中国科学技术大学

硕士学位论文

基于特征点选取的鲁棒分布外检测算法研究

作者姓名： 王中月

学科专业： 控制科学与工程

导　　师： 赵云波 教授

完成时间： 二〇二五年十月二十七日

University of Science and Technology of China

A dissertation for master's degree

Research on Robust Out-of-Distribution Detection Based on Feature Point Selection

Author: Zhongyue Wang

Speciality: Control Science and Engineering

Supervisor: Prof. Yun-Bo Zhao

Completion date: October 27, 2025

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：王中月　　签字日期：2025年10月29日

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

控阅的学位论文在解密后也遵守此规定。

☑公开　☐控阅（　　年）

作者签名：王中月　　导师签名：

签字日期：2025年10月29日　　签字日期：2025年10月29日

摘要

在开放场景中部署深度学习系统时，模型不可避免地会遭遇与训练分布不一致的输入数据，即分布外样本。常规深度学习模型难以处理此类样本，且往往会呈现过度自信的情况，这对模型的安全性构成严重威胁。分布外检测器能够区分分布外样本，在一定程度上增强了模型安全性。然而，在开放场景下，分布外检测器容易遭受恶意攻击。攻击者可通过制作对抗样本绕过检测器，进而危及整个系统安全。因此，设计具备鲁棒性的分布外检测器，对于保障开放场景中深度学习系统的稳定运行具有重要作用。

鲁棒分布外检测要求在面对对抗样本时，仍能精准实现分布外检测，是一项具有挑战性的任务。常规的鲁棒分类器借助对抗训练可获取一定鲁棒性，但缺乏有效的分布外检测能力；传统分布外检测器虽能较好区分干净样本，却难以抵御对抗攻击。部分研究尝试将对抗训练与分布外检测中的离群值暴露方法简单融合，然而效果不佳。由此可见，鲁棒分布外检测无法单纯依靠方法迁移或简单融合来实现，必须综合考量对抗性与分布外因素，协同优化，这无疑增加了研究的难度。

本文对鲁棒分布外检测进行了系统的研究，设计了针对分布外检测的鲁棒性评估方法，分别在有额外数据和无额外数据的场景提出了针对性的鲁棒分布外检测方法，改进了现有方法的不足，取得了最佳的性能。具体研究工作如下：

(1) 针对现有工作对分布外检测器的鲁棒性评估不够完善和强力的问题，提出了基于检测分数的分布外检测自适应攻击方法。该方法针对分布外检测器以检测分数为攻击目标，对分布内和分布外扰动从不同的方向优化求解，同时制作对抗分布内样本和对抗分布外样本以进行全面的攻击。实验表明，现有分布外检测方法面对该攻击时 AUROC 指标会大幅下降超过 70%。该工作揭示了分布外检测算法的脆弱性，并为分布外检测的鲁棒性提供了有效的评估方法。

(2) 针对有额外数据集场景下现有工作使用额外数据时会引入大量无用信息的问题，提出了基于有效点选取的鲁棒分布外检测方法。该方法通过特征聚集和马氏距离对额外数据集进行有效的选取以避免无用信息的引入，同时将有效点融入神经 ODE 去噪器的训练以增强模型的鲁棒性。实验表明，在有额外数据集的场景，该方法在 CIFAR10 数据集上性能比现有方法提升 2.5%，在 CIFAR100 数据集上提升 4.8%，面对攻击时性能下降不超过 5%。该工作为具有额外数据场景下的鲁棒分布外检测提供了有效的方法。

(3) 针对无额外数据集场景下现有工作生成虚拟分布外数据时未考虑对抗性且具有较强的分布假设的问题，提出了基于虚拟离群点生成的鲁棒分布外检测

方法。该方法通过对抗性传播将特征点驱使向不鲁棒区域，并利用 k-近邻距离进行边界点鉴别和虚拟离群点采样以获得对抗性的虚拟离群点。为了保证鲁棒性并削弱过度自信，该方法还将 logit 归一化融入神经 ODE 的训练过程。实验表明，在无额外数据集的场景，该方法在 CIFAR10 数据集上性能比现有方法提升 3.7%，在 CIFAR100 数据集上提升 6.6%，面对攻击时性能下降不超过 4%。该工作为无额外数据场景下的鲁棒分布外检测提供了有效的方法。

关键词：深度学习；开放世界；鲁棒性；鲁棒分布外检测；分布外检测

ABSTRACT

When deploying deep learning systems in open scenarios, models will inevitably encounter input data that is inconsistent with the training distribution, that is out-of-distribution (OOD) samples. Conventional deep learning models have difficulty when handling OOD samples and often exhibit overconfidence, which poses a serious threat to the model's security. OOD detectors can distinguish the OOD samples, enhancing the model security to a certain extent. However, in open scenarios, OOD detectors are vulnerable to malicious attacks. Attackers can bypass the detectors by creating adversarial samples, thereby endangering the security of the entire system. Therefore, designing a robust OOD detector is important to ensure the stability of deep learning systems in open scenarios.

Robust OOD detection requires accurate OOD detection even in the face of adversarial samples, which is a challenging task. Conventional robust classifiers can obtain a certain degree of robustness through adversarial training but lack effective OOD detection capabilities; traditional OOD detectors can distinguish clean samples well but are difficult to resist adversarial attacks. Some studies have attempted to simply combine adversarial training with the outlier exposure method in OOD detection, but the results are not satisfactory. It can be seen that robust OOD detection cannot be achieved simply by method transfer or simple combination. Instead, adversarial and OOD factors must be comprehensively considered and optimized collaboratively, which undoubtedly increases the difficulty of the research.

This paper conducts a systematic study on robust OOD detection, designs a robustness evaluation method for OOD detection, proposes targeted robust OOD detection methods for scenarios with and without additional data, improves the deficiencies of existing methods and achieves the best performance. The specific work is as follows:

(1) Aiming at the problem that the existing work on the robustness evaluation of OOD detectors is not perfect and powerful enough, an adaptive attack method for OOD detection based on detection scores is proposed. This method targets the detection scores of OOD detectors, optimizes and solves the in-distribution and OOD perturbations from different directions, and creates both adversarial in-distribution samples and adversarial OOD samples for comprehensive attacks. Experiments show that when existing OOD detection methods face this attack, the AUROC index will drop significantly by more than 70%. This work reveals the vulnerability of OOD detection algorithms

and provides a powerful evaluation method for the robustness of OOD detection.

(2) Aiming at the problem that existing works introduce a large amount of useless information when using additional data in scenarios with additional datasets, a robust out-of-distribution detection method based on the selection of effective points is proposed. This method effectively selects the additional datasets through feature aggregation and Mahalanobis distance to avoid the introduction of useless information. Meanwhile, the effective points are integrated into the training of the neural ODE denoiser to enhance the robustness of the model. Experiments show that in scenarios with additional datasets, the performance of this method is improved by 2.5% compared with existing methods on the CIFAR10 dataset and by 4.8% on the CIFAR100 dataset. When facing attacks, the performance degradation does not exceed 5%. This work provides an effective method for robust out-of-distribution detection in scenarios with additional data.

(3) Aiming at the problems that existing works do not consider the adversarial nature and have strong distribution assumptions when generating virtual out-of-distribution data in scenarios without additional datasets, a robust out-of-distribution detection method based on the generation of virtual outlier values is proposed. This method drives the feature points towards non-robust regions through adversarial propagation, and uses the k-nearest neighbor distance for boundary point discrimination and virtual outlier sampling to obtain adversarial virtual outliers. To ensure robustness and weaken overconfidence, this method also integrates logit normalization into the training process of the neural ODE. Experiments show that in scenarios without additional datasets, the performance of this method is improved by 3.7% compared with existing methods on the CIFAR10 dataset and by 6.6% on the CIFAR100 dataset. When facing attacks, the performance degradation does not exceed 4%. This work provides an effective method for robust out-of-distribution detection in scenarios without additional data.

KEY WORDS: Deep Learning, Open World, Robustness, Robust Out-of-Distribution Detection, Out-of-Distribution Detection

目录

第 1 章　绪论 1
1.1　研究背景和意义 1
1.2　国内外研究现状 2
1.2.1　分布外检测 2
1.2.2　对抗鲁棒性 6
1.2.3　鲁棒分布外检测 10
1.3　研究问题、内容及创新点 11
1.3.1　研究问题总结 11
1.3.2　研究内容和创新点 12
1.4　论文组织结构 13
第 2 章　相关知识及工作准备 15
2.1　分布外检测 15
2.1.1　分布外检测相关定义 15
2.1.2　常见分布外检测方法 15
2.2　对抗攻击和对抗防御 19
2.2.1　对抗攻击 19
2.2.2　对抗防御 20
2.3　数据集与评价指标 22
2.3.1　基准数据集介绍 22
2.3.2　评价指标 25
2.4　本章小结 25
第 3 章　基于检测分数的分布外检测自适应攻击方法 26
3.1　引言 26
3.2　问题描述 27
3.3　算法设计 28
3.4　实验与分析 30
3.4.1　实验环境及设置 30
3.4.2　分数自适应攻击有效性验证实验 31
3.4.3　超参数分析实验 33
3.4.4　可视化分析 36
3.5　本章小结 37

第 4 章　有额外数据场景基于有效点选取的鲁棒分布外检测方法 38
4.1　引言 38
4.2　问题描述 39
4.3　算法设计 40
4.3.1　模型框架 41
4.3.2　基于李雅普诺夫稳定性理论的神经 ODE 去噪器 42
4.3.3　基于原型学习的特征聚集投影 43
4.3.4　基于马氏距离的有效点选取 45
4.3.5　分布外评价分数 46
4.4　实验与分析 46
4.4.1　数据集与评价指标 46
4.4.2　实验设置 47
4.4.3　基准数据集鲁棒分布外检测对比实验 47
4.4.4　消融实验 49
4.4.5　超参数分析实验 50
4.4.6　可视化分析 52
4.5　本章小结 54
第 5 章　无额外数据场景基于虚拟离群点生成的鲁棒分布外检测方法 55
5.1　引言 55
5.2　问题描述 56
5.3　算法设计 57
5.3.1　模型框架 57
5.3.2　对抗虚拟分布外数据点合成 58
5.3.3　神经 ODE 去噪器训练 60
5.3.4　分布外评价分数 62
5.4　实验与分析 62
5.4.1　数据集与评价指标 62
5.4.2　实验设置 62
5.4.3　基准数据集鲁棒分布外检测对比实验 63
5.4.4　消融实验 64
5.4.5　超参数分析 65
5.4.6　可视化分析 67
5.5　本章小结 69

第 6 章　总结与展望 70
6.1　本文工作总结 70
6.2　未来工作展望 71
参考文献 72
致谢 79
在读期间取得的科研成果 80

插图清单

图 1.1 本文章节组织框架图 13

图 2.1 MCM 方法示意图[23] 17
图 2.2 对抗样本示意图[37] 19
图 2.3 Defense-GAN 示意图[57] 22
图 2.4 ASODE 示意图[60] 22
图 2.5 CIFAR 数据集示例图 23
图 2.6 Tiny-images 数据集示例图 23
图 2.7 SVHN 数据集示例图 23
图 2.8 LSUN 数据集示例图 24
图 2.9 iSUN 数据集示例图 24
图 2.10 Texture 数据集示例图 24
图 2.11 Places365 数据集示例图 24

图 3.1 分数自适应攻击示意图 29
图 3.2 不同扰动边界下的 AUROC 性能下降曲线图 34
图 3.3 不同攻击步数下的 AUROC 性能曲线图 35
图 3.4 不同迭代步长下的 AUROC 性能曲线图 35
图 3.5 不同攻击情况下的频率图 36

图 4.1 鲁棒分布外检测示意图 38
图 4.2 对抗离群值暴露方法示意图 39
图 4.3 EPSD 方法框架图 41
图 4.4 特征聚集示意图 44
图 4.5 有效点选取示意图 45
图 4.6 不同积分时间的 AUROC 性能曲线图 50
图 4.7 不同候选点数量的 AUROC 性能曲线图 51
图 4.8 不同扰动边界下的 AUROC 性能曲线图 52
图 4.9 不同数据集攻击前后的频率图 53
图 4.10 特征聚集前后各类别 UMAP 可视化图 53
图 4.11 特征聚集前后分布内外数据 UMAP 可视化图 54

图 5.1 APNOS 方法框架图 58
图 5.2 对抗传播示意图 59

图 5.3 虚拟分布外数据点采样示意图 60
图 5.4 不同距离基数的 AUROC 性能曲线图 65
图 5.5 不同采样方差的 AUROC 性能曲线图 66
图 5.6 不同边界点数的 AUROC 性能曲线图 66
图 5.7 不同数据集攻击前后的频率图 67
图 5.8 边界点检测 UMAP 可视化图 68
图 5.9 虚拟离群点 UMAP 可视化图 68

附表清单

表 3.1 实验环境 30
表 3.2 远-分布外数据实验 32
表 3.3 近-分布外数据实验 32
表 3.4 攻击前后数据集 AUROC 变化表 33
表 3.5 攻击前后数据集 FPR95 变化表 33

表 4.1 CIFAR10 数据集上攻击前后 AUROC 变化表 48
表 4.2 CIFAR100 数据集上攻击前后 AUROC 变化表 49
表 4.3 模块消融数据表 49

表 5.1 CIFAR10 数据集上攻击前后 AUROC 变化表 63
表 5.2 CIFAR100 数据集上攻击前后 AUROC 变化表 64
表 5.3 模块消融数据表 65

