet

相比于图像分类与目标检测，图像分割任务下的对抗攻击算法较少，且主要以基于代理模型的攻击算法为主.此外，图像分类与目标检测任务下攻击成功意味着通过添加扰动使图像分类错误或者目标未能检测到，但对于图像分割任务而言，评估其是否攻击成功，则并非是目标的检测有误，而是对应像素是否被覆盖，因此评估指标多采用 *mIOU* 或者其他对比覆盖面积的指标.类似于目标检测，图像分割对抗攻击所用指标的选择角度较多，并且各类算法采用的目标模型细节上有所区别，包括训练过程，模型结构的修改等，导致最后模型的性能有所区别，难以直接对比算法性能.

8 其他图像分析任务中的黑盒攻击

上述三类任务是目前主流的计算机视觉任务，其在多种工业场景均有涉及.此外，一些特定任务或场景中，也会涉及计算机视觉技术或与其他技术相结合以实现复杂的实际任务，包括目标跟踪、3D 检测、图像检索以及人脸识别等.由于这些任务目前高度依赖于深度神经网络，因此针对图像的黑盒攻击对于这些任务同样威胁巨大.

8.1 基于代理模型的黑盒攻击技术

文献[187]探索 3D 目标检测任务中的对抗攻击，针对摄像机 LiDAR 融合 3D 目标检测模型提出了一种简单的基于多视点相关性的对抗性攻击方法，利用生成式网络辅以图像分割模型来生成对抗样本.文献[188]聚焦于合成不存在人的高度逼真图像，同时在黑盒攻击场景下，利用原模型与目标模型间的决策边界相似性生成迁移性高的对抗样本.文献[189]主要针对图像检索任务中的深度哈希网络提出一种原型监督对抗网络(Prototype-supervised Adversarial Network，ProS-GAN)，同样利用了生成式模型，并将语义表示与原始图像馈送到生成器中用于之后的目标攻击.文献[190]针对于图像检索任务，其预先使用递归模型窃取方法获取的替代模型并进行白盒攻击，之后采用新损失函数的梯度信息为攻击提供方向.文献[191]针对自动驾驶场景，通过划分数据集以模拟现实世界的开放集环境.同时为提高对抗样本的迁移性，将动态卷积结合在代理模型，提高代理模型与目标模型的相似度.文献[192]主要针对跨模态检索任务，数据类型涉及图像与文本，通过模型来构建两者之间的关联.该文献提出一种针对深度跨模态哈希搜索的黑盒攻击，通过向目标检索系统发送查询请求来构建代理模型，并设计一种多模态代理驱动的对抗样本生成模型，其中对抗损失与量化损失分别用于构建对抗样本以及近似目标模型.文献[193]关注物理世界中的图像分割对抗攻击，并在自动驾驶领域中进行实验.攻击者可基于该文献提出的攻击框架，在物理世界放置简单对象即可欺骗语义分割模型.文献[194]针对人脸识别任务生成对抗样本，该文献基于 GAN 思想，引入替代模型来生成迁移性高的对抗样本，并设计了一种正则化模块以进一步增强对抗样本的可迁移性.

8.2 基于元启发式的黑盒攻击技术

文献[195]针对人脸识别中的对抗攻击任务，基于协方差自适应进化策略提出一种进化攻击算法，该算法对搜索方向的局部几何结构进行建模，以减少搜索空间的维数.

8.3 基于直接搜索的黑盒攻击技术

文献[196]研究自动驾驶系统中存在的对抗攻击问题，主要针对激光雷达中的感知任务，以探索当前基于激光雷达的感知架构的一般缺陷，并利用激光雷达点云中被忽略的遮挡模式来生成对抗样本.文献[6]则将简单黑盒攻击(Simple Black-box Attack，SimBA)与自动驾驶领域相结合，提出一种改进的简单黑盒攻击算法.文献[197]主要针对目标跟踪任务，与目标检测类似，目标跟踪主要是对物体在连续帧的位置进行定位和识别，该文献的对抗攻击主要基于当前帧与历史帧的预测 *IoU* 分数顺序生成扰动，通过降低 *IoU* 分数，所提出的攻击方法相应地降低了时间相干边界框(即物体运动)的准确性.文献[198]针对于手写签名验证任务提出相应的对抗攻击算法，并在对抗扰动搜索区域选择上仅考虑前景区域(即笔画)，通过迭代更新扰动的位置和优化它们的强度来实现对抗攻击.文献[199]则聚焦于人脸相关任务，其主要在特征域搜索对抗扰动，通过操纵面部特征来对人脸识别系统进行对抗攻击.类似地，文献[200]也针对人脸识别任务，通过构建自动编码器组，并在低维嵌入潜空间中找寻对抗样本.

8.4 基于零阶优化的黑盒攻击技术

文献[201]针对动作识别任务提出一种基于自适应分组策略的黑盒攻击方法，与目标跟踪类似，动作识别需要根据连续帧中人物的空间位置以及行为

的时序信息来判断动作类别. 该文献中的攻击算法首先利用帧间冗余性对连续帧分组，之后通过相似性度量来识别组内关键帧并估计关键帧梯度，最后在组内共享梯度信息以生成各帧的对抗样本. 文献[202]提出一种针对单目标跟踪任务的离散掩蔽黑盒攻击，通过基于动量的噪声生成器来构建扰动较大的对抗样本，同时采用 Sign-OPT 黑盒攻击[115]优化扰动范围以减小噪声幅度.

