分类号： TP311.5 单位代码: 10335

密 级： 无 学 号:



硕士学位论文



**中文论文题目 :基于多尺度特征网格群动态优化与熵控**

**对抗网络驱动的目标检测攻击方法研究**

**英文论文题目：Research on Object Detection Attack Methods**

**Driven by Dynamic Optimization of Multi-scale Feature Grid Clusters and Entropy-controlled Adversarial Networks**

申请人姓名：指导教师： 合作导师：

专业学位类别： 工程硕士 专业学位领域： 电气信息技术 所在学院:

### 论文提交日期 年 月 日

2

### 摘 要

随着计算机视觉技术的迅速发展，目标检测在自动驾驶、智能监控和医疗影像分析等领域的应用愈加广泛。然而，基于深度学习的目标检测模型在面对对抗样本时的脆弱性逐渐显露，严重影响了其安全性和可靠性。本文针对目标检测任务中的对抗样本生成方法进行了深入研究，提出了一种基于多尺度特征网格群动态优化的对抗攻击方法（DOG），以及其改进版本 DOGs。

DOG 方法借鉴了 C&W 对抗样本生成架构、DAG 模型激活值选取和 EA 粒子群优化策略，显著提高了对抗样本的生成效率、攻击成功率和泛化能力。针对目标识别模型（如 YOLOv5、SSD300、RetinaNet 等），在白盒条件下，DOG 方法的攻击成功率达到 97%以上，而在黑盒条件下也能保持 51%以上的攻击成功率，平均生成时间约为 2.4 秒。此外，DOG 方法生成的对抗样本不仅对目标检测模型有效，还对基于 ImageNet 数据集的图像分类模型（如 VGG19、ResNet50 和 Inception V3）造成了显著的攻击效果，黑盒条件下攻击成功率不低于 47%。

为进一步提升生成效率，DOGs 方法结合图像信息熵和对抗网络特性，通过训练特殊的 GAN 网络，实现了对抗样本生成时间的压缩，生成效率提升至 0.8 秒/张，同时保持了较好的攻击效果。实验结果表明，所提方法在对抗性、转移性和生成效率上均达到了良好的平衡，为目标检测模型及图像分类模型的安全性研究提供了新的思路和方法。

**关键词：**目标检测，对抗样本，信息熵，对抗网络，泛化性强，对抗性强

### Abstract

With the rapid development of computer vision technology, object detection has become increasingly widespread in fields such as autonomous driving, intelligent surveillance, and medical image analysis. However, the vulnerability of deep learning- based object detection models to adversarial samples has gradually emerged, severely affecting their safety and reliability. This paper conducts an in-depth study on the methods of generating adversarial samples in object detection tasks and proposes an adversarial attack method based on dynamic optimization of multi-scale feature grid clusters (DOG), as well as its improved version, DOGs.

The DOG method draws inspiration from the C&W adversarial sample generation architecture, DAG model activation value selection, and the EA particle swarm optimization strategy, significantly enhancing the efficiency of adversarial example generation and the success rate of attacks. For object detection models such as YOLOv5, SSD300, and RetinaNet, under white-box conditions, the attack success rate of the DOG method exceeds 97%, while maintaining over 51% under black-box conditions, with an average generation time of about 2.4 seconds. Furthermore, the adversarial examples generated by the DOG method not only effectively target object detection models but also cause significant attack effects on image classification models based on the ImageNet dataset, such as VGG19, ResNet50, and Inception V3, achieving a black-box attack success rate of no less than 47%.

To further enhance generation efficiency, the DOGs method combines image information entropy and adversarial network characteristics, training a specialized GAN network to compress the generation time of adversarial examples to 0.8 seconds per image while maintaining good attack effectiveness. Experimental results indicate that the proposed methods achieve a good balance in terms of adversarial effectiveness, transferability, and generation efficiency.

**Key Words：**Object Detection,Adversarial Samples,Information Entropy,

Adversarial Networks,Strong Generalization,Strong Adversariality

目 录

[摘 要 i](#_bookmark0)

[Abstract ii](#_bookmark1)

1. [绪论 3](#_bookmark2)
   1. [研究背景及意义 3](#_bookmark3)
   2. [国内外研究现状 4](#_bookmark4)
      1. [目标检测方法研究现状 5](#_bookmark5)
      2. [传统图像对抗样本生成方法 8](#_bookmark6)
      3. [面向目标检测任务的对抗攻击方法研究现状 16](#_bookmark7)
   3. [研究内容及组织结构 18](#_bookmark8)
      1. [研究内容 18](#_bookmark9)
      2. [关键技术 18](#_bookmark10)
      3. [章节安排 19](#_bookmark11)
2. [相关模型、算法及数据集介绍 20](#_bookmark12)

[2.1 引言 20](#_bookmark13)

* 1. [图像分类网络 20](#_bookmark14)
     1. [VGG19 网络 20](#_bookmark15)
     2. [ResNet50 网络 21](#_bookmark16)
     3. [Inception V3 网络 22](#_bookmark17)
  2. [目标检测网络 23](#_bookmark18)
     1. [YOLOv5 网络 23](#_bookmark19)

[2.3.2 SSD300 网络 25](#_bookmark20)

* + 1. [RetinaNet 网络 27](#_bookmark21)
    2. [FCOS 网络 28](#_bookmark22)
    3. [MaskRCNN 网络 30](#_bookmark23)
    4. [五种网络对比 31](#_bookmark24)
  1. [目标检测任务对抗攻击方法 33](#_bookmark25)
     1. [DAG 稠密对抗生成算法 33](#_bookmark26)
     2. [R-AP 鲁棒对抗扰动算法 33](#_bookmark27)
     3. [EA 蒸发攻击算法 34](#_bookmark28)
  2. [常用数据集 35](#_bookmark29)
     1. [ImageNet 数据集 36](#_bookmark30)
     2. [MS COCO 数据集 36](#_bookmark31)
  3. [评价指标 36](#_bookmark32)
     1. [图像分类任务攻击有效性评价指标 36](#_bookmark33)
     2. [目标检测任务的性能评价指标 37](#_bookmark34)
     3. [目标检测任务攻击有效性评价指标 38](#_bookmark35)
  4. [本章小结 39](#_bookmark36)

1. [基于多尺度特征网格群动态优化的目标检测对抗攻击方法 41](#_bookmark37)

[3.1 引言 41](#_bookmark38)

* 1. [DOG 攻击方法 42](#_bookmark39)
     1. [符号及问题描述 42](#_bookmark40)
     2. [DOG 攻击方法 42](#_bookmark41)
     3. [多尺度特征网格群 44](#_bookmark42)
     4. [损失函数 44](#_bookmark43)
     5. [FOCG 最优诱偏类别动态选择 46](#_bookmark44)
     6. [BRM 背景重置 46](#_bookmark45)
     7. [算法实现 47](#_bookmark46)
  2. [实验与结果分析 49](#_bookmark47)
     1. [实验基础搭建 49](#_bookmark48)
     2. [基于 SSD300 预训练网络的攻击有效性验证 50](#_bookmark49)
     3. [损失函数及功能模块消融实验 52](#_bookmark50)
     4. [迁移性验证实验 54](#_bookmark51)
     5. [和其他攻击算法的对比实验 56](#_bookmark52)
  3. [本章小结 57](#_bookmark53)

1. [融合图像信息熵和对抗网络特性的目标检测对抗攻击方法 59](#_bookmark54)

[4.1 引言 59](#_bookmark55)

* 1. [DOGs 攻击方法 59](#_bookmark56)
     1. [图像关键区域识别问题 59](#_bookmark57)
     2. [基于图像信息熵的关键区域识别方法 60](#_bookmark58)
     3. [DOGs 攻击方法及系统架构 63](#_bookmark59)
     4. [损失函数 66](#_bookmark60)
     5. [算法实现 67](#_bookmark61)
  2. [实验与结果分析 69](#_bookmark62)
     1. [实验基础搭建 69](#_bookmark63)
     2. [DOGs 攻击有效性验证 70](#_bookmark64)
     3. [DOGs 攻击迁移性验证 71](#_bookmark65)
  3. [本章小结 72](#_bookmark66)

1. [总结与展望 73](#_bookmark67)
   1. [工作总结 73](#_bookmark68)

[5.2 展望 74](#_bookmark69)

[参考文献 75](#_bookmark70)

# 1 绪论

## 研究背景及意义

目标检测任务在计算机视觉领域的应用日益广泛，涵盖了自动驾驶、智能监控、医疗影像分析等多个关键领域[1]。这些技术的发展大幅提高了生产效率和生活质量，为人类社会带来了诸多便利。来自 Fortune Business Insights 的《自动驾驶汽车市场规模、份额、趋势报告[2030]》表明，预计该市场将从 2023 年

的 19,211 亿美元增长到 2030 年的 136,324 亿美元，预测期内复合年增长率为 32.3%（见图 1-1）。而在这一领域中，目标检测技术是实现车辆自动感知周边环境的核心技术。如 YOLO（You Only Look Once）[2]和 Mask R-CNN[3]等经典网络已经在不同应用中的得以广泛应用，并展现出其强大的实时识别、检测能力。在智能监控中，目标检测被广泛用于分析城乡监控视频，提高公共安全。我国的“天网工程”就是一个例子，该系统利用目标检测技术实时监控和解析图像，显著提升了社会治安管理。目前，在医疗领域，目标检测技术在疾病诊断，如癌症检测方面展现了巨大的潜力，通过提高诊断的准确性和效率，推动了医疗行业的革新。

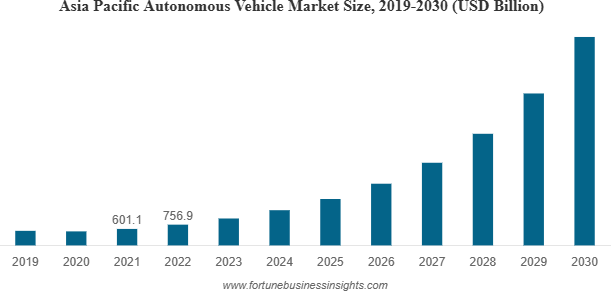


图 1-1 亚太地区自动驾驶汽车市场规模（2019-2030）

然而，随着目标检测技术的快速发展，其基于深度学习模型的安全性问题逐渐显露。最早追溯到 2013 年，Szegedy 等人发现深度神经网络在推理过程中会被经过特定扰动处理的图像误导，这些图像被称为对抗样本[4]。一些实验证实，添

加噪声数据可以导致知名模型如 Faster R-CNN 在识别公共交通标志时出现识别错误[5]，这可能导致自动驾驶系统产生严重的误判。近年来，针对这些安全挑战，研究人员开始探索生成对抗网络（GAN）[6]、对抗训练等技术，以提高模型对对抗样本的鲁棒性。

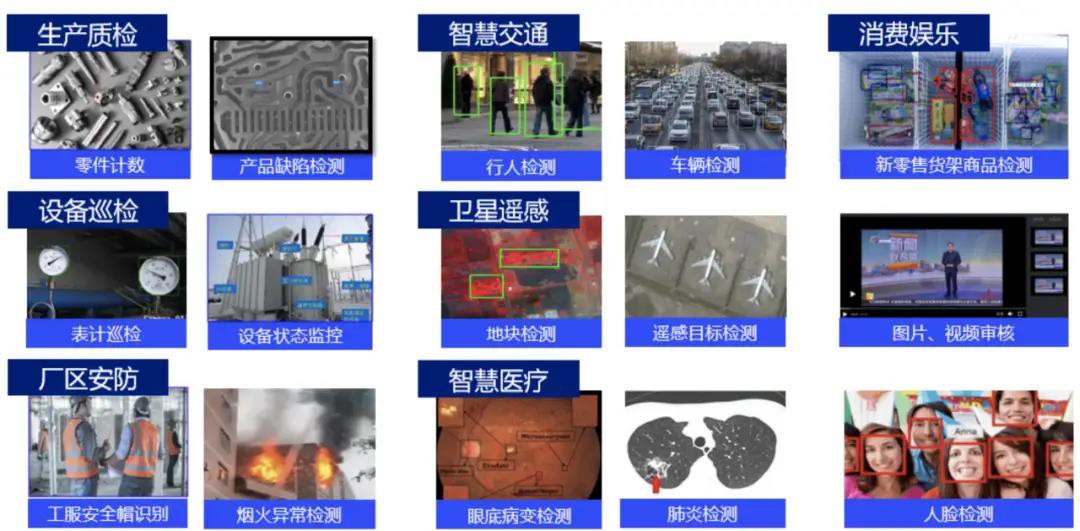


图 1-2 目标检测技术应用

因此，针对目标检测任务的对抗样本研究具有重要的理论和实际意义。从理论角度来看，深入研究对抗样本可以帮助我们更全面地理解深度学习模型的脆弱性及其内在工作机制；从实际应用角度，它有助于在高度依赖自动化和智能系统的社会中确保安全和稳定运行，同时能够进一步加强人工智能在诸如生产质检、设备巡检、厂区安防、交通管理等传统领域的拓展运用（见图 1-2）。通过制定有效的防御策略和增强模型鲁棒性，可以提高目标检测技术在实际应用中的安全性和可靠性，推动其在各个领域的深度应用和发展。此外，这一研究还可以为未来相关领域提供宝贵的参考和指导。例如，围绕该领域开发的新型对抗防御算法可以在金融、国防等其他敏感行业中推广应用。

## 国内外研究现状

随着人工智能的不断发展，高度自主化智能体的需求日益扩大，目标检测技术在多个应用领域的精度和效率得到了显著提升。得益于大规模图像数据集的训练，诸如 YOLO（You Only Look Once）、SSD（Single Shot MultiBox Detector）

[7]和 Faster R-CNN[8]等深度学习模型已成为目标检测研究的主流。这些模型通过高效的特征提取和处理，能够在复杂场景中实现实时目标检测，满足了物联网、自动驾驶及监控等领域的需求。

与此同时，针对目标检测任务的攻击算法也在不断迭代升级。传统的图像对抗样本生成方法，如基于梯度的方法（如 FGSM[9]和 PGD[10]）和优化方法（如 C&W 攻击[11]），虽然在分类任务中取得了一定成效，但在目标检测任务中的适应性和有效性往往不足。因此，针对目标检测的特定攻击方法应运而生，以便在保持模型输出的同时有效扰动检测结果。

本节将分别对目标检测方法、传统图像对抗样本生成方法以及面向目标检测任务的攻击方法进行综述。通过对这些领域的研究现状进行梳理，我们将重点分析当前技术的进展、存在的挑战以及未来的研究方向，为深入理解目标检测的安全性和鲁棒性提供参考。

### 目标检测方法研究现状

目标检测是计算机视觉领域的一项核心技术，主要用于识别图像中特定目标及其位置边界框[12]。目标检测不仅关注目标的存在与类型，还要求准确定位目标在图像中的位置，这通常通过边界框表示。随着深度学习技术的迅猛发展，目标检测方法逐步从传统基于人工特征的检测方法转向基于深度学习的自动检测。

* + - 1. 传统目标检测方法。传统的目标检测方法通常经历三个关键步骤：目标候选区域确定、目标特征提取与区域分类判定。这类方法依赖于手动提取的特征，其检测效果受制于特征的准确性。常用的特征提取方法包括梯度直方图（HOG）、局部二值模式（LBP）、哈尔特征、边缘检测及颜色直方图等，分类器则包括支持向量机（SVM）、随机森林和逻辑回归等[13]。然而，这些传统方法在计算上耗时，且提取的特征难以适应复杂的环境，导致性能受限。
      2. 基于深度学习的目标检测方法。深度学习的兴起带来了目标检测方法的革命，现代深度学习方法通过卷积神经网络（CNN）等结构，使特征提取和分类过程能够在同一框架内自动完成，从而极大地提高了检测的速度和准确性。根据实现阶段的不同，基于深度学习的目标检测方法可分为单阶段模型和双阶段模型。

1. 单阶段模型。单阶段模型通过一个阶段同时完成目标检测和分类。它快速预测输入图像中的目标边界框及对应的类别标签，具有较高的推理速度，适合于实时目标检测需求。代表性算法包括：

YOLO 系列（如 YOLOv3、YOLOv4、YOLOv5）：采用回归预测的方法，将目标检测转化为一个回归问题，从而实现快速处理。YOLO 通过在输出层直接预测目标的类别和框坐标，以实现高效目标检测。



图 1-3 基于 yolov7-tiny 的车辆识别示例

SSD（Single Shot MultiBox Detector）[14]：通过多尺度特征图进行目标检测，在不同尺度上同时生成多个边界框，从而实现高效的目标识别。SSD 结合了不同层的特征图，以支持对大小不一的目标进行检测。

FCOS（Fully Convolutional One-Stage Object Detection）[15]：一种完全卷积的单阶段目标检测器，摒弃了传统的 Anchor 机制。FCOS 通过在每个像素上预测目标的类别和边界框偏移，采用中心度量（centerness）来增强对真正目标框的关注。该方法的优点在于不需要预定义的锚框，因此减少了超参数的调试复杂性，并且在许多任务中实现了竞争力的性能。

RetinaNet[16]：结合了单阶段模型的高效性和双阶段模型的准确性。它使用了一种称为 Focal Loss 的损失函数，该函数旨在解决类别不平衡的问题，能够更

加关注难以检测的样本。RetinaNet 采用特征金字塔网络（FPN）来提取多尺度特征，使得模型在检测小目标时表现更为出色。

1. 双阶段模型。双阶段模型在检测过程中分为两个独立阶段：生成候选区域和分类与框回归。第一阶段生成候选区域，第二阶段则对这些区域进行分类和精确定位。典型模型包括：

Faster R-CNN：结合了区域建议网络和卷积神经网络的优势，首先生成候选区域，然后对每个区域进行目标分类和边界框回归。

Mask R-CNN：在 Faster R-CNN 的基础上扩展，增加了对目标分割的支持，能够同时进行目标检测和实例分割。

近年来，新一代 YOLO（YOLOv9、YOLOv10）和 EfficientDet 等模型在速度与精度之间达成新的平衡，成为研究热点。单阶段检测器如 YOLO 和 SSD 的全图回归思想简化了目标检测过程，非常适合实时应用，如无人驾驶和视频监控。

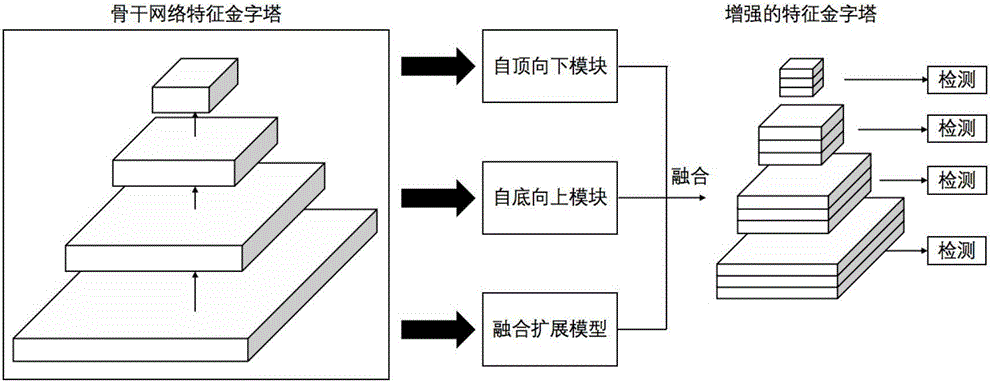
* + - 1. 现代技术的发展。在模型结构方面，特征金字塔网络（FPN）[17]被引入目标检测模型以增强对多尺度物体的检测能力。特征金字塔网络能够有效捕捉不同尺寸的目标，提升小目标的检测性能。同时，Transformer 架构的引入（如 DETR）通过全局自注意力机制改变了传统检测方法的范式，提高了复杂场景下的性能，开启了新的研究方向。针对特定应用的优化探索也在持续进行，包括面向边缘设备的模型压缩与加速技术、基于图像内容的自适应检测技术以及跨域目标检测，推动目标检测技术在移动设备和物联网环境中的高效应用。

图 1-4 特征金字塔结构

尽管目标检测技术在精度和速度上取得了显著进展，但仍然面临诸多挑战，如复杂背景下的检测困难、遮挡物体的识别、不同光照条件下的鲁棒性以及实时 处理的持续需求等。这些问题需要通过新的算法设计与更复杂的网络架构来克服。因此，未来的目标检测研究将继续聚焦于提升精度、适应性和速度，以满足不断 增长的应用需求与挑战。

### 传统图像对抗样本生成方法

对抗样本图像生成方法是近年来机器学习和计算机视觉领域的一个重要研究方向。对抗样本是指通过对原始输入数据进行微小的扰动，使得深度学习模型产生错误分类的样本。

对于图像对抗样本的研究可以追溯到 2014 年，当时 Szegedy 等人提出了 L- BFGS 方法。紧接着在 2015 年，Goodfellow 等人提出了单步攻击方法 FGSM（Fast Gradient Sign Method），并在此基础上衍生出多种改进方法。同年，也开始出现用于防御对抗样本的策略，如去噪声数据预处理方法。自此，攻防对抗进入了激烈的碰撞阶段。

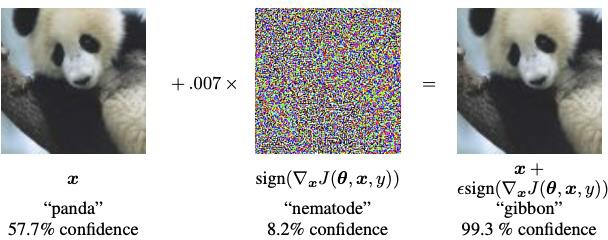


图 1-5 基于 FGSM 攻击方法将熊猫误识别为长臂猿

在 2016 年和 2017 年，攻防对抗方法大量涌现。攻方提出诸如 DeepFool、 JSMA、C&W 和 UAP 等多种攻击方法，这些方法在攻击效果（例如攻击成功率、定向攻击能力和泛化能力）与攻击依托条件（如是否需要提前掌握目标模型结构、权重和训练集数据，是否需要目标模型梯度反馈权限）之间寻求平衡与独特性。同时，防御方法如基于 JPEG 压缩的数据预处理和梯度遮蔽等纷纷被提出，并首次探讨了对抗样本在物理世界中的实现与影响。

从 2018 年至 2021 年，研究人员在前述攻击方法的基础上进一步提出了单像素攻击、PGD、AutoZOO 等新方法，同时利用生成式对抗网络（GAN）生成对抗样本，例如 advGAN[18]、WGAN[19]等。这些方法的基本思想是通过训练一个固定的网络，使得任何良性图像输入都能生成其相应的对抗样本。基于 GAN 的对抗样本生成方法，对于那些引入输入梯度或特征分析等防御措施的网络，依然能够保持有效的攻击特性。

自 2021 年以来，研究的重点逐渐由数字域对抗转向物理域对抗。在这一背景下，衍生出针对物理域攻击的两类方法。第一类是构造特殊的对抗性贴图，附着在眼镜、衣服或其他物体上，以实现图像的错误分类，张等人[20]提出了一种新的对抗性贴图方法，该方法能够在动态场景中有效的改变目标检测和分类的结果;第二类是扰动现实化转换，即将干扰源以光线、光阵列或雨滴等现实世界中可见但具有随机特性的方式呈现，将这些现实对抗扰动作为掩膜添加到图像输入端，形成具有对抗性的全新图像，技术包括利用激光（Adversarial Laser Beam）[21]、对 抗 雨 滴 （ AdvRD ） [22] 以 及 风 格 化 转 换 （ GE-AdvGAN ） [23] 等 。



图 1-6 对抗雨滴(AdvRD)生成效果对比，a 为真实雨滴图，b 为生成的对抗雨滴图

下面从基于梯度下降、基于迭代优化、基于迁移学习、基于梯度估算、基于决策、基于 GAN 等 6 类方法入手，梳理总结常见的经典图像对抗样本生成方法。

(1)基于梯度下降。基于梯度的方法是最早提出的一类对抗样本生成方法，其基本思想是通过对模型损失函数𝐿(𝑦, 𝑦𝑡)进行梯度计算来生成对抗扰动，获得对抗样本图像。下面介绍四种常用方法：

A.FGSM 。快速梯度符号方法（ Fast Gradient Sign Method, FGSM ）由

Goodfellow 等人在 2014 年提出。其基本原理是沿着损失函数对输入图像的梯度方向添加一个小的扰动，以使深度学习模型产生错误的分类结果。具体来说，FGSM通过计算输入图像梯度的符号信息，确定扰动的方向，然后将一个小扰动值乘以梯度的符号信息添加到输入图像中。这样，即使在输入图像上进行微小的修改，也可以导致模型错误分类。

B.BIM[24]。基于迭代的方法（Basic Iterative Method, BIM）是对 FGSM 的改进。BIM 通过多次迭代地应用 FGSM 来生成对抗样本。在每次迭代中，BIM 都会计算当前图像的梯度，并沿着梯度方向添加小的扰动。通过多次迭代，BIM 可以生成更强的对抗样本。

C.PGD[25]。投射梯度下降（Projected Gradient Descent, PGD）是另一种基于梯度的对抗样本生成方法。PGD 在每次迭代中计算梯度并添加扰动后，还会将生成的对抗样本投射回原始输入图像的 ε-邻域内。这样可以确保生成的对抗样本不会偏离原始图像太远，从而提高对抗样本的可感知性。

D.MIM[26]。动量迭代方法（MIM）。动量迭代方法（Momentum Iterative Method, MIM）是对 BIM 的进一步改进。MIM 在每次迭代中引入了动量项，以加速对抗样本的生成过程。具体来说，MIM 在计算梯度时会考虑前几次迭代的梯度信息，从而使得生成的对抗样本更加稳定和有效。

1. 基于迭代优化。对抗样本的生成本质上是寻找最佳的对抗扰动。其基本思想是通过直接优化某个目标函数，选择最优来生成对抗样本。最早雏形出现在 Szegedy 等提出的 Box-constrained L-BFGS 方法中。
2. Carlini&Wagner 攻击（C&W）。C&W 由 Carlini 和 Wagner 在 2017 年提出，也是最为经典的基于优化的攻击方法。C&W 方法定义了不同的目标函数，通过实验数据集优化一个目标函数来生成对抗样本。该目标函数包括两个部分：一个是对抗样本与原始图像之间的距离，另一个是对抗样本被错误分类的置信度。通过最小化这个目标函数，C&W 攻击可以生成高效的对抗样本。

该方法，具有以下特点和优势。**高效性**：C&W 攻击生成的对抗样本对深度学习模型具有较高的攻击成功率，并且对原始样本的扰动较小，具有高效的攻击性

能。**鲁棒性**：C&W 方法考虑了生成对抗样本的错误分类置信度，生成的对抗样本能够对抗各种防御机制，具有较强的鲁棒性。**通用性**：C&W 攻击方法适用于不同类型的深度学习模型和任务，表现出通用性强的特点，具有广泛的应用前景。**攻击强度可控**：C&W 方法可以通过调节目标函数中的参数来控制攻击的强度，从而在保持对抗性的同时尽可能减小对原始样本的扰动。

1. One-pixel[27]。One-Pixel 方法是一种基于优化的对抗样本生成方法，由 Su等人在 2017 年提出。与其他基于优化的方法不同，One-Pixel 方法的核心思想是仅修改输入图像中的极少像素，即只修改一个像素，使深度学习模型产生错误分类结果。通过最小程度的扰动，One-Pixel 方法展示了对抗攻击的潜力，并突出了深度学习模型对输入数据中微小变化的敏感性。