第 1 章　绪论

本章首先介绍了鲁棒分布外检测的背景与研究意义，然后阐述了分布外检测、对抗鲁棒性和鲁棒分布外检测的研究现状，最后总结了本文的主要工作并说明了全文结构安排。

1.1　研究背景和意义

在如今的数字化时代，人工智能已经取得了丰硕的成果，其应用范围广泛、影响深远，几乎渗透了社会中的各个角落。我国政府高度重视人工智能技术，《新一代人工智能发展规划》[①]明确指出要加快人工智能技术的研发与应用，将人工智能上升为国家战略，强调其在提升国家竞争力、推动经济转型升级方面的关键作用。深度学习算法作为人工智能中的关键技术，在人脸识别、自动驾驶、医学影像分析、工业缺陷检测等领域取得广泛应用[1-2]。

然而，大多数的深度学习算法都基于独立同分布假设，即假定训练数据和测试数据符合于同一分布，但这一假设对于在开放场景中部署的深度学习系统而言却难以满足。通常将与训练数据服从于同一分布的数据称为“分布内样本”，将与训练数据不服从于同一分布的数据称为“分布外样本”。在开放场景中，模型不可避免的会遇到训练过程中未见过的分布外数据，然而大多数模型在面对未知时却会表现出“不懂装懂”和“过度自信”的现象，对未知会产生一个错误预测并给予较高的置信度[3]。深度学习模型的这种现象对于在开放场景下部署的安全关键系统而言是危险的。因此，对安全关键系统而言正确区分已知与未知是极其重要的，区分出未知后可以进行拒绝并交由其他机制（如人工）处理从而避免一定的安全问题的产生。

分布外检测是一种分离分布外与分布内数据的技术[4-7]，符合开放世界的需求，能提升深度学习系统的安全性，在现实中有广阔的应用前景。例如，在自动驾驶领域中，当检测到训练中未见过的事件时，驾驶系统应当发出警报，并将控制权交给司机; 在医学图像领域，通常有监督方法的泛化能力有限，训练数据并不能包含所有可能的病理表现，当测试数据偏离训练数据的分布时，经常会导致错误和过度自信的预测，这对于患者来说是灾难性的，而医学图像分布外检测可以通过分析数据分布和检测潜在的错误样本来提前捕获可能预测错误的样本，然后交给医生处理，极大地保证了患者的安全性。

然而，近年来有研究表明，现有的分布外检测器极易受到对抗性输入的影

[①]国务院. 新一代人工智能发展规划. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

响。所谓的对抗性输入指的是对输入增加一些人肉眼难以分辨的噪声，这些噪声不会改变人对于输入的判断，却能使得深度学习系统的判断产生错误。大多数对抗鲁棒性领域的研究关注的是图像分类器，但这在分布外检测中也是重要且需要受到关注的。现有的分布外检测器只能较好的区分良性的输入，但在开放场景中，模型还会受到恶意攻击者的威胁。攻击者可以通过制作对抗性输入使得分布外检测器产生错误的判断从而攻破分布外检测系统，进一步威胁了安全关键系统的安全性与稳定性。例如，攻击者可以恶意破坏一个路标，使得分布外检测器将未见过的未知识别为已知，并“自以为是”地给出决策而不是发出警报或告知司机，从而引发安全问题。

在这样的背景下，鲁棒分布外检测开始逐渐受到研究者的关注[8-10]。鲁棒分布外检测是对抗鲁棒性和分布外检测这两个领域的结合。在开放世界中，攻击者可以对分布内输入施加特定噪声使得分布外检测器识别为分布外，从而引发大量的系统警报影响系统正常运行；同样，攻击者也可以对分布外输入施加噪声使得其被识别为分布内，从而躲避分布外检测器引发更严重的安全问题。因此在鲁棒分布外检测中，分布外检测器不仅需要正确区分分布内和分布外输入，还要具有抵御对抗扰动的能力，使得对抗扰动的施加并不会对分布外检测器的判断产生影响。鲁棒分布外检测更加贴合实际，同时考虑到了开放场景中分布外样本和对抗性攻击的存在，进一步为模型的安全性提供了保障。

综上所述，为了构建更为鲁棒安全的深度学习系统，需要研究深度学习模型如何区分已知与未知，并考虑到现实场景中攻击者的存在。鲁棒分布外检测同时考虑到分布外输入和对抗性攻击的存在，而不是顾此失彼，更加具有现实意义。鲁棒分布外检测的研究有利于开放世界中的深度学习理论的发展，并能为开放场景中系统的安全性提供支持，对于推进深度学习技术在开放世界中的部署应用是迫切且重要的。

1.2　国内外研究现状

鲁棒分布外检测是一个结合分布外检测和对抗鲁棒性的研究领域。因此，本节首先介绍分布外检测的研究现状，这些研究均是在干净样本的前提下进行；之后，为了引出对抗鲁棒性的问题，本节介绍了对抗鲁棒性领域的发展；最后，本节概括了鲁棒分布外检测的发展历程和现有成果。

1.2.1　分布外检测

早期，Nguyen 等发现深度神经网络会对一些人类完全无法识别的图像产生非常高的置信度，比如一个训练好的深度神经网络分类器会以 99.99% 的置信

度将白噪声分辨为狮子，初步揭示了深度学习模型“过度自信”的问题[3]。之后，Hendrycks 等首次提出了分布外检测的问题。他们发现神经网络在面对训练分布以外的数据时，模型会以高置信度给出错误的预测[4]。他们还提出了以最大 softmax 分数作为指标区分分布内和分布外样本作为该领域的基线方法。在此之后，该领域便受到了越来越多研究者的关注。

发展至今，分布外检测已有大量的研究成果。本文将这些方法分为三大类：基于事后处理的方法，基于额外数据的方法和基于模型训练的方法。基于事后处理的方法不修改原模型，而是在原模型的基础上计算各种分数作为指标，以该指标进行区分。基于额外数据的方法通过收集额外的辅助分布外数据集帮助模型进行正则化训练。基于模型训练的方法不采用额外的分布外数据集，而是通过构造特殊的特征空间或者生成虚拟的分布外数据进行训练。本小节将分别概括介绍这些方法。

1. **基于事后处理的方法**

基于事后处理的方法主要使用预训练模型的特征或输出 (如 logits、特征层激活值和梯度等) 直接构建分布外检测分数，无需额外训练或仅需轻量微调。其核心思想是通过统计差异区分分布内样本与分布外样本。

最初的基于事后处理的方法是 Hendrycks 等人[4]提出的 MSP，该方法使用 softmax 输出的最大值作为指标，设定阈值对输入进行区分，其思想是理想的分类器应该将分布外输入以较为平均的概率判断为各个类别。但其后的研究表明，softmax 分数会对分布外样本产生过度自信的现象。对此，Liang 等人[5]提出了改进的方法 ODIN。ODIN 使用温度缩放对 softmax 值进行修改，并且该方法还对输入添加微小的对抗扰动将预测概率向最大概率类别进一步增加的方向驱使，从而使得网络对分布内和分布外的输出概率产生更明显的差异。

除此之外，一些方法并没有局限于使用 softmax 分数进行检测。Liu 等人[11]提出使用能量分数能从理论上与输入的概率密度一致，相比于传统的 softmax 分数能更好的区分分布内与分布外样本，不易受到过度自信问题的影响。Lee 等人[12]提出 MDS，该方法统计训练集中各个类别在特征空间中的类条件高斯分布，然后计算测试输入在特征空间与各个类条件高斯分布的马氏距离，以最小马氏距离作为检测分布内与分布外的指标。与马氏距离不同，KNN 方法[13]使用 k-近邻距离作为分布外检测指标，该方法不依赖于特征空间中的分布假设，比马氏距离更具适用性。ViM 方法[14]将类别无关的特征空间分数与类别相关的 logits 分数相结合，通过将特征在训练样本特征构成的主空间的正交补空间上投影得到代表虚拟分布外类的额外 logit，再用匹配系数与原始 logits 匹配，经 softmax 后该虚拟 logit 对应的概率作为分布外的指标。与使用特征空间信息或者网络输出来进行检测的方法不同，Huang 等人[15]提出的 GradNorm 利用梯度空间的信

息，从 softmax 输出和均匀分布的 KL 散度反向传播获得梯度的向量范数并以此检测分布外输入。

一些其他的方法不是设计新的检测指标，而是通过修改网络激活方式对置信度进行矫正。Sun 等人[16]观察到分布内样本的平均激活接近恒定，而分布外样本的平均激活在不同的神经元之间存在显著差异并且偏向于具有尖锐的正值，这种高激活值无法在输出中表现出来，从而使得模型对分布外数据产生高置信度的结果，由此提出了 ReAct 检测方法。ReAct 对隐藏单元中过大的激活值进行修正截断，从而获得了更为合理的置信度分数。Djurisic 等人[17]提出的 ASH 对网络靠后的层的激活进行处理，移除网络在该层的很大一部分 (如百分之 90)，对剩余部分进行简化或轻微调整。ASH 认为过参数化的网络可能在学习表示上过度了，对当前任务而言生成了大量的冗余特征，可能使得网络对见过和未见过的数据的区分能力变差，因此减少这种冗余可以提升分布外检测的能力。

基于事后处理的方法无需重新训练模型，可直接利用现有模型进行分布外检测，使用更加方便。但是该类方法性能会受到现有模型的制约，由于现有模型在训练期间可能未对分布外样本进行优化，使得分布外检测性能欠佳。

2. **基于额外数据的方法**

基于额外数据的方法除了利用分布内数据之外，还利用了额外的分布外信息来辅助检测。其中，有的方法的分布外信息来源于直接收集一定量的分布外样本，并以此对模型进行正则化训练；有的方法的分布外信息来源于事先利用大量数据训练的预训练模型，如 CLIP[18]等。

通过收集额外分布外数据的一类方法称为基于离群值暴露 (Outlier Exposure，OE) 的方法。离群值暴露首先由 Hendrycks 等人[19]提出，他们观察到通过收集额外的分布外数据集，并将额外分布外数据与均匀分布的 KL 散度作为分布外正则化项，可以使得模型泛化到未见过的分布外数据。Yu 等人[20]使用一个双头深度卷积神经网络并最大化两个分类器之间的差异，还利用未标记的数据进行无监督训练，使用这些未标记的数据来最大化两个分类器的决策边界之间的差异，以将分布外样本推到分布内样本的流形之外，以此来检测分布外输入。UGD[21]在 OE 的基础上，利用额外未标记数据并通过无监督深度聚类任务探索未标记数据的语义表达，并利用辅助任务生成的分组信息区分分布内和分布外样本。Zhang 等人[22]提出 MixOE，使用 Mixup 和 CutMix 等数据增强方法对分布内和分布外样本的混合集进行数据增强以获得接近分布内样本的“虚拟”离群值，以此在更大的分布外区域内诱导正则化，并观察到对粗粒度和细粒度分布外样本都具有明显的影响。

另一些方法则利用大量数据预训练的模型来获取分布外信息。Ming 等人[23]提出 MCM，利用了 CLIP 模型将输入图像编码后与各个类别的文本编码

进行匹配，将 CLIP 模型转换为图像分类器，以最大概念匹配分数作为区分分布外的指标，实现了零样本的分布外检测。Wang 等人[24]通过向 CLIP 模型中加入“否定”逻辑提出了 CLIPN，CLIPN 在 CLIP 中添加了可学习的“否定”提示词及“否定”文本编码器，通过图像-文本二元相反损失和文本语义相反损失两种损失函数训练 CLIPN 将图像与“否定”提示词建立联系，融合否定语义进行分布外检测。同样，Nie 等人[25]观察到 CLIP 模型对于“否定”语义的缺失，利用 COOP 在分布内数据上训练得到每个分布内类别的正提示，然后固定正提示，通过负提示-图像分离损失、负-正提示距离损失、负-负提示距离损失三个学习目标，在分布内数据上学习每个分布内类别的负提示，使负提示远离分布内图像且不同类别负提示间有区分度，最后通过计算图像与正负提示的相似度得出分布外得分。

基于额外数据的方法由于引入了真实的分布外信息，因此具有较好的分布外检测性能。然而，该类方法对于额外数据有着较高的要求，需要额外数据具有多样性，并且与分布内数据无语义重叠的同时还需要有一定的相关性。若额外数据质量较差，该类方法的性能会受到较大影响。然而，符合条件的额外数据在一些场景下难以获得，从而限制了该类方法的适用性。

3. **基于模型训练的方法**

与基于额外数据的方法不同，基于模型训练的方法不依赖于额外的分布外数据，而是通过将输入映射到特定的特征空间或者通过合成虚拟分布外数据等方法来进行分布外检测。

一些方法通过设计损失获取特别的特征空间增强分布外检测能力。Hendrycks 等人[26]利用无监督学习增加模型的鲁棒性和分布外能力，通过对模型增加一个辅助的自监督任务预测图像的旋转角度可以使模型获得更好的不确定性估计。Tack 等人[27]提出将对比学习框架 SimCLR 运用到分布外检测中的方法 CSI，该方法将数据增强样本视作分布偏移样本，利用对比损失将原始样本与其分布偏移样本区分，并在此基础上增加辅助任务预测输入应用了哪个偏移转换操作，最后基于此训练框架设计了新的检测分数用于检测分布外。CIDER[28]引入一种新颖的表示学习框架用于分布外检测，将输入映射到超球面嵌入空间并计算各个类别原型，设计类内聚集损失和类间离散损失将每个类别嵌入聚集到类别原型周围并使各个类别原型尽量远离，以此得到良好的嵌入表示并用 K-近邻距离进行分布外检测。PALM[29]在 CIDER 的基础上进行改进，为了增加类别原型的表示能力使用混合原型代表单一类别，获得更强的分布外检测性能。

一些方法合成虚拟的分布外数据代替真实分布外数据以解决某些场景下额外数据难以获得的问题。Lee 等人[30]利用 GAN[31]生成的图片作为虚拟的分布外样本，使用预测概率与均匀分布的 KL 散度损失作为正则化项矫正分布外的置