针对于其他计算机视觉任务，其黑盒攻击方式与三类主流视觉任务类似，因此本节从黑盒优化角度对攻击算法分类. 上述文献涉及多种图像信息，包括人脸图像、3D 图像、激光雷达点云图像等，因此在模型的处理方式上均有不同，同样对于黑盒攻击方式也存在差异. 以人脸识别为例，其主要目标是捕捉用户的人脸信息，通过数据库比对以检索当前用户的信息并做出判定，因此人脸的轮廓、五官的位置与外观等信息对于模型判定至关重要. 据此，不同的视觉任务中的黑盒攻击则主要依据这些关键信息来对应生成对抗样本.

9　总结与展望

综上所述，本文对图像分析领域中的黑盒对抗攻击研究现状进行了总结，同时从三类图像主流任务以及黑盒优化两种角度对文献进行分析和汇总. 首先给出了黑盒优化相关概念；其次从不同图像任务角度出发，按照四类黑盒优化算法对相关研究进行划分和介绍，即分为基于元启发式、基于代理模型、基于直接搜索以及基于零阶优化的黑盒攻击算法，同时总结各类优化算法的特点、优势以及难点，为之后的相关研究提供优化方向；最后对这些算法的性能以及实验相关内容进行汇总，为未来从事该方向研究的人员提供相关评估指标和方案.

现有的黑盒攻击技术已取得一些成绩，但在一些方面仍然需要继续探索，包括攻击的查询次数过多、黑盒攻击跨领域跨任务能力、其他实际图像分析任务中的黑盒攻击等问题与方向.

9.1　降低查询次数以提升攻击效率

黑盒攻击场景下难以获取模型的梯度信息或者模型内在网络结构等关键信息，因此实际任务中攻击者仅能够通过访问目标模型或第三方应用平台来获取最后的输出结果，利用这些查询结果生成对抗样本. 上述四类黑盒攻击算法对于查询结果的使用各不相同，基于元启发式黑盒攻击利用查询结果更新候选解集，基于代理模型黑盒攻击利用查询结果构建目标模型的近似模型，基于直接搜索黑盒攻击利用查询结果来优化搜索方向，基于零阶优化黑盒攻击利用查询结果来估计目标模型的梯度. 查询次数越多，各类算法得到的对抗样本更优，进而攻击成功率越大，从现有研究的评估结果分析，多数有目标攻击的查询次数超过 500 次. 但现实任务中相关图像分析领域的应用或网站限制了用户短时间内的查询次数，因此未来在这一方面的研究方向在于如何高效利用查询结果，在保证攻击性能的同时以减少查询次数.

9.2　提升黑盒攻击跨领域跨任务能力

黑盒攻击跨领域跨任务的适应性是未来研究的主要热点之一，适应性要求黑盒攻击在面对不同的图像分析任务时都能有较好的表现. 现有的研究多数局限于单一任务场景，仅在单一任务中评估黑盒攻击算法的性能，对于该算法在其他领域下的攻击性能评估不足且研究较少，因此未来的研究方向需要考虑对抗样本在其他任务上的适应性. 对现有图像分析相关研究进行分析后发现，虽然各任务在最终目标、评估指标以及数据标注信息等方面均不同，但各任务下 DNN 模型都是通过对图像数据信息的高度抽象与分析来实现任务目标. 因此提升黑盒攻击跨领域跨任务能力的关键在于如何有效地利用图像信息，以及如何建立多任务间的联系. 目前一部分研究关注表示学习以及 DNN 可解释性，这些方向对于理解图像数据和 DNN 模型内部机理有一定的启示作用，因此未来可尝试将表示学习、DNN 可解释性与黑盒攻击有机结合，探索更为高效、适应性强的黑盒攻击算法.

9.3　探索多场景多任务中的黑盒攻击

本文在介绍黑盒攻击时选取了三类主流的图像分析任务，即图像分类、目标检测以及图像分割任务. 但从文献收集情况来看，相比于目标检测以及图像分割任务，目前多数研究关注于图像分类任务. 这主要由于早期图像分析任务多以分类为主，并且图像分类可作为部分图像分析任务的基础任务，如分类任务中的 DNN 模型常作为目标检测或图像分割任务中的预训练骨干网络，用于前期特征提取工作，极大提升整体模型的训练效率与精度. 近几年实际应用领域如自动驾驶、人脸识别、智慧医疗等越来越依赖更细粒度的图像分析任务，如目标检测以及图

像分割等任务.而且针对图像分类任务的黑盒攻击算法对于其他图像分析任务并不完全适用,因此需要对以更细粒度的图像分析技术为基础的黑盒攻击算法展开进一步研究,进而探索不同场景不同任务下的黑盒攻击算法.

参 考 文 献

[1] Afrasiyabi A, Larochelle H, Lalonde J, Gagne C. Matching feature sets for few-shot image classification//Proceedings of the 2022 IEEE International Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans, USA, 2022: 9004-9014

[2] Bar A, Wang X, Kantorov V, et al. DETReg: Unsupervised pretraining with region priors for object detection//Proceedings of the 2022 IEEE International Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans, USA, 2022: 14585-14595

[3] Yang Z, Wang J, Tang Y, et al. LAVT: Language-aware vision transformer for referring image segmentation//Proceedings of the 2022 IEEE International Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans, USA, 2022: 18134-18144

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

[5] Zhao Y, Zhu H, Liang R, et al. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security(CCS). London, UK, 2019: 1989-2004

[6] Kumar K N, Vishnu C, Mitra R, Mohan C K. Black-box adversarial attacks in autonomous vehicle technology//Proceedings of the 49th IEEE Applied Imagery Pattern Recognition Workshop(AIPR). 2020: 1-7