图 1-7 基于 One-Pixel 方法的攻击效果示例

具体来说，One-Pixel 方法通过对输入图像的每个像素应用一个离散的色域进行优化，以找到使模型产生错误分类结果的最小扰动。这一过程可以通过优化算法（如遗传算法、模拟退火算法等）来实现，在保证只修改一个像素的情况下，使得深度学习模型无法正确分类输入图像。

该方法，具有以下特点和优势。**高效性**：One-Pixel 方法只修改一个像素，相比其他对抗样本生成方法所需修改多个像素的情况下，具有更高的效率和计算效率。**低维度搜索**：由于只修改一个像素，One-Pixel 方法的搜索空间较小，使得寻找对抗扰动的任务更为可行。**避免感知性扰动**：One-Pixel 方法在生成对抗样本时避免了大规模的感知性扰动，可以生成更加不引人瞩目的对抗样本。**挑战性**：One-Pixel 方法提出了一个具有挑战性的问题，即如何在保持对抗攻击成功的同时，最小化像素的修改量。

D.DeepFool[28]。DeepFool 是一种基于优化的对抗样本生成方法，由 Seyed- Mohsen Moosavi-Dezfooli 等人于 2016 年提出。该方法旨在通过最小化具有线性决策边界的模型中每个类别的决策边界之间的距离，以生成对抗样本。DeepFool方法通过迭代计算最小化距离的方向，使原始样本朝着被错误分类的方向移动，生成能够欺骗深度学习模型的对抗样本。

该方法，具有以下特点和优势。**高效性**：DeepFool 方法生成的对抗样本通常需要较少的迭代次数，具有较高的效率和速度。**小幅度扰动**：DeepFool 生成的对抗样本对原始样本的扰动较小，保持了样本的自然性和可识别性。**通用性**： DeepFool 方法适用于线性决策边界的模型，并且可以应用于不同类型的深度学习模型和任务。**鲁棒性**：DeepFool 攻击可以绕过某些防御机制，生成的对抗样本对抗不同形式的模型防御，具有一定的鲁棒性。**强对抗性**：DeepFool 方法能够生成高度对抗性的对抗样本，欺骗各种类型的深度学习模型。

1. 基于迁移学习。攻击者利用白盒攻击的方法对目标模型的替代模型发起攻击，生成具有迁移性的对抗样本并成功攻击目标模型。因此，基于迁移学习的方法，核心是找到合适的替代模型，以及提高对抗样本的迁移性。下面介绍三种典型方法：
2. Transferability of Adversarial Examples[29]。是一种基于迁移学习的对抗样本生成方法。该方法利用一个预训练的模型来生成对抗样本，然后将这些对抗样本应用于另一个目标模型。由于对抗样本在不同模型之间具有一定的迁移性，这种方法可以在不需要访问目标模型的情况下对其进行攻击。

该方法，具有以下特点和优势。**无需访问目标模型**：该方法不需要直接访问目标模型，只需利用预训练模型生成对抗样本即可。**高效性**：通过预训练模型生成对抗样本，减少了对目标模型的依赖，提高了攻击效率。**强迁移性**：生成的对抗样本在不同模型之间具有较好的迁移性，能够成功攻击多个目标模型。

1. Ensemble Adversarial Training[30]。是另一种基于迁移学习的对抗样本生成方法。该方法通过在训练过程中引入多个预训练模型来生成对抗样本，并将这些对抗样本用于训练目标模型。通过这种方式，目标模型可以学习到更多的对

抗样本特征，从而提高其对抗攻击的鲁棒性。

该方法，具有以下特点和优势。**鲁棒性提升**：通过引入多个预训练模型生成对抗样本，目标模型能够学习到更多的对抗样本特征，提高其对抗攻击的鲁棒性。**多样性**：利用多个预训练模型生成的对抗样本具有多样性，能够更全面地覆盖对抗样本空间。**增强防御能力**：目标模型在训练过程中接触到更多的对抗样本，提高了其防御对抗攻击的能力。

C.SMBEA (Serial-Mini-Batch-Ensemble-Attack)[31]是一种基于迁移学习的 对抗样本生成方法。该方法通过在多个小批量数据上进行序列化的对抗样本生成，并结合多个预训练模型的攻击结果，生成具有更高迁移性的对抗样本。

该方法，具有以下特点和优势。**序列化处理**：通过在多个小批量数据上进行序列化处理，逐步生成对抗样本，提高了对抗样本的质量和迁移性。**集成攻击**：结合多个预训练模型的攻击结果，生成的对抗样本具有更高的迁移性和攻击成功率。**高效性**：利用小批量数据进行处理，减少了计算资源消耗，提高了攻击效率。

1. 基于梯度估算。基于梯度估算的对抗样本生成方法是一类常用的对抗攻击方法，通过对目标模型的损失函数关于输入样本的梯度进行估算，生成对抗性扰动以误导模型。以下介绍两种典型方法：

A.ZOO (Zeroth Order Optimization)[32]。ZOO 是一种基于梯度估算的对抗样本生成方法，由 Chen et al.在 2017 年提出。ZOO 基于零阶优化，即不需要访问模型的具体参数，只需要查询模型输出并估计其梯度，即可生成对抗样本。ZOO 通过迭代优化输入样本以最小化损失函数，生成具有欺骗性的对抗样本。

该方法，具有以下特点和优势。**零阶优化**：不需要模型参数，只需查询模型输出和估算梯度即可生成对抗性扰动。**强鲁棒性**：ZOO 方法针对各种类型的深度学习模型都具有较强的攻击鲁棒性。**高效性**：通过基于梯度的优化方法，生成的对抗样本效果较好且高效。

B.AutoZoom[33]。AutoZoom 是基于梯度估算的对抗样本生成方法，由 Yang et al.在 2020 年提出。该方法利用梯度信息动态调整对抗性扰动的大小，并通过自适应学习控制对抗样本的生成过程。AutoZoom 方法能够有效地生成对抗样本，欺

骗深度学习模型。

该方法，具有以下特点和优势。**动态调整**：根据梯度信息动态调整对抗性扰动的大小，提高对抗样本生成的效率和成功率。**自适应学习**：采用自适应学习控制生成过程，使对抗样本更具迷惑性。**鲁棒性**：AutoZoom 方法生成的对抗样本具有较强的攻击鲁棒性，能够成功欺骗深度学习模型。

1. 基于决策。基于决策的对抗样本生成方法是一类利用模型的决策信息进行对抗攻击的方法。这些方法通过优化决策相关的指标或者约束，生成对抗性扰动以欺骗目标模型。以下介绍三种典型方法：

A.CMA-ES (Covariance Matrix Adaptation Evolution Strategy)[34]。CMA- ES 是一种基于决策的对抗样本生成方法，通过进化策略优化模型输出结果以生成对抗样本。CMA-ES 方法根据目标函数的梯度信息，动态地调整协方差矩阵以实现更好的收敛性能，在对抗攻击中具有较高的成功率。

该方法，具有以下特点和优势：**进化策略**：采用进化策略进行对抗样本生成，能够有效地优化模型输出结果。**动态调整**：根据梯度信息动态调整协方差矩阵，提高收敛性能和攻击成功率。**有效性**：CMA-ES 方法在生成对抗样本时表现出较高的效率和稳定性。

B.CAB (Conditional Adversarial Bits)[35]。CAB 是一种基于决策的对抗样本生成方法，由 Yuan et al.在 2019 年提出。该方法通过引入条件对抗比特表示样本的不同类别和决策信息，生成具有欺骗性的对抗样本。CAB 方法在对抗攻击中能够成功地误导深度学习模型。

该方法，具有以下特点和优势。**条件对抗比特**：采用条件对抗比特表示不同类别和决策信息，增强对抗样本的欺骗性。**高成功率**：CAB 方法在对抗攻击中取得了较高的成功率，并且能够针对不同模型进行攻击。**灵活性**：根据不同的决策信息生成对抗样本，具有较强的灵活性和泛化能力。

C.GeoDA (Geometry-inspired Distribution Alignment)[36]。GeoDA 是一种 基于决策的对抗样本生成方法，通过几何启发式原则对模型的输出分布进行对齐，生成具有迷惑性的对抗样本。GeoDA 方法利用分布对齐的方式生成对抗性扰动，

欺骗深度学习模型。

该方法，具有以下特点和优势。**分布对齐**：利用几何启发式原则对模型的输出分布进行对齐，生成迷惑性高的对抗样本。**模型欺骗**：GeoDA 方法能够成功地欺骗深度学习模型，并对模型的决策产生影响。**提升成功率**：通过几何启发式的分布对齐方法，提高了生成对抗样本的成功率和效果。

1. 基于 GAN 网络。基于生成对抗网络（GAN）的对抗攻击方法利用生成器和判别器之间的博弈来生成具有欺骗性的对抗样本，以迷惑深度学习模型。以下介绍五种典型的基于 GAN 的对抗攻击方法：

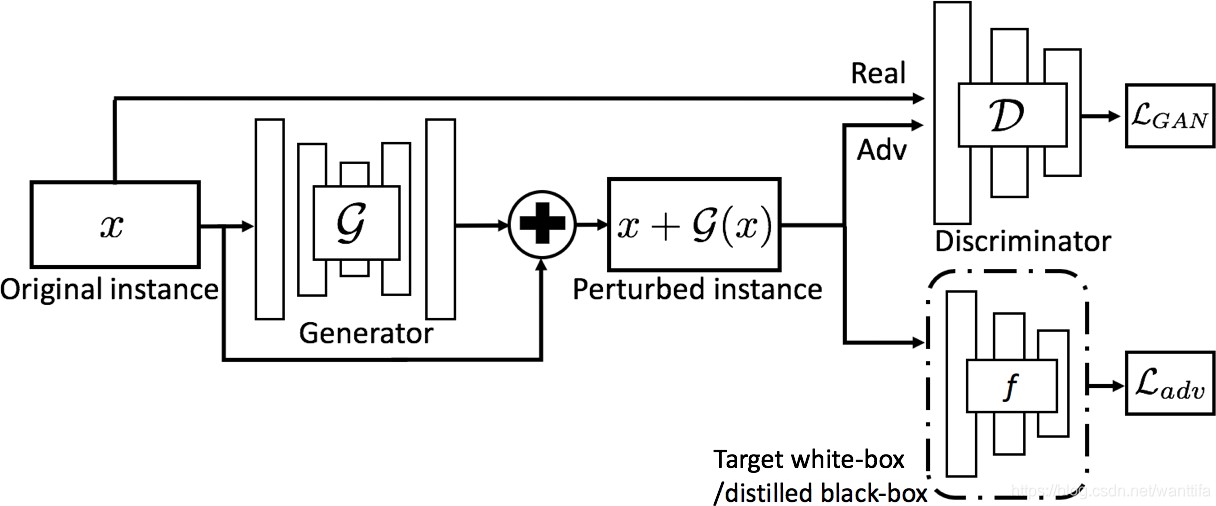
A.Adv-GAN (Adversarial Generative Adversarial Network)。adv-GAN 是一种基于对抗生成对抗网络的对抗攻击方法，由 Kurakin et al.在 2016 年提出。在 adv-GAN 中，对抗攻击者利用生成器生成伪造的对抗样本来欺骗目标模型，同时判别器用于评估对抗样本的真实性，构成了一种博弈关系。通过训练生成器和判别器来生成有效的对抗性扰动。

图 1-8 Adv-GAN 网络结构图

该方法，具有以下特点和优势。**博弈性质**：利用生成器和判别器之间的对抗训练，生成具有欺骗性的对抗样本。**效果优秀**：adv-GAN 方法在生成对抗样本时具有较高的成功率和鲁棒性。**适用广泛**：适用于各种类型的深度学习模型，能够成功欺骗不同类型的目标模型。

B.WGAN (Wasserstein Generative Adversarial Network)。WGAN 是一种基于 Wasserstein 距离的生成对抗网络，由 Arjovsky et al.在 2017 年提出。WGAN

通过引入 Wasserstein 距离作为生成器和判别器之间的损失函数，提高了对抗训练的稳定性和效果。在对抗攻击中，WGAN 方法能够生成具有更高迷惑性的对抗样本。

该方法，具有以下特点和优势。**Wasserstein 距离**：通过 Wasserstein 距离作为损失函数，提高了对抗训练的稳定性和收敛性。**更高迷惑性**：WGAN 方法生成的对抗样本具有更高的迷惑性和欺骗模型的能力。**对抗性**：WGAN 在对抗攻击中能够挑战模型的鲁棒性，成功地误导目标模型。

### 面向目标检测任务的对抗攻击方法研究现状

由于目标检测不仅需要识别物体的存在与类型，还需要精确确定物体的位置，这使得图像分类的对抗攻击方法无法直接应用于目标检测任务。为了应对这一挑 战，研究者们开始探索专门针对目标检测的对抗样本生成方法。

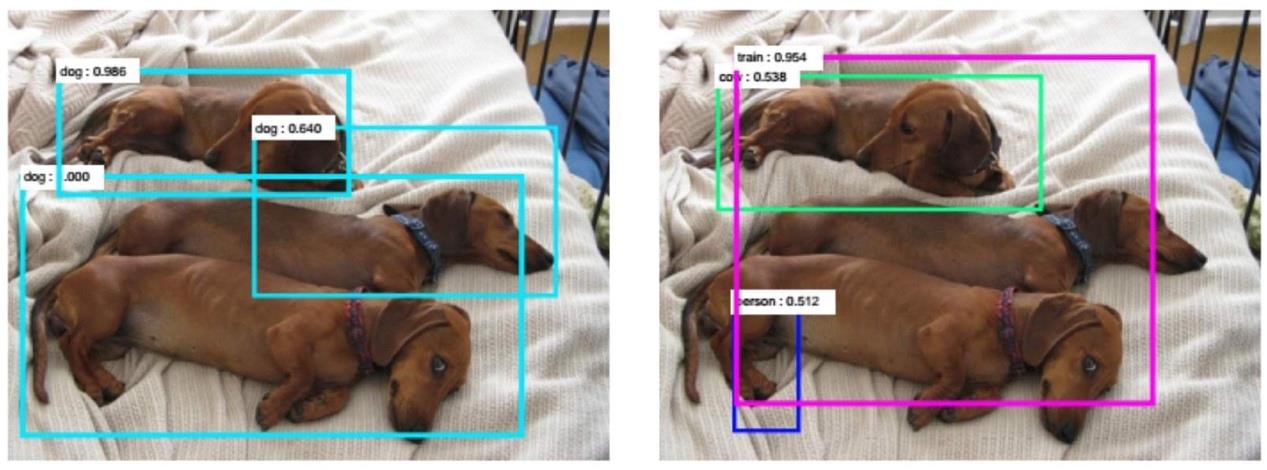


图 1-9 目标检测攻击示例

针对目标检测的对抗攻击方法可以根据对目标像素修改的范围分为全局扰动攻击和局部扰动攻击，同时根据攻击的目标检测器类型划分为针对单阶段模型攻击、针对双阶段模型攻击和全兼容攻击三种类型。

1. 针对单阶段模型攻击。在单阶段模型中，Song 等人受到物理攻击 RP2 的启发，提出了 PhyAttack 方法[37]，通过添加对抗性损失函数并利用梯度下降法，成功降低停车标志的检测分数，欺骗了 YOLOv2 检测器。Wang 设计的 Daedalus 方法[38]则针对单阶段检测器的非极大值抑制（NMS）模块实施攻击，导致 YOLO 网络生成了误报和漏报的结果。Liao 等人首次针对 Anchor-free 类模型 CenterNet 进行攻击[39]，依据重要像素区域的高级语义信息实施类别攻击，有效偏离原类，并

成功将攻击迁移到传统的单阶段及双阶段检测模型。

1. 针对双阶段模型攻击。2017 年，Lu 等人提出了 DFool 方法[40]，通过对 “停车标志”图像添加扰动来误导相应的目标检测器，这是目标检测领域首篇提出的对抗样本生成文章。同年，Xie 等人提出了密集对抗生成方法（DAG）[41]，该方法针对两阶段网络的区域提议网络（RPN）组件生成攻击，机制为对每个候选区域分配对抗标签并执行梯度上升策略。这种方法在实际攻击中表现出色，但由于要求对每个候选框进行迭代，计算开销较大。随后，Chen 等人提出的 ShapeShifter是针对 Faster R-CNN 的第一个目标攻击方法[42]。ShapeShifter 借鉴了图像分类中的对抗攻击方法，如 C&W 和期望变换（EOT），通过对停车标志进行攻击，成功欺骗了 Faster R-CNN，但由于需要对整个停车标志进行修改，导致成本较高。此外，Li 等人提出的 RAP 攻击[43]通过设计将分类损失和位置损失结合在一起的损失函数，利用目标检测的信息进行攻击。Zhang 等人提出的 CAP 方法[44]则充分利用上下文信息，通过破坏图像中的上下文区域来增强攻击力，达到更佳的攻击效果。

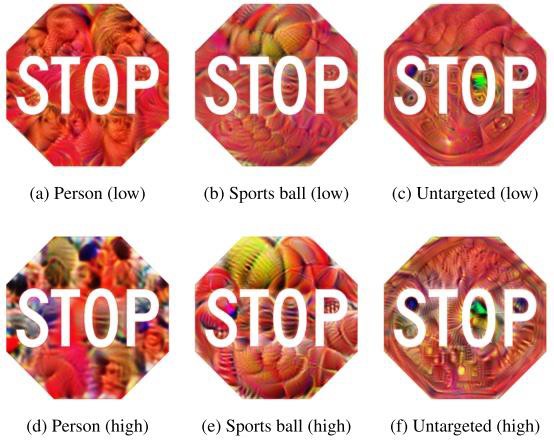


图 1-10 基于 ShapeShifter 方法生成的 6 种对抗样本

1. 全兼容攻击。Wei 提出的 UEA（Unified and Efficient Adversary）方法[45]结合了生成对抗网络（GAN）和高级分类损失及底层特征损失，通过实时生成对抗样本，实现了对视频的即时攻击，成功率较高。Wu 等人改进通用对抗扰动攻击（UAP），提出 G-UAP[46]，通过选择一批图像进行攻击，该方法能够有效降低前景物体的置信度得分，从而学习到通用扰动以欺骗更广泛的图像。Chow 等人提出基于迭代的 TOG 方法[47]，能够同时攻击两阶段与单阶段检测器，通过设定的损失函数不断迭代，达到高效的攻击效果，尽管在迁移性方面依然有所不足。

面对目标检测模型在对抗攻击下的脆弱性，研究者们持续探索多维度的对抗攻击方法，旨在揭示深度学习模型的潜在威胁。虽然已有多种对抗攻击方法被提出来改善目标检测器的抗攻击能力，但仍需要在攻击的多样性、兼容性和隐蔽性方面加强研究，以推动目标检测模型在实际应用中的安全性与可靠性。此外，针对对抗样本的防御机制和鲁棒性的研究也应成为未来工作的重点，以构建更为安全的智能视觉系统。

## 研究内容及组织结构

### 研究内容

当前，针对目标检测的对抗样本生成方法，虽然在指定环境条件下能够达到不错的攻击效果，但是基于白盒条件的攻击方法其生成的对抗样本转移性较差，而基于黑盒条件的攻击方法其生成的对抗样本过程资源消耗较大，且整体攻击效果也差强人意。本文针对白盒条件攻击转移能力差、黑盒条件生成周期长的问题，提出了基于多尺度特征网格群动态优化的目标检测对抗方法 DOG （ Dynamic Optimization of Multi-Scale Feature Grid Cluster），并在 DOG 方法的基础上，为进一步提高生成效率，提出了基于 GAN 网络的 DOGs 方法。DOG 方法利用目标检测模型均存在类似特征金字塔网络（FPN）结构来获得多尺度特征的特点，借鉴 C&W 对抗样本生成架构、DAG 白盒模型激活值选取、R-AP 物体标记重置以及 EA粒子群优化的思路，基于特殊构造的损失函数，利用任意一种预训练白盒模型作为攻击目标，通过有限次迭代获得图像对抗样本，该样本除对所选白盒模型具有极高攻击性的同时，还能对多种图像分类、目标检测模型造成不错的黑盒攻击效果。DOGs 方法利用 DOG 对抗噪声生成方法，采取类似 GAN 网络的训练方式训练 U- Net 结构生成器，使其能够生成与 DOG 方法相类似的对抗噪声，进一步压缩对抗噪声的生成时间，提高对抗效率。

### 关键技术

1. 多尺度特征网格群提取技术。基于目标检测模型均采用类似特征金字塔结构（FPN）来提取图像多尺度特征的共性特点，截取模型该结构输出作为重点识

别、攻击对象，以确保对抗攻击具有较好的攻击迁移性。

1. 最优诱骗类别动态选择技术。提出了一种基于交叉熵损失的最小、最大诱偏类别动态选择方法 FOCG(Find the Optimal Category Group)，通过将攻击目标动态调整为能够使得输出结果向造成最小或最大的交叉熵方向移动的诱偏目标类别，以达到使用最小或最大扰动实现攻击的目的，通过不断动态优化以达成最佳攻击效果。
2. 目标类别集背景化重置技术（BRM）。提出了一种控制迭代攻击方位始终保持在主要方向的处理技术。实现主攻方向能够随迭代持续优化，非主攻方向始终保持为背景状态，以确保迭代优化算法的高度精确性。
3. 基于图像信息熵的关键区域识别技术。从图像自身固有属性出发，以信息熵理论为基础，融合了卷积的思路，构造图像信息熵图，采取特殊的标准化方式转换为关键区域权重矩阵。

### 章节安排

全文由 5 章组成，其结构如下：

第 1 章首先阐述了课题的研究背景和意义，然后介绍了国内外目标检测方法、对抗攻击方法，最后就本文的研究内容和关键技术进行简要概述。

第 2 章介绍了本文会涉及的图像分类网络模型、目标检测模型、进行实验对比的目标检测攻击算法、常用数据集和评价指标。

第 3 章介绍了一种基于多尺度特征网格群动态优化的目标检测对抗方法 DOG

（Dynamic Optimization of Multi-Scale Feature Grid Cluster）。

第 4 章介绍了基于 advGAN 网络模型的攻击方法 DOGs，他是 DOG 方法的免迭代优化直出版本。

第 5 章总结展望，对本文的研究成果进行总结与分析，对未来的研究方向进行简要探讨。

# 2 相关模型、算法及数据集介绍

## 引言

本章主要对图像分类任务的三种经典模型和目标检测任务的四种经典网络模型，以及针对目标检测任务的三种对抗攻击方法进行介绍，并给出了三种常用数据集和评价标准。

## 图像分类网络

根据后续实验需求，本小节将简要介绍 VGG19、ResNet50 和 Inception V3 三种模型。

### VGG19 网络

VGG19 是一种经典的卷积神经网络模型，通常用于图像分类任务。该模型由 19 层神经网络组成，包括 16 个卷积层和 3 个全连接层。VGG19 通过堆叠多个 3x3尺寸的卷积核来增加网络深度，从而提高模型性能。网络结构见图 2-1。

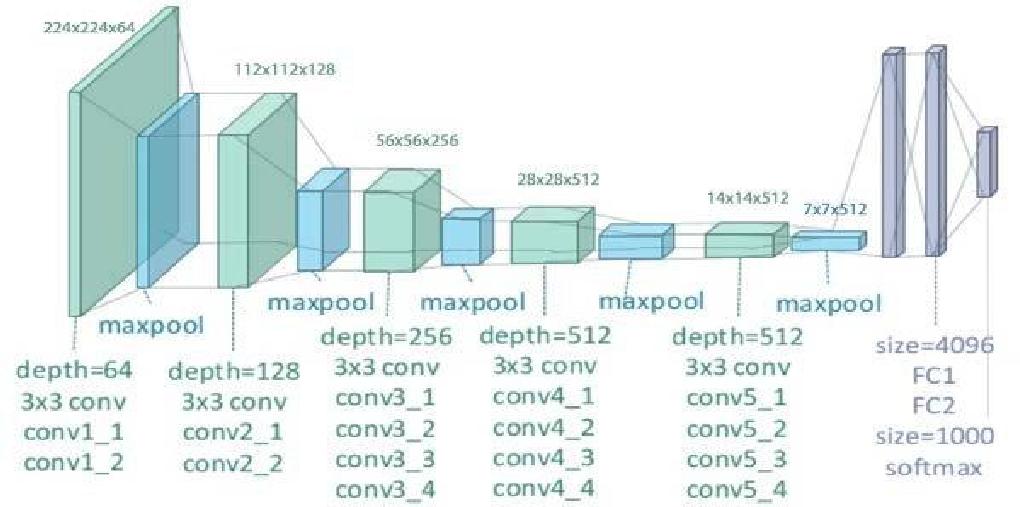


图 2-1 VGG19 结构图

* + - 1. 性能：在 ImageNet 数据集[49]上，VGG19 模型达到了 90.88%的 top-5 识别准确率和 72.38%的 top-1 识别准确率。
      2. 特点：网络结构较深，参数较多，因此具有较高的计算复杂度和内存占用。
      3. 输入图像预处理：图像输入模型前，首先采用双线性插值算法将图像短边调整至 256 像素，随后进行中心裁剪，裁剪至 224×224 的尺寸大小，接着，将

像素值缩放至[0.0, 1.0]范围内，并进行归一化处理，均值为[0.485, 0.456,

0.406]，标准差为[0.229, 0.224, 0.225]，最终送入模型进行判断。

### ResNet50 网络

ResNet50 是另一个常用的深度残差网络模型。该模型使用了残差块的概念，通过引入快捷连接（skip connections）来解决深度网络中的梯度消失和梯度爆炸问题，从而使得更深层次的网络训练变得可行。ResNet50 由 50 层神经网络组成，其中包括 48 个残差层和 2 个全连接层。网络结构见图 2-2。

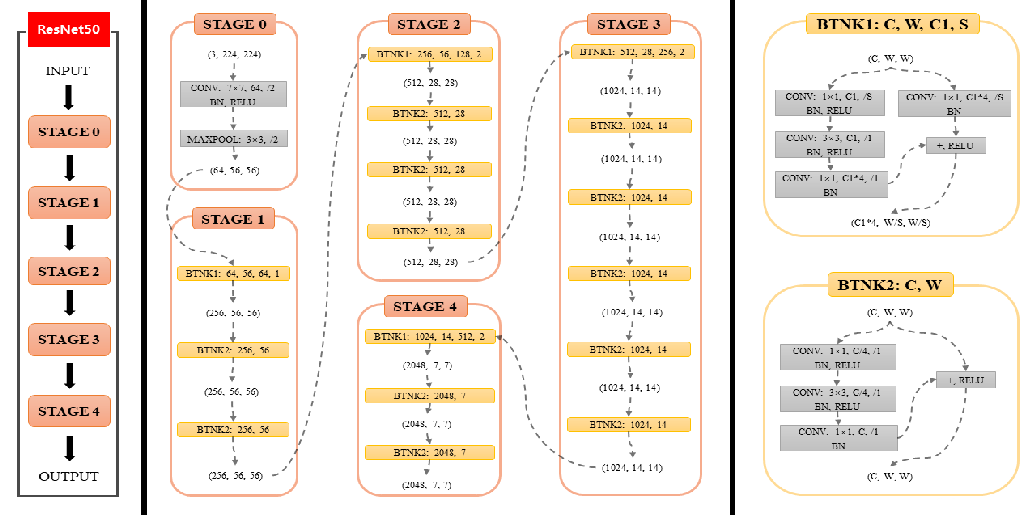


图 2-2 ResNet50 结构图

1. 性能：在 ImageNet 数据集上，ResNet50 模型达到了 92.88%的 top-5 识别准确率和 76.13%的 top-1 识别准确率。
2. 特点：具有较深的网络结构，能够更好地捕捉图像中的高级特征，性能优于 VGG 系列模型。
3. 输入图像预处理：图像输入模型前，首先采用双线性插值算法将图像短边调整至 256 像素，随后进行中心裁剪，裁剪至 224×224 的尺寸大小。接着，将