信度。同样，ARPL[32]也利用 GAN 生成的图片作为虚拟分布外样本，除此之外 ARPL 还引入对抗互反点代表未知空间，使分布内样本的嵌入远离对抗互反点，使生成的虚拟分布外样本的嵌入靠近对抗互反点，利用嵌入到对抗互反点的距离作为判别分数。直接生成虚拟图片比较困难且不够多样，因此一些其他方法选择在特征空间中生成虚拟特征代表分布外样本。Kong 等人[33]提出 openGAN 以生成虚拟特征，让判别器判别特征而不是图像，使得 GAN 的训练过程更加稳定且分布外检测性能更好。VOS[34]将特征空间建模为类条件高斯分布，在高斯分布的低似然区域采样作为虚拟离群值，并基于能量分数设计损失区分分布内和分布外。Tao 等人[35]认为将特征空间建模为高斯分布是一个强大且有限制性的假设，因此利用非参数化的采样方法采样获取虚拟离群值，具有更强的灵活性和通用性。

基于模型训练的方法不依赖额外数据，具有较强的适用性。然而，该类方法缺失真实的分布外信息，对特征空间的设计或虚拟分布外数据的合成通常依赖于一定的先验信息，若先验与真实分布外信息产生偏差则会对性能产生影响。因此该类方法性能通常不如基于额外数据的方法。

1.2.2　对抗鲁棒性

深度学习模型的脆弱性首先由 Szegedy 等人[36]揭露，他们发现对输入施加极低水平的扰动会导致模型给出截然不同的结果。所施加的扰动称为对抗扰动，这样的扰动不会影响人眼的判断，却会使得模型给出高置信度的错误判断。从此，对抗鲁棒性便受到了研究者的关注。对抗鲁棒性领域的研究通常可分为两类：一类研究如何添加扰动使得模型展现出更严重的脆弱性，即对抗攻击；一类研究如何获取更为鲁棒的模型使得对对抗扰动有抵抗能力，即对抗防御。本小节分别概括这两类研究。

1. 对抗攻击

对抗攻击的目的是对输入施加一定量的微小扰动，使得在人眼的判断不发生改变的前提下让深度学习模型的输出产生改变。原始样本添加扰动后被称为对抗样本。对抗样本是悬在深度学习头上的“达摩克里斯之剑”，在对安全性有着较高要求高的场景下是不容忽视的。自 Szegedy 等人揭露深度学习的脆弱性以来，各种各样的攻击方法层出不穷。这些方法按照是否可以获得模型的参数信息分为白盒攻击和黑盒攻击。

（1） 白盒攻击

白盒攻击指攻击者可以获得深度学习模型结构和参数，并以此生成对抗样本。最初的白盒攻击方法是 Goodfellow 等人[37]提出的快速梯度符号法 (Fast Gradient Sign Method，FGSM)。FGSM 是一种基于梯度的一步攻击方法，通过损失

梯度上升来获取欺骗模型的扰动。之后的很多攻击方法都受到了 FGSM 的启发。Kurakin 等人[38]提出 BIM，在 FGSM 的基础上使用多次迭代，每次添加一个小的扰动直到达到预设的最大扰动，期间使用裁剪函数确保扰动不超过最大扰动。PGD[39]也采用了类似的多次迭代，与 BIM 不同的是 PGD 采用更为合理的投影操作来确保扰动不越界。PGD 因其卓越的效果已经成为验证模型鲁棒性的通用方法。DeepFool[40]是一种非目标攻击算法，利用线性分类器的决策边界将图像投影到最近的分类超平面来寻找导致错误分类的最小失真。Carlini 等人[41]提出 CW 攻击，这是一种基于优化的攻击方法，攻击成功率高且可定制性强，可以根据不同的场景选择不同的距离度量函数来生成对抗样本，但具有较高的计算复杂度。Auto Attack[42]是一个集成了多种攻击方法的对抗攻击工具包，主要包含 APGD-CE、APGD-DLR、FAB 和 Square Attack 等攻击方法，通过多样化的无参数攻击使得模型的脆弱性得到更为可靠的评估，为研究人员评估和改善模型的鲁棒性提供了全面的框架，被广泛运用于之后的鲁棒性评估研究中。

白盒攻击所生成的对抗样本极为强大，能够近似体现模型在最坏情况下对抗鲁棒性的真实水平。不过，白盒攻击与实际场景并不相符。在现实中，攻击者往往难以完整获取模型的结构及参数信息，因此无法实施白盒攻击。鉴于此，白盒攻击一般仅用于评估模型的鲁棒性。

（2） 黑盒攻击

黑盒攻击假定攻击者无法获得模型的结构参数等信息，只能以黑盒的方式使用模型。黑盒攻击从原理上来说一般可分为基于查询的攻击和基于迁移的攻击。

基于查询的攻击通过向目标黑盒模型发送一系列查询，并根据模型的响应结果逐步调整输入，以构建能够欺骗模型的对抗样本。Chen 等人[43]提出 ZOO，在黑盒场景下利用有限差分等零阶优化算法来估计梯度，通过在不同的维度上施加微小扰动并查询模型输出来近似计算梯度值，之后通过迭代和梯度上升来逐步调整对抗样本。Brendel 等人[44]提出了一种基于决策边界的查询攻击方法，攻击者从一个初始点开始，通过不断向目标模型查询，沿着对抗和非对抗之间的决策边界逐步调整输入，朝着能够改变模型决策的方向移动，最终找到对抗样本。Ilyas 等人[45]聚焦于查询次数和信息有限的场景，利用模型输出的置信度分数，采用基于进化策略的方法估计梯度方向，在每次迭代中，通过对当前样本进行随机扰动并评估模型响应，筛选出更优的扰动方向以更新对抗样本，在减少查询次数的同时提高攻击效率和成功率，为解决黑盒环境中资源受限情况下的模型攻击问题提供有效方案。

基于迁移的攻击利用对抗样本在不同模型间的迁移性，先在一个与目标模型有一定相似性的替代模型（白盒模型）上生成对抗样本，然后将这些样本应用

到目标黑盒模型上，尝试欺骗目标模型。Szegedy 等人[36]在研究对抗扰动时发现针对某一模型生成的对抗样本也能对其他模型起到攻击的作用，初步揭露了对抗攻击的迁移性。Moosavi 等人[46]发现了通用对抗扰动的存在，即存在一个单一的、相对较小的扰动向量，能够在不同的图像上一致地导致深度神经网络做出错误分类。他们提出了一种迭代算法来生成通用对抗扰动，该方法基于在多个自然图像上的迭代优化，在每次迭代中，对当前的扰动进行更新使模型错误分类的图像数量增加，通过不断迭代调整扰动，逐渐收敛到一个通用的对抗扰动。Papernot 等人[47]提出了一种实用的黑盒攻击方法，该方法结合查询的思想，通过多次查询获取数据源，然后利用数据源训练代理模型，通过攻击代理模型生成源模型的对抗样本。Wang 等人[48]通过引入聚合梯度来获取特征的重要性，并利用特征的重要性来指导搜索对抗性示例以破坏关键特征从而达到攻击的目的。

从实际的角度来看，黑盒攻击比白盒攻击更具有现实意义。但由于缺乏模型内部信息使得黑盒攻击的设计和实施较为困难，攻击者只能通过有限的输入输出信息来推断模型的行为而无法获得真实的梯度信息，所以攻击成功率通常不如白盒攻击，容易进行防御。

2. 对抗防御

对抗防御的目的与对抗攻击相对，即使得模型的决策尽可能地不受对抗扰动的影响，提高深度学习模型的鲁棒性。对抗防御的方法可分为基于检测的防御、基于对抗训练的防御和基于对抗净化的防御。

（1） 基于检测的防御

基于检测的防御指设计专门的检测机制识别输入样本是否为对抗样本，一旦检测到对抗样本，可以采取相应的措施，如拒绝处理该样本或对其进行修正。基于检测的方法大多利用干净样本和对抗样本在模型中的某些差异性指标来进行区分。Ma 等人[49]提出局部内在维度 (LID) 来检测对抗样本，他们发现对抗性扰动会影响对抗性区域的 LID 特征，因此 LID 可以表现干净样本与对抗样本之间的区别。Huang 等人[50]提出了一种模型无关的方法，通过对输入进行多次随机扰动并观察模型输出的变化情况，计算输出的统计量，如均值、方差等，根据设定的阈值判断输入是否为对抗样本。Cintas 等人[51]则使用无监督的方法，将输入数据输入到自动编码器中进行编码和解码并计算重建误差，同时对自动编码器隐藏层的激活值进行子集扫描，分析激活值的分布和变化情况，综合重建误差和激活值子集的特征判断输入数据是否为对抗样本。

（2） 基于对抗训练的防御

基于对抗训练的防御在模型的训练过程中使用对抗样本，使得对抗样本的测试误差得以减少，目前已经成为公认有效的防御方法。Goodfellow 等人[37]使用 FGSM 生成对抗样本，并将对抗样本添加到正则化损失中进行正则化训练，该方

法对单步攻击具有良好的效果，但却仍然容易受到多步迭代攻击的欺骗。Madry 等人[39]在此基础上使用 PGD 攻击产生对抗样本，这样生成的对抗样本在干净样本周围更加具有对抗性，取得了更好的对抗训练结果。但因为这些方法使用线性函数来近似损失函数，会在决策面上的数据点附近产生尖锐的曲率，因此依然容易受到多步攻击的影响。Huang 等人[52]将对抗训练看作一个最小最大问题，对抗样本的生成作为内部最大化问题的求解，使用对抗样本更新网络参数作为外部最大化问题的求解。Cai 等人[53]指出仅使用对抗样本进行训练会导致对抗样本的过拟合，他们将课程训练的思想引入对抗训练提出了课程对抗训练 CAT。CAT 从少量步骤开始逐渐添加 PGD 迭代步骤直到模型对当前攻击具有较高精度，因为在训练早期使用弱攻击，因此这种方法具有较好的干净数据泛化能力。由于对抗样本的产生是一个耗时的过程，因此一些方法研究如何在对抗训练中减少时间成本。Shafahi 等人[54]提出 Free-AT，通过循环更新模型参数时计算的梯度信息来消除生成对抗性示例的开销成本。基于此，Wong 等人[55]提出 Fast-AT，使用随机初始化与 FGSM 相结合，发现与基于 PGD 的方法一样有效，且计算成本得以显著地降低。

（3） 基于对抗净化的防御

与基于对抗训练的方法不同，基于对抗净化的防御通过消除输入中的对抗噪声从而将其转化为干净样本用于模型预测。Song 等人[56]提出一种对抗净化的方法 PixelDefend，该方法利用 PixelCNN 来学习自然图像的分布，将可能含有对抗扰动的输入图像通过生成模型进行重构，使其尽可能符合自然图像的分布状态，将对抗样本图片重构为符合训练图片的分布状态，以去除或减轻对抗扰动的影响，从而在输入给分类模型之前对对抗样本进行净化，提高模型对对抗样本的鲁棒性。一些其他方法使用生成对抗网络 GAN 来学习干净数据的分布，并以此净化对抗噪声。Samangouei 等人[57]提出 Defense-GAN，利用 GAN 模型对未扰动的干净样本的分布进行建模，在推理时从 GAN 模型中生成与输入图像接近的图像用于分类器输入，从而达到去除噪声的效果。与生成模型的像素级去噪不同，一些方法在特征级别上进行对抗去噪。Liao 等人[58]提出 HGD，他们认为标准去噪器存在误差放大效应，会使残留对抗噪声最终导致错误分类，因此 HGD 将损失函数定义为干净图像和去噪图像激活的目标模型输出之间的差值，通过这种方式训练去噪器达到特征层面的去噪而非修改像素。一些基于神经差分方程 (NODE) 的防御方法也可以看作在特征层面进行去噪。TisODE[59]引入一个时不变的神经 ODE，通过稳态正则化来约束输入输出的变化从而减少扰动对于特征的修改。ASODE[60]受启发于动力系统的李雅普诺夫稳定性理论，通过对 NODE 对应的线性时变系统施加约束，使每个干净实例成为缓慢时变系统的渐近稳定平衡点，则如果对抗实例在这个点的邻域中，渐近稳定性将降低对抗噪声，使对

抗实例接近干净的实例从而起到对抗净化的作用。SODEF[61]同样使用李雅普诺夫稳定性理论，使用正则化器使得提取的特征点位于平衡点附近。

1.2.3　鲁棒分布外检测

鲁棒分布外检测的目的是在具有对抗攻击的情况下进行稳健的分布外检测。在这样的设定下，分布外检测器通常需要检测四种输入：干净分布内数据、对抗分布内数据、干净分布外数据、对抗分布外数据。在鲁棒分布外检测中，干净分布内样本和对抗分布内样本都应该转发给主分类器，而干净分布外样本和对抗分布外样本都应该被检测后拒绝或触发报警系统后交由其他机制处理。对抗分布内数据的目的是欺骗检测器将其识别为分布外输入，从而频繁地引发系统报警影响系统正常运行；对抗分布外样本的目的则是欺骗检测器将其识别为分布内输入，从而躲避警报系统产生更严重的安全后果。鲁棒分布外检测因其对开放场景中部署的深度学习系统的安全性至关重要，近年来逐渐受到研究者的关注。