[7] Li Y, Xu X, Xiao J, et al. Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. IEEE Internet of Things Journal, 2021, 8(8): 6337-6347

[8] Wang J, Liu A, Bai X, Liu X. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. IEEE Transactions on Image Processing, 2022, 31: 598-611

[9] Li C, Wang L, Ji S, et al. Seeing is living? Rethinking the security of facial liveness verification in the deepfake era//Proceedings of the 31st USENIX Security Symposium. Boston, USA, 2022: 2673-2690

[10] Sun X, Cheng G, Li H, et al. Exploring effective data for surrogate training towards black-box attack//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans, USA, 2022: 15334-15343

[11] Zanddizari H, Zeinali B, Chang J M. Generating black-box adversarial examples in sparse domain. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(4): 795-804

[12] Zhang Z, Huang S, Liu X, et al. Adversarial attacks on YOLACT instance segmentation. Computers and Security, 2022, 116: 102682

[13] Pardalos P M, Rasskazova V, Vrahatis M N. Black Box Optimization, Machine Learning, and No-Free Lunch Theorems. Gewerbestrasse, Switzerland: Springer, 2021

[14] Liu S, Chen P, Kailkhura B, et al. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. IEEE Signal Processing Magazine, 2020, 37(5): 43-54

[15] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations(ICLR). Banff, Canada, 2014: 1-10

[16] Ren K, Zheng T, Qin Z, Liu X. Adversarial attacks and defenses in deep learning. Engineering, 2020, 6(3): 346-360

[17] Li Xin-Jiao, Wu Guo-Wei, Yao Lin, et al. Progress and future challenges of security attacks and defense mechanisms in machine learning. Journal of Software, 2021, 32(2): 406-423(in Chinese)
(李欣姣, 吴国伟, 姚琳等. 机器学习安全攻击与防御机制研究进展和未来挑战. 软件学报, 2021, 32(2): 406-423)

[18] Zhang Tian, Yang Kui-Wu, Wei Jiang-Hong, et al. Survey on detecting and defending adversarial examples for image data. Journal of Computer Research and Development, 2022, 59(6): 1315-1328(in Chinese)
(张田, 杨奎武, 魏江宏等. 面向图像数据的对抗样本检测与防御技术综述. 计算机研究与发展, 2022, 59(6): 1315-1328)

[19] Liu Xi-Meng, Xie Le-Hui, Wang Yao-Peng, Li Xu-Ru. Adversarial attacks and defenses in deep learning. Chinese Journal of Network and Information Security, 2020, 6(5): 36-53(in Chinese)
(刘西蒙, 谢乐辉, 王耀鹏, 李旭如. 深度学习中的对抗攻击与防御. 网络与信息安全学报, 2020, 6(5): 36-53)

[20] Chen Meng-Xuan, Zhang Zhen-Yong, Ji Shou-Lin, et al. Survey of research progress on adversarial examples in images. Computer Science, 2022, 49(2): 92-106(in Chinese)
(陈梦轩, 张振永, 纪守领等. 图像对抗样本研究综述. 计算机科学, 2022, 49(2): 92-106)

[21] Pan Wen-Wen, Wang Xin-Yu, Song Ming-Li, Chen Chun. Survey on generating adversarial examples. Journal of Software, 2020, 31(1): 67-81(in Chinese)
(潘文雯, 王新宇, 宋明黎, 陈纯. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67-81)

[22] Bhambri S, Muku S, TulasiA, Buduru A B. A survey of black-box adversarial attacks on computer vision models. arXiv preprint arXiv: 1912.01667. https://arxiv.org/abs/1912.01667, 2020

[23] Machado G R, Silva E, Goldschmidt R R. Adversarial machine learning in image classification: A survey toward the defender's perspective. ACM Computing Surveys, 2023, 55(2): 8:1-8:38

[24] Li Y, Cheng M, Hsieh C-J, Lee T C M. A review of adversarial attack and defense for classification methods. The American Statistician, 2022, 76(4): 329-345

[25] Yann L. The MNIST database of handwritten digits. https://yann.lecun.com/exdb/mnist/, 1998

[26] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.222.9220, 2009

[27] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, USA, 2009: 248-255

[28] Wang S, Shi Y, Han Y. Universal perturbation generation for black-box attack using evolutionary algorithms//Proceedings of the 24th International Conference on Pattern Recognition (ICPR). Beijing, China, 2018: 1277-1282

[29] Mosli R, Wright M, Yuan B, Pan Y. They might NOT be giants crafting black-box adversarial examples using particle swarm optimization//Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS). Guildford, UK, 2020: 439-459

[30] Jiang W, Li H, Xu G, et al. Physical black-box adversarial attacks through transformations. IEEE Transactions on Big Data, 2023, 9(3): 964-974

[31] Wierstra D, Schaul T, Glasmachers T, et al. Natural evolution strategies. Journal of Machine Learning Research, 2014, 15(1): 949-980

[32] Ilyas A, Engstrom L, Athalye A, Lin J. Query-efficient black-box adversarial examples. http://arxiv.org/abs/1712.07113, 2018

[33] Chen P, Zhang H, Sharma Y, et al. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). Dallas, USA, 2017: 15-26

[34] Du Y, Fang M, Yi J, et al. Towards query efficient black-box attacks: An input-free perspective//Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec). Toronto, Canada, 2018: 13-24