像素值重缩放至[0.0, 1.0]范围内，并进行归一化处理，均值为[0.485, 0.456,

0.406]，标准差为[0.229, 0.224, 0.225]，最终送入模型进行判断。

### Inception V3 网络

Inception V3 是基于 Inception 架构的一种深度卷积神经网络。该模型利用了并行的多卷积核卷积模块（Inception 模块），使其能够捕捉不同尺度的图像特征，同时通过非对称分层卷积来降低计算量。网络结构见图 2-3。

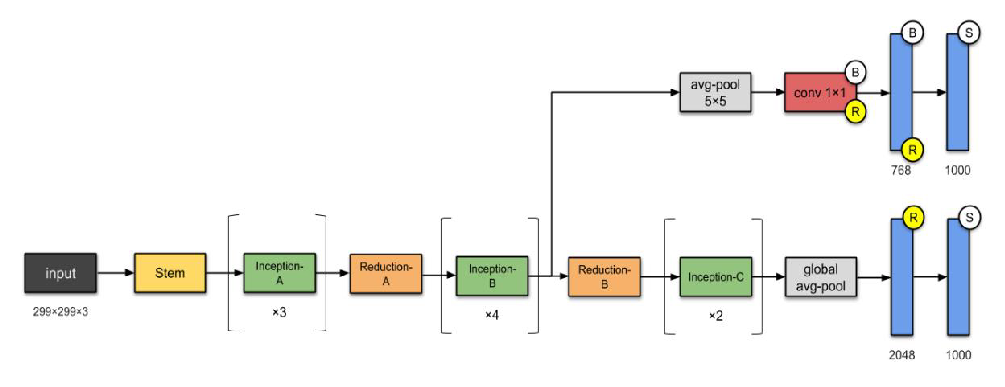


图 2-3 Inception 模块结构图

1. 性能：在 ImageNet 数据集上，Inception V3 模型达到了 93.45%的 top- 5 识别准确率和 77.30%的 top-1 识别准确率。
2. 特点：通过卷积拆分降低了计算成本，具体为将 5×5 卷积调整为两个 3×3 卷积，将 n×n 卷积调整为 n×1 和 1×n 两个卷积；通过多尺度特征提取和较高的拟合能力，提高了模型的性能和精度。
3. 输入图像预处理：图像输入模型前，首先采取双线性中间插值算法将图像短边调整至 342 像素，而后中心裁剪至 299×299 的尺寸大小，接着，将像素值缩放至[0.0, 1.0]范围内，经归一化处理后送入模型进行判断，均值为[0.485, 0.456, 0.406]、标准差为[0.229, 0.224, 0.225]。

从以上介绍不难看出，三种网络模型各有特色，VGG19 采取最直接的堆叠小 卷积层的方式增加深度从而提高模型性能，而 ResNet50 则通过在层与层之间加 入残差连接使得模型可以叠加更多层数从而提高性能，而 Inception V3 更多的 体现了巧妙组合带来的模型性能提升。三种网络模型代表了三种不同的设计思路，因此本文选用这三种图像分类模型来检验基于目标检测任务生成的图像对抗样 本，能否同样对图像分类任务模型造成影响。

在进行多模型对比实验室，还需要关注模型在输入图像尺寸和数值上的差异，这三类模型，数值像素缩放、归一化处理完全一致，但在图像尺寸上，vgg19 和 ResNet50 一致，都是 256 像素裁剪为 224×224 的尺寸，而 Inception V3 则是从 342 像素裁剪为 299×299 尺寸，后续实验应进行适当的统一化处理。

## 目标检测网络

本小节将对四种单阶段网络（YOLOv5s、SSD300、FCOS、RetinaNet）、一种双阶段网络(MaskRCNN)进行详细介绍。

### YOLOv5 网络

YOLOv5 是一种在 YOLOv4 基础上进一步改进的单阶段目标检测算法，通过在多个关键环节的改进，使其性能得到显著提升，同时在实用性和易用性方面也得到了进一步拓展。具体来说，YOLOv5 引入了 Mosaic 数据增强、自适应锚框计算、自适应图片缩放等创新技术，这些改进使得模型在处理不同类型的目标时更加灵活和高效。根据网络模型深度（depth\_multiple）和模型宽度（width\_multiple）参数的不同，YOLOv5 系列分为四个版本：YOLOv5s、YOLOv5m、YOLOv5l 和 YOLOv5x。每个版本在模型的深度和宽度上有所不同，以适应不同的应用场景和计算资源需求。本文所使用的模型为 YOLOv5s，在 COCO val2017 数据集[50]上 mAP 值达到 37.2，其在速度和精度之间达到了良好的平衡，适合本文目标检测任务需求。各版本的参数和评分情况见图 2-4，展示了不同模型在 COCO 数据集上的性能表现，包括 mAP（平均精度均值）、推理速度和模型大小等指标。

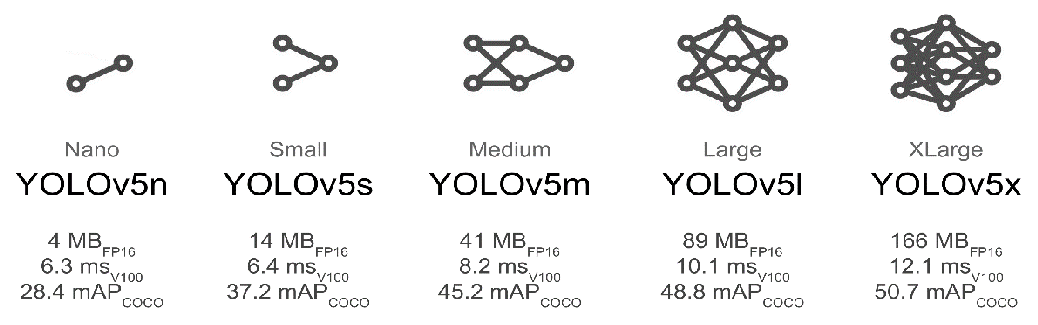


图 2-4 YOLOv5 各版本参数体量及性能对比

YOLOv5s 主要由输入端、基础网络（backbone）、衔接网络（neck）、输出端（head）四个模块组成。网络结构见图 2-5。

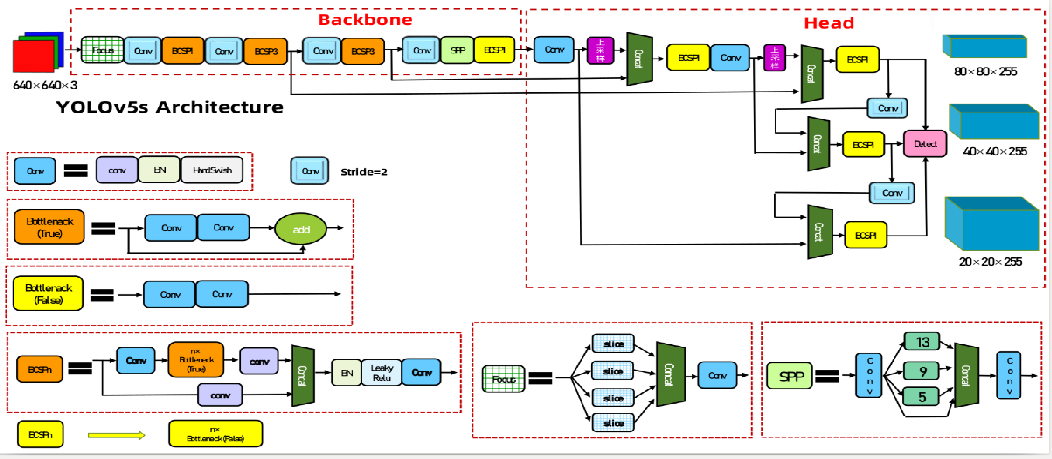


图 2-5 YOLOv5s 结构图

1. 输入端。输入端负责接收输入的图像，YOLOv5s 的网络输入图像大小为 640x640 像素。在这一阶段，主要进行图像预处理，包括将输入图像缩放到网络所需的尺寸，并进行归一化处理，以确保数据在训练过程中的一致性和稳定性。为了提升模型的训练速度和检测精度，YOLOv5s 采用了 Mosaic 数据增强技术，该技术通过将多张图像拼接成一张，增加了数据的多样性。此外，YOLOv5s 还引入了自适应锚框计算和自适应图片缩放方法，使得模型能够根据不同的数据集动态调整锚框的大小和数量，从而更好地适应目标的尺度变化。
2. 基准网络。基准网络是用于提取通用特征表示的模块，通常由性能优异的分类器构成。在 YOLOv5s 中，基准网络采用 CSPDarknet53 结构，这是一种高效的特征提取网络，能够在保持较低计算复杂度的同时提取丰富的特征信息。此外， YOLOv5s 还引入 Focus 结构，该结构通过对输入图像进行切片操作，增强了特征的表达能力和计算效率，使得模型在特征提取阶段表现更加出色。
3. Neck 网络。Neck 网络位于基准网络和头网络之间，主要用于进一步提升特征的多样性和鲁棒性。YOLOv5s 在Neck 部分同样使用了SPP（Spatial Pyramid Pooling）模块和 FPN（Feature Pyramid Network）+PAN（Path Aggregation Network）结构，这些模块的结合使得不同尺度的特征能够有效融合，从而提高了

模型对各种目标的检测能力。尽管 YOLOv5 与 YOLOv4 在 Neck 结构上有相似之处，但在实现细节上有所不同，以适应 YOLOv5 的整体设计理念。

1. Head 输出端。Head 部分负责生成最终的目标检测结果。根据不同的检测任务，输出端的分支数量可能会有所不同，通常包括一个分类分支和一个回归分支。在 YOLOv5s 中，输出层采用了 CIOU（Complete Intersection over Union）损失函数，替代了 YOLOv4 中的 GIOU（Generalized Intersection over Union）损失函数。CIOU 损失在考虑重叠面积的同时，还综合考虑了中心点距离和长宽比等因素，从而进一步提升了目标框的定位精度和检测效果。通过这些设计， YOLOv5s 能够在各种目标检测任务中实现高效且准确的性能。

### SSD300 网络

SSD（Single Shot MultiBox Detector）是一种单阶段目标检测算法，首次由 Wei Liu 等人于 2016 年提出。通过采用特征金字塔结构和 Prior Box 等技术，使其在确保检测效率的同时，能够同时处理不同尺度的目标，从而实现快速且高效的目标检测。具体来说，SSD 利用特征金字塔结构，能够从不同层的特征图中提取信息，从而提高对小目标和大目标的检测精度。此外，SSD 引入了 Prior Box的概念，这些预选框在目标检测中起到了关键作用，使得模型能够灵活适应不同目标的尺寸和形状。每个 Prior Box 通过 softmax 分类和坐标回归来预测目标的位置，这种方法使得 SSD 在处理复杂场景时表现出色。SSD 还支持多尺度输入图片和特征图，进一步增强了模型的适应能力，无论是处理大尺寸的目标还是小尺寸的目标，SSD 都能有效进行检测，确保其在各种应用场景中的实用性。通过采用 Batch Normalization 和卷积技术，SSD 不仅提高了模型的训练速度，还增强了检测的准确性。根据输入图像的尺寸不同，SSD 分为 SSD300 和 SSD512 两个版本。SSD300 以 300x300 像素的输入图像进行训练和推理，适合对速度要求较高的应用场景。而 SSD512 则使用 512x512 像素的输入图像，能够提供更高的检测精度，适合需要更细致目标识别的任务。本文所使用的模型为 SSD300，在 COCO val2017 数据集上 mAP 值达到 25.1, 其在处理速度和检测精度之间达到了良好的平衡, 适合本文目标检测任务需求。

为与 YOLOv5s 模型进行直观对比， SSD300 同样按照输入端、基础网络

（backbone）、衔接网络（neck）、输出端四个模块来拆分。

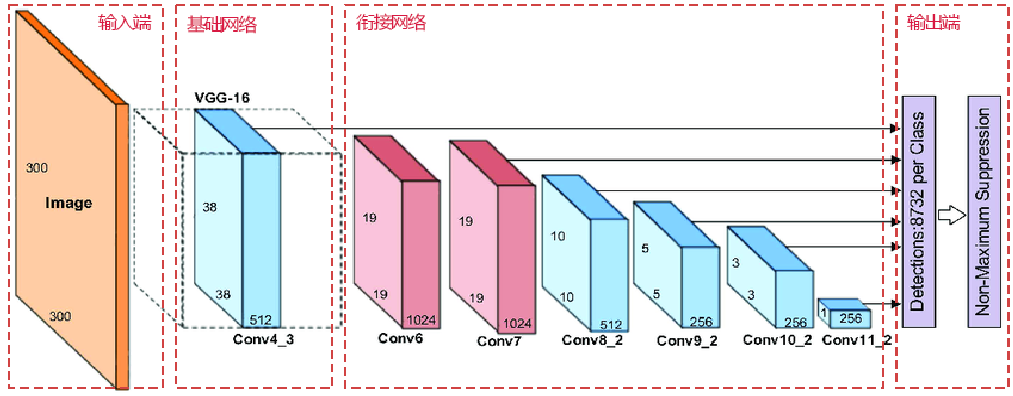


图 2-6 SSD300 结构图

1. 输入端。输入端负责接收输入图像。在 SSD 网络中，输入图像的大小对于性能有直接影响。对于 SSD300，输入图像的尺寸为 300x300 像素；而对于 SSD512，输入图像的大小为 512x512 像素。在这一阶段，模型对输入图像进行预处理，包括图像缩放、归一化和数据增强，以提高模型的训练效率。SSD 采用了随机裁剪、水平翻转等数据增强方法，增加了图像的多样性，提高了模型的鲁棒性。
2. 基准网络（Backbone）。基准网络是特征提取的核心模块。在 SSD 中，通常采用 VGG16 作为特征提取网络，VGG16 是一个经典的卷积神经网络，能够在保持较高精度的同时提取丰富的特征信息。通过将 VGG16 与多尺度特征图相结合， SSD 能够在不同的层上生成候选边界框，适应不同尺寸的目标。
3. Neck 网络。在 SSD 中，Neck 部分主要用于特征融合。SSD 在不同层级的特征图上生成多个边界框，并结合不同尺度的特征信息。通过在不同的层上进行特征组合，SSD 模型能够有效捕捉到各种大小目标的信息。具体而言，SSD 使用了额外的卷积层，以帮助合并来自不同层的特征，增强了模型对各种目标的检测能力。
4. 输出端（Head）。输出端负责生成最终的目标检测结果。在 SSD 中，对于每一个候选框，模型会输出类别概率（即在这个框中存在特定目标的可能性）和边界框的回归信息。SSD 使用了 softmax 函数来处理类别概率，并结合 Smooth L1

损失函数以优化边界框的回归精度。

### RetinaNet 网络

RetinaNet 是一种单阶段目标检测算法，首次由 Tsung-Yi Lin 等人于 2017年提出。该算法引入了 Focal Loss，以解决类别不平衡问题，特别是在负样本远多于正样本的情况下。RetinaNet 结合了特征金字塔网络（FPN），能够有效处理不同尺度的目标，从而在精度和速度之间取得了良好的平衡，适用于多种目标检测任务。具体来说，RetinaNet 利用特征金字塔网络（FPN）从不同层级的特征图中提取信息，从而有效改善对小目标和大目标的检测能力。与传统的基于 anchor的方法不同，RetinaNet 在特征图上生成多个 anchors，并通过 Focal Loss 来优化这些 anchors 的预测，解决了类别不平衡的问题。Focal Loss 是 RetinaNet 的核心创新，旨在降低易分样本的权重，增强难分样本的贡献，从而提高模型的学习效果。通过对每个预测框的得分进行加权，Focal Loss 使得模型更加关注那些难以检测的样本，提高了在复杂背景下的鲁棒性。结合多尺度特征图的强大能力， RetinaNet 能够在单个前向推理中获取目标类别及其位置，从而实现高效精准的检测。本文所用的 RetinaNet 模型（使用 ResNet50 作为基础网络）在 COCO val2017数据集上达到了较高的 36.4mAP 值，展现了在处理速度和检测精度之间的良好平衡，适用于各种实际应用。

为了方便理解与比较，RetinaNet 模型同样按照输入端、基础网络（Backbone）、衔接网络（Neck）和输出端（Head）四个模块进行拆分，结构见图 2-7。

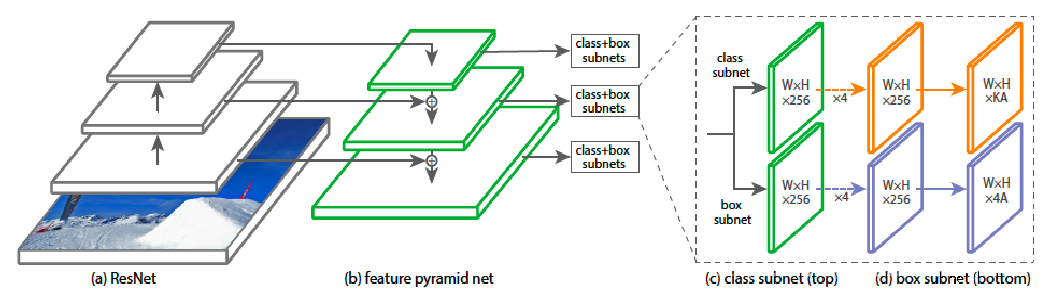


图 2-7 RetinaNet 结构图

1. 输入端。在 RetinaNet 网络中，输入图像的尺寸会影响特征图的大小和网

络输出的结果。然而，RetinaNet 的设计具备灵活性，允许根据不同的任务需求和计算资源调整输入图像的尺寸。虽然 RetinaNet 通常支持多种输入图像尺寸，但在使用之前，图像需要经过预处理，包括缩放、归一化和数据增强。这些步骤旨在提高模型的训练效果。具体而言，RetinaNet 采用了如随机裁剪和水平翻转等数据增强方法，以增加图像的多样性，从而提升模型的鲁棒性。

1. 基础网络（Backbone）。基础网络是 RetinaNet 的核心模块，通常采用 ResNet 或其他经典卷积神经网络作为基础网络。通过在多个尺度上获取特征， RetinaNet 可以有效适应不同尺寸的目标。这些特征图将为后续的目标检测提供重要的信息支持。
2. 衔接网络（Neck）。在 RetinaNet 中，衔接网络（Neck）部分负责特征融合，主要聚合来自不同层级的特征图信息，以提高对各种规模目标的检测能力。 RetinaNet 采用特征金字塔网络（FPN）结构，通过自顶向下的特征融合和自底向上的特征提取，实现对多尺度特征的有效组合，以便更好地捕捉小、中、大目标的特征信息。这种逐层融合的方式使得 RetinaNet 在处理不同目标尺寸时更加灵活和高效。
3. 输出端（Head）。输出端负责生成最终的目标检测结果。在 RetinaNet 中，每个 anchor 会生成一组类别分数和回归信息。RetinaNet 采用 softmax 函数来处理类别概率，同时使用 Focal Loss 来优化每个预测框的得分。通过结合 Smooth L1 损失函数以优化框的回归精度，RetinaNet 模型确保在整体性能上的强大表现。

### FCOS 网络

FCOS（Fully Convolutional One-Stage Object Detection）是一种单阶段目标检测算法，首次由 Tianzhi Huang 等人于 2019 年提出。该算法采用无 Anchor的设计，完全基于像素的预测，简化了传统目标检测模型的复杂性。FCOS 通过使用特征金字塔网络（FPN）实现多尺度预测，并引入中心度（Centerness）分支来提高目标检测的准确性。该方法在处理不同尺度的目标时表现优异，尤其在小目标检测方面具有显著优势。具体来说，FCOS 利用特征金字塔网络从不同层级的特征图中提取信息，从而有效改善对小目标和大目标的检测能力。与传统的基于

anchor 的方法不同，FCOS 直接在特征图每个像素上进行目标检测，消除了 anchor 的设计和匹配复杂性。FCOS 引入的中心度消除机制对于提高预测的质量至关重要，中心度通过对每个预测框的得分进行加权，使得离目标中心更近的点获得更高的 权重，从而提高模型在复杂背景中的鲁棒性。通过结合多尺度特征图的强大能力， FCOS 能够在单个前向推理中获取目标类别及其位置，从而实现高效精准的检测。本文所用的 FCOS 模型（使用 ResNet50 作为基础网络）在 COCO val2017 数据集 上达到了一定的 39.2mAP 值，其在处理速度和检测精度之间很好地平衡，适用于 各种实际应用。

为了方便理解与比较，FCOS 模型同样按照输入端、基础网络（Backbone）、衔接网络（Neck）和输出端（Head）四个模块进行拆分，结构见图 2-8。

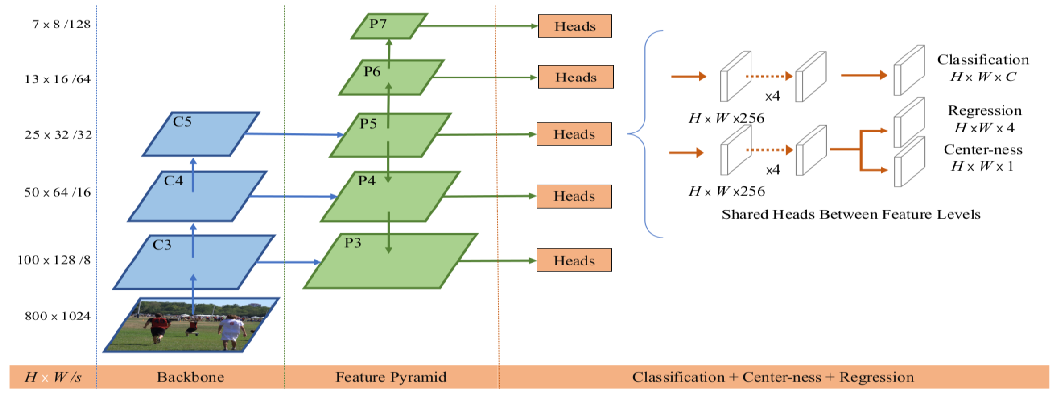


图 2-8 FCOS 结构图

1. 输入端。输入端负责接收输入图像。FCOS 网络对输入图像尺寸的要求相对灵活，通常支持多种输入图像尺寸，但输入图像需进行预处理，包括缩放、归一化和数据增强，以提高模型训练效果。FCOS 采用随机裁剪、水平翻转等数据增强方法，增加了图像多样性，提升模型的鲁棒性。
2. 基础网络（Backbone）。基础网络是 FCOS 的核心模块，通常采用 ResNet 或其他经典卷积神经网络作为基础网络。通过在多个尺度上获取特征，FCOS 可以 有效适应不同尺寸的目标。这些特征图将为后续的目标检测提供重要的信息支持。
3. 衔接网络（Neck）。在 FCOS 中，衔接网络（Neck）部分负责特征融合，主要聚合来自不同层级的特征图信息，以提高对各种规模目标的检测能力。FCOS 采

用特征金字塔网络（FPN）结构，与 SSD 通过特征图直接拼接叠加的方法不同， FCOS 通过自顶向下的特征融合和自底向上的特征提取，实现对多尺度特征的有效组合，以便更好地捕捉小、中、大目标的特征信息。这种逐层融合的方式使得 FCOS在处理不同目标尺寸时更加灵活和高效。

1. 输出端（Head）。输出端负责生成最终的目标检测结果。在 FCOS 中，每个像素点会生成一组类别分数和回归信息。FCOS 采用 softmax 函数来处理类别概率，同时使用中心度机制来优化每个预测框的得分。通过结合 Smooth L1 损失函数以优化框的回归精度，FCOS 模型确保在整体性能上的强大表现。

### MaskRCNN 网络

MaskR-CNN 是一种两阶段目标检测算法，首次由 Kaiming He 等人于 2017 年提出。该算法在 Faster R-CNN 的基础上进行了扩展，增加了一个掩码分支，用于实现实例分割。MaskR-CNN 通过区域提议网络（RPN）生成候选框，并对每个框进行分类、回归以及掩码预测。该方法不仅能够进行高精度的目标检测，还能为每个目标生成精确的分割掩码，适用于需要高精度检测和分割的复杂场景。尽管计算复杂度较高，Mask R-CNN 在精度上表现出色，广泛应用于各种计算机视觉任务。

Mask R-CNN 网络可以分为以下几个主要模块：输入端、基础网络（Backbone）、区域提议网络（RPN）、ROI Align、输出端（Head）和损失函数。结构见图 2-9。

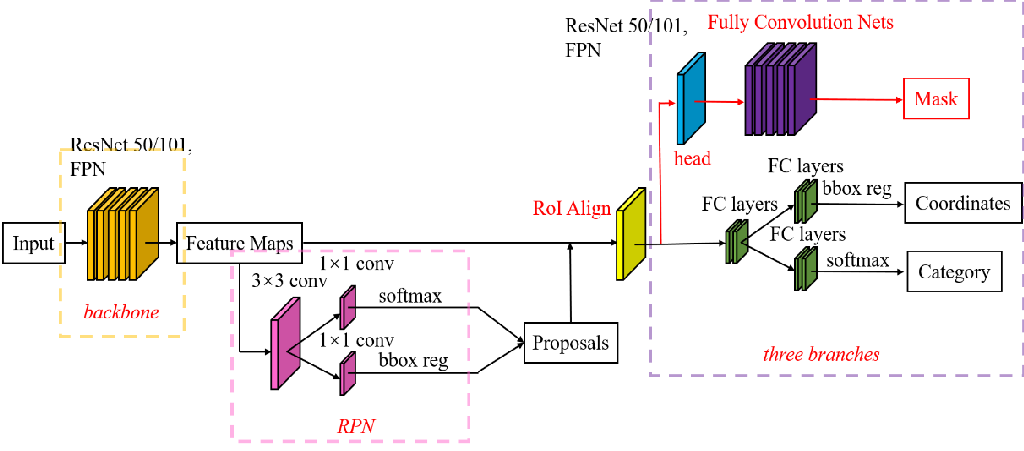


图 2-9 Mask R-CNN 结构图

1. 输入端。输入端负责接收输入图像，并进行必要的预处理。Mask R-CNN

支持多种输入图像尺寸，通常需要对输入图像进行缩放、归一化和数据增强，以提升模型的训练效果。常用的数据增强方法包括随机裁剪、水平翻转和颜色变换，这些方法增加了图像的多样性，提升了模型的鲁棒性。