Hein 等人[62]发现 ReLU 网络存在对远离训练分布的数据产生高置信度的问题，为了解决这个问题他们提出 ACET，将分布内对抗样本运用到分布外检测的置信度正则化中。大多数方法为了使分布外检测器具有抵抗对抗分布外样本的能力，采用了结合离群值暴露和对抗训练的方法。Sehwag 等人[63]分析了开放场景下分布外数据的鲁棒性问题，并提出 RATIO 将对抗分布外样本纳入到对抗训练框架中，RATIO 使用分类损失生成对抗分布外样本，使模型具有一定抵御对抗分布外样本的能力，但未考虑到对抗分布内样本。Chen 等人[8]首次规范化地提出了鲁棒分布外检测的概念，并指出需要同时考虑对抗分布内样本和对抗分布外样本，使用额外分布外数据集，通过最大化分布外样本输出和均匀分布之间的 KL 散度施加扰动产生对抗样本。与此类似，ATOM[9]也采用辅助数据集进行对抗训练，为了对额外分布外数据的信息进行高效利用同时减少对抗训练成本，ATOM 选择性地对信息丰富的异常值进行采样，而不是随机选择。但是 ATOM 的对抗性评估不是完全标准的，因为它只针对分布外测试样本。Azizmalayeri 等人[10]提出先前工作的评估都是基于小扰动或弱攻击，在扰动大小 $\epsilon = 8/255$ 的设定下先前方法的表现都比随机检测要差，因此他们提出了抵御更强扰动幅度的方法 ATD。ATD 采用 GAN 的生成器生成虚拟特征作为分布外特征，并利用额外分布外数据和虚拟特征训练判别器判别分布外样本。RODEO 方法[64]考虑到基于离群值暴露的方法对额外分布外数据的质量依赖，通过文本编码器和多模态图像生成器使生成样本与分布内样本的语义接近并增强辅助图像的多样性，以此提高生成样本的质量。Mirzaei 等人[65]考虑到在一些情况下额外分布外数据是难以获得的，并且仅依靠辅助数据集来生成对抗分布外数据会导致模型偏向特定的分布外实例，从而影响检测器在推理过程中泛化到未见过的分布外数据

的能力。他们提出 AROS，在分布内嵌入空间的低似然区域采样获取虚拟离群值，并结合了李雅普诺夫理论在神经微分方程中关于鲁棒性的成果，作为第一个不依赖额外分布外数据的方法取得了良好的效果。

1.3　研究问题、内容及创新点

1.3.1　研究问题总结

随着深度学习技术的不断发展，如何安全可靠的在开放场景中部署深度学习模型变得愈发重要。鲁棒分布外检测是对抗性设定下的分布外检测，同时考虑到开放场景中攻击者和分布外输入的存在，是深度学习模型的一道安全防线，对推进深度学习模型稳健安全的部署至关重要。近年来鲁棒分布外检测领域逐渐受到研究者的关注，各种理论方法层出不穷，但仍然存在未解决的问题。基于对研究现状的阐述，本文总结了现有研究存在的问题如下：

（1） 分布外检测器鲁棒性评估不够完善

现有方法所产生的对抗样本不够全面和强力，使得分布外检测器的鲁棒性没有得到较好的评估。一些方法没有同时考虑对抗分布内样本和对抗分布外样本的存在，使得对抗样本不够全面。一些方法在生成对抗样本时不是使用分布外评分函数而是使用分类损失或 KL 散度，且仅在较小扰动边界下生成，使得对抗样本不够强力。这些不足都使得分布外检测器的鲁棒性没有得到较好的评估。

（2） 辅助分布外数据集没有得到较好利用

在具有辅助分布外数据集的情况下，一些方法采用了离群值暴露的策略，在训练期间随机选取辅助分布外样本进行训练，而随机选取会导致被选的辅助分布外样本与分布内样本相关性不够，导致引入大量无效的信息，使模型关注不重要的特征从而对模型训练过程产生误导。ATOM[9] 利用信息量进行选取，但其训练期间只产生对抗分布外样本，导致模型的对于对抗分布内样本的鲁棒性不够。因此，在具有额外分布外数据集的情况下，如何有效利用辅助分布外数据集并获得具有较好对抗鲁棒性的模型是未完善解决的问题。

（3） 无额外数据集情况下生成分布外数据的方法欠佳

一些情况下，符合要求的额外辅助分布外数据集是难以获得的，因此需要研究无额外分布外数据场景下的鲁棒分布外检测方法。目前只有 AROS[65] 在该项设定下设计鲁棒分布外检测方法，但是该方法将特征空间假设为类条件高斯分布并以此生成分布外数据，具有较强的分布假设，并不能保证满足。而且该方法利用干净样本生成虚拟分布外数据，并未更加关注到不鲁棒的区域。因此，在无额外数据集的情况下，如何仅利用分布内数据集设计更适用和更有效的鲁棒分布外检测方法是待解决的问题。

1.3.2　研究内容和创新点

针对以上问题，本文的主要研究内容如下：

（1） 基于检测分数的自适应攻击方法

针对分布外检测器鲁棒性评估不够完善的问题，本文提出了一种基于检测分数的自适应攻击方法，以产生更为全面和强力的对抗样本用于分布外检测鲁棒性的评估。该方法以检测分数为攻击目标，针对分布内样本最小化检测分数产生分布内对抗样本，针对分布外样本最大化其检测分数产生分布外对抗样本，在测试过程中同时考虑分布内和分布外对抗样本以获得更全面的鲁棒性评估。同时，本文还分析了自适应攻击的扰动边界、攻击步长和迭代步数对分布外检测性能的影响，为使用该方法用于验证时的参数设置提供参考。

该方法的创新点在于：生成对抗样本时的攻击目标随着不同的分布外检测分数和不同的输入类型 (分布内输入和分布外输入) 自适应的调整，使得产生对分布外检测具有更强更具针对性的对抗样本；在评估模型的分布外对抗鲁棒性时，同时攻击分布内样本和分布外样本，以获得更全面的鲁棒性评估。

（2） 有额外数据集场景基于有效点选取的鲁棒分布外检测方法

针对在有额外数据集的场景下辅助分布外数据集没有得到较好利用的问题，本文提出了一种基于有效点选取的鲁棒分布外检测方法，通过选取有效点进行训练以获得更为紧凑的决策边界并减少无用信息的输入，以增强鲁棒分布外检测能力。该方法通过原型学习进行特征聚集投影获取类别原型，然后利用马氏距离筛选辅助分布外数据中的与类别原型具有高相似性的点进行训练，以减少无关信息的引入。为了获得更好的模型鲁棒性，该方法结合基于李雅普诺夫理论的神经微分方程模块，利用辅助分布外样本和分布内样本训练对抗去噪模块以同时对分布内和分布外产生对抗鲁棒性。

该方法的创新点在于：创新性地结合了原型学习和马氏距离用于辅助分布外数据的有效点选取，使得能够在特征层面根据分布外数据和分布内数据的相似性选取更有效的辅助分布外数据进行训练；创新性地将有效点选取运用于神经 ODE 去噪器模块，通过结合李雅普诺夫稳定性理论获得较强的鲁棒分布外检测能力。

（3） 无额外数据集场景下基于虚拟离群点生成的鲁棒分布外检测方法

针对在无额外数据集场景下生成分布外数据的方法欠佳的问题，本文提出了一种基于虚拟离群点生成的鲁棒分布外检测方法，通过对抗性传播和非参数采样生成不鲁棒区域的虚拟离群点作为辅助分布外样本进行训练，以获得无额外数据集场景下更好的鲁棒分布外检测能力。该方法通过对抗性传播将易受攻击的分布内数据驱使为边界点，然后在边界点周围进行非参数化的采样和筛选

获得虚拟离群点，以虚拟离群点进行训练填补了额外分布外数据的缺失。为了进一步增强更好的鲁棒性与分布外检测能力，该方法还将 logit 归一化结合至神经微分方程去噪器中，以削弱模型过度自信的问题。

该方法的创新点在于：创新性地通过对抗性传播和非参数化采样获取脆弱的虚拟离群值，使得模型更加关注易受攻击的区域并去除了分布假设的限制；创新性的将 logit 归一化操作与神经 ODE 去噪器结合并用于鲁棒分布外检测中，使得在保证鲁棒性的同时削弱过度自信问题。

1.4　论文组织结构

本文的章节安排及组织结构如图1.1所示。全文一共六个章节，每个章节的具体内容安排如下：

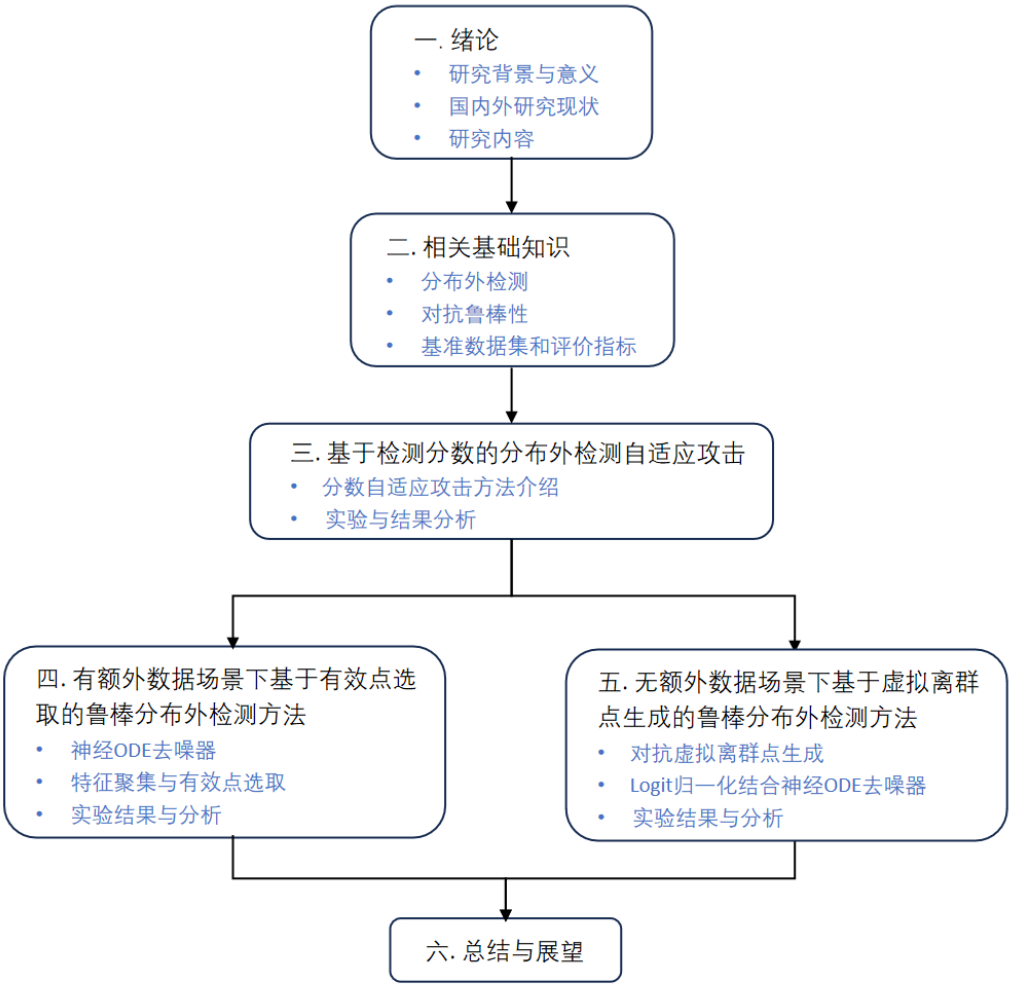


图 1.1　本文章节组织框架图

第一章为绪论。该章节主要介绍鲁棒分布外检测的研究背景与研究意义，并结合国内外研究现状对该领域的发展历程与现有状况进行了阐述。该章节最后

基于研究现状总结了研究问题，并给出了本文主要的研究内容。

第二章为相关知识及工作准备。该章节首先介绍分布外检测问题的描述及常见的分布外检测方法。其次该章节概述了和对抗鲁棒性相关的对抗攻击和对抗防御方法。最后该章节介绍了本文实验中所使用的基准数据集及评价指标为后续实验作准备。

第三章为基于检测分数的分布外检测自适应攻击方法。该章节描述了针对分布外检测的攻击方法的问题设定，详细介绍了所提出的分数自适应攻击方法，并通过实验分析自适应攻击对分布外检测的影响。该章节揭示了现有分布外检测算法的脆弱性，且规范了分布外检测鲁棒性的评估方法，为后续章节作铺垫。

第四章为有额外数据场景基于有效点选取的鲁棒分布外检测方法。该章节针对有额外辅助分布外数据的场景提出了算法设计并进行实验分析。算法设计中介绍了模型框架，并对框架中的神经 ODE 去噪器和有效点选取步骤进行详细地阐述。实验部分与多种基于额外数据的方法进行对比，并通过超参数分析和可视化进一步分析方法特点。

第五章为无额外数据场景基于虚拟离群点生成的鲁棒分布外检测方法。该章节针对与第四章不同的无额外辅助分布外数据场景，提出了算法设计并进行实验分析。算法设计中介绍了模型框架，并对框架中的对抗虚拟分布外离群点合成和结合 logit 归一化的神经 ODE 去噪器训练步骤进行详细地阐述。实验部分与多种不使用额外数据的方法进行对比，并通过超参数分析和可视化进一步分析方法特点。

第六章为总结和展望。该章节对本文研究工作进行了总结，并在此基础上展望了该研究领域可能的发展方向。

第 2 章　相关知识及工作准备

本章概述了本文研究所需的相关基础知识及研究工作准备。首先，介绍了分布外检测的相关定义及常见分布外检测技术的基本原理；其次，阐述了对抗攻击和对抗防御的通用方法；最后，为了对后续实验作准备，介绍了本文使用到的基准数据集及性能评价指标。

2.1　分布外检测

本节介绍分布外检测技术的一些基础知识。首先介绍分布外检测问题的相关定义，然后介绍不同类别下常见的分布外检测方法。

2.1.1　分布外检测相关定义

将模型训练所使用的数据集表示为 $\mathcal{D}_{in}$, 其中包含 N 个输入-标签对 (x_{in}, y_{in})，x_{in} 表示输入训练样本，$y_{in} \in Y_{id}$ 表示与之对应的标签，(x_{in}, y_{in}) 服从联合分布 $P_{X_{id},Y_{id}}$。对具有 K 个类别的分类任务而言 $Y_{id} := \{1, 2, \ldots, K\}$。在开放场景下，模型会遇到分布不服从于 $P_{X,Y_{id}}$ 的输入，即产生了分布外输入 (x_{ood}, y_{ood})。其中 $x_{ood} \sim P_{X_{id}}$ 表示协变量偏移，比如光照，图像风格等的变化。$y_{ood} \sim P_{Y_{id}}$ 表示语义偏移，即输入所属的类别不存在于训练集的 K 个类别中。分布外检测任务特指在语义偏移下，将所有标签属于 Y_{ood} 的样本检测出来，其中 $Y_{ood} \cap Y_{id} = \varnothing$。标签属于分布外标签集合 Y_{ood} 的样本称为分布外样本。