[35] Ilyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018: 2142-2151

[36] Alzantot M, Sharma Y, Chakraborty S, et al. GenAttack: Practical black-box attacks with gradient-free optimization//Proceedings of the Genetic and Evolutionary Computation Conference (GECCO). Prague, Czech Republic, 2019: 1111-1119

[37] Huang L, Yu T. TAGA: A transfer-based black-box adversarial attack with genetic algorithms//Proceedings of the 2022 Genetic and Evolutionary Computation Conference (GECCO). Boston, USA, 2022: 712-720

[38] Wang L, Yang K, Wang W, et al. MGAAttack: Toward more query-efficient black-box attack by microbial genetic algorithm//Proceedings of the 28th ACM International Conference on Multimedia (ACM MM). Seattle, USA, 2020: 2229-2236

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

[40] Jere M, Hitaj B, Ciocarlie G F, Koushanfar F. Scratch that! An evolution-based adversarial attack against neural networks. http://arxiv.org/abs/1912.02316, 2019

[41] Li C, Wang H, Zhang J, et al. An approximated gradient sign method using differential evolution for black-box adversarial attack. IEEE Transactions on Evolutionary Computation, 2022, 26(5): 976-990

[42] Ghosh A, Mullick S S, Datta S, et al. A black-box adversarial attack strategy with adjustable sparsity and generalizability for deep image classifiers. Pattern Recognition, 2022, 122: 108279

[43] Wei X, Guo Y, Li B. Black-box adversarial attacks by manipulating image attributes. Information Sciences, 2021, 550: 285-296

[44] Liu H, Zhao B, Ji M, et al. GreedyFool: Multi-factor imperceptibility and its application to designing a black-box adversarial attack. Information Sciences, 2022, 613: 717-730

[45] Deng Y, Zhang C, Wang X. A multi-objective examples generation approach to fool the deep neural networks in the black-box scenario//Proceedings of the 4th IEEE International Conference on Data Science in Cyberspace (DSC). Hangzhou, China, 2019: 92-99

[46] Wang J, Yin Z, Jiang J, et al. PISA: Pixel skipping-based attentional black-box adversarial attack. Computers and Security, 2022, 123: 102947

[47] Williams P N, Li K. Black-box sparse adversarial attack via multi-objective optimisation CVPR proceedings//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada, 2023: 12291-12301

[48] Kuang X, Liu H, Wang Y, et al. A CMA-ES-based adversarial attack on black-box deep neural networks. IEEE Access, 2019, 7: 172938-172947

[49] Li Z, Cheng H, Cai X, et al. SA-ES: Subspace activation evolution strategy for black-box adversarial attacks. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(3): 780-790

[50] Qiu H, Custode L L, Iacca G. Black-box adversarial attacks using evolution strategies//Proceedings of the Genetic and Evolutionary Computation Conference (GECCO). Lille, France, 2021: 1827-1833

[51] Huang Li-Feng, Zhuang Wen-Zi, Liao Yong-Xian, Liu Ning. Black-box adversarial attack method based on evolution strategy and attention mechanism. Journal of Software, 2021, 32(11): 3512-3529(in Chinese)
(黄立峰, 庄文梓, 廖泳贤, 刘宁. 一种基于进化策略和注意力机制的黑盒对抗攻击算法. 软件学报, 2021, 32(11): 3512-3529)

[52] Dong X, Zhang W, Yu N. CAAD 2018: Powerful none-access black-box attack based on adversarial transformation network. http://arxiv.org/abs/1811.01225, 2018

[53] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks//Proceedings of the 2018 AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 2687-2695

[54] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 2818-2826

[55] Cui W, Li X, Huang J, et al. Substitute model generation for black-box adversarial attack based on knowledge distillation//Proceedings of the IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates, 2020: 648-652

[56] Duan M, Li K, Xie L, et al. Towards multiple black-boxes attack via adversarial example generation network//Proceedings of the ACM Multimedia Conference (ACM MM). Virtual Event, China, 2021: 264-272

[57] Yuan Z, Zhang J, Jia Y, et al. Meta gradient adversarial attack//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 7728-7737

[58] Zhang C, Benz P, Karjauv A, et al. Investigating top-k white-box and transferable black-box attack//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 15064-15073

[59] Zhu J, Dai F, Yu L, et al. Attention-guided transformation-invariant attack for black-box adversarial examples. International Journal of Intelligent Systems, 2022, 37(5): 3142-3165

[60] Huang L, Wei S, Gao C, Liu N. Cyclical adversarial attack pierces black-box deep neural networks. Pattern Recognition, 2022, 131: 108831

[61] Lin C, Han S, Zhu J, et al. Sensitive region-aware black-box adversarial attacks. Information Sciences, 2023, 637: 118929

[62] Xiao C, Li B, Zhu J, et al. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden, 2018: 3905-3911

[63] Feng Y, Wu B, Fan Y, et al. Boosting black-box attack with partially transferred conditional adversarial distribution//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 15074-15083

[64] Yuan J, He Z. Consistency-sensitivity guided ensemble black-box adversarial attacks in low-dimensional spaces//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 7758-7766

[65] Wang W, Qian X, Fu Y, Xue X. DST: Dynamic substitute training for data-free black-box attack//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 14341-14350

[66] Gao X, Tan Y, Jiang H, et al. Boosting targeted black-box attacks via ensemble substitute training and linear augmentation. Applied Sciences, 2019, 9(11): 1-14

[67] Ding K, Liu X, Niu W, et al. A low-query black-box adversarial attack based on transferability. Knowledge-Based Systems, 2021, 226: 107102