1. 基础网络（Backbone）。基础网络是 Mask R-CNN 的核心模块，通常采用 ResNet 或其他经典卷积神经网络作为基础网络。基础网络负责从输入图像中提取多层次的特征图，这些特征图为后续的目标检测和实例分割提供了重要的信息支持。Mask R-CNN 通过在多个尺度上获取特征，能够有效适应不同尺寸的目标。
2. 区域提议网络（RPN）。区域提议网络（RPN）负责生成候选区域提议。 RPN 通过滑动窗口的方式在特征图上生成多个 anchors，并对每个 anchor 进行分类和回归，输出目标的边界框。RPN 的设计使得 Mask R-CNN 能够实现端到端的训练，提升了目标检测的效率。
3. ROI Align。ROI Align 是 Mask R-CNN 的关键创新之一，作为对传统 ROI Pooling 的改进。ROI Align 通过对每个候选区域进行精确的特征对齐，避免了由于取整操作导致的特征信息损失。这一机制确保了在进行目标分类和边界框回归时，特征图的精确性，从而提高了模型的预测精度，尤其在小物体检测中表现突出。
4. 输出端（Head）。输出端负责生成最终的目标检测和实例分割结果。在 Mask R-CNN 中，每个候选区域会生成一组类别分数、边界框回归信息和掩码预测。模型采用 softmax 函数处理类别概率，同时结合 Focal Loss 来优化每个预测框的得分。掩码分支由两个卷积神经网络组成，为每个 ROI 输出一个二进制掩码，允许模型为每个类生成掩码，而不会在类之间进行竞争。
5. 损失函数。Mask R-CNN 使用多任务损失函数来优化模型性能，包括分类损失、边界框损失和掩码损失。分类损失和边界框损失与 Faster R-CNN 相同，而掩码损失则用于优化每个 ROI 的掩码生成，确保模型在实例分割任务中的表现。

### 五种网络对比

通过以上介绍可以看出，这五种目标检测网络在处理大小目标时都采用了特征金字塔结构（FPN），实现了多尺度的检测分析。在识别方法上，YOLOv5s 和 FCOS

网络基于回归，而其他三种网络则依赖于建议机制。在结构组成上，尽管单阶段模型的基本架构相似，它们在一些细节层面上存在些微差异。除了这些共通的特点，各模型还有其独特之处，具体如下：

1. SSD300 针对快速有效的目标检测而设计。该模型采用了简单而直接的结构，使用 6 个特定的卷积层（包括 conv4\_3、conv7、conv8\_2、conv9\_2、conv10\_2和 conv11\_2）进行特征提取，而不是对所有卷积层进行卷积预测。通过引入多尺度的特征图，SSD 能够在不同的层次上检测不同尺寸的物体，特别增强了对小物体的检测能力。同时，SSD 使用卷积层替换了 YOLO 模型中的全连接层，在保持模型轻量化的同时，提高了处理速度，从而兼顾了检测的准确性和效率。
2. RetinaNet 针对类别影响的权重一致性而设计。该模型引入了损失函数 Focal Loss，用于增强模型对难以分类样本的关注，同时抑制对容易分类样本的影响。Focal Loss 通过调整标签的识别权重，使得正样本和负样本的贡献在模型训练过程中得到了合理的平衡，从而改善模型在不同类别上的性能。
3. FCOS 针对打破传统锚框机制检测模式而设计。通过取消传统的锚框

（anchor）机制，FCOS 能够在像素级别上进行目标定位和识别，更加直接地捕捉物体的特征。这种无锚框的设计使得模型在运行时更加灵活，同时显著减少了超参数的设置和调优复杂性。此外，FCOS 通过对每个像素的特征进行评估，可以有效应对不同尺寸和形状的目标，从而提高了检测的准确性与鲁棒性。

1. YOLOv5s 针对融合提升而设计。该模型采用了多种技术，包括自适应锚框计算、自适应图像缩放、Mosaic 数据增强和混合数据增强（MixUp），以及特征金字塔网络（FPN），以确保在推理速率和预测精度之间达到良好的平衡。YOLOv5s的这种设计使其在实时应用中表现出色，尤其适合需要高效处理的场景。
2. Mask R-CNN 针对高判别率和实例分割任务而设计。该模型不仅能够检测目标，还能够为每个检测到的物体生成高质量的分割掩膜。通过引入区域建议网络（RPN）和掩膜分支，Mask R-CNN 实现了精确的物体识别和分割，使得它在复杂场景下的表现尤为突出。

综上可知，选择这五类目标检测网络进行对抗样本攻击有效性和迁移性验证，

具备较强的代表性。这将为深入研究目标检测模型在易受攻击性和防御机制方面的表现提供丰富的数据支持与理论基础。

## 目标检测任务对抗攻击方法

本小节，主要介绍目标检测领域经典的几种对抗攻击，详细解释攻击的原理和方式，在并且在后几章的实验中会用到这些方法进行分析比较。

### DAG 稠密对抗生成算法

DAG（Dense Adversary Generation）是一种针对两阶段网络的白盒攻击策略。根据两阶段检测器通过区域建议网络（RPN）筛选出含有物体的候选框的特点， Xie 等人提出了在一组像素或者目标框集上优化目标损失函数来实施攻击的方法 DAG。其攻击表达式如公式 2.1 所示：

𝑟𝑚 = − ∑𝑡𝑛∈𝑇𝑚 𝛻𝑋𝑚𝑙𝑜𝑠𝑠𝑚

𝑙𝑜𝑠𝑠𝑚 = 𝑓𝑙𝑛(𝑋𝑚, 𝑡𝑛) − 𝑓𝑙′ (𝑋𝑚, 𝑡𝑛) (2.1)

𝑛

𝑋𝑚+1

= 𝑋𝑚

+ 𝛼

‖𝑟𝑚‖∞

𝑟𝑚

公式中，𝑋𝑚表示第𝑚次迭代得到的图像，初始值为输入图像，𝑓(·)是目标检测器输出的结果函数,𝑙𝑜𝑠𝑠𝑚是定义的损失函数，𝑡𝑛是目标图片中的一个候选目

标， 𝑙𝑛是该候选目标的真实类别，𝑙′ 是设定的错误类别标签，𝑟

是计算得到的对

𝑛 𝑚

抗梯度，𝛼则是迭代学习率。

DAG 的攻击逻辑为：给网络预测正确的 ROI 区域，随机指定一个错误标签，通过反向传播和多次迭代，提升错误类别的得分。然后，针对目标函数相对于输入图片的梯度进行计算，获得的对抗梯度通过无穷范数归一化后作为累计攻击噪声逐步添加到图像中去，当所有正样本均预测出错或达到最大迭代次数后完成该次任务。

DAG 的攻击方法属于分类损失攻击，通过为目标从非正确标签中随机赋予误导标签，多次迭代后使得类别置信度向着错误标签方向转移，最终使得检测器对输入图片的大部分 ROI 都发生错误。

### R-AP 鲁棒对抗扰动算法

R-AP（Robust Adversarial Perturbation）是一种用于攻击基于深度提议的

目标检测器和实例分割算法，该方法针对区域提议网络(RPN)这一常见组件，通过结合标签损失和新颖的形状损失设计损失函数，并利用基于梯度的迭代算法进行优化，使得生成的对抗样本在黑盒条件下能够有效降低模型性能。R-AP 不仅干扰标签预测、还干扰形状回归、影响边界框的准确度，Li 等人通过实验进一步验证了其所提方法的有效性，在 MS COCO 2014 数据集上有效地攻陷了六种目标检测器和两种实例分割算法。其攻击表达式如公式 2.2 所示：

𝑝̂𝑡 = 𝛻𝐼𝑡 (𝐿𝑙𝑎𝑏𝑒𝑙 + 𝐿𝑠ℎ𝑎𝑝𝑒)

𝐿𝑙𝑎𝑏𝑒𝑙(𝑋; 𝐹) = ∑𝑚

𝑗=1

𝑧𝑗𝑙𝑜𝑔(𝑠𝑗) ， 𝐿𝑠ℎ𝑎𝑝𝑒(𝑋; 𝐹) = ∑𝑚‖𝑏𝑜𝑥𝑗 − 𝑏𝑜𝑥𝑏𝑖𝑎𝑠‖

(2.2)

𝑋 = 𝐶𝑙𝑖𝑝(𝑋

𝑗

2

− 𝜆𝑅𝐴𝑃 𝑃̂ ), 𝑠 . 𝑡. 𝑃𝑆𝑁𝑅(𝑋 ) ≥ 𝜀

𝑡+1

𝑡 ‖𝑃̂𝑡‖2 𝑡 𝑡

公式中，𝑋𝑡表示第𝑡次迭代得到的图像，初始值为输入图像，𝐹是目标检测器输出的结果函数,𝐿𝑙𝑎𝑏𝑒𝑙、𝐿𝑠ℎ𝑎𝑝𝑒分别定义了基于标签类别和分界框的损失函数。𝑧𝑗是目标图片中的第 j 个候选目标，𝑧𝑗 = 1说明候选目标包含物体，𝑧𝑗 = 0说明不包含物体，对这部分区域不进行计算，𝑠𝑗是目标图片中第 j 个候选目标的置信度。

𝑏𝑜𝑥𝑗是第 j 个候选目标的预测分界框值，包含边框的中心坐标、边框、宽和高，

𝑏𝑜𝑥𝑏𝑖𝑎𝑠是人为设定的较大分界框偏移量。𝐶𝑙𝑖𝑝(·)为裁剪函数，使得生成的对抗样本不超过图像规定的上下限。 PSNR 为峰值信噪比，扰动越小，PSNR 越高。

RAP 方法设计了一种同时包含位置损失和分类损失的组合损失函数。通过降低 RPN 网络生成的建议框的置信度，使得目标检测器将图像中的物体错误分类为背景，从而达成无法识别目标的目的。对于那些经过修改仍能够被识别的图像， RAP 方法进一步通过干扰位置参数，使网络错误地定位物体的实际位置，从而实现攻击效果。

### EA 蒸发攻击算法

EA（Evaporate Attack）[48]是由 Wang 等人提出的一种同时能够针对回归型和 建议型两种模型的黑盒攻击方法。通过生成多个初始化对抗样本构建初始粒子群，使用 GA-PSO（遗传算法与粒子群优化结合）算法对粒子群进行优化，寻找能够欺 骗目标检测模型的对抗样本。具体实现步骤如下：

1. 初始化粒子群。通过向原始图像添加大量噪声，将图像粒子初始化在远离原始图像的距离，从而确保添加噪声的图像粒子都满足对抗标准，这样使得我们的后续攻击更容易。见公式 2.3：

𝑥′ = 𝑥 + 𝜖 ⋅ 𝑧*，* 𝑠. 𝑡. 𝑧 ∼ 𝑁(0, 𝜎2) (2.3)

𝑖 𝑐

其中，𝑥′是粒子i，𝑥原始图像，𝑧是随机生成的高斯噪声，𝜖是限制噪声的超

𝑖

参数。

1. 粒子群优化（PSO）。在粒子群优化（PSO）算法中，速度迭代是一个关键步骤，它决定了粒子在搜索空间中的移动方向和速度。根据粒子的历史最佳位置、全局最佳位置、原图投影参考点和原始注入噪声来更新粒子的速度，从而引导粒子朝向更加优秀的解。实现公式见 2.4：

𝑓𝑖𝑡𝑛𝑒𝑠𝑠𝑖 = 𝛼 ⋅ 𝑠𝑐𝑜𝑟𝑒 − 𝑑(𝑥𝑖, 𝑥)

𝑥(𝑡+1) = 𝑥(𝑡) + 𝑣(𝑡+1) (2.4)

𝑣(𝑡+1) = 𝜇

𝑖

⋅ 𝑣(𝑡) + 𝜇

𝑖

⋅ 𝑧 ⋅ (𝑃

𝑖

− 𝑥(𝑡)) + 𝜇

⋅ 𝑧 ⋅ (𝐺

− 𝑥(𝑡))

𝑖 1 𝑖 2

𝑏𝑒𝑠𝑡𝑖 𝑖 3

𝑏𝑒𝑠𝑡 𝑖

+𝜇4

⋅ (𝑈𝑏𝑒𝑠𝑡

− 𝑥(𝑡)) + 𝜇

⋅ 𝑧

公式中，𝑣(𝑡)是粒子i 在时间步t 的速度，𝑥(𝑡)粒子i 在时间步t 的位置信息，

𝑖

5

𝑖 𝑖

z 是随机生成的高斯噪声，𝑃𝑏𝑒𝑠𝑡𝑖是粒子i 的个体最佳位置，𝐺𝑏𝑒𝑠𝑡为全局最佳位置，即所有粒子中最佳的位置，𝑈𝑏𝑒𝑠𝑡为原始图像投影的参考点，𝜇1、𝜇2、𝜇3、𝜇4、𝜇5是各项的权重系数，控制不同因素对粒子速度更新的影响，𝑓𝑖𝑡𝑛𝑒𝑠𝑠𝑖是粒子 i 的适应度，用于衡量该粒子的欺骗性，𝑃𝑏𝑒𝑠𝑡𝑖和𝐺𝑏𝑒𝑠𝑡 基于适应度进行判定。

1. 选择最终攻击样本。如果全局最佳粒子𝐺𝑏𝑒𝑠𝑡 满足对抗标准且与原始图

像之间的距离小于阈值θ，则返回该粒子作为生成的对抗样本。

EA 算法通过构建初始高误导率的粒子群，通过多次查询和 PSO 算法迭代优化，从粒子群中选出最佳粒子作为最终的攻击样本。

## 常用数据集

本小节分别介绍基于图像分类任务的 ImageNet 数据集和着眼目标识别任务的 COCO 数据集。

### ImageNet 数据集

ImageNet 是由斯坦福大学和普林斯顿大学的研究团队于 2009 年推出的一个大规模图像数据集，广泛用于计算机视觉中的图像分类任务。数据集包含超过 1400 万张图像，其中超过 100 万张图像经过人工标注，涉及 1000 个物体分类。图片的尺寸多样，一般不小于 256\*256，使用 XML 格式存储标注信息，包括图像 ID、目标类别、边界框（bounding box）等。

### MS COCO 数据集

MS COCO（Microsoft Common Objects in Context）数据集是 2014 年微软团队提供的一个可以用来进行目标检测、语义分割、人体关键点检测等任务的通用大规模数据集。数据集中包含 33 万张图片，超过 20 万张经过标注，涉及 80 个物体分类和 91 个事务分类。图片的尺寸多样，一般不小于 640\*480，使用 JSON 格式保存标注信息，包括但不限于图片 ID 编码、目标类别、目标分界框、目标置信度等。COCO 包含了自然图片和生活中常见的图片，平均每张图片包含 3.5 个类别和 7.7 个物体，比同类型数据集具有更难的识别环境，如 PASCALVOC 数据集超过 50%的图片只有一个物体。

## 评价指标

本小节主要讲述图像分类、目标检测常用的攻击有效性评价指标，以及针对目标检测任务的性能评价指标。

### 图像分类任务攻击有效性评价指标

在图像分类任务中，一般使用攻击成功率(Attack Success Rate, ASR)指标来评价攻击效果。其定义为攻击样本被模型误分类的样本数量与总样本数量的比值，计算公式如下：

𝐴𝑆𝑅 = 𝑆𝑎𝑑𝑣⁄

𝑆

𝑎𝑙𝑙

(2.5)

其中，𝑆𝑎𝑑𝑣为被误分类的样本数量，𝑆𝑎𝑙𝑙为样本总数，𝐴𝑆𝑅值趋近于 1 说明攻击效果越好。

### 目标检测任务的性能评价指标

1. 平均精度均值（mean Average Precision，mAP）。在目标检测任务中，通常使用平均精度均值（mAP）对模型性能进行评估。其定义为平均精度均值 AP(Average Precision)的平均值，它是[目标检测算法](https://so.csdn.net/so/search?q=%E7%9B%AE%E6%A0%87%E6%A3%80%E6%B5%8B%E7%AE%97%E6%B3%95&spm=1001.2101.3001.7020)的主要评估指标。具体计算步骤为：

①计算每个类别的 AP（Average Precision）。首先，将模型输出预测结果按照置信度降序进行排列。然后，根据对抗样本预测框与原样本预测框的 IOU（一般设置大于 0.5）来判断是否为 TP（True Positive）或 FP（False Positive）。接着，根据不同的置信度阈值，计算不同召回率（Recall）下的精准率（Precision）。最后，求经过插值操作后的 P-R 折线与 X 轴包络的面积，得到该类别的 AP 值。公式可表示为：

𝐴𝑃 = ∑𝑛−1(𝑟𝑖+1 − 𝑟𝑖)𝑃𝑖𝑛𝑡𝑒𝑟𝑝(𝑟𝑖+1)

𝑖=1

(2.6)

其中，𝑟1，𝑟2 … … 𝑟𝑛是按升序排列的 Precision 插值段第一个插值处对应的

recall 值，𝑃𝑖𝑛𝑡𝑒𝑟𝑝(𝑟𝑖+1)为第𝑖 + 1个插值精确度，可表示为：

𝑃𝑖𝑛𝑡𝑒𝑟𝑝(𝑟) = max 𝑟, (2.7)

𝑟′≥𝑟

②计算所有类别的平均 AP。将所有类型的 AP 值求平均，得到 mAP 值。公式可表示为：

𝑚𝐴𝑃 =

𝐾

𝑖=1

∑

𝐴𝑃𝑖

(2.8)

𝐾

1. 精确率(precision)和召回率(recall)。除 mAP 指标外，精确率和召回率同样也是重要的性能评价指标。

①精确率：表示为分类正确的正样本个数与分类后判别为正样本个 数的比值，衡量的是一个分类器分出来的正样本的确是正样本的概率，公式如式下：

𝑃 = 𝑇𝑃

𝑇𝑃+𝐹𝑃

(2.9)

②召回率：表示为分类正确的正样本数与真正的正样本数的比值，衡量 的是一个分类器能把所有的正样本都找出来的能力，在通常情况下，精确率越高 则召回率越低，公式如式下：

𝑅 = 𝑇𝑃

𝑇𝑃+𝐹𝑁

(2.10)

以上公式中， 𝑇𝑃表示真阳性，𝐹𝑃表示假阳性，𝐹𝑁表示假阴性，在目标检测任务中，通过计算 IOU 值与阈值之间的大小关系给出判定，若大于阈值则为判断为真阳性，反之为假阳性，若无法区别该类型则判断为假阴性。为确保评价标准及参考系的一致性，针对基于 coco 数据集的目标检测任务模型的评价，采取动态 IoU 的方式计算 AP 和 mAP, 本文主要计算基于所有大小目标的三类 mAP，分别为

𝑚𝐴𝑃𝐴𝑃@0.5/𝑚𝐴𝑃𝐴𝑃@0.5:0.95/𝑚𝐴𝑃𝐴𝑃@0.75。[AP@0.75](mailto:AP@0.75) 代表用 0.75 的 IoU 计算 AP 值。

### 目标检测任务攻击有效性评价指标

1. 攻击成功率 （Attack Success Rate，ASR）。根据目标检测任务的特点及其性能评价指标，在结合图像分类任务攻击有效性评价指标的基础上，提出类似评价标准。该评价标准，使用攻击前后的 mAP 值的下降幅度占攻击前 mAP 值的百分比，来评价攻击效果的成功率。公式如下：

𝐴𝑆𝑅𝑜𝑑

= 𝑚𝐴𝑃𝑜𝑟𝑎𝑙−𝑚𝐴𝑃𝑎𝑡𝑡𝑎𝑐𝑘

𝑚𝐴𝑃𝑜𝑟𝑎𝑙

(2.11)

1. 分类置信度偏差（Classification Confidence Bias，CCB）。其定义为攻击样本对模型各目标分类置信度均值影响的偏差之和，通过计算攻击前后同类型多个目标的置信度均值偏差，再将所有类型偏差求和，以评估针对目标识别任务的对抗样本攻击效果，CCB值越大说明攻击效果越好，计算公式如下：

CCB = 1

∑𝑐𝑙𝑎𝑠𝑠( 1 ∑𝑛𝑖

′

𝑝 (𝑥|𝑖) − ∑ 𝑖

1 𝑛

𝑛

𝑝′

(𝑥

|𝑖))

(2.12)

𝑐𝑙𝑎𝑠𝑠

𝑖=1

𝑛𝑖

𝑗=1 𝑗

′ 𝑗′=1

𝑖

𝑗′

𝑎𝑑𝑣

其中，𝑥和𝑥𝑎𝑑𝑣分别为输入样本和对抗样本，𝑐𝑙𝑎𝑠𝑠是计划用于判断置信度偏差的类型集合， 这里使用输入样本的目标识别类型作为类型集合， 𝑝𝑗(𝑥|𝑖) 和

𝑝′ ′(𝑥𝑎𝑑𝑣|𝑖)分别为输入样本、对抗样本分类为 i 的各自第𝑗和𝑗′个目标的识别置信度，公式中，若𝑛′ = 0说明对抗样本无法识别出类型 i，说明针对该类型的攻击成

𝑗

𝑖

𝑛

′

功，此时规定1

∑𝑛𝑖 𝑝 (𝑥|𝑖) − 1 ∑ 𝑖

𝑝′ ′(𝑥

|𝑖) = 1，由于𝑐𝑙𝑎𝑠𝑠源自输入样本，

固𝑛′ ≠ 0。

𝑖

𝑛

𝑛𝑖

𝑗=1 𝑗

′ 𝑗′=1 𝑗

𝑖

𝑎𝑑𝑣

其值域范围(−𝑝𝑡, 1],其中𝑝𝑡为模型分类输出的置信度下限（即当类别置信度高于𝑝𝑡时输出该类别），值越大说明造成的攻击影响越大，值为 1 时，说明原目标

类别经过对抗处理后，置信度已低于 0。

𝐶𝐶𝐵指标更侧重于对目标的误分类能力的评价，不考虑对边界框大小、位置的偏移影响，而𝐴𝑆𝑅𝑜𝑑将边界框的变化也引入到攻击能力的评价体系，前者对于后者的评价更加严格，后者对比前者的评价更加全面，所以，一般情况下，𝐶𝐶𝐵的值会低于𝐴𝑆𝑅𝑜𝑑的评价结果。

（2）攻击迁移率（Attack Mobility Rate，AMR）。其定义为黑盒条件下的攻击成功率与白盒条件下的攻击成功率的比值，以评估攻击方法由白盒转移到黑盒状态的迁移效果。𝐴𝑀𝑅值越大说明攻击的迁移效果越好,计算公式如下：

## 本章小结

𝐴𝑀𝑅 = 𝐴𝑆𝑅𝑏𝑙𝑎𝑐𝑘

𝐴𝑆𝑅𝑤ℎ𝑖𝑡𝑒

(2.13)

本章主要介绍了图像分类和目标检测任务中的多种经典模型、算法及数据集，并探讨了针对目标检测任务的对抗攻击方法及其评价标准。

在图像分类部分，详细介绍了三种经典的卷积神经网络模型：VGG19、ResNet50和 Inception V3。VGG19 通过堆叠多个 3x3 卷积层来增加网络深度，尽管其计算复杂度较高，但在 ImageNet 数据集上取得了 90.88%的 top-5 识别准确率。 ResNet50 则引入了残差块和快捷连接，解决了深度网络中的梯度消失问题，达到了 92.88%的 top-5 识别准确率，表现优于 VGG 系列。Inception V3 利用并行的多卷积核模块，能够捕捉不同尺度的图像特征，并在计算成本上进行了优化，取得了93.45%的top-5 识别准确率。这三种模型各具特色，代表了不同的设计思路，为后续实验提供了基础。

在目标检测部分，介绍了五种经典网络模型，包括 YOLOv5、SSD、FCOS、 RetinaNet 和 Mask R-CNN。YOLOv5 作为单阶段检测算法，通过引入 Mosaic 数据增强和自适应锚框计算等技术，实现了高效的目标检测。SSD 则通过特征金字塔结构，能够快速处理不同尺度的目标，适合实时应用。RetinaNet 通过 Focal Loss解决类别不平衡问题，提升了模型在复杂背景下的鲁棒性。FCOS 则采用无锚框设计，简化了目标检测过程，提高了小目标的检测能力。Mask R-CNN 在 Faster R-

CNN 的基础上扩展，增加了实例分割功能，适用于需要高精度检测的场景。

此外，本章还探讨了针对目标检测的对抗攻击方法，包括 DAG、R-AP 和 EA 等算法，分析了它们的攻击原理和实现方式。最后，介绍了图像分类和目标检测任务的评价指标，如攻击成功率、平均精度均值（mAP）等，为后续研究提供了理论基础和数据支持。通过对这些模型和方法的综合分析，本章为深入研究目标检测模型的易受攻击性和防御机制奠定了基础。

# 基于多尺度特征网格群动态优化的目标检测对抗攻击方法

## 引言

现有针对目标检测的对抗样本生成方法，能够在指定环境条件下达到不错的攻击效果，但白盒攻击转移能力差、黑盒样本生成周期长的壁垒问题却很难往前突破，即便是基于 GAN 网络和基于 patch 的对抗算法，也仅仅是将资源消耗由前端转移到了后端，或是通过降低对抗性的方法来抵消资源的消耗。比如，DAG 算法，针对白盒条件的基于 PASCALVOC 2007 数据集的 Faster-RCNN 模型进行攻击，可使得 mAP 由 70%下降至 5%，成功率 [ASR@0.50](mailto:ASR@0.50) 约 92.85%，但将对抗样本转移至 SSD300 网络后，mAP 只能由 68%下降至 64%，成功率 [ASR@0.50](mailto:ASR@0.50) 约 5.88%、转移率 [AMR@0.50](mailto:AMR@0.50) 约 6.33%，生成对抗样本平均耗时约 9.3s;R-AP 算法，针对黑盒条件下的基于 MS COCO 2014 数据集的 Faster-RCNN 模型进行攻击，可使得 mAP 由 59.5%下降至 31.3%，成功率 [ASR@0.50](mailto:ASR@0.50) 约 47.3%；EA 算法，同样为黑盒攻击，针对 YOLOv3模型成功率 [ASR@0.50](mailto:ASR@0.50) 约 84%，针对 Faster-RCNN 模型成功率 [ASR@0.50](mailto:ASR@0.50) 约 48%，但生成对抗样本平均需要查询 35,000 次，耗时较长。以上结论采用的 PSNR 阈值均为 25dB。

针对以上问题，本章借鉴 C&W 对抗样本生成架构、DAG 白盒模型激活值选取、 R-AP 物体标记重置以及 EA 粒子群优化的思路，提出了一种基于多尺度特征网格群动态优化的目标检测对抗方法 DOG（Dynamic Optimization of Multi-Scale Feature Grid Cluster）。DOG 方法的总体思路为：迭代优化对抗噪声，使得每次迭代后，前次加入对抗噪声的图像能够向着经过目标类别诱偏后的多尺度特征网格群移动，该网格群由多尺度候选区域类别概率分布经过原背景重置、最优类别替换处理后构成，每次迭代都会进行一次最优类别计算，并进行物体类别置换。实验结果表明，基于该方法可使基于 MS COCO 2017 数据集的 YOLOv5s、YOLOv5m、 SSD300、RentinaNet、FCOS、MaskRCNN 等 6 类目标检测网络，在白盒条件下 [ASR@0.50](mailto:ASR@0.50)≥97%、CCB≥75%，在黑盒条件下 [ASR@0.50](mailto:ASR@0.50)≥51%、CCB≥48%、[AMR@0.50](mailto:AMR@0.50)