分布外检测依赖于某一特定的检测分数 $S(x)$，该分数由分布外检测器 $G(x)$ 产生，通常通过特殊的计算方式产生或者由训练模型获得。训练好分布外检测器后，对于每一个输入 x，分布外检测器 $G(x)$ 会产生与之关联的检测分数 $S(x)$，之后设置阈值 λ 作为判断是否是分布外样本的标准。常见的指标是将检测分数 $S(x)$ 高于阈值 λ 的样本判断为分布内样本 (ID)，将检测分数 $S(x)$ 低于阈值 λ 的样本判断为分布外样本 (OOD)。即：

$$G(x) = \begin{cases} ID, & \text{如果 } S(x) > \lambda \\ OOD, & \text{如果 } S(x) < \lambda \end{cases} \tag{2.1}$$

2.1.2　常见分布外检测方法

本小节介绍常见的分布外检测方法，按照不同的类别包括基于事后处理的方法、基于额外数据的方法和基于模型训练的方法。

1. **基于事后处理的分布外检测方法**

基于事后处理的分布外检测方法不需要重新训练模型，而是使用预训练模型的特征或输出构建分布外检测分数。基于事后处理的方法较为简单，即插即用，利用统计差异来进行分布外检测。常见的基于事后处理的分布外检测方法包括 MSP[4]、ODIN[5]、energy[11]和 MDS[12]。

MSP 即最大 softmax 概率，是分布外检测的基线方法。MSP 的核心是基于一个实验观察，即在输出概率最大的类别上，分布内样本的 softmax 值相比于分布外样本较高。但后来有研究表明 softmax 值会对分布外样本产生过度自信的现象，因此该方法有一定的局限性。其用公式表示如下：

$$S(x)=\max_k\ p(y=k|x)=\max_k\frac{\exp\left(f_k(x)\right)}{\sum_{j=1}^{K}\exp\left(f_j(x)\right)} \tag{2.2}$$

ODIN 对 MSP 进行修改，通过温度系数对输出概率进行缩放操作，缓解了过度自信问题。具体来说，通过将最后一层的输出除以一个较大温度缩放系数 T, 便可对 softmax 值产生平滑的效果，处理后的分数为：

$$S(x;T)=\max_k\ \frac{\exp\left(f_k(x)/T\right)}{\sum_{i=1}^{N}\exp\left(f_j(x)/T\right)} \tag{2.3}$$

除此之外，为了进一步增加分布外和分布内的分数差异 ODIN 还采用了对抗的思想，将输入向分数增加的方向驱使，并观察到通过这种操作会增加分布外检测分数的差异性，预处理方式为：

$$\tilde{x}=x-\varepsilon\text{sign}(-\nabla_{\boldsymbol{x}}\log S(\boldsymbol{x};\boldsymbol{T})) \tag{2.4}$$

其中 ε 为扰动大小，输入 x 往检测分数梯度上升的方向变化。

energy 即能量分数，使用了能量模型的理论，通过预训练的分类器计算能量分数对输出的不确定性进行估计。能量分数基于能量模型，计算简单，相比于 softmax 不易受到过度自信问题的影响，计算公式为：

$$E(x;f)=-T\cdot\log\sum_{i}^{K}e^{f_i(\mathbf{x})/T} \tag{2.5}$$

MDS 使用马氏距离区分分布外样本。该方法将输出空间假设为类条件高斯分布，因此只需要对各个类别的高斯分布建模，衡量输入到各个分布的距离即可得到最大的输出概率 $\max_k\ p(y=k|x)$。类条件高斯分布取决于均值和协方差矩阵，对于某一类别 c 计算方法为：

$$\widehat{\mu}_c=\frac{1}{N_c}\sum_{i:y_i=c}f(\mathbf{x}_i),\ \widehat{\Sigma}=\frac{1}{N}\sum_{c}\sum_{i:y_i=c}\left(f(\mathbf{x}_i)-\widehat{\mu}_c\right)\left(f(\mathbf{x}_i)-\widehat{\mu}_c\right)^{\top} \tag{2.6}$$

其中 N_c 是标签 c 的训练样本的数量。获得类条件高斯分布的表示以后，对于任意输入样本 x，可计算马氏距离分数：

$$M(\mathbf{x}) = \min_c \left(f(\mathbf{x}) - \hat{\mu}_c\right)^\top \hat{\Sigma}^{-1} \left(f(\mathbf{x}) - \hat{\mu}_c\right) \tag{2.7}$$

若 $M(x)$ 越大则代表与最相邻的高斯分布越接近，越应该被判断为分布内样本。MDS 不仅仅可以在最后一层进行计算，也可以计算中间任意一层的高斯分布然后利用马氏距离进行衡量，是一种在特征空间利用距离度量的方法。

2. **基于额外数据的分布外检测方法**

基于额外数据的方法不仅仅使用了分布内数据，还利用了一些额外的分布外数据的信息来辅助分布外检测。总的来说有两类，一类基于离群值暴露，直接使用额外的分布外数据集辅助模型进行正则化训练；一类直接利用经大量数据预训练的模型中所包含的分布外信息进行检测。常见的基于额外数据的方法有 OE[19]，MixOE[22] 和 MCM[23]。

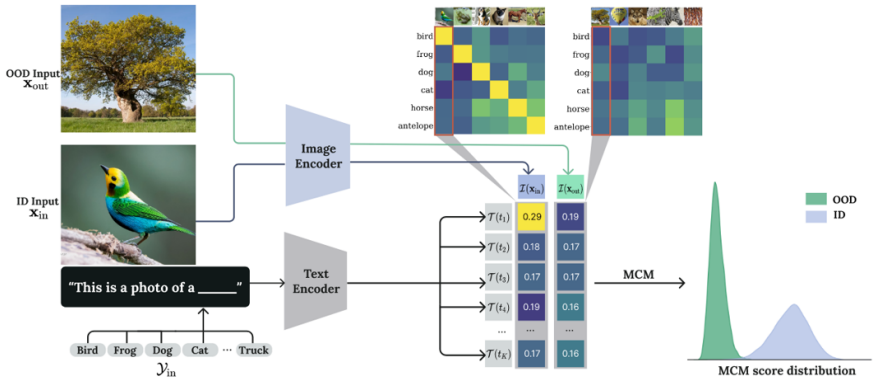


图 2.1　MCM 方法示意图[23]

OE 即离群值暴露 (Outlier Exposure)。该方法的基本思想是原有分类器对于分布外信息的缺失使得其产生过度自信的现象，因此通过收集一些额外的分布外数据辅助模型训练即可使模型泛化到一般的分布外数据上。OE 将收集到的额外分布外数据集与与均匀分布的 KL 散度作为正则化项以减轻分布外的过度自信问题，其训练目标为最小化以下损失：

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}}[-\log f_y(x)] + \lambda \mathbb{E}_{x\sim\mathcal{D}_{\text{out}}^{\text{OE}}}[H(U; f(x))] \tag{2.8}$$

其中 $\mathcal{D}_{\text{out}}^{\text{OE}}$ 为额外收集的分布外数据集，H 表示交叉熵,U 表示均匀分布。

MixOE 为 OE 的改进版，通过在数据集上进行混合操作产生更多样的离群值，即用于训练的离群值表示为：

$$\tilde{x} = \min(x_{\text{in}}, x_{\text{out}}, \lambda) \tag{2.9}$$

其中 $\lambda \in [0,1]$ 用于控制每个样本对混合操作的贡献，分布内样本的系数越大表示混合产生的离群值越接近分布内。混合操作包括 Mixup 和 CutMix 等数据增强方法。

MCM 利用预训练的多模态模型 clip 中的分布外知识来进行分布外检测。通过将输入图像编码后与各个类别的文本编码进行匹配，把 CLIP 模型转换为图像分类器，以最大概念匹配分数作为区分分布外的指标。其方法示意图如图2.1。

3. **基于模型训练的分布外检测方法**

基于模型训练的方法不使用额外的分布外数据，通过设计特殊的训练策略以获得对分布外检测有益的模型。一般有两类方法，一类利用无监督学习等策略获得具有区分分布内和分布外样本的特征表示空间；一类通过合成虚拟的离群值代表分布外样本。常见的额基于模型训练的方法包括 CSI[27]、CIDER[28]、VOS[34]和 NOPS[35]。

CSI 将对比学习的思想结合到分布外检测中。利用对比学习可以使得模型获得更有区分度的特征表示，也更有利于分布外检测。与常规的对比学习不同，CSI 将数据增强后的样本退离原始样本，并观察到这样可以对分类器进行置信度校准。

CIDER 也利用到对比学习的思想。不同的是，CIDER 不是简单的将增强样本推离原始样本，而是将特征映射到一个超球面空间上，在该空间上计算各个类别的类别原型。为了使得各个类别汇聚在类别原型周围，CIDER 设计了类内聚集损失：

$$\mathcal{L}_{\text{comp}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathbf{z}_i^{\top}\boldsymbol{\mu}_{c(i)}/\tau\right)}{\sum_{j=1}^{C}\exp\left(\mathbf{z}_i^{\top}\boldsymbol{\mu}_j/\tau\right)} \tag{2.10}$$

其中 z_i 为超球面空间的嵌入表示，μ_c 表示类别 c 的类别原型，τ 为用于缩放的温度系数。该损失可以将不同类别的样本远离类别原型，将同一类别的样本汇聚到类别原型周围。此外，为了使各个类别尽量分散，CIDER 还采用了类间离散损失：

$$\mathcal{L}_{\text{dis}} = \frac{1}{C}\sum_{i=1}^{C}\log\frac{1}{C-1}\sum_{j=1}^{C}\mathbb{1}\{j\neq i\}e^{\boldsymbol{\mu}_i^{\top}\boldsymbol{\mu}_j/\tau} \tag{2.11}$$

即让不同类别的类别原型尽量远离。经过类内聚集和类间离散可以获得具有区分度的特征空间，在特征空间中计算距离指标即可区分分布外样本。

VOS 是一种利用分布内样本合成虚拟分布外样本的方法。VOS 将特征空间假设为类条件高斯分布，即：

$$p_\theta(h(\mathbf{x},\mathbf{b})|y=k) = \mathcal{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}) \tag{2.12}$$

其中 μ_c 表示类别 c 的高斯分布均值，Σ 为协方差矩阵。通过在类条件高斯分布

的低似然区域采样，即可将采样点作为虚拟离群值。采样集合表示为：

$$\mathcal{V}_k = \{\mathbf{v}_k | \frac{1}{(2\pi)^{m/2}|\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{v}_k - \hat{\boldsymbol{\mu}}_k)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{v}_k - \hat{\boldsymbol{\mu}}_k)\right) < \epsilon\} \tag{2.13}$$

其中 v_k 为被选中的采样点,ϵ 表示采样阈值。

2.2　对抗攻击和对抗防御

本节介绍对抗鲁棒性的相关知识，包括对抗攻击和对抗防御。首先介绍对抗攻击的相关概念及常见的对抗攻击方法原理，然后从防御者的角度介绍常见的对抗防御方法。

2.2.1　对抗攻击

基于深度学习的模型在近年越来越多的被运用到各种安全关键应用上。自 Szegedy 等人发现对抗样本以来，对抗攻击逐渐受到关注，并引发了人们对安全关键应用中使用的深度学习安全性产生了担忧。

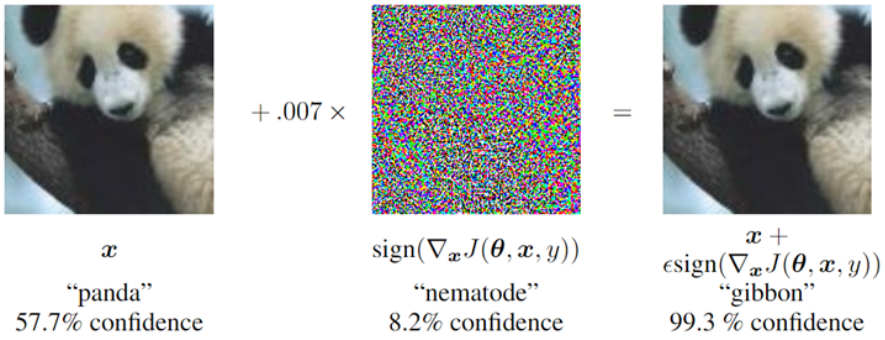


图 2.2　对抗样本示意图[37]

直观上来说，对抗攻击即原始输入被添加人为不可见的扰动后，会导致模型的决策变为错误决策。通常将原始输入称为干净样本，将扰动称为对抗扰动，将添加扰动后的输入称为对抗样本。如图2.2所示，左边为干净样本，中间为对抗扰动，右边为对抗样本。可以观察到在人眼看来干净样本和对抗样本的语义并没有差别，而分类器却会以高置信度错误判断对抗样本。

对抗攻击较为形式化的数学定义如下：

$$\begin{aligned} &\min_{\delta} \ \mathcal{L}(f(x_{adv}), y) \\ &\text{s.t. } \|\delta\| \leqslant \epsilon, \ x_{adv} = x + \delta \end{aligned} \tag{2.14}$$

其中 δ 表示对抗扰动，为了满足不修改人眼所见的语义信息，扰动大小被限制于 δ 内。$\mathcal{L}$ 表示损失函数，用于表示模型对于对抗样本的输出 $f(x_{adv})$ 与真实标签之间的差异。