[68] Huan Z, Wang Y, Zhang X, et al. Data-free adversarial perturbations for practical black-box attack//Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Singapore, 2020: 127-138

[69] Duan M, Li K, Deng J, et al. A novel multi-sample generation method for adversarial attacks. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(4): 112:1-112:21

[70] Yang C, Kortylewski A, Xie C, et al. PatchAttack: A black-box texture-based attack with reinforcement learning//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 681-698

[71] Wei X, Guo Y, Yu J, Zhang B. Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 9041-9054

[72] Fu J, Sun J, Wang G. Boosting black-box adversarial attacks with meta learning//Proceedings of the 2022 41st Chinese Control Conference (CCC). Hefei, China, 2022: 1-12

[73] Yin F, Zhang Y, Wu B, et al. Generalizable black-box adversarial attack with meta learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access), 2023: 1-13

[74] Ji Y, Ding J, Chen Z, et al. Simulator attack+for black-box adversarial attack//Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP). Bordeaux, France, 2022: 636-640

[75] Hu C, Xu H, Wu X. Substitute meta-learning for black-box adversarial attack. IEEE Signal Processing Letters, 2022, 29: 2472-2476

[76] Shukla S N, Sahu A K, Willmott D, Kolter J Z. Black-box adversarial attacks with Bayesian optimization. http://arxiv.org/abs/1909.13857, 2019

[77] Ru B, Cobb A D, Blaas A, Gal Y. BayesOpt adversarial attack//Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia, 2020: 1-16

[78] Moon S, An G, Song H O. Parsimonious black-box adversarial attacks via efficient combinatorial optimization//Proceedings of the 36th International Conference on Machine Learning (ICML). 2019: 4636-4645

[79] Tran H, Lu D, Zhang G. Exploiting the local parabolic landscapes of adversarial losses to accelerate black-box adversarial attack//Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel, 2022: 317-334

[80] Zhang S, Gao H, Shu C, et al. Black-box Bayesian adversarial attack with transferable priors. Machine Learning, 2022, 111(10): 1-18

[81] Fan Y, Wu B, Li T, et al. Sparse adversarial attack via perturbation factorization//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 35-50

[82] Yu M, Sun S. Natural black-box adversarial examples against deep reinforcement learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI). Virtual Event, 2022: 8936-8944

[83] Huang Z, Huang Y, Zhang T. CorrAttack: Black-box adversarial attack with structured search. https://arxiv.org/abs/2010.01250, 2020

[84] Andriushchenko M, Croce F, Flammarion N, Hein M. Square attack: A query-efficient black-box adversarial attack via random search//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 484-501

[85] Dai Z, Liu S, Li Q, Tang K. Saliency attack: Towards imperceptible black-box adversarial attack. ACM Transactions on Intelligent Systems and Technology, 2023, 14(3): 45:1-45:20

[86] Pomponi J, Scardapane S, Uncini A. Pixle: A fast and effective black-box attack based on rearranging pixels//Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN). Padua, Italy, 2022: 1-7

[87] Chen Z, Li B, Wu S, et al. Query-efficient decision-based black-box patch attack. IEEE Transactions on Information Forensics and Security, 2023, 18: 5522-5536

[88] Chen W, Zhang Z, Hu X, Wu B. Boosting decision-based black-box adversarial attacks with random sign flip//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 276-293

[89] Croce F, Andriushchenko M, Singh N D, et al. Sparse-RS: A versatile framework for query-efficient sparse black-box adversarial attacks//Proceedings of the 2022 AAAI Conference on Artificial Intelligence (AAAI). 2022: 6437-6445

[90] Bu L, Zhao Z, Duan Y, Song F. Taking care of the discretization problem: A comprehensive study of the discretization problem and a black-box adversarial attack in discrete integer domain. IEEE Transactions on Dependable and Secure Computing, 2022, 19(5): 3200-3217

[91] Yan Z, Guo Y, Liang J, Zhang C. Policy-driven attack: Learning to query for hard-label black-box adversarial examples//Proceedings of the 9th International Conference on Learning Representations (ICLR). Austria, 2021: 1-15

[92] Huang Y, Zhou Y, Hefenbrock M, et al. Universal distributional decision-based black-box adversarial attack with reinforcement learning. https://doi.org/10.48550/arXiv.2211.08384, 2022

[93] Zhou L, Cui P, Zhang X, et al. Adversarial eigen attack on black-box models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 15254-15262

[94] Guo C, Gardner J R, You Y, et al. Simple black-box adversarial attacks//Proceedings of the 36th International Conference on Machine Learning (ICML). 2019: 2484-2493

[95] Rahmati A, Moosavi-Dezfooli S, Frossard P, Dai H. GeoDA: A geometric framework for black-box adversarial attacks//Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 8443-8452

[96] Maho T, Furon T, Merrer E L. SurFree: A fast surrogate-free black-box attack//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 10430-10439

[97] Shi Y, Han Y, Hu Q, et al. Query-efficient black-box adversarial attack with customized iteration and sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2226-2245

[98] Kim B C, Yu Y, Ro Y M. Robust decision-based black-box adversarial attack via coarse-to-fine random search//Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA, 2021: 3048-3052

[99] Liu Hao, Zhang Ze-Hui, Xia Xiao-Fan, Gao Tie-Gang. A fast black box boundary attack algorithm based on geometric detection. Journal of Computer Research and Development, 2023, 60(2): 435-447(in Chinese)
(刘昊，张泽辉，夏晓帆，高铁杠. 一种基于几何探测的快速黑盒边界攻击算法. 计算机研究与发展，2023，60(2)：435-447)