≥51%，对抗样本生成平均耗时约 2.4s(最小 0.9s、最大 4.3s)；同时经过验证，

生成的对抗样本能够对基于 ImageNet 数据集的 VGG19、ResNet50、Inception V3等三类图像分类网络，造成一定的攻击效果，ASR≥47%、CCB≥49%，以上测试结果所有使用的对抗样本 PSNR≥25dB。基本上在通用性、对抗性、转移性、时效性上做到了多方向均衡。

在本章中，我们将详细介绍 DOG 方法的设计原理、关键组件的作用及相互关联，以及在 9 类模型中的实际对抗效果。

## DOG 攻击方法

### 符号及问题描述

设𝐼0为原始输入图像，其中包含𝑛个目标共有𝑘个物体类别，G(·)表示多尺度候选区域类别概率分布输出函数, 𝐼𝑡为第 t 次迭代后的输出图像，他由𝐼0和第 t 次迭代后的噪声𝑁𝑡叠加后经𝑐𝑙𝑎𝑚𝑝(·)函数处理后获得。𝐺(𝐼0)表示给定输入图像𝐼0生成包含多个候选区域的类别概率分布矩阵，一般形状为(𝑆, 𝐶),其中 S 根据所选白盒模型的不同而不同，一些网络还与输入图像 I 的尺寸大小有关，如 SSD 模型 S=8732,C 为模型实际可分类数量，一般与训练使用的数据集有关，本文基于目标检测的模型，都是基于 MS COCO 数据集训练的，故而 C=91，其中包括 1 到 90 号物体类别和 0 号背景类别。

### DOG 攻击方法

DOG（Dynamic Optimization of Multi-Scale Feature Grid Cluster）攻击以𝐺(𝐼0~𝑡)作为媒介，利用了 Adam 优化器对噪声𝑁𝑡进行迭代优化，使得原图识别类别𝑇𝑜𝑟𝑎𝑙逐步向诱偏类别𝑇𝑖𝑛𝑑𝑢𝑐𝑒𝑡转移，其间加入最优诱偏类别动态选择，可按照最小损失或最大损失选择最合适的诱偏类别，设置了 PSNR 过限、全部目标类别识别出错和达到最大迭代次数三种停止迭代条件，在保证峰值信噪比 PSNR≥ 𝜀（一般不低于20dB，本章实验设置为25dB）的条件下尽可能多的诱偏𝑇𝑜𝑟𝑎𝑙，使模型最终识别出错。DOG 攻击方法可分解为攻击初始化、噪声迭代优化、输出对抗噪声三个步骤，其结构原理如图 3-1 所示，下面将分别对这三个步骤逐一进行介绍。

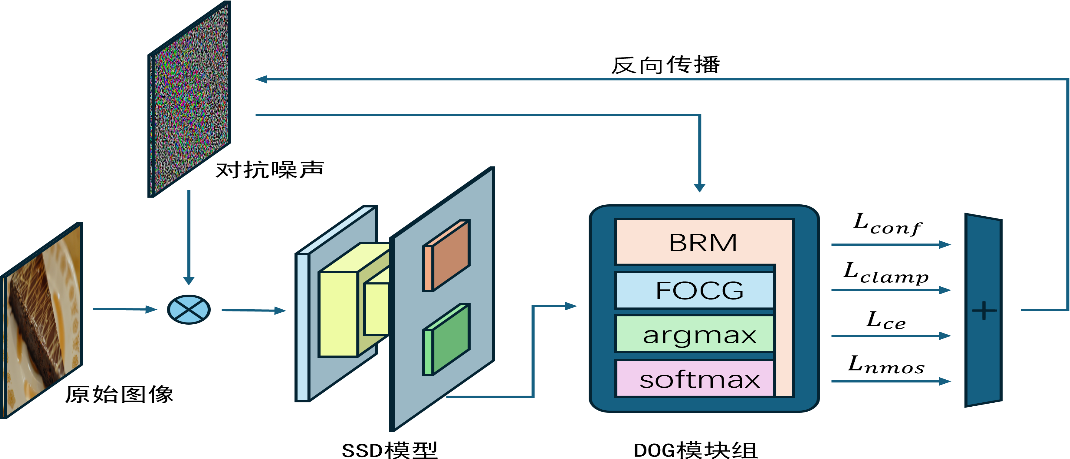


图 3-1 DOG 攻击方法结构原理图

* + - 1. 攻击初始化。该步骤主要完成输入参数的准备及执行初始化，初始背景重置矩阵的构建，以及噪声的初始化等工作。**一是**注入起算参数。主要包括 5 项参数，分别是输入图像𝐼0、进行白盒攻击的模型𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(·)、与模型相对应的能够同步提取类 FPN 结构输出的𝐺(·)跟踪接口、待攻击目标类别集合𝑠𝑒𝑡(𝑘)、用于调整正则项𝐿𝑐𝑜𝑛𝑓损失的权重参数𝑐（见 3.2.4（1）置信度对比损失𝐿𝑐𝑜𝑛𝑓）。**二是**执行初始化，计算获得𝑇𝑜𝑟𝑎𝑙 = 𝑎𝑟𝑔𝑚𝑎𝑥{𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼0)]}，𝑁0 = 𝑍𝑟𝑒𝑜\_𝑙𝑖𝑘𝑒(𝐼0),设置将对抗噪声𝑁0采用 Adam 优化器进行优化。**三是**构建初始背景重置矩阵𝑅𝐹𝑜𝑟𝑎𝑙。构建形状同𝑇𝑜𝑟𝑎𝑙的 0-1 矩阵，其中𝑇𝑜𝑟𝑎𝑙元素值在𝑠𝑒𝑡(𝑘)中的设置为 1，其他则设置为 0，起到屏蔽非关键类别的目的，具体见 6.2.6 背景重置。
      2. 噪声迭代优化。该步骤为对抗噪声迭代生成的关键环节，主要完成对抗图像的生成、前向传播及模型监控参数提取、构造诱偏类别集、构造迭代背景重置、损失计算及反向传播等。**一是**对抗图像的生成。将𝐼𝑡 = 𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡−1)，t为迭代轮次。**二是**前向传播及模型监控参数提取。执行𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(𝐼𝑡),抓取𝐺(𝐼𝑡)，计算获得 t 轮次预测类别集𝑇𝑝𝑟𝑒𝑡 = 𝑎𝑟𝑔𝑚𝑎𝑥{𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼𝑡)]}。**三是**构造诱偏类别集𝑇𝑖𝑛𝑑𝑢𝑐𝑒𝑡，定义 FOCG 方法能够根据当前对抗样本输出结果动态寻找最优的诱偏类别集，具体见 3.2.5 FOCG 最优诱偏类别动态选择。**四是**损失计算及反向传播。基于设定的损失函数计算四类损失，并将损失反向传播，更新下一轮的对抗噪声，具体见 3.2.4 损失函数。
      3. 输出对抗噪声。经过多次迭代后，如果满足 PSNR < 𝜀（本实验设置为 25dB）或𝑠𝑒𝑡(𝑇𝑜𝑟𝑎𝑙 × 𝑅𝐹𝑜𝑟𝑎𝑙))不再包含𝑠𝑒𝑡(𝑘)或达到最大迭代上限（本实验设置为

5 次），则退出循环，输出最终对抗噪声𝑁𝑓𝑖𝑛𝑎𝑙 = 𝐼𝑓𝑖𝑛𝑎𝑙 − 𝐼0,需要注意的是，虽然引入了噪声越界偏差损失𝐿𝑐𝑙𝑎𝑚𝑝但最后一轮迭代的对抗噪声也会存在超过图像设定值域范围的情况，不能把生成的对抗噪声直接输出作为生成结果，需要将最后一次的经过处理后的对抗样本与原图像求差才是最终的对抗噪声。

### 多尺度特征网格群

现有目标检测模型按照执行阶段可分为单阶段型和双阶段型，按照推理策略可分为回归型和建议型，按照定位方式可分为有锚框和无锚框。虽然模型类型多样、结构差异较大，但为了尽可能准确的捕捉图像中的目标，均采用了多尺度特征金字塔结构（FPN）来提取图像特征，将这些不同尺度的特征拼接成一个更大的特征群进行集中推理和计算。本章所提出的方法具有较好的转移性，便是基于这个共性特点。

本小节所提出的多尺度特征网格群，既由 FPN 结构输出的二维张量𝐺(𝐼0~𝑡)，两个维度分别代表候选框（或者候选中心点）的数量和模型的输出类别数量，通过监测每次输入噪声对其造成的影响，并通过损失函数进行梯度反向传播，不断优化对抗噪声。

### 损失函数

我们的目标是寻找添加到图像 I 上的引起最小偏差的最快迭代扰动，以使得目标类型判别出错。这里定义的损失函数包括置信度对比损失𝐿𝑐𝑜𝑛𝑓、噪声越界偏差损失𝐿𝑐𝑙𝑎𝑚𝑝、交叉熵损失𝐿𝑐𝑒、噪声强度损失𝐿𝑛𝑚𝑜𝑠。其公式如 3.1：

𝐿 = 𝑚𝑎𝑥(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠) (3.1)

其中，c 为置信度对比损失的权重系数,由于𝐿𝑐𝑜𝑛𝑓是一个张量，其他三类损失都是标量，相加会进行自动广播处理，最后取最大值作为最终的损失函数值，这里选取最大值的原因有三，一是 DOG 攻击方法中采用了 BRM 背景重置策略，会把非必要元素置零，而类别标签值在𝑠𝑒𝑡(𝑘)中的进行保留，且保留元素占比不超过 10%（标签 1 除外），基于全元素进行计算会带来很大的误差；二是𝐿𝑐𝑜𝑛𝑓的值不会

超过 1，使其相对于其他三个损失对总体损失的影响较小（一般是 8 倍以上的关系），取最大值可以保证在很少的迭代周期内𝐿𝑐𝑜𝑛𝑓便能达到趋于零的效果，实验证明，经过 1 到 2 轮次迭代该项目值就会回归至10−4数量级；三是首次计算时，由于对抗噪声𝑁0为全零张量，使得𝐿𝑐𝑙𝑎𝑚𝑝和𝐿𝑛𝑚𝑜𝑠均为零值，此时主要对𝐿𝑐𝑜𝑛𝑓和

𝐿𝑐𝑒进行修正，采用最大值可以在权重占比最高的时候快速得到反馈补偿。

1. 置信度对比损失𝐿𝑐𝑜𝑛𝑓。表示了模型预测的输出与目标输出之间的差异，通过缩小之间的差距，实现预测结果逐步向预定变化方向移动的目的，作为损失函数的正则项使用。其公式如 3.2：

𝐿𝑐𝑜𝑛𝑓 = 𝑟𝑒𝑙𝑢(𝑍𝑖 − 𝑍𝑡) (3.2)

其中，𝑍𝑖为针对当前输入图像的最大预估类别的概率张量（softmax 输出中类的概率），𝑍𝑡为针对当前输入图像的计划诱偏类别的概率张量（softmax 输出中类的概率），𝑟𝑒𝑙𝑢(·)为 relu 激活函数，能够把结果小于 0 的部分置 0，大于 0 的保持不变，只有当 𝑍𝑖大于𝑍𝑡 时，损失才会产生，使得噪声迭代方向始终朝着目标类别移动。为进一步提升目标转移的精度和稳定性，𝑍𝑖、𝑍𝑡都进行了背景重置操作，详见 3.2.6 BRM 背景重置。

1. 噪声越界偏差损失𝐿𝑐𝑙𝑎𝑚𝑝。通过迭代法生成的噪声在叠加到原图像上时会出现超出图像设定值域范围的情况（超出[0,1]区间），这里定义了𝐿𝑐𝑙𝑎𝑚𝑝使得生成的对抗噪声叠加到原图像上后，尽可能不超出值域范围。其公式如 3.3：

𝐿𝑐𝑙𝑎𝑚𝑝 = ‖𝐼0 + 𝑁𝑡 − 𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡)‖2 (3.3)

其中，𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡)既为𝐼𝑡+1, 𝑐𝑙𝑎𝑚𝑝(·)函数能够把值调制到[0,1]区间。

1. 交叉熵损失𝐿𝑐𝑒。计算当前轮次模型输出的类别概率与诱偏类别之间的差异。 其公式如 3.4：

𝐿𝑐𝑒 = − ∑𝐶

𝑚=1

y𝑚log (𝑙𝑚)

(3.4)

其中，y𝑚是𝐺(𝐼𝑡)的输出矩阵元素，𝑙𝑚是当前轮次设定的诱偏类别𝑇𝑖𝑛𝑑𝑢𝑐𝑒𝑡的元素，𝐶是模型类别总数。

1. 噪声强度损失𝐿𝑛𝑚𝑜𝑠。这一项计算了对抗噪声的均方值，表示对抗扰动的强度。其公式如 3.5：

𝐿𝑛𝑚𝑜𝑠

= − 1

𝑁𝑠

𝑁𝑠 2

𝑚=1 𝑚

∑

𝑛

(3.5)

其中，𝑁𝑠是对抗噪声𝑁的元素个数，𝑛𝑚是第 m 个元素的值。

### FOCG 最优诱偏类别动态选择

在进行每一轮次迭代前，需要确认诱偏类别𝑇𝑖𝑛𝑑𝑢𝑐𝑒𝑡 ，使得对抗噪声𝑁𝑡按照偏离原类别𝑇𝑜𝑟𝑎𝑙的方向移动，逐步降低原类别置信度，提升诱偏类别置信度，最终实现图像目标的错误判断。对于无目标攻击而言，如何提升攻击效果和攻击效率，选择合适的诱偏类别相当重要。

本小节，提出了一种基于交叉熵损失的最小、最大诱偏类别动态选择方法 FOCG(Find the Optimal Category Group)。具体为，每次迭代前计算模型类 FPN组件输出𝐺(𝐼0~𝑡)与𝑇𝑔𝑟𝑜𝑢𝑝𝑠𝑡之间的交叉熵，（其中，𝑇𝑔𝑟𝑜𝑢𝑝𝑠𝑡指代除去𝑇𝑜𝑟𝑎𝑙以外的单类别集合，以 COCO 数据集 90 个事件类别为例，当𝑇𝑜𝑟𝑎𝑙包含类别 1、2 时，𝑇𝑔𝑟𝑜𝑢𝑝𝑠𝑡即为分别只包含 3 到 90 的 87 个单类别集合体），按照交叉熵值从小到大进行排序，若选最小诱偏类别则取第一个作为本轮次最优诱偏类别，若选最大诱偏类别则取最后一个作为最优诱偏类别，每次迭代都会重新选择最优解进行本轮此损失计算。

最小诱偏类别动态选择方案，即每次迭代都选择距离当前输出结果最近的且与𝑇𝑔𝑟𝑜𝑢𝑝𝑠𝑡一致的类别作为诱偏类别，该方案旨在控制对抗噪声引起的 PSNR 值波动，牺牲生成速度换取较好的对抗图像质量。

最大诱偏类别动态选择方案，与最小方案刚好相反，以对抗图像质量换取生成速度的提升，由于在 DOG 方法加入了 PSNR 判断和迭代次数判断的退出迭代条件，对于目标较多（大于 5 个）、特别是人物较多的图像，采取该方案会导致个别类别 AP 下降过快，影响 mAP 和 ASR 指标数据。

根据 DOG 方法的结构特点，基于该诱偏类别选择方法也可实现有目标攻击，但同最大诱偏类别动态选择方案所描述的一样，会带来指标数据迅速下降可能。

### BRM 背景重置

从 3.2.2 可知，使用多尺度特征网格群作为白盒攻击的切入点、输出变化的反馈层，优点显而易见，就是多模型的通用性，同时它的缺点也暴露无遗，就是

根据所选白盒模型的不同、输入图片尺寸的不同，其资源消耗各有不同，比如使用 SSD 模型，需要对 8732 个候选框进行计算，但如果选用 FCOS 模型，其候选区域个数提升到万级别，选用 RetinaNet 模型，其候选框个数提升到了十万级别，这将使得交叉熵损失𝐿𝑐𝑒过大，收敛速度会加快，容易过拟合，需要用更小的学习率和更多的迭代次数进行训练，才能获得较好的对抗样本。

故此，引入背景重置 BRM(Background Reset Matrix)。背景重置顾名思义，就是将图像的网格群一一映射的类别群𝑇𝑜𝑟𝑎𝑙/𝑡中的值不在𝑠𝑒𝑡(𝑘)的替换为背景，保留原图目标类别标签的操作。具体实现方法就是，复制一个与初始类别群𝑇𝑜𝑟𝑎𝑙完全相同 𝑅𝐹𝑐𝑜𝑝𝑦，将𝑅𝐹𝑐𝑜𝑝𝑦中值在𝑠𝑒𝑡(𝑘)中的设置为 1，其他设置为 0，构建背景重置矩阵𝑅𝐹𝑜𝑟𝑎𝑙，第一次迭代将𝑅𝐹𝑜𝑟𝑎𝑙复制给𝑅𝐹𝑟𝑒𝑠𝑒𝑡，进行𝐿𝑐𝑜𝑛𝑓、𝐿𝑐𝑒计算前，将计算输出参数乘以𝑅𝐹𝑟𝑒𝑠𝑒𝑡进行背景重置。而后每进行一次迭代，需要判断𝑇𝑝𝑟𝑒𝑡 ×

𝑅𝐹𝑜𝑟𝑎𝑙中有无不在𝑇𝑜𝑟𝑎𝑙中的类别产生（其中，𝑇𝑝𝑟𝑒𝑡 = 𝑎𝑟𝑔𝑚𝑎𝑥[𝐺(𝐼t)],既第 t 轮次

输出类别集），并按照有则保持无则置零的原则，基于𝑅𝐹𝑜𝑟𝑎𝑙重新构建𝑅𝐹𝑟𝑒𝑠𝑒𝑡，用于本轮次损失计算。

### 算法实现

本小节，将对以上提及的对 DOG 算法、背景重置、最优诱偏类别动态选择算法实现进行补充。

1. DOG 攻击算法。文字描述详见 3.2.2 DOG 攻击方法描述。

|  |
| --- |
| 算法 3.1: DOG（Dynamic Optimization of Multi-Scale Feature Grid Cluster） |
| 输入：原图像𝐼0，白盒模型𝑎𝑡𝑡𝑐𝑚𝑜𝑑𝑒𝑙(·)，FPN 输出函数𝐺(·)，待攻击目标类别集合  𝑠𝑒𝑡(𝑘)，𝐴𝑑𝑚(·)优化器，𝐵𝑅𝑀(·)，𝐹𝐶𝑂𝐺(·)，𝑜𝑛𝑒\_ℎ𝑜𝑡(·)，成像质量阈值𝜀，最大迭代次数𝑡𝑠，噪声迭代学习率𝑙𝑟, 置信度正则参数𝑐*，*  输出：最终对抗噪声𝑁𝑓𝑖𝑛𝑎𝑙  初始化：𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(𝐼0)，𝑇𝑜𝑟𝑎𝑙 = 𝑎𝑟𝑔𝑚𝑎𝑥{𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼0)]}*，*𝑁0 = 𝑍𝑟𝑒𝑜\_𝑙𝑖𝑘𝑒(𝐼0)，  𝑜𝑝𝑡𝑖𝑚𝑖𝑧𝑒𝑟 = 𝐴𝑑𝑚(𝑁0, 𝑙𝑟)，𝑅𝐹𝑜𝑟𝑎𝑙 = 𝐵𝑅𝑀 (𝑇𝑜𝑟𝑎𝑙*，*𝑠𝑒𝑡(𝑘))  1. 𝑤ℎ𝑖𝑙𝑒(1):  2. 𝐼𝑡 = 𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡)  3. 𝑁𝑜𝑢𝑡 = 𝐼𝑡 − 𝐼0 |

4. 𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(𝐼𝑡)，𝑇𝑝𝑟𝑒𝑡 = 𝑎𝑟𝑔𝑚𝑎𝑥[𝐺(𝐼𝑡)]

5. 𝒊𝒇 𝑠𝑒𝑡(𝑇𝑝𝑟𝑒𝑡 ) 𝑛𝑜𝑡 𝑖𝑛 𝑠𝑒𝑡(𝑘) 𝑜𝑟 𝑡𝑖𝑚𝑒𝑠 > 𝑡𝑠 𝒕𝒉𝒆𝒏 𝑏𝑟𝑒𝑎𝑘 //退出条件

6. 𝒊𝒇 𝑃𝑆𝑁𝑅 ≤ 𝜀 𝒕𝒉𝒆𝒏 𝑟𝑜𝑙𝑙 𝑏𝑎𝑐𝑘 //回滚条件

7. 𝑇𝑖𝑑𝑐𝑡 = 𝐹𝐶𝑂𝐺 (𝐺(𝐼𝑡)*，*𝑠𝑒𝑡(𝑘)) //选择本轮次诱偏类别𝑇𝑖𝑑𝑐𝑡

8. 𝑅𝐹𝑟𝑒𝑠𝑒𝑡 = 𝐵𝑅𝑀 (𝑇𝑝𝑟𝑒𝑡 *，*𝑠𝑒𝑡(𝑘)) ∗ 𝑅𝐹𝑜𝑟𝑎𝑙

9. 𝑍𝑡 = 𝑠𝑢𝑚(𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼𝑡)] ∗ 𝑜𝑛𝑒\_ℎ𝑜𝑡(𝑇𝑖𝑑𝑐𝑡 ), 𝑑𝑖𝑚 = 1) ∗ 𝑅𝐹𝑟𝑒𝑠𝑒𝑡 //诱偏类别概率

10. 𝑍𝑖 = 𝑠𝑢𝑚(𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼𝑡)] ∗ 𝑜𝑛𝑒\_ℎ𝑜𝑡(𝑇𝑝𝑟𝑒𝑡 ), 𝑑𝑖𝑚 = 1) ∗ 𝑅𝐹𝑟𝑒𝑠𝑒𝑡 //预测类别概率

11. 𝐿𝑐𝑜𝑛𝑓 = 𝑟𝑒𝑙𝑢(𝑍𝑖 − 𝑍𝑡)

12. 𝐿𝑐𝑙𝑎𝑚𝑝 = ‖𝑁𝑡 − 𝑁𝑜𝑢𝑡‖2

13. 𝐿𝑐𝑒 = 𝐶𝑟𝑜𝑠𝑠EntropyLoss(𝐺(𝐼𝑡) ∗ 𝑅𝐹𝑟𝑒𝑠𝑡. 𝑣𝑖𝑒𝑤(−1,1), 𝑇𝑖𝑑𝑐𝑡𝑖 ∗ 𝑅𝐹𝑟𝑒𝑠𝑡 )

14. 𝐿𝑛𝑚𝑜𝑠 = 𝑚𝑒𝑎𝑛(𝑁𝑡 ∗∗ 2)

15. 𝐿 = 𝑚𝑎𝑥(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠)

*16.* 𝐿. 𝑏𝑎𝑐𝑘𝑤𝑎𝑟𝑑, 𝑜𝑝𝑡𝑖𝑚𝑖𝑧𝑒𝑟. 𝑠𝑡𝑒𝑝, 𝑜𝑝𝑡𝑖𝑚𝑖𝑧𝑒𝑟. 𝑧𝑒𝑟𝑜\_𝑔𝑟𝑎𝑑

*17.* 𝑁𝑓𝑖𝑛𝑎𝑙 = 𝑁𝑜𝑢𝑡

18. 𝑅𝑒𝑡𝑢𝑟𝑛 𝑁𝑓𝑖𝑛𝑎𝑙

1. FOCG 最优诱偏类别动态选择。该算法仅展示单次迭代的实现逻辑。

|  |
| --- |
| 算法 3.2: FOCG(find the optimal category group) |
| 输入：第 t+1 次迭代的输入图像𝐼𝑡，原图像类别集𝑇𝑜𝑟𝑎𝑙, 待攻击类别集合𝑠𝑒𝑡(𝑘)，FPN 输出函数𝐺(·)  输出：最优诱偏类别集𝑇𝑖𝑑𝑐𝑡  1. 𝑠𝑒𝑡1 = 𝑠𝑒𝑡(1~90) − 𝑠𝑒𝑡(𝑘) //获得备选类别集合，按 90 个分类计  2. 𝒇𝒐𝒓 𝑖 𝑖𝑛 𝑠𝑒𝑡1:  3. 𝑇𝑖𝑑𝑐𝑡𝑖 = 𝑠ℎ𝑎𝑝𝑒\_𝑙𝑖𝑘𝑒(𝑇𝑜𝑟𝑎𝑙, 𝑣𝑎𝑙𝑢𝑒𝑠 = 𝑖) //构建由 i 个类别组成的𝑇𝑖𝑑𝑐𝑡𝑖  4. 𝑙𝑜𝑠𝑠𝑠𝑒𝑡1 += 𝐶𝑟𝑜𝑠𝑠EntropyLoss(𝐺(𝐼𝑡), 𝑇𝑖𝑑𝑐𝑡𝑖 )  5. 𝑇𝑖𝑑𝑐𝑠𝑒𝑡1 += 𝑇𝑖𝑑𝑐𝑡𝑖  6. 𝑇𝑖𝑑𝑐𝑠𝑒𝑡1 = 𝑜𝑟𝑑𝑒𝑟((𝑙𝑜𝑠𝑠𝑠𝑒𝑡1,𝑇𝑖𝑑𝑐𝑠𝑒𝑡1 ), 𝑘𝑒𝑦 = 𝑙𝑜𝑠𝑠𝑠𝑒𝑡1,, 𝑎𝑠𝑐𝑒𝑛𝑑𝑖𝑛𝑔)[1] //以 loss 值排序  7. 𝑖𝑓 min\_𝑤𝑎𝑦*：*  8. 𝑇𝑖𝑑𝑐𝑡 = 𝑇𝑖𝑑𝑐𝑠𝑒𝑡1 [0]  9. 𝑒𝑙𝑖𝑓 max\_𝑤𝑎𝑦*：*  𝑇𝑖𝑑𝑐𝑡 = 𝑇𝑖𝑑𝑐𝑠𝑒𝑡1 [−1] |

1. BRM 背景重置。该算法仅展示单次迭代的实现逻辑。

|  |
| --- |
| 算法 3.1: BRM(Background Reset Matrix) |
| 输入：原图像𝐼0，第 t+1 次迭代的输入图像𝐼𝑡，原图像类别集𝑇𝑜𝑟𝑎𝑙, FPN 输出函数𝐺(·)输出：第 t+1 次迭代的背景重置矩阵𝑅𝐹𝑟𝑒𝑠𝑒𝑡  10. 𝑅𝐹𝑜𝑟𝑎𝑙 = [𝒊𝒇 𝐺(𝐼0) 𝑖𝑛 𝑇𝑜𝑟𝑎𝑙 𝒕𝒉𝒆𝒏 1 𝒆𝒍𝒔𝒆 0]  11. 𝑅𝐹𝑟𝑒𝑠𝑒𝑡 = 𝑅𝐹𝑜𝑟𝑎𝑙  12. 𝒇𝒐𝒓 𝑖 𝑖𝑛 𝑟𝑎𝑛𝑔𝑒(𝑅𝐹𝑜𝑟𝑎𝑙. 𝑠𝑖𝑧𝑒(0)): //每轮迭代重复执行  13. 𝒊𝒇 𝑅𝐹𝑜𝑟𝑎𝑙[𝑖] == 1:  14. 𝒊𝒇 𝑎𝑟𝑔𝑚𝑎𝑥[𝐺(𝐼𝑡)] ∗ 𝑅𝐹𝑜𝑟𝑎𝑙[𝑖] 𝑛𝑜𝑡 𝑖𝑛 𝑇𝑜𝑟𝑎𝑙:  15. 𝑅𝐹𝑟𝑒𝑠𝑒𝑡 [𝑖] = 0 |