对抗攻击的核心就在于如何找到满足要求的扰动 δ，施加于干净样本上获得对抗样本 x_{adv}。目前有许许多多的对抗攻击算法，本文主要介绍一些常见的具有代表性的算法，包括 FGSM[37]、BIM[38]、PGD[39]和 CW[41]。

FGSM 快速符号梯度法，由 Goodfellow 等人提出的一种基于梯度的一步攻击方法。该方法将干净样本往损失函数的符号梯度上升的方向进行扰动，从而得到使损失函数增加的对抗样本。该方法计算简单，能够较快的生成对抗样本，但由于采用单步攻击因此攻击强度较弱。FGSM 用公式表达如下：

$$x_{adv} = x + \varepsilon \cdot sign(\nabla_x \mathcal{L}(f(x), y)) \tag{2.15}$$

其中 ε 用于控制扰动大小，采用符号梯度有利于快速计算。

BIM 在 FGSM 的基础上使用多次迭代，每次添加一个小的扰动知道达到预设的最大扰动。BIM 采用的小扰动多步迭代策略使得其寻找对抗样本的过程更加精细，牺牲计算时间以获得更强大的对抗样本。每次迭代中，BIM 为了防止扰动超过最大扰动采用剪切函数对超过扰动的样本进行剪切。迭代公式如下：

$$x_{adv}^{t+1} = \text{Clip}\left\{ x_{adv}^{t} + \alpha \cdot sign(\nabla_{x_{adv}^{t}} \mathcal{L}(f(x_{adv}^{t}), y)) \right\} \tag{2.16}$$

PGD 也采用多步扰动的策略，并使用投影来保证扰动不越界。PGD 是一种公认有效的对抗攻击生成方法，其产生的对抗样本强力，被广泛用于对抗训练的内部最大化求解上。PGD 与 BIM 不同的点主要在于采用更为合理的投影操作来控制扰动，其迭代公式如下：

$$x_{adv}^{t+1} = \text{Proj}\left\{ x_{adv}^{t} + \alpha \cdot sign(\nabla_{x_{adv}^{t}} \mathcal{L}(f(x_{adv}^{t}), y)) \right\} \tag{2.17}$$

CW 是一种基于优化的攻击方法，将对抗样本的生成转化为求解一个对抗样本为变量，损失函数为优化目标的优化问题。用公式表达为：

$$\begin{aligned} &\min_{\delta} \ \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \\ &\text{s.t. } x + \delta \in [0,1]^n \end{aligned} \tag{2.18}$$

2.2.2　对抗防御

对抗攻击的目标是欺骗模型，对抗防御的目标则是防止模型被对抗样本所欺骗。对抗防御对于模型的鲁棒性和安全性至关重要。到目前为止，对抗防御尚未发展完善，但也已经诞生了许多具有指导意义的对抗防御方法。对抗防御方法有基于检测的防御、基于对抗训练的防御和基于对抗净化的防御。本章主要介绍最常用的对抗训练方法和本文所涉及的基于对抗净化的方法。

1. 对抗训练方法

对抗训练方法是一种有效的缓解对抗攻击的方法，其基本思想是将对抗样本加入到模型的训练过程中，与干净样本混合或者仅使用对抗样本优化网络参数，从而起到对抗性正则化的效果，使网络学习到更为鲁棒的特征。形式上，对抗训练可看作一个最小最大问题的求解，用公式表示如下：

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta} \mathcal{L}(x+\delta, y, \theta)] \tag{2.19}$$

其中，内部最大化问题的求解目标是寻找使损失函数最大化的对抗扰动，其最优解无法直接获得，通常使用上文提到的 PGD、FGSM 等对抗样本生成算法来近似获得。外部最小化求解则类似于一般的模型训练，采用反向传播和随机梯度下降算法更新网络参数。对抗训练具有良好的抗对抗攻击效果，但计算复杂度较高，训练时间较长。对抗训练的效果依赖于内部最大化问题的求解，而现有的能够近似最优解的对抗样本生成方法需要多步的梯度下降迭代，因此对抗训练的时间通常是常规训练的几十倍。

此外，有研究观察到经过对抗训练的模型虽然具有良好的对抗鲁棒性，但在干净样本上的准确率却下降了。因此也有一些研究试图平衡对抗鲁棒性和干净样本的准确率。比如 TRADES[66]在对抗训练的框架下进行了修改，优化以下的损失函数：

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim D}[L(f(x,\theta), y)] + \lambda \cdot \mathbb{E}_{x\sim D, x'\sim D'}[D(f(x,\theta), f(x',\theta))] \tag{2.20}$$

其中第一项是标准的监督学习损失，通常使用交叉熵损失等，用于衡量模型在原始数据分布上的准确性。第二项是差异损失，用于对其干净样本和对抗样本的输出，D 是一个衡量分布差异的函数，如 KL 散度等。λ 是一个调节参数，用于调节干净损失和对抗损失的权重比例。因此，通过设置合适的调节参数，可在干净样本准确性和对抗样本鲁棒性之间进行平衡。

2. 对抗净化方法

除了对抗训练外，一些使用对抗净化的方法也受到关注。对抗净化方法的思想是通过消除输入中的对抗扰动从而将其转化为干净样本用于模型预测。对抗净化方法常见的有通过随机化操作修改图像使得对抗扰动失效的方法和训练一个去噪器对扰动进行去噪的方法。

利用随机化策略对图像进行预处理，可修改图像的像素信息从而消除对抗扰动的有效性。Xie 等人[67]通过随机调整输入样本的大小来减轻对抗扰动的影响，并对调整后的样本进行填充，将图像经过处理后再交给分类器，通过这样的策略可降低对抗样本的攻击性。另一些方法训练对抗扰动去噪器消除输入的对抗扰动，如 Defense-GAN[57]，利用 GAN 模型对干净图像建模，之后利用生成器

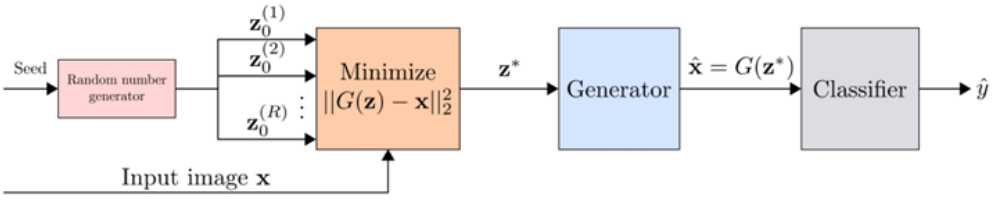


图 2.3　Defense-GAN 示意图[57]

净化图像 $\hat{x}$ 再交由分类器判别，其算法框架如2.3图所示。与使用生成模型在图像层面进行去噪不同，也有一些方法在特征层面进行对抗净化。如 HGD[58]将损失函数定义为干净图像和去噪图像激活的目标模型输出之间的差值：

$$L = ||f_l(\hat{x}) - f_l(x)|| \tag{2.21}$$

通过这种方式通过这种方式可以将干净样本和对抗样本在特征空间中的距离缩小，从而起到特征层面的去噪效果。还有一些方法利用神经微分方程的理论，如 TisODE[59]、ASODE[60]等，这些方法利用神经微分方程模块作为中间层进行去噪，使输入输出在模型的平衡点附近，其框架如图2.4所示。

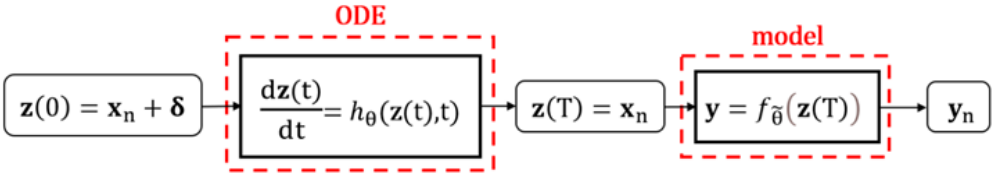


图 2.4　ASODE 示意图[60]

2.3　数据集与评价指标

2.3.1　基准数据集介绍

本节介绍后续章节中使用到的基准数据集，包括 CIFAR10、CIFAR100[68]、Tiny-images[69]、SVHN[70]、LSUN[71]、iSUN[72]、Texture[73]和 Places365[74]。

（1） CIFAR10 和 CIFAR100

CIFAR10 和 CIFAR100 数据集是图像分类任务中常用的基准数据集，在分布外检测中常作为分布内数据集。数据集示例如图2.5所示。

CIFAR10 与 CIFAR100 数据集都包含 60000 张图像，其中 50000 张作为训练集，其余 10000 张为测试集，每张图像都为 32×32 像素，具有 RGB 三个通道。其中 CIFAR10 一共有 10 个不同类别，每个类别 6000 张图像。CIFAR100 具有 100 个不同类别，每个类别有 600 张图像，提供了更复杂更细粒度的分类挑战。

图 2.5　CIFAR 数据集示例图

（2） Tiny-images

Tiny-images 全称 80 Million Tiny Images，在分布外检测中常作为辅助分布外数据集或者作为 CIFAR 数据集的近-分布外数据集。数据集示例如图2.6所示。

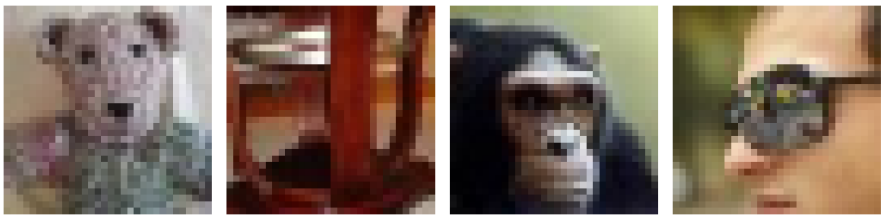


图 2.6　Tiny-images 数据集示例图

Tiny-Images 包含 79302017 张 32×32 像素的彩色图像，这些图像是从万维网上提取并按比例缩小获得。Tiny-Images 数据集的标注采用 WordNet 词汇数据库进行松散标注，标注信息较为宽泛和粗糙，存在一定的标注噪声，所以仅适合用作分布外数据。

（3） SVHN

SVHN 是一个用于图像识别和数字分类的大型数据集，在分布外检测中常作为分布外数据集。SVHN 由来自谷歌街景图像中的门牌号数字组成，包含超过 60 万张带有数字标注的图像，图像中的数字以自然场景中的形式呈现，具有复杂的背景和多种字体。数据集示例如图2.7所示。

图 2.7　SVHN 数据集示例图

（4） LSUN

LSUN 是为大规模场景分类和理解所设置的数据集，它包含有 10 个场景类别，例如餐厅、卧室、会议室和户外教堂等，每个类别样本有 12 万至 300 万左右。LSUN 数据集分辨率率不固定，在分布外检测中通常将其裁剪至与 CIFAR 数据集一致的分辨率，裁剪后的示例如图2.8。

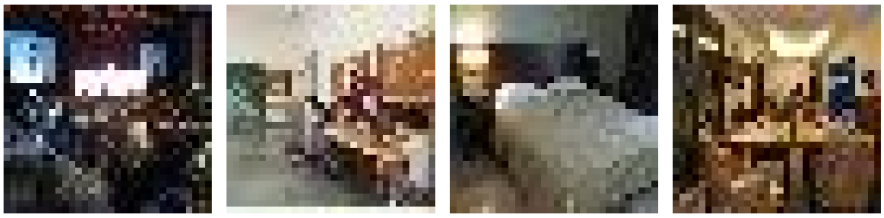


图 2.8　LSUN 数据集示例图

（5） iSUN

iSUN 是一种用于场景理解的视线追踪数据集，通过摄像头收集测试者注视点从而用视线追踪标签标记，一共有 8925 张图像。数据集示例如图2.9。

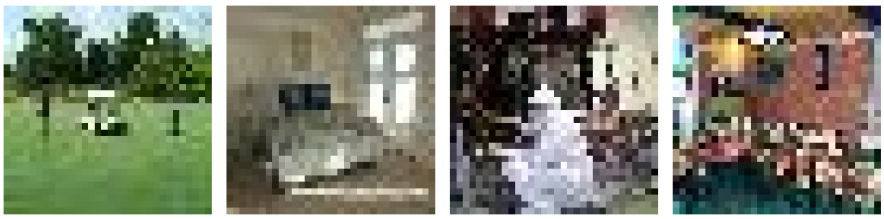


图 2.9　iSUN 数据集示例图

（6） Texture

Texture 是一个纹理图像集合，包含 5640 张图像，按照受人类感知启发分为 47 个类别，每个类别有 120 张图像。图像尺寸在 300×300 到 640×640 之间，且至少 90% 的区域展示了类别属性对应的纹理。数据集示例如图2.10。

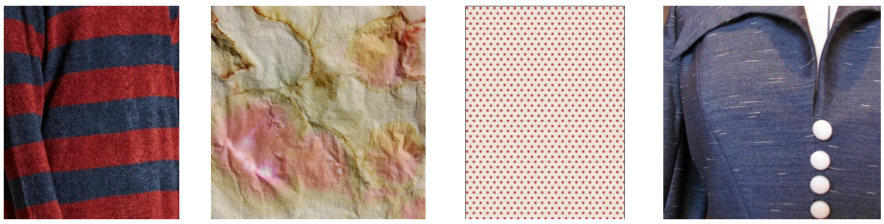


图 2.10　Texture 数据集示例图

（7） Places365

Places365 是一个专注于场景理解的大规模图像数据集，它包含了超过 1000 万张高分辨率图像，涵盖了 365 个不同的场景类别，具有丰富的多样性和真实场景的复杂性，具有详细的标注信息。数据集示例如图2.11。