[100] Li J, Ji R, Liu H, et al. Projection & probability-driven black-box attack//Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 359-368

[101] Zhang Q, Zhang C, Li C, et al. Practical no-box adversarial attacks with training-free hybrid image transformation. arXiv preprint arXiv:2203.04607. https://doi.org/10.48550/arXiv.2203.04607, 2022

[102] Bai Y, Wang Y, Zeng Y, et al. Query efficient black-box adversarial attack on deep neural networks. Pattern Recognition, 2023, 133: 109037

[103] Wang L, Zhang H, Yi J, et al. Spanning attack: Reinforce black-box attacks with unlabeled data. Machine Learning, 2021, 109(12): 2349-2368

[104] Shukla S N, Sahu A K, Willmott D, Kolter J Z. Simple and efficient hard label black-box adversarial attacks in low query budget regimes//Proceedings of the 27th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). Singapore, 2021: 1461-1469

[105] Na D, Ji S, Kim J. Unrestricted black-box adversarial attack using GAN with limited queries//Proceedings of the 2022 European Conference on Computer Vision Workshops (ECCV Workshops). Tel Aviv, Israel, 2022: 467-482

[106] Yuan S, Zhang Q, Gao L, et al. Natural color fool: Towards boosting black-box unrestricted attacks//Proceedings of the 36th Conference and Workshop on Neural Information Processing Systems (NIPS). 2022: 7546-7560

[107] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, USA, 2015: 1-11

[108] Lin J, Xu L, Liu Y, Zhang X. Black-box adversarial sample generation based on differential evolution. Journal of Systems and Software, 2020, 170: 110767

[109] Zhang Y, Shin S-Y, Tan X, Xiong B. A self-adaptive approximated-gradient-simulation method for black-box adversarial sample generation. Applied Sciences, 2022, 13(3): 1298-1321

[110] Narodytska N, Kasiviswanathan S P. Simple black-box adversarial attacks on deep neural networks//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 1310-1318

[111] Bhagoji A N, He W, Li B, Song D. Practical black-box attacks on deep neural networks using efficient query mechanisms//Proceedings of the 15th European Conference on Computer Vision (ECCV). Glasgow, UK, 2018: 158-174

[112] Tu C, Ting P, Chen P, et al. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks//Proceedings of the 2019 AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019: 742-749

[113] Zhao P, Liu S, Chen P, et al. On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 121-130

[114] Zhao P, Chen P, Wang S, Lin X. Towards query-efficient black-box adversary with zeroth-order natural gradient descent //Proceedings of the 2020 AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 6909-6916

[115] Cheng M, Singh S, Chen P H, et al. Sign-OPT: A query-efficient hard-label adversarial attack//Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia, 2020: 1-16

[116] Dong Y, Cheng S, Pang T, et al. Query-efficient black-box adversarial attacks guided by a transfer-based prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9536-9548

[117] Du J, Zhang H, Zhou J T, et al. Query-efficient meta attack to deep neural networks//Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia, 2020: 1-15

[118] Zhao C, Ni B, Mei S. Explore adversarial attack via black box variational inference. IEEE Signal Processing Letters, 2022, 29: 2088-2092

[119] Chen J, Zhou D, Yi J, Gu Q. A Frank-Wolfe framework for efficient and effective adversarial attacks//Proceedings of the 2020 AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 3486-3494

[120] Gragnaniello D, Marra F, Verdoliva L, Poggi G. Perceptual quality-preserving black-box attack against deep learning image classifiers. Pattern Recognition Letters, 2021, 147: 142-149

[121] Carlini N, Wagner D A. Towards evaluating the robustness of neural networks//Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P). San Jose, USA, 2017: 39-57

[122] LeCun Y, Boser B E, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural Computation, 2021, 1(4): 541-551

[123] Lin M, Chen Q, Yan S. Network in network//Proceedings of the 2nd International Conference on Learning Representations (ICLR). Banff, Canada, 2014: 1-10

[124] Tramer F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses//Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada, 2018: 1-20

[125] Springenberg J T, Dosovitskiy A, Brox T, Riedmiller M A. Striving for simplicity: The all convolutional net//Proceedings of the 3rd International Conference on Learning Representations (ICLR Workshop). San Diego, USA, 2015: 1-14

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

[127] Girshick R B. Fast R-CNN//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1440-1448

[128] Ren S, He K, Girshick R B, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[129] Dai J, Li Y, He K, Sun J. R-FCN: Object detection via region-based fully convolutional networks//Proceedings of the 2016 Annual Conference on Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 379-387

[130] Lin T, Goyal P, Girshick R B, et al. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327

[131] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector//Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands, 2016: 21-37

[132] Zhang S, Wen L, Bian X, et al. Single-shot refinement neural network for object detection//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 4203-4212

[133] Liu S, Huang D, Wang Y. Receptive field block net for accurate and fast object detection//Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 404-419

[134] Redmon J, Divvala S K, Girshick R B, Farhadi A. You only look once: Unified, real-time object detection//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 779-788

[135] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger //Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 6517-6525

[136] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767. http://arxiv.org/abs/1804.02767, 2018

[137] Ge Z, Liu S, Wang F, et al. YOLOX: Exceeding YOLO series in 2021. https://arxiv.org/abs/2107.08430, 2021

[138] Law H, Deng J. CornerNet: Detecting objects as paired keypoints//Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 765-781

[139] Law H, Teng Y, Russakovsky O, Deng J. CornerNet-Lite: Efficient keypoint based object detection//Proceedings of the 31st British Machine Vision Conference (BMVC). UK, 2020: 1-15