## 实验与结果分析

为了证明本章所提出的 DOG 算法的有效性、通用性、转移性，本节设置相应实验进行验证。首先就实验平台、实验数据、模型选择、评估指标等事项进行说明。其次，为验证 DOG 算法的有效性，这里对基于 MS COCO 数据集的预训练模型 SSD300做生成对抗样本实验，所提方法在白盒对抗中的各类别的 AP 值变化和总体 mAP 值变化。然后，为说明各损失函数组成部分和 FOCG 及 BRM 两个模块的作用，在前序实验的基础上进行各模块的消融实验对比分析。接着，为验证算法的转移性，将基于 SSD300 模型生成的对抗样本，在其他 3 类图像分类网络和 5 类目标检测网络中进行黑盒攻击验证。最后，为验证算法的高效性，与其他 3 类对抗算法进行对比实验。以上所涉及模型、数据集、对抗算法、评价指标均在第二章有详细介绍，本节中将不再进行说明。

### 实验基础搭建

1. 操作平台。

Windows 10 操作系统。

1. 硬件设备。

CPU：Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz GPU：NVIDIA GeForce RTX 2080 Ti

RAM：64.0 GB

1. 软件工具。

IDE：PyCharm 2024.1.2 (Professional Edition)

Python 版本：3.10

深度学习框架：PyTorch 2.3.1+cu121

1. 数据集选择。均使用 MS COCO 数据集的验证集进行实验。由于本章仅进行对抗验证，不涉及模型训练，在功能验证、转移性评价等所有涉及使用数据集的环节，均使用验证集中图像进行实验，使用验证集而不使用测试集的原因有三点，**一是**在训练模型时需要验证集参与阶段效果评估，故模型对于验证集数据的判断相比测试集更加准确，能够更好的反应 DOG 算法效果；**二是**验证集的标签相对于测试集更加完善，对于 AP 和 mAP 的评价更加具有说服力，如果使用无标签的照片，需要使用其他模型（如精度相对较高的 MaskRCNN）进行标签生成或人工标绘，而后在此基础上进行效果评价；**三是**验证集图像，一般按照类别均衡性原则进行选择，以覆盖所有样例，故使用验证集进行测试，可以规避随机样例带来的其他隐形问题。另外，由于验证集数据庞大，这里从前中后各取 200 张图像共

计 600 张图像进行试验，根据所选模型的不同，实际参与对抗样本生成的图片会

少于 600 张（以 SSD 模型为例，实际仅有 566 张图片参与对抗评估，且 80 个目标类别仅有 toaster 类别未覆盖到），因为在实验结构设计中，把模型无法从原图中的识别出目标的图片进行剔除，进一步提升实验准确性。

1. 模型选择。图像分类：基于 ImageNet 数据集的 VGG19、ResNet50、 Inception V3 预训练模型。目标检测：基于 MS COCO 数据集的 YOLOv5、SSD300、 RetinaNet、FCOS、MaskRCNN 预训练模型。
2. 评估指标。见 2.6 评价指标。

### 基于 SSD300 预训练网络的攻击有效性验证

为了验证本章提出的攻击方法的有效性，将 MS COCO 验证集中的 600 张干净图片进行基于 SSD300 预训练网络的白盒 DOG 攻击，共生成 566 张对抗样本（有 34 张干净原图 SSD 模型无法识别出目标，进行剔除操作）。分别将 566 张干净图

片和 566 张对抗图片送入 SSD300 网络进行推理预测与评估,其中，成像质量阈值

𝜀 = 25，最大迭代次数𝑡𝑠 = 5，噪声迭代学习率𝑙𝑟 = 0.05，置信度正则参数𝑐 = 1， FOCG 策略为最大诱偏类别动态选择方案。进行攻击前后效果对比见表 3-1、3-2，图 3-2。

从上表可知，把样例 1 中的熊攻击为长颈鹿，把样例 2 中的 6 个人、1 个风

筝攻击为 3 个人，且人的边界框也出现了较大偏差，把样例 3 中的蛋糕攻击为鸟，

|  |  |
| --- | --- |
|  | 表 3-1 基于 SSD 模型白盒攻击前后目标检测效果对比 |
| **原图** | 棕色的熊  描述已自动生成 小孩们在草地上放风筝  中度可信度描述已自动生成 盘子里有蛋糕  描述已自动生成 绿色的杯子  描述已自动生成 |
|  | 1 bear 6 person、1 kite 1 cake 1 cup、1 dining table |
| **对抗噪声** | 背景图案  描述已自动生成 图片包含 游戏机, 布  描述已自动生成 绿色的叶子  低可信度描述已自动生成 背景图案  描述已自动生成 |
| **对抗样本** | 棕色的熊  描述已自动生成 许多小孩在草地上放风筝  中度可信度描述已自动生成 桌子上的食物  低可信度描述已自动生成 桌子上的叉子  中度可信度描述已自动生成 |
|  | 1 giraffe 3 person 1 bird None |

把样例 4 中的杯子和餐桌攻击为了背景，由于采用了多尺度特征网格群作为迭代输入，故而对抗噪声在一定程度也呈现为网格状。

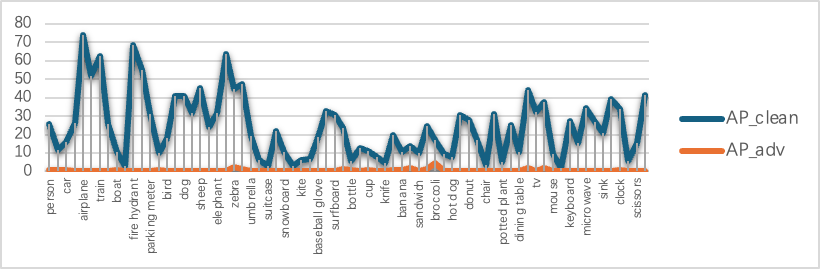


图 3-2 基于 SSD 模型白盒攻击前后类别 AP 值对比

如图 3-2 所示，纵坐标为 AP 标尺，横坐标为类别标签，攻击前后各类别受到了极大的抑制，如 airplane 类别 AP 值由 73.9%攻击到了 0%，所有类别攻击后的

AP 值最大不超过 4.6%（类别标签为 broccoli），绝大部分小于 0.5%。

表 3-2 基于 SSD 模型白盒攻击指标评估结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **TC(s)** | **CCB** |
| SSD  (白盒) | 0.50:0.95 | 0.2196 | 0.0048 | 97.80% | AVG:2.487  MIN:0.948 MAX:4.478 | 75.32% |
| 0.50 | 0.3135 | 0.0112 | 96.42% |
| 0.75 | 0.2494 | 0.0036 | 98.56% |

如表 3-2 所示，[采取该方法生成的对抗样本成功率 ASR@0.50](mailto:閲囧彇璇ユ柟娉曠敓鎴愮殑瀵规姉鏍锋湰鎴愬姛鐜嘇SR@0.50)：0.95 约 96.80%、 [ASR@0.50](mailto:ASR@0.50) 约 96.42%、[ASR@0.50](mailto:ASR@0.50) 约 98.56%，TC 均值约为 2.487s,CCB 值约 75.32%。

### 损失函数及功能模块消融实验

为进一步分析本章提出的损失函数组成及 FOCG 不同策略和 BRM 背景重置对生成对抗样本的影响，本小结区分损失函数、是否启用 FOCG 策略及 BRM 背景重置功能模块三个部分独立展开消融实验，比较不同情况下，生成的对抗样本在生成耗时、aP、mAP、ASR 及 CCB 等指标参数的影响。

三类消融实验将采用与 3.3.2 相同的模型架构，基于 SSD 预训练模型进行白盒攻击，从而生成 566 张对抗样本，参数设置，采用成像质量阈值𝜀 = 25，最大迭代次数𝑡𝑠 = 5，噪声迭代学习率𝑙𝑟 = 0.05，置信度正则参数𝑐 = 1的设置方式。

1. 针对损失函数的消融实验。根据 3.2.4 可知，本章提出的损失函数由置信度对比损失𝐿𝑐𝑜𝑛𝑓、噪声越界偏差损失𝐿𝑐𝑙𝑎𝑚𝑝、交叉熵损失𝐿𝑐𝑒、噪声强度损失

𝐿𝑛𝑚𝑜𝑠等四部分组成，该项实验，FOCG 策略为最大诱偏类别方案，并启用 BRM 模块，表 3-3 展示了依次增加不同部分后的各评价指标情况。

表 3-3 不同损失函数组成下的攻击指标评估结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **LOSS** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **TC(s)** | **CCB** |
|  | 0.50:0.95 | 0.2196 | 0.0070 | 96.81% | AVG:2.532 MIN:0.984  MAX:4.754 |  |
| 𝐿𝑐𝑜𝑛𝑓 | 0.50 | 0.3135 | 0.0139 | 95.56% | 73.64% |
| 0.75 | 0.2494 | 0.0059 | 97.60% |
|  |  |
|  | 0.50:0.95 | 0.2196 | 0.0058 | 97.36% | AVG:2.742 MIN:1.063 MAX:5.062 |  |
| 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 | 0.50 | 0.3135 | 0.0126 | 95.96% | 73.80% |
| 0.75 | 0.2494 | 0.0042 | 98.31% |
|  |  |
| 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝  + 𝐿𝑐𝑒 | 0.50:0.95 | 0.2196 | 0.0048 | **97.80%** | AVG:2.672 MIN:1.064 MAX:4.871 | 75.32% |
| 0.50 | 0.3135 | 0.0112 | 96.42% |
| 0.75 | 0.2494 | 0.0036 | 98.56% |
| 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝  + 𝐿𝑛𝑚𝑜𝑠 | 0.50:0.95 | 0.2196 | 0.0058 | 97.35% | AVG:2.550 MIN:0.989 MAX:4.763 | 73.70% |
| 0.50 | 0.3135 | 0.0127 | 95.96% |
| 0.75 | 0.2494 | 0.0042 | 98.30% |
| 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝  + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠 | 0.50:0.95 | 0.2196 | 0.0048 | 97.39% | **AVG:2.487 MIN:0.948**  **MAX:4.478** | **75.32%** |
| 0.50 | 0.3135 | 0.0112 | **96.46%** |
| 0.75 | 0.2494 | 0.0036 | **98.74%** |

从上表可知，通过不断增加损失函数的组成，对抗样本的各项指标均呈现逐

步变优趋势，其中，引入𝐿𝑐𝑙𝑎𝑚𝑝损失项，使得 TC 值平均增加 [0.21s,mAP@0.5](mailto:0.21s%2CmAP@0.5) 值下降约 1.3‰，[ASR@0.5](mailto:ASR@0.5)、[CCB@0.5](mailto:CCB@0.5) 值分别提升约 0.4%、0.16%；引入𝐿𝑐𝑒损失项，使得各项指标均得到提升，TC 值平均降低 0.07s, [mAP@0.5](mailto:mAP@0.5) 值下降约 1.4‰， [ASR@0.5](mailto:ASR@0.5)、[CCB@0.5](mailto:CCB@0.5) 值分别提升约 0.46%、1.52%;引入𝐿𝑛𝑚𝑜𝑠损失项，使得各项指标均得到提升，TC 值平均降低 0.185s，[ASR@0.5](mailto:ASR@0.5) 值提升约 0.5%, [mAP@0.5、](mailto:mAP@0.5銆丆CB@0.5) [CCB@0.5](mailto:mAP@0.5銆丆CB@0.5) 指标有提升但变化较小。

总体来看，本章提出的损失函数结构，能够有效提升对抗样本的攻击成功率

[ASR@0.5](mailto:ASR@0.5) 约 1%，在生成平均耗时上降低了约 0.045s。

1. 针对 FOCG 策略选择的消融实验。根据 3.2.5 可知，FOCG 模块主要用于确定每次迭代的诱偏类别集合，其策略可分为最小诱偏类别动态选择和最大诱偏类别动态选择两种，表 3-4 展示了分别采取最小、最大方案和背景类别后的评价指标情况，重点对比分析启用 FOCG 对于攻击模型性能的提升作用。

表 3-4 不同 FOCG 策略下的攻击指标评估结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **FOCG** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **TC(s)** | **CCB** |
| **启用最小**诱偏策略 | 0.50:0.95 | 0.2196 | 0.0033 | **98.48%** | AVG:2.716 MIN:1.008  MAX:5.486 | **75.37%** |
| 0.50 | 0.3135 | 0.0079 | **97.47%** |
| 0.75 | 0.2494 | 0.0024 | **99.02%** |
| **启用最大**诱偏策略 | 0.50:0.95 | 0.2196 | 0.0048 | 97.39% | AVG:2.487  MIN:0.948 MAX:4.478 | 75.32% |
| 0.50 | 0.3135 | 0.0112 | 96.46% |
| 0.75 | 0.2494 | 0.0036 | 98.74% |
| **不启用** FOCG,使  用类别 0 作为诱偏类别 | 0.50:0.95 | 0.2196 | 0.0058 | 97.36% | **AVG:2.152 MIN:0.825 MAX:4.290** | 74.07% |
| 0.50 | 0.3135 | 0.0128 | 95.90% |
| 0.75 | 0.2494 | 0.0046 | 98.14% |

从上表可知，启用 FOCG 策略能够有效提升对抗样本的攻击成功率，选择最小诱偏策略时，成功率最高，成功率 [ASR@0.50](mailto:ASR@0.50) 值提升约 1.12%，CCB 提升约 1.3%，但对抗样本平均生成时间增加也最多，TC@avg 增加约 0.564s;选择最大诱偏策略，成功率 [ASR@0.50](mailto:ASR@0.50) 值提升约 0.56%，CCB 提升约 1.25%，TC@avg 增加约 0.335s，比最小诱偏策略增加耗时要少。在对攻击成功率要求较高的任务场景，可采取 FOCG 的最小诱偏策略实施攻击，在对时间要求较高的任务场景，可以采取 FOCG的最大诱偏策略实施攻击。

1. 针对启用 BRM 背景重置功能的消融实验。根据 3.2.6 可知，设计 BRM 背景重置模块的主要目的是为了关注主要目标类别，实行针对性攻击。表 3-5 分别

展示了 SSD、FCOS、RetinaNet 三类模型在启用 BRM 和不启用 BRM 条件下的评价指标情况。FCOG 策略选用最小诱偏策略。

表 3-5 不同 FOCG 策略下的攻击指标评估结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **BRM** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **TC(s)** | **CCB** |
|  |  | 0.50:0.95 | 0.2196 | 0.0033 | **98.48%** | AVG:2.716 |  |
| 0.50 | 0.3135 | 0.0079 | **97.47%** |
|  | 启用 | MIN:0.977 | **75.37%** |
| SSD |  | 0.75 | 0.2494 | 0.0024 | **99.02%** | MAX:5.397 |  |
| 不启用 | 0.50:0.95 | 0.2196 | 0.0046 | 97.89% | **AVG:2.497 MIN:0.922**  **MAX:4.308** | 74.89% |
| 0.50 | 0.3135 | 0.0091 | 97.10% |
| 0.75 | 0.2494 | 0.0051 | 97.93% |
|  |  | 0.50:0.95 | 0.3705 | 0.0294 | **92.06%** | AVG:14.404 |  |
| 0.50 | 0.5093 | 0.0795 | **89.80%** |
|  | 启用 | **MIN:6.711** | **68.53%** |
| FCOS |  | 0.75 | 0.4006 | 0.0275 | **93.14%** | **MAX:20.314** |  |
| 不启用 | 0.50:0.95 | 0.3720 | 0.0512 | 86.24% | **AVG:12.875** MIN:7.863 MAX:20.846 | 67.73% |
| 0.50 | 0.5116 | 0.0795 | 84.47% |
| 0.75 | 0.4021 | 0.0564 | 85.97% |
|  |  | 0.50:0.95 | 0.3257 | 0.0099 | **96.90%** | AVG:23.143 |  |
| 0.50 | 0.4445 | 0.0166 | **96.27%** |
|  | 启用 | MIN:14.490 | **81.33%** |
| Retina |  | 0.75 | 0.3508 | 0.0102 | **97.10%** | MAX:33.487 |  |
| Net | 不启用 | 0.50:0.95 | 0.3257 | 0.0115 | 96.44% | **AVG:22.895 MIN:12.515**  **MAX:27.323** | 80.39% |
| 0.50 | 0.4445 | 0.0186 | 95.79% |
| 0.75 | 0.3508 | 0.0113 | 96.75% |

从上表可知，启用 BRM 模块后，由于每次迭代计算需要重新计算重置矩阵，故在 TC 指标上会有一定增加（增幅约 8%），其他主要指标参数如 ASR、CCB 都有不同程度的提升。

### 迁移性验证实验

本小结采用与 3.3.2 相同架构，基于 SSD 预训练模型生成 566 张对抗样本，再将原图和对抗样本分别送入 3 类图像分类预训练网络和 5 类目标检测网络进行参数评估。参数值设定为：成像质量阈值𝜀 = 20/25/30/35，最大迭代次数𝑡𝑠 = 5，噪声迭代学习率𝑙𝑟 = 0.05，置信度正则参数𝑐 = 1，由于该实验不考量生成时间指标，这里 FOCG 策略选用最小诱偏类别动态选择方案。表 3-6、3-7 分别展示当成像质量阈值𝜀 = 25时不同模型的评价指标情况；图 3-3、3-4 分别展示了 IoU 为

0.5 的条件下𝜀值从 20 到 35 的不同模型 ASR、AMR 和 CCB 的曲线变化图。

表 3-6 𝜀 = 25时，6 类目标识别模型（含白盒 SSD）的攻击指标评估结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **AMR** | **CCB** |
| SSD  (白盒) | 0.50:0.95 | 0.2196 | 0.0033 | 98.48% | 100% | 75.37% |
| 0.50 | 0.3135 | 0.0079 | 97.47% | 100% |
| 0.75 | 0.2494 | 0.0024 | 99.02% | 100% |
| YOLOv5s | 0.50:0.95 | 0.2889 | 0.1037 | 64.13% | 65.12% | 56.43% |
| 0.50 | 0.3878 | 0.1437 | 62.98% | 64.61% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **AMR** | **CCB** |
| (黑盒) | 0.75 | 0.3266 | 0.1165 | 64.33% | 64.97% |  |
| YOLOv5m  (黑盒) | 0.50:0.95 | 0.3618 | 0.1703 | 52.93% | 53.75% | 41.83% |
| 0.50 | 0.4705 | 0.2245 | 52.28% | 53.64% |
| 0.75 | 0.3934 | 0.1913 | 51.37% | 51.88% |
| FCOS  (黑盒) | 0.50:0.95 | 0.3772 | 0.1662 | 55.94% | 56.80% | 42.07% |
| 0.50 | 0.5190 | 0.2462 | 52.55% | 53.91% |
| 0.75 | 0.4077 | 0.1736 | 57.41% | 57.98% |
| RetinaNet (黑盒) | 0.50:0.95 | 0.3262 | 0.1353 | 58.53% | 59.43% | 48.58% |
| 0.50 | 0.4452 | 0.1975 | 55.62% | 57.06% |
| 0.75 | 0.3513 | 0.1470 | 58.14% | 58.71% |
| MaskRCNN (黑盒) | 0.50:0.95 | 0.3682 | 0.1676 | 54.48% | 55.32% | 43.69% |
| 0.50 | 0.5360 | 0.2665 | 50.28% | 51.58% |
| 0.75 | 0.4102 | 0.1812 | 55.83% | 56.38% |



图 3-3 IoU 为 0.5，不同𝜀值下的 5 类目标检测模型 ASR、AMR、CCB 的曲线变化图

从以上结果可以发现，本章提出的 DOG 算法所生成的对抗样本,对于其他 5 类具有代表性的目标检测模型有较好的攻击迁移性，当𝜀 = 25dB时，在基于 MS CO CO 数据集的攻击对抗验证中，攻击迁移率 AMR 均大于 51.58%，AMR@0.50:0.95 不低于 53.75%(YOLOv5m)、最高 65.12%(YOLOv5s)，[AMR@0.50](mailto:AMR@0.50) 不低于 51.58%(MaskR CNN)、最高 64.61%(YOLOv5s), [AMR@0.75](mailto:AMR@0.75) 不低于 51.88%(YOLOv5m)、最高 64.9

7%(YOLOv5s)；随着𝜀的提升 ASR、CCB 表现出了一定的下降趋势,特别当𝜀 = 30*、*35时，值几乎不变，说明在较高的𝜀退出条件下，需要使用更低的学习率和更高的迭代退出次数进行拟合，同时不难发现，转移率受到𝜀变化的影响较小，说明该方法在转移性上有较好的稳定性。

表 3-7 𝜀 = 25时，3 类图像分类模型的攻击指标评估结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ASR** | **CCB** | **Model** | **ASR** | **CCB** | **Model** | **ASR** | **CCB** |
| **VGG19** | 67.17% | 72.05% | **ResNet50** | 47.17% | 49.43% | **Inception V3** | 49.50% | 53.23% |

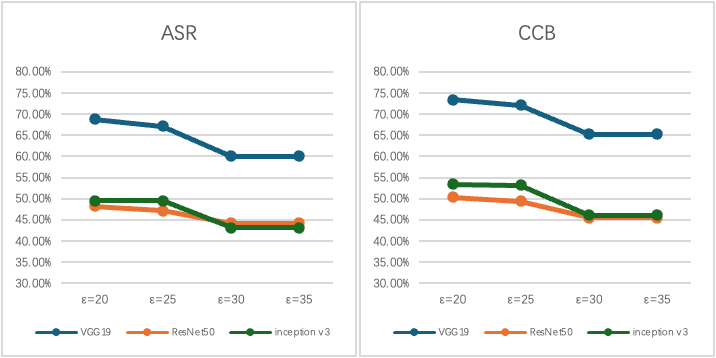


图 3-4 IoU 为 0.5，不同 值下的 3 类图像分类模型 ASR、CCB 的曲线变化图

从以上结果可以发现，本章提出的 DOG 算法所生成的对抗样本,对于 3 类图像分类模型（基于 ImageNet 的预训练模型）也有较好的迁移性，针对 VGG19 模型的攻击效果最佳，ASR、CBB 分别为 67.50%和 72.71%，高攻击成功率的原因应该与生成对抗样本的白盒模型选取有关，由于本章对抗样本生成模型为 SSD300，其基础网络为 VGG16，与 VGG19 模型具有同源性；同目标检测模型转移性评估结果类似，随着𝜀的提升 ASR、CCB 表现出了一定的下降趋势，但在𝜀 = 30*、*35时，趋于平稳，主要原因还是取决于学习率和最大迭代退出次数有关。

总的来说，DOG 算法，在对抗样本𝑃𝑆𝑁𝑅 ≥ 25dB的条件下，能够对绝大部分目标识别模型和图像分类模型造成有效攻击，攻击成功率分别不低于 51%和 47%，具有较高的攻击转移。

### 和其他攻击算法的对比实验

为进一步验证本章提出 DOG 方法的有效性，本章选取目标检测领域较为经典的 DAG、R-AP、EA 三种对抗攻击方法（由于 EA 生成单张对抗样本至少需要查询 5千次，约 500s,这里仅进行了针对 SSD 和 YOYOv5s 的对比），在 MS COCO 数据集上对 6 类目标检测模型进行对比实验，评价指标包括攻击成功率 [ASR@0.5](mailto:ASR@0.5)、攻击转移率 [AMR@0.5](mailto:AMR@0.5) 和平均生成照片时间 TC@avg。

表 3-8 𝜀 = 25时，6 类目标识别模型（含白盒 SSD）在 3 种攻击下的对比结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Method** | **Clean mAP** | **Adv mAP** | **ASR** | **AMR** | **TC** |
| SSD  (白盒，EA  为黑盒) | DAG | 0.3135 | 0.0249 | 92.05% | 100% | 9.324s |
| R-AP | 0.0225 | 92.83% | 3.772s |
| EA | 0.1029 | 67.16% | 546.175s |
| Ours(DOG) | 0.0079 | **97.47%** | **2.716s** |
| YOLOv5s  (黑盒) | DAG | 0.3878 | 0.3677 | 5.18% | 5.63% |  |
| R-AP | 0.3734 | 3.72% | 4.01% |
| EA | 0.1777 | 54.16% | - |
| Ours(DOG) | 0.1435 | **62.98%** | **64.61%** |
| YOLOv5m  (黑盒) | DAG | 0.4705 | 0.4594 | 2.35% | 2.55% |
| R-AP | 0.4508 | 4.18% | 4.5% |
| EA | - | - | - |
| Ours(DOG) | 0.2245 | **52.28%** | **53.64%** |
| FCOS  (黑盒) | DAG | 0.5190 | 0.5080 | 2.11% | 2.29% |
| R-AP | 0.4997 | 3.71% | 3.99% |
| EA | - | - | - |
| Ours(DOG) | 0.2463 | **52.55%** | **53.91%** |
| RetinaNet (黑盒) | DAG | 0.4452 | 0.4284 | 3.78% | 4.11% |
| R-AP | 0.4248 | 4.58% | 4.93% |
| EA | - | - | - |
| Ours(DOG) | 0.1976 | **55.62%** | **57.06%** |
| MaskRCNN (黑盒) | DAG | 0.5360 | 0.5098 | 4.88% | 5.30% |
| R-AP | 0.2823 | 47.32% | 50.97% |
| EA | - | - | - |
| Ours(DOG) | 0.2665 | **50.28%** | **51.58%** |

从以上结果可知，本章提出的 DOG 攻击方法，对比 DAG、R-AP、EA 三种攻击方法，在攻击成功率、攻击转移率、生成耗时三类指标上均有较好的表现。

## 3.4 本章小结

本章针对现有基于目标检测任务的对抗样本生成方法存在白盒攻击转移能力差、黑盒样本生成周期长的壁垒问题，通过综合借鉴 C&W 对抗样本生成架构、DAG模型激活值选取、R-AP 物体标记重置，DOG 方法除保留传统的置信度对比损失外，通过引入了噪声越界偏差损失、交叉熵损失、噪声强度损失，增加 FOCG 最优诱偏类别动态选择模块，提高了算法的攻击效果和效率。通过进行算法有效性、功能必要性、攻击迁移性验证实验，表明 DOG 方法与其他传统方法相比，能够以约 2s一张图的生成效率生成对抗样本，且该样本不仅能在白盒条件下达到 98%以上攻击成功率，黑盒条件下成功率也可达到 50%以上，且能对图像分类模型进行攻击成功率不低于 47%。在白盒攻击能力上，性能优于 DAG 稠密对抗生成算法，在黑盒迁移能力上，远超 R-AP 鲁棒对抗扰动算法和 EA 蒸发攻击算法。

# 融合图像信息熵和对抗网络特性的目标检测对抗攻击方法

## 引言

第三章介绍了一种基于多尺度特征网格群动态优化的目标检测对抗攻击方法，无论是白盒攻击还是黑盒攻击，无论是针对图像分类模型攻击还是目标检测模型攻击都取得了不错的成绩，做到了攻击效果、对抗范围和生成效率三者的最佳均衡。但 DOG 攻击方法存在生成效率高度依赖白盒基准模型的问题，为解决这一问题，进一步提升生成速度和图像生成质量，本章将在第三章 DOG 攻击算法的基础上，提出一种融合图像信息熵和对抗网络特性的目标检测对抗攻击方法 DOGs