图 2.11　Places365 数据集示例图

2.3.2　评价指标

分布外检测器将每一个输入判别为 ID 和 OOD 两个类，因此可采用二分类中的指标作为模型性能的判断标准。本文采用分布外检测中两个广泛使用的评价指标 FPR95(95% 真阳性率下的假阳性率) 和 AUROC(受试者工作特征曲线下面积)。

在二分类任务中，检测结果可分为四类：真正例 (TP)、假正例 (FP)、真负例 (TN)、假负例 (FN)。真正例率表示所有正例中被正确预测为正的比例，公式表示为：

$$TPR = \frac{TP}{TP + FN} \tag{2.22}$$

假正例率表示所有负例中被错误预测为正例的比例，公式表示为：

$$FPR = \frac{FP}{FP + TN} \tag{2.23}$$

FPR95 表示在真正例率为 95% 时的假正例率，即取 $TPR = 95\%$ 时 FPR 的值作为指标。该项指标越低表示分布外检测性能越好。

AUROC 表示 ROC 曲线的积分面积，是一个阈值无关的度量。ROC 曲线描述了 TPR 和 FPR 之间的关系，其中 FPR 作为横坐标，TPR 作为纵坐标，由不同的阈值设定下绘制所得。因此，AUROC 的定义为：

$$AUROC = \int_0^1 TPR(FPR)\, dFPR \tag{2.24}$$

可以将 AUROC 解释为正例被分配比负例更高的检测分数的概率，其取值于 0 到 1 之间，低于 0.5 表示该检测器性能不如随机猜测。在分布外检测中，越高的 AUROC 代表越强的检测性能。

2.4　本章小结

本章介绍了和鲁棒分布外检测相关的基础知识和工作准备。首先，简要介绍了分布外检测，包括分布外检测的问题定义和三种常见的分布外检测方法；其次，介绍了鲁棒性相关的内容，包括对抗攻击和对抗防御，给出了常见的对抗样本生成方法和两种不同原理的对抗防御方法；最后，介绍了常用的基准数据集和评价指标。

第 3 章　基于检测分数的分布外检测自适应攻击方法

本章主要研究对抗攻击对现有分布外检测方法所产生的威胁，揭示分布外检测方法的脆弱性，为分布外检测的鲁棒性评估提供方法。为了针对分布外检测器产生更为强力的对抗样本用于评测分布外检测算法的鲁棒性，本章提出了基于检测分数的自适应攻击方法，并研究了扰度幅度、攻击步数和迭代步长对攻击强度的影响。

3.1　引言

分布外检测任务在开放场景下将分布内样本和分布外样本进行区分，可将后续模型不能处理的分布外样本提前筛选出来，对系统的安全性至关重要。然而，常规的分布外检测算法只研究干净样本下分布外检测器区分分布外样本的能力，或者仅考虑一些常规的协变量偏移 (如光照、风格变化等)，并没有考虑到对抗样本的存在。若可制作出能够使分布外检测器的判断出现大量错误的对抗样本，则会对开放场景下的分布外检测模型造成巨大威胁。若攻击者对输入施加恶意的对抗噪声导致分布外检测器的判断出现大量错误，会导致大量的分布内样本被拒绝而大量的分布外样本被错误的输入给后续模型，严重影响系统的安全性。

在对抗鲁棒性领域常规的对抗样本都是针对分类器产生，即攻击者通过对输入施加对抗噪声产生对抗样本使得分类器将输入归类到错误的类别。而针对于分布外检测器的对抗样本的攻击目标则不同，其目的是使得分布外检测器将分布内输入识别为分布外或者将分布外输入识别为分布内。在这种情况下，有两种对抗样本的存在：对抗分布内样本和对抗分布外样本。对抗分布内样本的语义是分布内，攻击者施加噪声后使得分布外检测器将其检测为分布外，从而将该输入错误的拒绝。对抗分布外样本的语义是分布外，攻击者施加噪声后使得分布外检测器将其检测为分布内，从而将该输入继续传入后续模型，而分布外输入是后续模型不能处理的，从而引发安全问题。

已有一部分研究[8-9,62-63]探索分布外检测中对抗样本的影响，并利用对抗样本评测分布外检测器的鲁棒性，但这些方法所产生的对抗样本不够全面和强力，使得分布外检测器的鲁棒性没有得到较好的评估。这些方法有的没有同时考虑对抗分布内样本和对抗分布外样本的存在，有的在生成对抗样本时不是使用分布外评分函数而是使用分类损失，有的仅在较小的扰动下评估分布外检测器的鲁棒性。这些缺失都使得分布外检测器的鲁棒性没有得到较好的评估。

针对上述问题，本章提出了基于检测分数的分布外检测自适应攻击方法，用于产生针对分布外检测器更强力的对抗样本以进行分布外检测鲁棒性评估。具体来说，分布外检测器在判别时依赖于分布外检测分数，通常对分布外输入给予较低的检测分数，对分布内输入给予较高的检测分数。为了使分布外检测器判别出现错误，应该基于检测分数通过施加扰动使输入的检测分数向错误的方向变化。因为在分布外检测中有多种检测分数，所以在产生扰动要自适应的使用不同的损失。此外，为了使得攻击更加强力，在不修改图像语义信息的前提下使用较大的扰动幅度，并通过多步迭代进行求解。

3.2　问题描述

模型训练所使用的分布内数据集表示为 $\mathcal{D}_{in} = \{(x_{in}, y_{in}) \mid i = 1, 2, \cdots, N\}$，其中 (x_{in}, y_{in}) 服从联合分布 $P_{X_{id}, Y_{id}}$。在开放场景下，模型会遇到分布外输入 (x_{ood}, y_{ood})，其中 $y_{ood} \sim P_{Y_{ood}}$ 且 $Y_{ood} \cap Y_{id} = \varnothing$，即产生了新的输入类别。分布外检测任务需要分布外检测器 $G(x)$ 对每一个输入 x 产生与之对应的分布外检测分数 $S(x)$，再根据检测分数设置阈值 λ 判断输入是否是分布外样本。用公式表示如下：

$$G(x) = \begin{cases} ID, & \text{如果 } S(x) > \lambda \\ OOD, & \text{如果 } S(x) < \lambda \end{cases} \tag{3.1}$$

即将检测分数 $S(x)$ 高于阈值 λ 的样本判断为分布内样本 (ID)，将检测分数 $S(x)$ 低于阈值 λ 的样本判断为分布外样本 (OOD)。

由于输入有分布内样本 x_{in} 和分布外样本 x_{ood} 两种形式，因此针对分布外检测器的对抗样本也有两种类别，即对抗分布内样本和对抗分布外样本。对抗分布内样本 x_{adv}^{id} 由分布内样本产生，其目的是在增加扰动后使得分布外检测器将其判别为分布外，添加的扰动称为对抗扰动。对于对抗分布内样本，其对抗扰动由以下公式产生：

$$\begin{aligned} &\min_{\delta} \mathcal{L}\left(G(x_{adv}^{id}), OOD\right) \\ &\text{s.t. } x_{adv}^{id} = x^{id} + \delta,\ \|\delta\| \leqslant \epsilon \end{aligned} \tag{3.2}$$

其中 x^{id} 表示分布内输入，δ 表示对抗扰动，其大小被 ϵ 约束以保证人眼所见的语义信息不发生改变，经扰动后分布外检测器更容易将其判别为分布外样本。与此类似，对抗分布外样本 x_{adv}^{ood} 目的是在增加扰动后使得分布外检测器将其判别为分布内，其对抗扰动通过以下公式产生：

$$\begin{aligned} &\min_{\delta} \mathcal{L}\left(G(x_{adv}^{ood}), ID\right) \\ &\text{s.t. } x_{adv}^{ood} = x^{ood} + \delta,\ \|\delta\| \leqslant \epsilon \end{aligned} \tag{3.3}$$

通常而言，分布外检测器检测到分布内输入后，会直接将其交给后续模型处理，而检测到分布外输入后则触发警报并交由其他机制处理 (如人工)。所以，对抗分布内样本会被错误的识别为分布外，从而引发大量的报警影响系统正常运行；而对抗分布外样本则会被识别为分布内，躲避了分布外检测器，在交由后续模型处理时引发安全问题。因此，在评估分布外检测器性能时，除了评估其对于干净样本的检测能力外，还需要制作对抗分布内样本和对抗分布外样本以评估其鲁棒性。而制作对抗样本的关键在于式3.2和式3.3中的损失 $\mathcal{L}$ 如何选择，以及通过怎样的方式获得最佳对抗扰动 δ^*。

3.3　算法设计

在针对于分类器的对抗样本的制作中，非目标攻击的攻击目的是使样本 x 偏离其真实标签 y，对于使用 softmax 概率的网络而言，对抗扰动的优化问题为：

$$\begin{aligned} &\min_{\delta} \frac{\exp\left(f_y(x_{adv})\right)}{\sum_{j=1}^{K}\exp\left(f_j(x_{adv})\right)} \\ &\text{s.t. } x_{adv}=x+\delta,\ \|\delta\|\leqslant\epsilon \end{aligned} \tag{3.4}$$

其中 $f_j(x)$ 为网络模型输出向量在维度 j 上的输出值。通过最小化以上问题，样本 x 的输出将偏离其真实标签。目标攻击的目的是使样本 x 的标签偏离至另一错误标签 y'，优化目标为：

$$\begin{aligned} &\min_{\delta} -\frac{\exp\left(f_{y'}(x_{adv})\right)}{\sum_{j=1}^{K}\exp\left(f_j(x_{adv})\right)} \\ &\text{s.t. } x_{adv}=x+\delta,\ \|\delta\|\leqslant\epsilon \end{aligned} \tag{3.5}$$

而针对分布外检测的攻击目标则与针对分类器的攻击目标不同。对于一个使用检测分数 $S(x)$ 的分布外检测器，为了使分布外检测器的判断产生错误，使输入的标签偏移与攻击的目标不一致，应该使其检测分数向错误的方向变化，如图3.1所示。具体来说，针对分布内样本 x^{id}, 其对抗扰动的获得应该是优化以下问题：

$$\begin{aligned} &\min_{\delta}\ S(x_{adv}^{id}) \\ &\text{s.t. } x_{adv}^{id}=x^{id}+\delta,\ \|\delta\|\leqslant\epsilon \end{aligned} \tag{3.6}$$

其中，通过最小化问题的求解获得对抗扰动 δ^* 后，将其添加到输入中后将会使检测分数 $S(x)$ 减小，一旦检测分数小于检测阈值 λ 后分布外检测器便会将其错误判断为分布外样本。与之对应，对抗分布外样本的优化目标为：

$$\begin{aligned} &\min_{\delta}\ -S(x_{adv}^{ood}) \\ &\text{s.t. } x_{adv}^{ood}=x^{ood}+\delta,\ \|\delta\|\leqslant\epsilon \end{aligned} \tag{3.7}$$

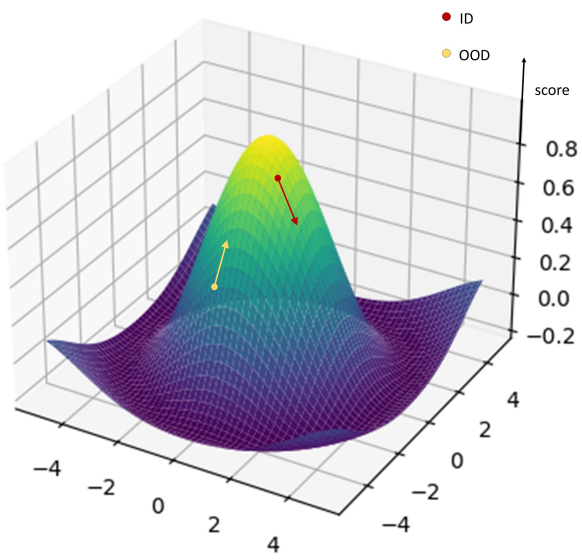


图 3.1　分数自适应攻击示意图

其中攻击目标是使使检测分数 $S(x)$ 增大，一旦检测分数大于检测阈值 λ 后分布外检测器便会将其错误判断为分布内样本。在下面的讨论中，只展示对抗分布内样本的公式，对抗分布外样本将优化目标变为相反数即可。

对于优化问题的求解，公认有效的方法是使用 PGD 攻击进行多次迭代。对于对抗分布内样本，其迭代公式为：

$$x_{adv}^{t+1} = \text{Proj}\left\{ x_{adv}^{t} + \alpha \cdot \text{sign}(\nabla_{x_{adv}^{t}} S(x_{adv}^{t})) \right\} \tag{3.8}$$

其中 t 表示攻击步数，α 表示迭代步长，Proj 表示投影操作用来保证迭代过程中满足扰动幅度约束。

分布外检测器可以使用多种检测分数，根据式3.6和3.7，为了形成有效的攻击需要根据不同的检测分数自适应的调整，且在实际制作对抗样本过程中，可以采用批量化的方法进行加速。如在使用 MSP 分数时，分布内对抗样本的优化目标为：

$$\min_{\delta} \sum_{i=1}^{M} \max_{k} \frac{\exp\left(f_k(x_{adv}^{i})\right)}{\sum_{j=1}^{K} \exp\left(f_j(x_{adv}^{i})\right)} \tag{3.9}$$

使用能量分数时，优化问题为：

$$\min_{\delta} \sum_{i=1}^{M} -T \cdot \log\left(\Sigma_{j=1}^{K} e^{f_j(\mathbf{x}_{adv}^{i})/T} \right) \tag{3.10}$$

使用马氏距离时，优化问题为：

$$\min_{\delta} \sum_{i=1}^{M} \min_{c} \left(f(\mathbf{x}_{adv}^{i}) - \hat{\mu}_c\right)^{\top} \hat{\Sigma}^{-1} \left(f(\mathbf{x}_{adv}^{i}) - \hat{\mu}_c\right) \tag{3.11}$$