[140] Duan K, Bai S, Xie L, et al. CenterNet: Keypoint triplets for object detection//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea, 2019: 6568-6577

[141] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 850-859

[142] Tian Z, Shen C, Chen H, He T. FCOS: Fully convolutional one-stage object detection//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 9626-9635

[143] Kong T, Sun F, Liu H, et al. FoveaBox: Beyound anchor-based object detection. IEEE Transactions on Image Processing, 2020, 29: 7389-7398

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

[145] Lyu H, Tan Y, Xue Y, et al. A CMA-ES-based adversarial attack against black-box object detectors. Chinese Journal of Electronics, 2021, 30(3): 406-412

[146] Xie C, Wang J, Zhang Z, et el. Adversarial examples for semantic segmentation and object detection//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 1378-1387

[147] Liu X, Yang H, Liu Z, et al. DPATCH: An adversarial patch attack on object detectors//Proceedings of the 2019 AAAI Conference on Artificial Intelligence (AAAI Workshop). Honolulu, USA, 2019: 1-8

[148] Brown T B, Mane D, Roy A, et al. Adversarial patch. http://arxiv.org/abs/1712.09665, 2017

[149] Zhang Q, Zhao Y, Wang Y, et al. Towards cross-task universal perturbation against black-box object detectors in autonomous driving. Computer Networks, 2020, 180: 107388

[150] Wang Y, Wang K, Zhu Z, Wang F. Adversarial attacks on faster R-CNN object detector. Neurocomputing, 2020, 382: 87-95

[151] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada, 2018: 1-23

[152] Wu Z, Lim S, Davis L S, Goldstein T. Making an invisibility cloak: Real world adversarial attacks on object detectors//Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 1-17

[153] Li D, Zhang J, Huang K. Universal adversarial perturbations against object detection. Pattern Recognition, 2021, 110: 107584

[154] Liang S, Wei X, Cao X. Generate more imperceptible adversarial examples for object detection//Proceedings of the ICML 2021 Workshop on Adversarial Machine Learning (ICML). 2021: 1-5

[155] Lapid R, Sipper M. Patch of invisibility: Naturalistic black-box adversarial attacks on object detectors. abs/2303.04238, 2023. doi:10.48550/ARXIV.2303.04238

[156] Lu Y, Jia Y, Wang J, et al. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 937-946

[157] Bao J. Sparse adversarial attack to object detection. arXiv preprint arXiv:2012.13692. https://arxiv.org/abs/2012.13692, 2020

[158] Liang S, Wu B, Fan Y, et al. Parallel rectangle flip attack: A query-based black-box attack against object detection// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 7677-7687

[159] Kuang X, Gao X, Wang L, et al. A discrete cosine transform-based query efficient attack on black-box object detectors. Information Sciences, 2021, 546: 596-607

[160] Everingham M, Gool L V, Williams C K I, et al. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338

[161] Everingham M, Eslami S M A, Gool L V, et al. The PASCAL visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1): 98-136

[162] Lin T, Maire M, Belongie S J, et al. Microsoft COCO: Common objects in context//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland, 2014: 740-755

[163] Codella N C F, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). http://arxiv.org/abs/1902.03368, 2019

[164] Fischer V, Kumar M C, Metzen J H, Brox T. Adversarial examples for semantic image segmentation//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France, 2017: 1-4

[165] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France, 2017: 1-14

[166] Arnab A, Miksik O, Torr P H S. On the robustness of semantic segmentation models to adversarial attacks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(12): 3040-3053

[167] Shen G, Mao C, Yang J, Ray B. AdvSPADE: Realistic unrestricted attacks for semantic segmentation. https://arxiv.org/abs/1910.02354, 2019

[168] Nakka K K, Salzmann M. Indirect local attacks for context-aware semantic segmentation networks//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 611-628

[169] He Y, Rahimian S, Schiele B, Fritz M. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 519-535

[170] Gu J, Zhao H, Tresp V, Torr P H S. Adversarial examples on segmentation models can be easy to transfer. https://arxiv.org/abs/2111.11368, 2021

[171] Zhang C, Benz P, Karjauv A, Kweon I S. Data-free universal adversarial perturbation and black-box attack//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 7848-7857

[172] Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 2574-2582

[173] Shibata H, Hanaoka S, Nomura Y, et al. On the matrix-free generation of adversarial perturbations for black-box attacks. https://arxiv.org/abs/2002.07317, 2020

[174] Li S, Huang G, Xu X, Lu H. Query-based black-box attack against medical image segmentation model. Future Generation Computer Systems, 2022, 133: 331-337

[175] Cui X, Chang S, Li C, et al. DEAttack: A differential evolution based attack method for the robustness evaluation of medical image segmentation. Neurocomputing, 2021, 465: 38-52

[176] Yin H, Wang R, Liu B, Yan J. On adversarial robustness of semantic segmentation models for automated driving// Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV). Aachen, Germany, 2022: 867-873

[177] Kang X, Song B, Du X, Guizani M. Adversarial attacks for image segmentation on multiple lightweight models. IEEE Access, 2020, 8: 31359-31370

[178] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 3213-3223

[179] Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision, 2019, 127(3): 302-321

[180] Yu F, Chen H, Wang X, et al. BDD100K: A diverse driving dataset for heterogeneous multitask learning//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 2633-2642

[181] Neuhold G, Ollmann T, Bulo S R, Kontschieder P. The mapillary vistas dataset for semantic understanding of street scenes//Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 5000-5009