（Dynamic Optimization of Multi-Scale Feature Grid Cluster-Speedup version），该方法旨在通过搭建并训练一个能够快速生成效果与 DOG 攻击算法对抗噪声基本一致的特殊 GAN 网络，通过将前端资源消耗压缩到后段模型训练中去，以提高、固化对抗噪声生产效率，通过采取滑窗采样的方式生成尺寸维度与原图像一致且不依赖于目标识别模型的图像信息熵，并按照特殊的标准化处理方式，获得全图关键区域权重矩阵，从而进一步实施重点攻击，提升对抗样本生成质量。该方法能够保证在 ASR 和 AMR 指标衰减不大于 10%的前提下，将噪声 PSNR 值平均提升 2dB，将生成效率提升 5-10 倍（约 0.8 秒/张）。

本章中，我们将详细介绍 DOGs 攻击方法的设计原理、关键组件的作用，信息熵图像及其关键区域权重的生成方法，以及在 9 类模型（3 类图像分类模型、6 类目标识别模型）中的实际对抗效果。

## DOGs 攻击方法

### 图像关键区域识别问题

本小节将针对图像关键区域识别技术和 DOG 攻击算法的短板缺陷入手，提出合适的问题解决思路。

* + - 1. 常见图像关键区域识别技术的共性问题。当前，图像关键区域的识别技术主要依赖深度学习模型，例如 CNN、FCN 等，通过端到端学习来提取目标类别的关键区域；利用 CAM、Grad-CAM 等方法生成类激活热图以确定重要区域；使用感

知敏感度分析方法逐步遮挡图像部分区域，以评估每个区域对目标类别识别的重要性。尽管这些方法在获取目标区域方面表现出色，但端到端推理模型需要耗费大量时间和资源进行重新训练，并且最终效果容易受模型结构、训练方法、训练集和超参数选择的影响。基于现有模型的推理需要提前获得模型结构访问权限或采用遮挡方式进行反复试验才能获取。

图像作为一个独立个体应该有其固有的特征属性，其关键区域识别矩阵或权 重矩阵原则上应该受识别模型的影响极小或不受影响，但由于现有识别方法都是 采取第三方判断方式实施，其识别“精度”和测量“准度”受第三方工具影响极 大，误差损失也主要来源于此。为进一步提升图像关键区域的识别提取能力，解 决第三方工具的依赖问题，本章提出了一种基于图像信息熵的关键区域识别技术。该技术以信息熵理论为基础，融合了卷积的思路，构造图像信息熵图，采取特殊 的标准化方式转换为值大于等于零的关键区域权重矩阵。

* + - 1. DOG 攻击方法的短板缺陷。第三章提到，DOG 攻击的对抗样本生成速率受所选白盒攻击模型影响极大，分别采取 SSD、FOCS、RentiaNet 等不同的白盒攻击模型，平均耗时分别为 2.5s、14.4s、23.1s，只能对静态图像造成有效攻击，且生成效率与白盒基准模型选择重度关联，难以应对复杂的、高动态的对抗场景，需要将平均耗时压缩到秒级以下并固化生成时间。

为把 DOG 攻击方法的高转移性、高对抗性有效应用到高动态移动目标场景中去，本章提出了 DOGs 攻击方法，基于 GAN 网络的生成能力，构造与 DOG 攻击算法类似的对抗噪声。

### 基于图像信息熵的关键区域识别方法

信息熵是由克劳德·香农（Claude Shannon）在信息论中引入的概念，用来量化信息的不确定性。信息熵本质上是对可能性的一种度量：一个事件的不确定性越大，我们从该事件的发生中获得的信息就越多。信息熵的一维数学定义为：

𝑯(𝑿) = −∑𝒏 𝒑(𝒙𝒊)log𝒑(𝒙𝒊) (4.1)

𝒊=𝟏

其中，𝑯(𝑿)表示随机变量𝑿的熵，𝒙𝒊表示随机变量的一个可能的取值, 𝒑(𝒙𝒊)

表示该取值的概率，而 n 是随机变量所有可能取值的数目。对数的底通常取 2，

这时熵的单位是比特（bit）。如果对数的底取自然对数 e，则熵的单位是纳特（nat）。在图像领域，信息熵用于描述图像的复杂度和信息含量，是图像中像素值分

布的一种统计度量，信息熵越高说明图像的复杂程度、信息含量越高。按照传统的信息熵求解方法，首先需要将 RGB 图像转换为只含有一层的灰度图像，而后再将该灰度图像延展为一维数据，经直方图数值统计、概率化转换，最后能够获得针对全图的单一信息熵值。



7.25 6.89 7.56 7.76

图 4-1 不同图像的信息熵示意图

采取传统方法获得的图像信息熵，只能反应图像的整体情况，很难据此获取图像中目标与背景、目标与目标、局部与局部之间的信息含量差异,如图 4-1 所示，无法从单一信息熵值推断出图片中推断出细节信息。特别在目标识别任务中，仅通过图像信息熵值来判断目标关键区域是没有意义的，需要获得图像局部区域或像素的熵值分布，这里称为信息熵图（EG，Entropy Graph），进而再与全图信息熵值进行比较以获取基于信息熵的目标关键区域范围。

为了解决这一问题，提出了借鉴卷积滑动窗口（Sliding Windows）和边缘填充（padding）思路的局部信息熵求解方法，该方法可以同时对图像 3 个通道进行并行求解，最终获得尺寸与原图像一致的信息熵图，再通过求解特定的权重转换阈值，将该信息熵图转换为图像信息熵权重矩阵以确定目标关键区域。具体执行步骤如下：

首先，为避免边缘像素信息的丢失，对原图像进行边缘填充，常见的填充方式包括固定值填充、镜像填充和周期性填，由于信息熵的计算元是某一元素值在一定范围内的出现频率，为充分保留原图的信息细节，避免出现边缘效应，本方法选用镜像填充方式补全边缘缺失（由于滑块尺寸为 3×3，镜像尺度设为 1）。

其次，利用尺寸为 3×3 的矩形框按照原图像位置自左至右、自上而下按步长

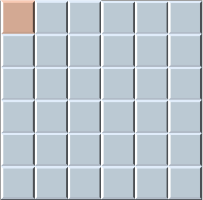
1 顺序滑动，每移动一次求取矩形框内的一维信息熵值，遍历全图后，获得和原图尺寸一致的图像信息熵矩阵。接着，计算全图信息熵均值entropyf，并以原图与滑动窗口面积比的二次方根作为标准化转换系数，获得权重转换阈值thread𝑣（用于区分图像像素值对图像目标识别是削弱还是加强，小于该值则为削弱、大于该值则为加强），设输入图像尺寸为(n, 3, h, w)，滑动窗口尺寸为(s𝑥, s𝑦)计算公式为：

thread𝑣 = entropyf × 2√ℎ × 𝑤 × 1/(s𝑥 × s𝑦) (4.2)

最后，将信息熵图像按照权重转换阈值进行标准化处理（信息熵值除以权重转换阈值），获得图像信息熵权重矩阵，当图像信息熵值小于thread𝑣说明该像素对图像目标识别影响偏弱，可削弱对抗噪声中相同位置的值的大小，反之则应当增强相应位置上的值。

这里使用GEM(∙)表示图像信息熵图的生产函数，𝑌𝑒𝑤为图像信息熵权重矩阵生成函数，其生成算法表示为：

𝑌𝑒𝑤 = GEM(X, s, p) × thread𝑣 (4.3)

其中，X 为输入图像样本，滑窗大小 s，滑窗步长 p，图 4-2 展示了滑窗大小为 3、滑窗步长为 1 的图像信息熵权重矩阵计算原理。

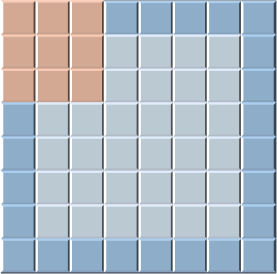


图 4-2 图像信息熵布尔图计算示意图

如若将𝑌𝑒𝑤按照阈值 1 进行二分转换，则可获得基于图像自身信息的关键区域识别矩阵。

如若将𝑌𝑒𝑤值进行 0 到 255 区间的归一化转换，则可获得基于图像自身信息

的信息熵图，表 4-1 展示了该技术构造图像信息熵布尔图的输出效果，区分三通道、单通道进行展示。

表 4-1 图像信息熵布尔图构造效果图

|  |  |
| --- | --- |
| **原图** | 狗站在地上  描述已自动生成 |
| **信息熵图**  **(三通道)** | 电脑萤幕画面  中度可信度描述已自动生成 |
| **信息熵布尔图** | 黑白色的照片  低可信度描述已自动生成 |

### DOGs 攻击方法及系统架构

DOGs 攻击策略基于 DOG 攻击方法，引入了 advGAN 架构的核心思路，构建了一种基于生成式网络架构的高级对抗攻击机制。该方法将DOG 攻击方法融合到GAN网络中，通过对抗训练，使生成器能够迅速而高效地生成对抗噪声。两种算法在针对单张图像的迭代优化过程中，损失函数的构造及优化迭代方法大致相同，其根本区别在于：DOG 方法采用“推理式”策略，对于每一张输入图像进行独立的迭代优化以产生具有攻击性的对抗噪声；而 DOGs 则采用“生成式”方法，通过大规模的图像生成训练，使得生成过程具有概括性，从而达到准确生成对抗噪声的最终目标。

DOGs 系统架构与 advGAN 架构相似，主要由图像信息熵特征提取模块、生成 器、判别器、受攻击白盒模型四部分组成（结构原理见图 4-3），由于两个系统的 攻击目的分别为目标识别和图像分类两种不同任务，故而在判别器选择上判别器 选择 SSD300 模型（未训练），受攻击白盒模型同样适用 SSD300 模型（已训练），生成器使用包含 6 个卷机层、2 个下采样层、2 个上采样层的 U-net 模型，这样可 使得该攻击架构能够适应不同尺寸的图像样式，进一步提高适用性。关键区域提 取模块按照 4.2.2 小节所述方法提取原始图像的关键区域和非关键区域权重矩阵，该矩阵仅在对抗训练期间使用，将生成的对抗噪声按照关键部分和非关键部分分

别进行最大化、最小化训练，使得生成器能够学习到关键区域划分的特点。DOGs攻击方法可以被细分为以下四个关键步骤。

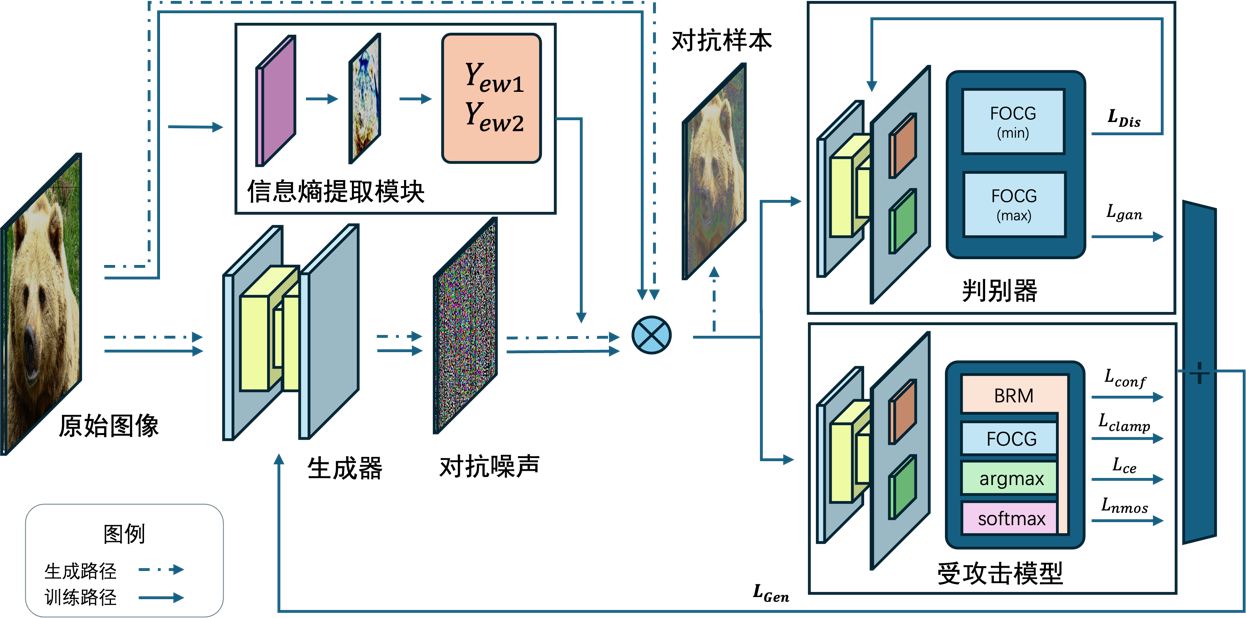


图 4-3 DOGs 攻击方法结构原理图

1. 模型初始化。本阶段主要涵盖输入参数的准备和生成网络的构建过程，包括四部分。**一是**数据集准备，收集并整理输入图像数据集及其对应标签，确保数据质量与多样性。**二是**构建生成器。生成器主要用于将白盒模型 FPN 输出的多尺度特征𝐺(𝐼0)转换为能够欺骗目标模型的对抗噪声𝑁𝑡，其输入尺寸同𝐺(𝐼0)、输出尺寸同输入原始图像𝐼0，生成器采用 U-Net 结构，通过编码器和解码器的组合进一步提升特征转移效率和精度。通过与判别器的对抗训练，生成器不断优化其输出，以提高生成对抗示例的质量和攻击效果。**三是**构建判别器。主要用于判别叠加对抗噪声后的对抗图像𝐼𝑡与原始输入图像𝐼0之间的差异。DOGs 判别器一般使用未训练的目标识别模型（本章使用 SSD300），与生成器同步优化，避免对抗训练模式崩溃，提高训练稳定性。**四是**构建受攻击模型。受攻击模型是基于 GAN 网络的对抗样本生成的核心部件，通过该模型实时获得对抗图像𝐼𝑡针对白盒模型的攻击效果，并通过特殊的损失函数反馈给生成器进行迭代优化。**五是**注入初始参数。核心参数共 10 项，分别是输入图像𝐼0、进行白盒攻击的模型𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(·)、与模型相对应的能够同步提取类 FPN 结构输出的𝐺(·)跟踪接口、待攻击目标类别集合

𝑠𝑒𝑡(𝑘)、以及生成器𝐺𝑒𝑛(·)和判别器𝐷𝑖𝑠(·)、关键区域权重矩阵𝑌𝑒𝑤1、非关键区域

权重矩阵𝑌𝑒𝑤2。

1. 训练生成器。由于判别器选用预训练模型，已经具备较高的判别能 力，故而在训练之初，首先对生成器进行训练，该步骤是生成对抗噪声的关键环节，主要完成对抗噪声的生成、对抗图像合成、前向传播及模型监控参数提取、构造 t 轮次诱骗类别集、损失计算与反向传播，主要执行环节与 DOG 攻击算法中的噪声迭代优化基本一致。**一是**对抗噪声的生成。执行𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(𝐼0),抓取𝐺(𝐼0)，向生成器𝐺𝑒𝑛(·)输入𝐺(𝐼0)获得当前输入图像第 t 次迭代噪声𝑁𝑡，既有

𝑁𝑡 = 𝐺𝑒𝑛(𝐺(𝐼0))，而后根据图像信息熵提取模块获得𝑌𝑒𝑤1和𝑌𝑒𝑤2，再分别与𝑁𝑡按位点乘获得带区域权重的对抗噪声。**二是**对抗图像的生成。将对抗噪声与原图像进行叠加并进行边界限制处理获得对抗图像，即𝐼𝑡 = 𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡)，t 为迭代轮次。三是前向传播及模型监控参数提取。执行𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(𝐼𝑡),抓取𝐺(𝐼𝑡)，计算获得 t 轮次预测类别集𝑇𝑝𝑟𝑒𝑡 = 𝑎𝑟𝑔𝑚𝑎𝑥{𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼𝑡)]}。**三是**构造诱偏类别集𝑇𝑖𝑛𝑑𝑢𝑐𝑒𝑡，定义 FOCG 方法能够根据当前对抗样本输出结果动态寻找最优的诱偏类别集。四是损失计算及反向传播。基于设定的损失函数计算五类损失（比较 DOG 攻击算法增加了基于判别器𝐷𝑖𝑠(·)的𝐿𝑔𝑎𝑛损失函数，并将损失反向传播，更新下一轮的对抗噪声；另外，为提升生成器对关键区域特征学习判断能力，使用𝑌𝑒𝑤1时，取𝑚𝑎𝑥(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠)，使用𝑌𝑒𝑤2时，则取最小 值。

1. 优化判别器。在这一阶段，重点放在提升判别器模型的能力。判别器对 生成器输出的对抗样本进行评估和反馈，通过优化判断结果，不断调整判别器的 参数，提高其对于对抗特征的识别能力，从而在训练过程中有效促进生成器的性 能进步。将一张图像进行最后一次迭代优化后的𝐿𝑔𝑎𝑛值记做该图的最终损失𝐿1𝑚𝑓，当𝐿1𝑚𝑓的衰减率小于 0.1dB 时，固定生成器，并使用当前输入图像进行判别器优 化。
2. 输出对抗噪声。经过多次对抗训练后，切换为生成模式，执行𝐼𝑡 − 𝐼0 =

𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝐺𝑒𝑛(𝐺(𝐼0))) − 𝐼0获得最终的对抗噪声𝑁𝑔𝑒𝑛。

### 损失函数

我们的目标是寻找添加到图像𝐼0上的引起最小偏差的最快迭代扰动，以使得目标类型判别出错。这里定义的损失函数包括置信度对比损失𝐿𝑐𝑜𝑛𝑓、噪声越界偏差损失𝐿𝑐𝑙𝑎𝑚𝑝、交叉熵损失𝐿𝑐𝑒、噪声强度损失𝐿𝑛𝑚𝑜𝑠和对抗损失𝐿𝑔𝑎𝑛。其公式如下：

𝐿 = 𝑚𝑎𝑥(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠) + 𝑀𝐼𝑁 𝑀𝐴𝑋 (𝐿𝑔𝑎𝑛) (4.4)

𝐺 𝐷

前四类损失定义方式同 DOG 攻击方法，详见 3.2.4，本小节重点就前四类损失的变化及对抗训练损失𝐿𝑔𝑎𝑛进行说明。

1. 置信度对比损失𝐿𝑐𝑜𝑛𝑓。表示了生成器的输出与目标输出之间的差异，通过缩小之间的差距，实现预测结果逐步向预定变化方向移动的目的，作为损失函数的正则项使用。其公式如下：

𝐿𝑐𝑜𝑛𝑓 = 𝑟𝑒𝑙𝑢(𝑍𝑖 − 𝑍𝑡) (4.5)

其中，𝑍𝑖 为基于判别器𝐷𝑖𝑠(·)生成对抗图像𝐼𝑡 的最大预估类别的概率张量

（softmax 输出中类的概率），𝑍𝑡为基于判别器𝐷𝑖𝑠(·)生成对抗图像𝐼𝑡的计划诱偏类别的概率张量（softmax 输出中类的概率），𝑟𝑒𝑙𝑢(·)为 relu 激活函数，能够把结果小于 0 的部分置 0，大于 0 的保持不变，只有当 𝑍𝑖大于𝑍𝑡 时，损失才会产生，使得噪声迭代方向始终朝着目标类别移动。

1. 噪声越界偏差损失𝐿𝑐𝑙𝑎𝑚𝑝。生成的对抗噪声在叠加到原图像上时会出现超出图像设定值域范围的情况（超出[0,1]区间），这里定义了𝐿𝑐𝑙𝑎𝑚𝑝使得生成的对抗噪声叠加到原图像上后，尽可能不超出值域范围。其公式如下：

𝐿𝑐𝑙𝑎𝑚𝑝 = ‖𝐼0 + 𝑁𝑡 − 𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡)‖2 (4.6)

其中，𝑐𝑙𝑎𝑚𝑝(·)函数能够把值调制到[0,1]区间。

1. 交叉熵损失𝐿𝑐𝑒。计算当前轮次模型输出的类别概率与诱偏类别之间的差异。 其公式如下：

𝐿𝑐𝑒 = − ∑𝐶

𝑚=1

y𝑚log (𝑙𝑚)

(4.7)

其中，y𝑚是𝐺(𝐼𝑡)的输出矩阵元素，𝑙𝑚是当前轮次设定的诱偏类别𝑇𝑖𝑛𝑑𝑢𝑐𝑒𝑡的元素，𝐶是模型类别总数。

1. 噪声强度损失𝐿𝑛𝑚𝑜𝑠。这一项计算了对抗噪声的均方值，表示对抗扰动

的强度。其公式如下：

𝐿𝑛𝑚𝑜𝑠

= − 1

𝑁𝑠

𝑁𝑠 2

𝑚=1 𝑚

∑

𝑛

(4.8)

其中，𝑁𝑠是对抗噪声𝑁的元素个数，𝑛𝑚是第 m 个元素的值。

1. 对抗损失𝐿𝑔𝑎𝑛。在生成对抗网络中，生成器和判别器通过对抗训练相互竞争，生成器希望最大化判别器对生成样本的预测概率，即希望𝑍𝑖尽可能接近 1。这意味着生成器希望生成的样本能够被判别器认为是真实的。判别器希望最大化对真实样本的预测概率，同时最小化对生成样本的预测概率，即希望 𝑍𝑖接近 0。故此，对抗损失𝐿𝑔𝑎𝑛可表述为：

𝐿𝑔𝑎𝑛 = 𝐸𝑥(log (𝑍𝑖)) + 𝐸𝑥(log (1 − 𝑍𝑖)) (4.9)

其中，𝑍𝑖 为基于判别器𝐷𝑖𝑠(·)生成对抗图像𝐼𝑡 的最大预估类别的概率张量

（softmax 输出中类的概率），𝐸𝑥(·)表示对真实样本的期望值。当进行生成器训练时𝐿𝑔𝑎𝑛取最小值，以提高其生成样本的质量，使得生成的样本更难以被判别器识别为假样本。当进行判别器训练时𝐿𝑔𝑎𝑛取最大值，以提高其区分真实样本和生成样本的能力。

### 算法实现

本小节，将对以上提及的 DOGs 算法、基于图像信息熵的关键区域识别算法实现进行补充。

（1）DOGs 攻击算法。下面将对 DOGs 攻击算法进行说明，文字描述详见 4.2.3 DOGs 攻击方法描述。

|  |
| --- |
| 算法 4.1: DOGs（Dynamic Optimization of Multi-Scale Feature Grid Cluster- Speedup version） |
| 输入：原图像𝐼0，白盒模型𝑎𝑡𝑡𝑐𝑚𝑜𝑑𝑒𝑙(·)，FPN 输出函数𝐺(·)，待攻击目标类别集合  𝑠𝑒𝑡(𝑘)，𝐴𝑑𝑚(·)优化器，𝐵𝑅𝑀(·)，𝐹𝐶𝑂𝐺(·)，𝑜𝑛𝑒\_ℎ𝑜𝑡(·)，成像质量阈值𝜀，最大迭代次数𝑡𝑠，生成器学习率𝑙𝑟𝑔, 判别器学习率𝑙𝑟𝑑，置信度正则参数𝑐，生成器𝐺𝑒𝑛(·)，判别器  𝐷𝑖𝑠(·)，关键区域权重矩阵𝑌𝑒𝑤1，非关键区域权重矩阵𝑌𝑒𝑤2  输出：无 |

初始化：𝑎𝑡𝑡𝑚𝑜𝑑𝑒𝑙(𝐼0)，𝑇𝑜𝑟𝑎𝑙 = 𝑎𝑟𝑔𝑚𝑎𝑥{𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐺(𝐼0)]}， 𝑜𝑝𝑡𝑔 = 𝐴𝑑𝑚(𝐺𝑒𝑛(·), 𝑙𝑟𝑔)，

𝑜𝑝𝑡𝑑 = 𝐴𝑑𝑚(𝐷𝑖𝑠(·), 𝑙𝑟𝑑)*，*𝑅𝐹𝑜𝑟𝑎𝑙 = 𝐵𝑅𝑀 (𝑇𝑜𝑟𝑎𝑙*，*𝑠𝑒𝑡(𝑘))，model = Gen

16. 𝐺𝑒𝑛. 𝑡𝑟𝑎𝑖𝑛, 𝐷𝑖𝑠. 𝑣𝑎𝑙

17. 𝒊𝒇 𝑚𝑜𝑑𝑒𝑙 𝑖𝑠 𝐷𝑖𝑠 𝒕𝒉𝒆𝒏 𝐺𝑒𝑛. 𝑣𝑎𝑙, 𝐷𝑖𝑠. 𝑡𝑟𝑎𝑖𝑛

18. 𝑤ℎ𝑖𝑙𝑒(1):

19. 𝑁𝑡 = 𝐺𝑒𝑛(𝐺(𝐼0)) × 𝑌𝑒𝑤1,𝑁′ = 𝐺𝑒𝑛(𝐺(𝐼0)) × 𝑌𝑒𝑤2

𝑡

20. 𝐼𝑡 = 𝑐𝑙𝑎𝑚𝑝(𝐼0 + 𝑁𝑡)

21. 𝑁𝑜𝑢𝑡 = 𝐼𝑡 − 𝐼0

22. 𝑇𝑝𝑟𝑒𝑡 = 𝑎𝑟𝑔𝑚𝑎𝑥[𝐷𝑖𝑠(𝐼𝑡)]

23. 𝒊𝒇 𝑠𝑒𝑡(𝑇𝑝𝑟𝑒𝑡 ) 𝑛𝑜𝑡 𝑖𝑛 𝑠𝑒𝑡(𝑘) 𝑜𝑟 𝑡𝑖𝑚𝑒𝑠 > 𝑡𝑠 𝒕𝒉𝒆𝒏 𝑏𝑟𝑒𝑎𝑘 //退出条件

24. 𝒊𝒇 𝑃𝑆𝑁𝑅 ≤ 𝜀 𝒕𝒉𝒆𝒏 𝑟𝑜𝑙𝑙 𝑏𝑎𝑐𝑘 //回滚条件

25. 𝑇𝑖𝑑𝑐𝑡 = 𝐹𝐶𝑂𝐺 (𝐷𝑖𝑠(𝐼𝑡)*，*𝑠𝑒𝑡(𝑘)) //选择本轮次诱偏类别𝑇𝑖𝑑𝑐𝑡

26. 𝑅𝐹𝑟𝑒𝑠𝑒𝑡 = 𝐵𝑅𝑀 (𝑇𝑝𝑟𝑒𝑡 *，*𝑠𝑒𝑡(𝑘)) ∗ 𝑅𝐹𝑜𝑟𝑎𝑙

27. 𝑍𝑡 = 𝑠𝑢𝑚(𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐷𝑖𝑠(𝐼𝑡)] ∗ 𝑜𝑛𝑒\_ℎ𝑜𝑡(𝑇𝑖𝑑𝑐𝑡 ), 𝑑𝑖𝑚 = 1) ∗ 𝑅𝐹𝑟𝑒𝑠𝑒𝑡//诱偏类别概率

28. 𝑍𝑖 = 𝑠𝑢𝑚(𝑠𝑜𝑓𝑡𝑚𝑎𝑥[𝐷𝑖𝑠(𝐼𝑡)] ∗ 𝑜𝑛𝑒\_ℎ𝑜𝑡(𝑇𝑝𝑟𝑒𝑡 ), 𝑑𝑖𝑚 = 1) ∗ 𝑅𝐹𝑟𝑒𝑠𝑒𝑡//预测类别概率

29. 𝐿𝑐𝑜𝑛𝑓 = 𝑟𝑒𝑙𝑢(𝑍𝑖 − 𝑍𝑡)

30. 𝐿𝑐𝑙𝑎𝑚𝑝 = ‖𝑁𝑡 − 𝑁𝑜𝑢𝑡‖2

31. 𝐿𝑐𝑒 = 𝐶𝑟𝑜𝑠𝑠EntropyLoss(Dis(𝐼𝑡) ∗ 𝑅𝐹𝑟𝑒𝑠𝑡 . 𝑣𝑖𝑒𝑤(−1,1), 𝑇𝑖𝑑𝑐𝑡𝑖 ∗ 𝑅𝐹𝑟𝑒𝑠𝑡)

32. 𝐿𝑛𝑚𝑜𝑠 = 𝑚𝑒𝑎𝑛(𝑁𝑡 ∗∗ 2)

33. 𝐿𝑔𝑎𝑛 = 𝐸𝑥(log (𝑍𝑖)) + 𝐸𝑥(log (1 − 𝑍𝑖))

34. 𝐿𝑔 = 𝑚𝑎𝑥(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠) + 𝑚𝑖𝑛(𝐿𝑔𝑎𝑛)

35. 𝐿𝑑 = 𝑚𝑎𝑥(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠) + 𝑚𝑎𝑥(𝐿𝑔𝑎𝑛)

36. 𝒊𝒇 𝑢𝑠𝑒 𝑌𝑒𝑤2 𝒕𝒉𝒆𝒏 𝐿𝑔 = 𝑚𝑖𝑛(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠) + 𝑚𝑖𝑛(𝐿𝑔𝑎𝑛)

37. 𝒊𝒇 𝑢𝑠𝑒 𝑌𝑒𝑤2 𝒕𝒉𝒆𝒏 𝐿𝑑 = 𝑚𝑖𝑛(𝑐 × 𝐿𝑐𝑜𝑛𝑓 + 𝐿𝑐𝑙𝑎𝑚𝑝 + 𝐿𝑐𝑒 + 𝐿𝑛𝑚𝑜𝑠) + max(𝐿𝑔𝑎𝑛)

38. 𝒊𝒇 𝑚𝑜𝑑𝑒𝑙 𝑖𝑠 𝐺𝑒𝑛 𝒕𝒉𝒆𝒏 𝐿𝑔. 𝑏𝑎𝑐𝑘𝑤𝑎𝑟𝑑, 𝑜𝑝𝑡𝑔. 𝑠𝑡𝑒𝑝, 𝑜𝑝𝑡𝑔. 𝑧𝑒𝑟𝑜\_𝑔𝑟𝑎𝑑

39. 𝒊𝒇 𝑚𝑜𝑑𝑒𝑙 𝑖𝑠 𝐷𝑖𝑠 𝒕𝒉𝒆𝒏 𝐿𝑑. 𝑏𝑎𝑐𝑘𝑤𝑎𝑟𝑑, 𝑜𝑝𝑡𝑑. 𝑠𝑡𝑒𝑝, 𝑜𝑝𝑡𝑑. 𝑧𝑒𝑟𝑜\_𝑔𝑟𝑎𝑑

40. 𝒊𝒇 𝑚𝑜𝑑𝑒𝑙 𝑖𝑠 𝐷𝑖𝑠 𝑎𝑛𝑑 𝑢𝑝𝑑𝑎𝑡𝑒 𝐷𝑖𝑠 𝑖𝑠 𝑜𝑣𝑒𝑟 𝒕𝒉𝒆𝒏 𝑚𝑜𝑑𝑒𝑙 = 𝐺𝑒𝑛

41. 𝒊𝒇 10𝑙𝑜𝑔10(𝐿𝑔−1 ⁄𝐿𝑔) × 𝑙𝑟𝑔 < 0.1 𝒕𝒉𝒆𝒏 𝑚𝑜𝑑𝑒𝑙 = 𝐷𝑖𝑠, 𝑟𝑜𝑙𝑙 𝑏𝑎𝑐𝑘 𝑡ℎ𝑒 𝑖𝑛𝑝𝑢𝑡 𝑖𝑚𝑎𝑔𝑒𝑠

在实际训练中还会增加回滚微调判别器的次数（设置为 5 次），5 次后将跳转到下一组图片执行生成器的训练。

## 实验与结果分析

为证明本章所提 DOGs 算法的有效性、通用性、转移性，本节设置相应实验进行验证。首先就实验平台、实验数据、模型选择、评估指标等事项进行说明。其次，为验证 DOG 算法的有效性，这里对基于 MS COCO 数据集的预训练模型 SSD300做生成对抗样本实验，所提方法在白盒对抗中的各类别的 AP 值变化和总体 mAP 值变化。然后，为验证算法的转移性，将基于 SSD 模型训练的生成器在其他 3 类图像分类网络和 5 类目标检测网络中进行黑盒攻击验证。以上所涉及模型、数据集、对抗算法、评价指标在第二章、第三章有详细介绍，本节中将不再进行说明。

### 实验基础搭建

1. 操作平台。

Windows 10 操作系统。

1. 硬件设备。

CPU：Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz GPU：NVIDIA GeForce RTX 2080 Ti

RAM：64.0 GB

1. 软件工具。

IDE：PyCharm 2024.1.2 (Professional Edition)

Python 版本：3.10

深度学习框架：PyTorch 2.3.1+cu121

1. 数据集选择。均使用 MS COCO 数据集的验证集进行实验。由于本章节涉及模型训练和评估，为与 DOG 攻击算法评价数据保持一致，在验证集使用方面与第三章相同，从 MS COCO 数据集的验证集前中后各取 200 张图像共计 600 张图像作为验证集，训练集则错位 100 张选取 1200 张作为训练集。
2. 模型选择。图像分类：基于 ImageNet 数据集的 VGG19、ResNet50、 Inception V3 预训练模型。

目标检测：基于 MS COCO 数据集的 YOLOv5、SSD300、RetinaNet、FCOS、MaskRCNN

预训练模型。

1. 评估指标。见 2.6 评价指标。

### DOGs 攻击有效性验证

参照 4.2.3 所示方法，使用 MS COCO 验证集中使用 1200 张训练集图像进行训练，而后用训练好的生成器，对 600 张验证集图像进行对抗攻击，并对其攻击性能进行分析评估。其中，成像质量阈值𝜀 = 25，最大迭代次数𝑡𝑠 = 5，学习率

𝑙𝑟𝑔 = 0.05、𝑙𝑟𝑑 = 0.01，置信度正则参数𝑐 = 1，FOCG 策略为最小诱偏类别动态选择方案。进行攻击前后效果对比见表 4-2，实验结果如图 4-4、表 4-2。

|  |  |
| --- | --- |
|  | 表 4-2 基于 SSD 模型白盒攻击前后目标检测效果对比 |
| **原图** | 棕色的熊  描述已自动生成 小孩在草地上放风筝  中度可信度描述已自动生成 |
|  | 1 bear 6 person、1 kite 1 cake 1 cup、1 dining table |
| **对抗噪声** | 猫的照片  中度可信度描述已自动生成   桌子上有杯子和刀  低可信度描述已自动生成 |
| **对抗样本** | 棕色的熊  描述已自动生成 |

从上表可知，把样例 1 中的熊、样例 3 中的蛋糕攻击为了背景，把样例 2 中的 6 个人、1 个风筝攻击为 3 个人、1 个风筝，且风筝的边界框也出现了较大偏差，把样例 4 中的杯子和餐桌攻击为了杯子，由于采用 GAN 网络作为系统基本架构，生成的对抗噪声较 DOG 攻击方法具有更强的连续性。

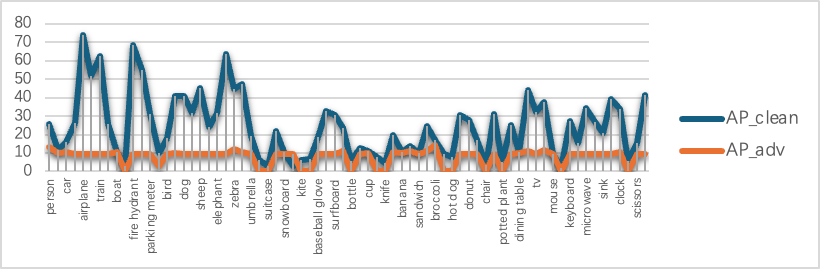


图 4-4 基于 SSD 模型白盒攻击前后类别 AP 值对比

如图 4-2 所示，纵坐标为 AP 标尺，横坐标为类别标签，攻击前后各类别受到了极大的抑制，如 airplane 类别 AP 值由 73.9%攻击到了 9.5%，所有类别攻击后的 AP 值最大不超过 13.2%(person)，绝大部分小于 9%。

表 4-3 基于 SSD 模型白盒攻击指标评估结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **TC(s)** | **CCB** |
| SSD  (白盒) | 0.50:0.95 | 0.2196 | 0.0266 | 87.90% | AVG:0.854  MIN:0.745 MAX:0.983 | 65.52% |
| 0.50 | 0.3135 | 0.0421 | 86.56% |
| 0.75 | 0.2494 | 0.0282 | 88.70% |

如表4-3 所示，[采取该方法生成的对抗样本成功率ASR@0.50](mailto:閲囧彇璇ユ柟娉曠敓鎴愮殑瀵规姉鏍锋湰鎴愬姛鐜嘇SR@0.50)：0.95 约87.90%、 [ASR@0.50](mailto:ASR@0.50) 约 86.56%、[ASR@0.50](mailto:ASR@0.50) 约 88.70%，TC 均值约为 0.854s,CCB 值约 65.52%。

### DOGs 攻击迁移性验证

利用 4.3.2 训练好的生成器，对其他 3 类图像分类预训练网络和 5 类目标检测网络进行参数评估。

表 4-4 5 类目标识别模型的攻击指标评估结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **IoU** | **Clean mAP** | **Adv mAP** | **ASR** | **AMR** | **CCB** |
| YOLOv5s  (黑盒) | 0.50:0.95 | 0.2889 | 0.1297 | 55.10% | 62.68% | 46.51% |
| 0.50 | 0.3878 | 0.1786 | 53.95% | 62.33% |
| 0.75 | 0.3266 | 0.1460 | 55.30% | 62.34% |
| YOLOv5m  (黑盒) | 0.50:0.95 | 0.3618 | 0.2031 | 43.85% | 49.89% | 31.85% |
| 0.50 | 0.4705 | 0.2668 | 43.30% | 50.02% |
| 0.75 | 0.3934 | 0.2270 | 42.30% | 47.69% |
| FCOS  (黑盒) | 0.50:0.95 | 0.3772 | 0.2004 | 46.86% | 53.31% | 32.22% |
| 0.50 | 0.5190 | 0.2929 | 43.56% | 50.32% |
| 0.75 | 0.4077 | 0.2103 | 48.40% | 54.57% |
| RetinaNet (黑盒) | 0.50:0.95 | 0.3262 | 0.1647 | 49.50% | 56.31% | 38.63% |
| 0.50 | 0.4452 | 0.2377 | 46.61% | 53.85% |
| 0.75 | 0.3513 | 0.1787 | 49.12% | 55.38% |
| MaskRCNN (黑盒) | 0.50:0.95 | 0.3682 | 0.2009 | 45.42% | 51.67% | 33.73% |
| 0.50 | 0.5360 | 0.3149 | 41.25% | 47.65% |
| 0.75 | 0.4102 | 0.2184 | 46.75% | 52.71% |

从以上结果可以发现，本章提出的 DOGs 算法所生成的对抗样本,对于其他 5类具有代表性的目标检测模型有较好的攻击迁移性，在基于 MS COCO 数据集的攻击对抗验证中，攻击迁移率 AMR 均大于 47.65%，AMR@0.50:0.95 不低于 49.89%(Y OLOv5m)、最高 62.68%(YOLOv5s)，[AMR@0.50](mailto:AMR@0.50) 不低于 47.65%(MaskRCNN)、最高 62. 33%(YOLOv5s), [AMR@0.75](mailto:AMR@0.75) 不低于 47.69%(YOLOv5m)、最高 62.34%(YOLOv5s)。

表 4-5 3 类图像分类模型的攻击指标评估结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ASR** | **CCB** | **Model** | **ASR** | **CCB** | **Model** | **ASR** | **CCB** |
| **VGG19** | 58.24% | 61.25% | **ResNet50** | 42.13% | 43.43% | **Inception V3** | 43.51% | 47.36% |

从以上结果可以发现，本章提出的 DOGs 算法所生成的对抗样本,对于 3 类图像分类模型（基于 ImageNet 的预训练模型）也有较好的迁移性，针对 VGG19 模型的攻击效果最佳，ASR、CBB 分别为 58.24%和 61.25%。

总的来说，DOGs 算法，能够对绝大部分目标识别模型和图像分类模型造成有效攻击，攻击成功率分别不低于 41%和 42%，具有较高的攻击转移性。

## 本章小结

本章针对 DOG 攻击算法攻击效率高度依赖白盒模型选择类型的问题，通过借鉴 advGAN 的思路，融合图像信息熵提取技术，构建了以 GAN 网络为基础架构的 DOGs 攻击算法，并采取与DOG 攻击算法相类似的迭代原则，训练对抗样本生成器。通过进行 DOGs 攻击算法有效性、迁移性实验，表明 DOGs 方法能够将对抗样本生成时间压缩至 1 秒以下，且不受训练使用白盒模型类型限制，通过该方法的生成的对抗样本依然能够保证较好的攻击效果，在白盒条件下攻击成功率约 86%，在黑盒条件下攻击成功率可达 40%以上，虽然攻击成功率较 DOG 攻击算法有约 10%的下降，但生成效率得到大幅提升、且生成时间稳定。

# 5 总结与展望

## 工作总结

本文从目标检测模型特点及其对抗攻击研究出发，首先介绍了基于深度学习网络架构的目标检测技术和传统图像对抗攻击方法的发展现状，并针对目标检测攻击方法所针对的不同类别模型进行分类介绍。接着，介绍了典型的图像分类模型、目标检测模型、目标检测任务攻击方法，详细阐述了不同模型、方法的共性特点和不同之处。同时指出目标检测领域对抗攻击面临的挑战，**一是**白盒条件下的攻击虽然具有有极高的对抗效果和生成效率，但其转移能力很差，**二是**黑盒条件下的攻击虽然在转移能力上表现不错，但其样本的生成效率较低，**三是**很难做到对抗效果、转移能力、生成效率的最佳均衡。本文从这些问题出发，开展如下主要工作：

* + 1. 针对现有目标检测的对抗样本生成方法，虽然能够在指定环境条件下达到不错的攻击效果，但白盒攻击转移能力差、黑盒样本生成周期长的壁垒问题却很难突破的特点。在借鉴 C&W 对抗样本生成架构、DAG 白盒模型激活值选取、R- AP 物体标记重置以及 EA 粒子群优化的思路，提出了一种基于多尺度特征网格群动态优化的目标检测对抗方法 DOG。实验结果表明，该方法可使基于 MS COCO 2017数据集的 YOLOv5s、YOLOv5m、SSD300、RentinaNet、FCOS、MaskRCNN 等 6 类目标检测网络，在白盒条件下 [ASR@0.50](mailto:ASR@0.50)≥97%、CCB≥75%，在黑盒条件下 [ASR@0.50](mailto:ASR@0.50)≥ 51%、CCB≥48%、[AMR@0.50](mailto:AMR@0.50)≥51%，对抗样本生成平均耗时约 2.4s(最小 0.9s、最大 4.3s)；同时经过验证，生成的对抗样本能够对基于 ImageNet 数据集的 VGG19、 ResNet50、Inception V3 等 3 类图像分类网络，造成一定的攻击效果，ASR≥47%、 CCB≥49%。基本上在通用性、对抗性、转移性、时效性上做到了多方向均衡。
    2. 为进一步提升基于 DOG 方法的对抗样本生成效率，以适应如自动驾驶的高动态多目标识别场景。提出了一种融合图像信息熵和对抗网络特性的目标检测对抗攻击方法 DOGs（Dynamic Optimization of Multi-Scale Feature Grid Cluster-Speedup version），该方法旨在通过搭建并训练一个能够快速生成效果

与 DOG 攻击算法对抗噪声基本一致的特殊 GAN 网络，通过将前端资源消耗压缩到后端模型训练中去，以提高对抗噪声生产效率，通过采取滑窗采样的方式生成尺寸维度与原图像一致且不依赖于目标识别模型的图像信息熵，并按照特殊的标准化处理方式，获得全图关键区域权重矩阵，从而进一步提升对抗样本图像质量。该方法能够保证在 ASR 和 AMR 指标衰减不大于 10%的前提下，将生成效率提升 5- 10 倍。

## 展望

本文从当前目标检测对抗攻击研究存在的三个方面挑战和高动态多目标快速识别的现实要求出发，提出了两种新的攻击方法，实验结果表明所提方法能够达成设定预想。本小节将在本文所提出的两种方法基础上进行延伸，并对后续工作进行展望。

1. 本文提出的 DOG 攻击方法，一方面解决了目标检测对抗样本的低生成时间和极高白盒条件攻击成功率同时达成的问题，另一方面解决了保持极高白盒攻击条件下对多型图形分类、目标识别网络的较高攻击转移能力。下一步将在损失函数优化、模块构建上进行拓展研究，进一步提升迭代优化效率和攻击转移能力。
2. 本文提出的 DOGs 攻击方法，把对抗样本的生成时间进一步压缩，解决了 DOG 算法对抗样本生成时间因白盒攻击模型选择不同、图像尺寸不同产生而较大偏差的问题，使得该方法在应对高动态多目标低时延场景具有更强的适用性。下一步将在模型架构、训练方法上进行迭代优化，在保持较好生成效率的同时，提升 DOGs 攻击方法的攻击效果，逐步接近 DOG 攻击方法。
3. 随着人工智能的快速发展，目标识别网络作为智能体的感知终端变得越发重要，从当前发展趋势来看诸如无人机巡查、自动驾驶等高动态、多目标、低时延的人工智能应用场景将会越发普遍，对其识别精度、准度、速度也提出了更高的要求，将传统方法通过对抗生成网络来模拟还原将会有很大的研究空间。下一步，将在静态场景攻击方法向高动态场景转移方面加大研究力度。

### 参考文献

1. 张晓闯. 计算机图像处理技术的发展与未来创新 [J]. 家庭影院技术,2023,316:76-81.DOI:10.3969/j.issn.1008-0945.2023.14.026.
2. Redmon, Joseph, et al. "You Only Look Once: Unified Real-Time Object Detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.
3. He, Kaiming, et al. "Mask R-CNN." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969.
4. Szegedy, Christian, et al. "Intriguing Properties of Neural Networks." Proceedings of the International Conference on Learning Representations (ICLR), 2014.
5. Chen S T, Cornelius C, Martin J, et al. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018: 52-68.
6. Goodfellow, Ian, et al. "Generative Adversarial Nets." Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 27-34.
7. Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." European Conference on Computer Vision (ECCV), 2016, pp. 21-37.
8. Ren, Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 91-99.
9. Goodfellow, Ian J., et al. "Explaining and Harnessing Adversarial Examples." Proceedings of the International Conference on Learning Representations (ICLR), 2015.
10. Madry, Aleksander, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." International Conference on Learning Representations (ICLR), 2018.
11. Carlini, Nicholas, and David Wagner. "Towards Evaluating the Robustness of Neural Networks." 2017 IEEE European Symposium on Security and Privacy (EuroS&P), 2017, pp. 39-57.
12. 陈憶悯等. "基于深度学习的两阶段目标检测算法综述." 信息与电脑

35.14(2023):112-114.

1. 姜文泽." 浅述计算机视觉中的图像特征及提取方法." 电脑知识与技术 15.16(2019):185-186. doi:10.14004/j.cnki.ckt.2019.2131.
2. Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." European Conference on Computer Vision (ECCV), 2016, pp. 21-37. Springer, doi:10.1007/978-3-319- 46448-0\_2.
3. Cai, Zhaowei, and Nuno Vasconcelos. "FCOS: Fully Convolutional One-Stage Object Detection." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1-10. IEEE, doi:10.1109/CVPR.2018.00100.
4. Lin, Tsung-Yi, et al. "Focal Loss for Dense Object Detection." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, 2020, pp. 318-327. IEEE, doi:10.1109/TPAMI.2019.2562024.
5. Lin, Tsung-Yi, et al. "Feature Pyramid Networks for Object Detection." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936- 944. IEEE, doi:10.1109/CVPR.2017.106.
6. Zhang, Y., et al. "AdvGAN: Generative Adversarial Networks for Adversarial Examples." IEEE International Conference on Computer Vision (ICCV), 2019, pp. 123-132. IEEE, doi:10.1109/ICCV.2019.00123.
7. Arjovsky, Martin, et al. "Wasserstein GAN." 34th International Conference on Machine Learning (ICML), 2017, pp. 214-223. PMLR, doi:10.5555/3305890.3305891.
8. Zhang, Y., Liu, H., and Wang, J. "Dynamic Adversarial Patches for Object Detection." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1-10. IEEE, doi:10.1109/CVPR52688.2022.00012.
9. Chen, X., Li, Y., and Zhang, L. "Laser-Based Physical Attacks on Deep Learning Models." IEEE International Conference on Computer Vision (ICCV), 2023, pp. 1-10. IEEE, doi:10.1109/ICCV53876.2023.00045.
10. Wu, A., Zhao, Z., and Liu, Y. "AdvRD: Adversarial Raindrop Attack on Object Detection in Real-World Conditions." European Conference on Computer Vision (ECCV), 2022, pp. 1-10. Springer, doi:10.1007/978-3-031-20048-3\_12.
11. Zhang, Y., and Wang, T. "Style Transfer for Physical Adversarial Examples." IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, 2023, pp. 123-135. IEEE, doi:10.1109/TNNLS.2023.1234567.
12. Kurakin, Alexey, Vladimir Golkov, and Christoph H. Lampert. "Adversarial Examples in the Physical World." arXiv, 2016, arXiv:1607.02533.
13. Madry, Aleksander, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." International Conference on Learning Representations (ICLR), 2018.
14. Dong, Yinpeng, et al. "Boosting Adversarial Attacks with Momentum." arXiv, 2018, arXiv:1710.60.
15. Su, Jin, et al. "One Pixel Attack for Fooling Deep Neural Networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1-6. IEEE, doi:10.1109/CVPR.2017.00149.
16. Moosavi-Dezfooli, Seyed-Mohsen, et al. "DeepFool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574-2582. IEEE, doi:10.1109/CVPR.2016.282.
17. Liu, Y., et al. "Delving into Transferable Adversarial Examples and Black-box Attacks." arXiv preprint arXiv:1611.02770, 2016.
18. Tramèr, Florian, et al. "Ensemble Adversarial Training: Attacks and Defenses." arXiv preprint arXiv:1705.07204, 2017.
19. Liu, Y., et al. "Serial-Mini-Batch-Ensemble-Attack for Transferable Adversarial Examples." arXiv preprint arXiv:2103.05009, 2021.
20. Chen, Jianbo, et al. "ZOO: Zeroth Order Optimization Based Black-box Attacks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1-9. IEEE, doi:10.1109/CVPR.2017.00041.
21. Yang, Yinpeng, et al. "AutoZoom: Auto-zooming for Black-Box Attacks." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 6, 2020, pp. 2075-2089. IEEE, doi:10.1109/TPAMI.2020.2973582.
22. Hansen, Nikolaus, et al. "CMA-ES/POI: A Successful Evolution Strategy on the 50 DIMENSIONAL BBOB TESTBED." Comparing Evolution Strategies for BBOB, 2012.
23. Yuan, Zeyuan, et al. "Adversarial Bits: Conditional Adversarial Examples." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6899-6907. IEEE, doi:10.1109/CVPR.2019.00706.
24. Zhang, T., et al. "GeoDA: Geometry-inspired Distribution Alignment for Transferable Adversarial Examples." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1618-1627. IEEE, doi:10.1109/CVPR46437.2021.00170.
25. Song, L., et al. "PhyAttack: Physical-World Adversarial Attack on the Object Detection Systems." IEEE Transactions on Cybernetics, vol. 50, no. 10, 2020, pp. 4170-4181. IEEE, doi:10.1109/TCYB.2020.2960155.
26. Wang, H., et al. "Daedalus: A Physical-Object Adversarial Attack on One-stage Detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5827-5836. IEEE, doi:10.1109/CVPR42600.2020.00583.
27. Liao, R., et al. "Attacking Anchor-Free Object Detection via Semantic Important Pixels." IEEE Transactions on Image Processing, vol. 30, 2021, pp. 3451-3464. IEEE, doi:10.1109/TIP.2021.3064839.
28. Lu, Y., et al. "DFool: DeepFool for Object Detection." IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1-9. IEEE, doi:10.1109/ICCV.2017.00056.
29. Xie, L., et al. "DAG: Dense Adversarial Generation for Object Detection." Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 4, 2020, pp. 3905-3912. AAAI, doi:10.1609/aaai.v34i04.5900.
30. Chen, X., et al. "ShapeShifter: A Targeted Attack on Object Detectors." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2405-2414. IEEE, doi:10.1109/CVPR42600.2020.00245.
31. Li Y , Tian D ,Mingching-Chang,et al.Robust Adversarial Perturbation on Deep Proposal-based Models[J]. 2018.DOI:10.48550/arXiv.1809.05962.
32. Zhang, Z., et al. "CAP: Context-Aware Attack on Object Detection." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, pp. 10719-10729. IEEE, doi:10.1109/ICCV48922.2021.01036.
33. Wei, Y., et al. "Unified and Efficient Adversary: A GAN-based Approach for Adversarial Attack on Video." IEEE Transactions on Image Processing, vol. 29, 2020, pp. 7492-7505. IEEE, doi:10.1109/TIP.2020.2982630.
34. Wu, Y., et al. "G-UAP: Generalized Universal Adversarial Perturbations for Deep Neural Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13974-13982. IEEE, doi:10.1109/CVPR42600.2020.01397.
35. Chow, M., et al. "TOG: A Two-Stage Targeted Object Generation Framework for Adversarial Attacks." European Conference on Computer Vision (ECCV), 2020, pp. 1-16. Springer, doi:10.1007/978-3-030-58604-8\_65.
36. Wang Y , Tan Y A , Zhang W ,et al.An adversarial attack on DNN-based black- box object detectors[J].Journal of Network and Computer Applications, 2020, 161:102634.DOI:10.1016/j.jnca.2020.102634.
37. Russakovsky, O., Deng, J., Su, H., et al. "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision, vol. 115, no. 3, 2015, pp. 211-252.
38. Lin, T.-Y., Maire, M., Belongie, S., et al. "Microsoft COCO: Common Objects in Context." Proceedings of the European Conference on Computer Vision, Berlin, 2014, pp. 740-755. Springer