[182] Behley J, Garbade M, Milioto A, et al. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 9296-9306

[183] Brostow G J, Shotton J, Fauqueur J, Cipolla R. Segmentation and recognition using structure from motion point clouds//Proceedings of the 10th European Conference on Computer Vision (ECCV). Marseille, France, 2008: 44-57

[184] Van Ginneken B, Stegmann M B, Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. Medical Image Analysis, 2006, 10(1): 19-40

[185] Pena-Betancor C, Gonzalez-Hernandez M, Fumero-Batista F, et al. Estimation of the relative amount of hemoglobin in the cup and neuroretinal rim using stereoscopic color fundus images. Investigative Ophthalmology & Visual Science, 2015, 56(3): 1562-1568

[186] Shiraishi J, Katsuragawa S, Ikezoe J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology, 2000, 174(1): 71-74

[187] Liu B, Guo Y, Jiang J, et al. Multi-view correlation based black-box adversarial attack for 3D object detection//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD). Virtual Event, Singapore, 2021: 1036-1044

[188] Carlini N, Farid H. Evading deepfake-image detectors with white- and black-box attacks//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 2804-2813

[189] Wang X, Zhang Z, Wu B, et al. Prototype-supervised adversarial network for targeted attack of deep hashing//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Virtual, 2021: 16357-16366

[190] Li X, Li J, Chen Y, et al. QAIR: Practical query-efficient black-box attacks for image retrieval//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Virtual, 2021: 3330-3339

[191] Wang Y, Zhang K, Lu K, et al. Practical black-box adversarial attack on open-set recognition: Towards robust autonomous driving. Peer-to-Peer Networking and Applications, 2023, 16(1): 295-311

[192] Zhu L, Wang T, Li J, et al. Efficient query-based black-box attack against cross-modal hashing retrieval. ACM Transactions on Information Systems, 2023, 41(3): 54:1-54:25

[193] Zhu Y, Miao C, Hajiaghajani F, et al. Adversarial attacks against LiDAR semantic segmentation in autonomous driving//Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys). Coimbra, Portugal, 2021: 329-342

[194] Dong J, Wang Y, Lai J, Xie X. Restricted black-box adversarial attack against deepfake face swapping. IEEE Transactions on Information Forensics and Security, 2023, 18: 2596-2608

[195] Dong Y, Su H, Wu B, et al. Efficient decision-based black-box adversarial attacks on face recognition//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 7714-7722

[196] Sun J, Cao Y, Chen Q A, Mao Z M. Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures//Proceedings of the 29th USENIX Security Symposium (USENIX Security). Boston, USA, 2020: 877-894

[197] Jia S, Song Y, Ma C, Yang X. IoU attack: Towards temporally coherent black-box adversarial attack for visual object tracking//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Virtual, 2021: 6709-6718

[198] Li H, Li H, Zhang H, Yuan W. Black-box attack against handwritten signature verification with region-restricted adversarial perturbations. Pattern Recognition, 2021, 111: 107689

[199] Wang R, Juefei-Xu F, Guo Q, et al. Amora: Black-box adversarial morphing attack//Proceedings of the 28th ACM International Conference on Multimedia (ACM MM). Virtual, 2020: 1376-1385

[200] Nguyen V N K, Terada T, Nishigaki M, Ohki T. Examining of shallow autoencoder on black-box attack against face recognition//Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Tokyo, Japan, 2021: 1775-1780

[201] Wei Z, Chen J, Zhang H, et al. Adaptive temporal grouping for black-box adversarial attacks on videos//Proceedings of the International Conference on Multimedia Retrieval (ICMR). Trento, Italy, 2022: 587-593

[202] Yin X, Ruan W, Fieldsend J E. DIMBA: Discretely masked black-box attack in single object tracking. Machine Learning (Open Access), 2022: 1-19

WU Yang, Ph. D. candidate. His main research interests include black-box adversarial attacks and robustness analysis of deep learning models.

LIU Jing, Ph. D., professor, Ph. D. supervisor. His main research interests include trustworthy artificial intelligence, software reliability engineering, cloud computing and big data analytics.

Background

The image domain task is one of the main computer vision tasks currently, which aims at automating the analysis and processing of image data through computer technology. As deep learning is widely used in various complex scenarios and tasks, deep neural networks are used to process complex tasks in the image domain. However, existing studies have found that adversarial samples with small perturbations can lead to an incorrect inference of neural networks, resulting in model performance degradation. Therefore, the problem of the robustness of deep learning has gradually been valued by researchers since adversarial samples were found.

The adversarial attack is one of the important technologies to study the internal principle of deep learning robustness, which can be divided into white-box and black-box attacks according to whether the model structure and parameter information can be obtained. However, it is difficult to obtain the specific model parameter in the real environment, so the black-box attack is more suitable for the actual task. Accordingly, this paper classifies and summarizes relevant literature from two perspectives: the currently popular image domain tasks and the existing black-box adversarial attack algorithms. By introducing the black-box optimization problem, the current research status is analyzed from the perspective of black-box optimization, including the introduction of attack methods and performance analysis. Finally, possible development trends of the future for black-box adversarial attack is discussed.

Our team has been committed to the research on adversarial attack and defense and has achieved certain research results in the combination of model interpretability and adversarial attack. Therefore, we have collected a large amount of information and literature in the process of related research, but most of them are mainly in English. We share and publish this paper in Chinese to facilitate domestic researchers to learn the black-box anti-attack technology, and also hope to draw the attention of more researchers.