综上，为了针对分布外检测器产生强力的对抗样本以评估其分布外检测的鲁棒性，需要同时考虑对抗分布内样本和对抗分布外样本，并根据分布外检测器的检测分数自适应的调整优化目标才可达到攻击效果。在求解优化问题时使用 PGD 多次迭代，并利用批次求解增加效率，具体流程如下:

算法 3.1　基于检测分数的自适应攻击算法

Input: 分布外检测器 $G(\cdot)$, 检测分数 $S(\cdot)$, 扰动边界 ϵ, 攻击步数 t, 迭代步长 α, 批次大小 b, 数据集 $D_{id} = \{(x_i^{id}, y_i^{id})\}$,$D_{ood} = \{(x_i^{ood}, y_i^{ood})\}$, $i = 1, \cdots, N$

Result: 对抗样本 x_{adv}^{id},x_{adv}^{ood}

for $n \leftarrow 1$ *to* $\lceil \frac{N}{b} \rceil$ **do**
  for $i \leftarrow 1$ *to* b **do**
    if x_i 为分布内数据 **then**
      $\mathcal{L}(x_i) = S(x_i)$
    else
      $\mathcal{L}(x_i) = -S(x_i)$
    end
  end
  for $j \leftarrow 1$ *to* t **do**
    $X_{batch}^{t} = \text{Proj}\left\{ X_{batch}^{t-1} + \alpha \cdot \text{sign}(\nabla_{X_{batch}^{t-1}} \mathcal{L}(X_{batch}^{t-1})) \right\}$
  end
end

3.4　实验与分析

本节设计实验验证本章所提方法的有效性，并以此揭示不考虑对抗样本的分布外检测方法的脆弱性。除此之外，本节还通过实验分析扰动边界、攻击步数和迭代步长的影响，并通过可视化分析展现自适应对抗攻击的影响。

3.4.1　实验环境及设置

本文中所有实验的环境均如下表：

表 3.1　实验环境

软硬件环境	版本型号
操作系统	Ubuntu 20.04.6 LTS
CPU	Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz
GPU	NVIDIA GeForce RTX 3090
开发语言	Python 3.10.11
开发框架	PyTorch 2.0.0

实验中使用的数据集包括：CIFAR10、CIFAR100、Tiny-images、mnist、SVHN、Texture 和 Places365。CIFAR10 作为分布内数据集。CIFAR100 与 Tiny-images 与

CIFAR10 数据集的相似度较大，作为近-分布外数据集。mnist、SVHN、Texture 和 Places365 由于与 CIFAR10 数据集相似度较小，将其设定为远-分布外数据集。性能指标采用分布外检测中常用的 AUROC 和 FPR95。

实验中使用常见的分布外检测方法用于验证对抗样本的强力性。所有方法模型来源于分布外检测方法库 OpenOOD[75]，并直接利用该库的默认配置参数训练获得分布外检测模型。OpenOOD 包含三类常见的分布外检测方法：基于事后处理的方法、基于离群值暴露的方法和基于模型训练的方法。本章从三个类别中各选取具有代表性的方法用于验证。其中选择基于事后处理的方法 MSP、Energy、MDS，选择基于离群值暴露的方法 OE，选择基于模型训练的方法 VOS。有关方法的介绍见2.1.2小节。

3.4.2　分数自适应攻击有效性验证实验

本小节利用提出的基于检测分数的自适应攻击方法攻击常见的分布外检测模型，以验证所提攻击方法的有效性并揭露这些分布外检测方法在面对对抗攻击时的脆弱性。自适应攻击的参数包含扰动边界 ϵ，攻击步数 t 和迭代步长 α。本小节实验将参数设置为 $\epsilon = 4/255$，$t = 5$，$\alpha = 2/255$。

本小节实验分别在远-分布外数据集和近-分布外数据集上，计算仅攻击分布内样本、仅攻击分布外样本和同时攻击分布内和分布外样本的平均 AUROC 指标，以展示不同对抗样本对分布外检测性能的影响。此外，还在同时攻击分布内和分布外样本的情况下测试了各个数据集攻击前后的 AUROC 和 FPR95 指标，以详细展现对抗攻击对分布外检测算法的影响。

（1） 远-分布外数据集实验

远-分布外数据集包含 mnist、SVHN、Texture 和 Places365，这些数据集与 CIFAR10 数据集相似度较小。实验分别在无攻击、仅攻击分布内样本、仅攻击分布外样本和同时攻击分布内和分布外样本的情况下，对比四个数据集测试结果 AUROC 指标的均值的百分比，该值越小表示分布外检测性能越差。具体实验结果如表3.2所示。

分析表中数据可知，无攻击情况下这些方法的分布外检测性能都较为良好，但在具有攻击的情况下性能都出现了较大幅度的下降。其中，在同时攻击分布内外样本时性能下降幅度最大的方法为 MDS，性能下降 89.76%；性能下降幅度最小的方法为 MSP，性能下降 70.25%。总体上来说攻击分布内的性能下降幅度比攻击分布外的下降幅度大，但对于 MSP 有所不同，因此不能简单的得出对于分布外样本的鲁棒性要比分布内样本的要强的结论。MDS 在同时攻击分布内和分布外时性能为 5 种方法中最差，这可能是由于 MDS 在特征层面计算，特征的维度较高所以对抗鲁棒性较差。表中还可观察到，同时攻击分布内外时性能下降幅

表 3.2　远-分布外数据实验

检测方法	无攻击	攻击分布内	攻击分布外	攻击分布内外
MSP	91.00	61.81	58.97	27.07
Energy	91.74	38.45	56.02	15.26
MDS	87.52	29.39	61.63	8.96
OE	98.13	50.05	88.21	20.99
VOS	90.15	34.80	64.02	18.38

度最大，说明在检验分布外模型的鲁棒性时不能单一的考虑对抗分布内样本或对抗分布外样本，需要同时攻击分布内外才能对模型的鲁棒性有更全面的评估。

（2） 近-分布外数据集实验

近分布外数据集包含 CIFAR100 和 Tiny-Images，这些数据集与 CIFAR10 数据集相似度较大，在近-分布内的情况下进行与（1）中相同的实验，具体实验结果如表所示：

表 3.3　近-分布外数据实验

检测方法	无攻击	攻击分布内	攻击分布外	攻击分布内外
MSP	87.68	56.70	56.23	21.68
Energy	86.93	31.61	53.48	10.55
MSD	82.39	21.23	54.61	8.06
OE	94.60	41.42	83.29	17.60
VOS	87.43	31.96	58.75	12.01

分析表3.3并与表3.2对比可知，近-分布外数据因为与分布内数据集相似度较高，区分较为困难，所以总体性能都比对应的远-分布外性能差。其中，在同时攻击分布内外样本时性能下降幅度最大的方法为 MDS，性能下降 90.22%；性能下降幅度最小的方法为 MSP，性能下降 75.27%。在每种攻击情况下近-分布外与远分布外的性能下降幅度相似，因此各种检测方法的鲁棒性与近-分布外和远-分布外没有强相关。所以，无论测试分布外数据集是近-分布外数据集还是远-分布外数据集，都需要在同时考虑对抗分布内样本和对抗分布外样本的情况下进行模型对抗鲁棒性验证。

（3） 攻击前后基准数据集对比实验

为了详细展现在各个数据集下各种方法的性能变化，本部分在同时攻击分布内和分布外的情况下，详细展示了各个数据集的指标。

表 3.4　攻击前后数据集 AUROC 变化表

检测方法	CIFAR100	Tiny-Images	mnist	SVHN	Texture	Places365
MSP	86.73/21.87	88.64/21.49	93.51/33.28	91.57/26.42	89.13/26.96	89.35/21.61
Energy	85.55/10.44	88.31/10.65	96.32/15.14	92.38/18.18	88.64/16.90	89.64/10.82
MSD	81.79/8.85	82.79/7.27	87.46/8.24	89.79/8.02	92.31/13.46	80.54/6.12
OE	90.05/15.45	99.14/19.75	98.88/13.72	99.58/31.09	97.40/21.61	96.67/17.55
VOS	86.24/11.95	88.63/12.07	95.29/28.97	88.71/16.27	86.87/16.14	89.75/12.12

攻击前后数据集 AUROC 变化如表3.4，其中表格中数据斜线左侧表示攻击前的 AUROC 指标，斜线右侧表示攻击后的 AUROC 指标，以直观地展现对抗攻击的影响。观察表格数据可知，不存在哪种方法在某一个数据集上对对抗攻击展现出较好的鲁棒性，进一步说明了分数自适应攻击方法对于分布外检测来说是强力的攻击方法。

分布外检测中也常用 FPR95 指标衡量模型的性能，其中 FPR95 越低表明性能越好。表3.5展示了攻击前后各种方法在各个数据集上的 FPR95 指标变化，其中表格中数据斜线左侧表示攻击前的 FPR95 指标，斜线右侧表示攻击后的 FPR95 指标。详细数据如表3.5所示。

表 3.5　攻击前后数据集 FPR95 变化表

检测方法	CIFAR100	Tiny-Images	mnist	SVHN	Texture	Places365
MSP	59.88/99.92	47.23/99.92	19.22/99.76	24.24/99.99	40.44/99.99	41.83/99.81
Energy	72.70/100.00	62.41/100.00	15.49/99.99	30.14/100.00	60.21/100.00	56.37/99.99
MSD	61.21/99.87	55.57/99.88	44.14/99.31	35.36/99.51	32.17/99.68	62.56/99.93
OE	37.67/99.90	2.94/99.60	3.33/98.09	1.22/96.09	12.63/99.93	14.24/99.60
VOS	68.38/99.96	59.33/99.96	18.69/98.48	48.44/99.98	69.21/99.98	53.28/99.89

可以观察到在攻击后的 FPR95 指标都很高，根据 FPR95 的定义，说明此时在保证较高的真正例率时，模型会将大量的负例预测错误预测为正，会产生大量的误判，更进一步验证了分数自适应对抗攻击的强力。

3.4.3　超参数分析实验

分数自适应攻击有三个重要的超参数，分别是扰动边界 ϵ、攻击步数 t 和迭代步长 α。其中扰动边界代表着扰动的取值范围，攻击步数表示在搜寻最优对抗扰动的过程中的搜寻次数，迭代步长表示每次进行梯度下降时的步长大小。为了

表明三个超参数对于攻击强度的影响，本小节通过控制变量的方法调整单一参数进行实验。实验在同时攻击分布内和分布外的情况下进行，并展现近-分布外数据集的 AUROC 结果均值。

（1） 扰动边界分析实验

在不同的扰动边界下，各种方法的 AUROC 下降曲线如图3.2所示。图中横轴表示扰动边界，纵轴表示在对应扰动边界下的 AUROC 百分比值，其中扰动边界为 0 时表示无攻击，以五种不同颜色的曲线展示不同方法的性能情况。

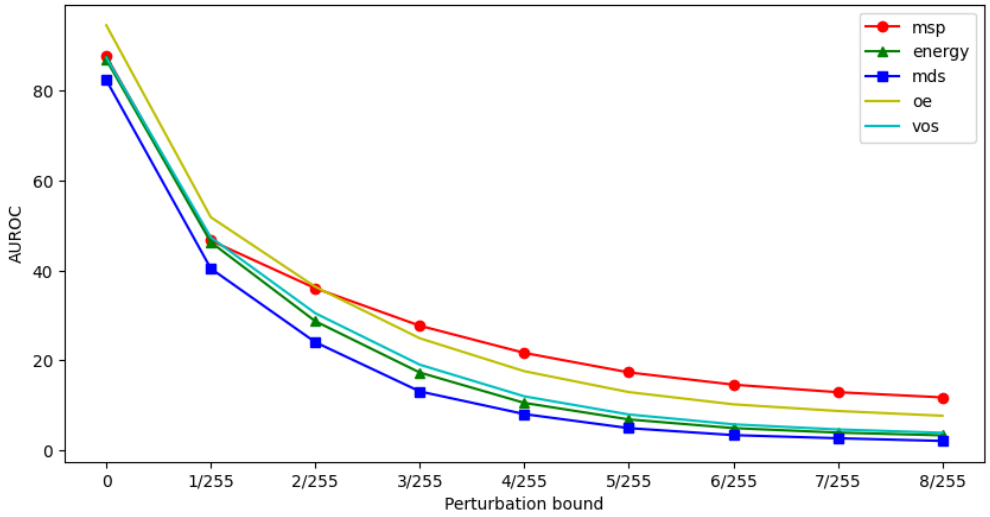


图 3.2　不同扰动边界下的 AUROC 性能下降曲线图

观察图像可知，随着扰动边界的增大，分布外检测性能呈下降趋势。这是由于扰动边界扩大，扰动的搜索空间也随之增大，更容易搜索到更为强力的对抗扰动。而且随着扰动边界的不断扩大，AUROC 性能下降速度逐渐变缓慢，且呈现收敛趋势。由此可知，增大扰动边界并不会使分布外检测性能展现线性性的下降，而且过大的扰动边界还会有使得图像语义信息发生变化的风险，违背了对抗样本的定义；而过小的扰动边界却使得产生的对抗样本不够强力，模型鲁棒性得不到较好评估。由图可知，扰动边界取在 4/255 至 8/255 较为合适。

（2） 攻击步数分析实验

在不同的攻击步数下，以 MSP 方法为例展现攻击步数对性能产生的影响，结果如图3.3所示, 图中数据点依次表示攻击步数为 1、2、5、10、25、50、100、200 和 500 的情况。

分析图像可知，随着攻击步数的增加，AUROC 呈现下降趋势。这是由于随着攻击步数的增加，扰动的搜索次数也增加，更可能搜索到更为强力的扰动。同样随着攻击步数的增加，曲线下降速率逐渐变缓，且最终收敛。这是由于搜索空间受到扰动边界的限制，对抗扰动在该空间内会趋于收敛，以此攻击步数的增加也不能获得更多强力的对抗扰动使得性能进一步下降。由此可知，在进行分数自

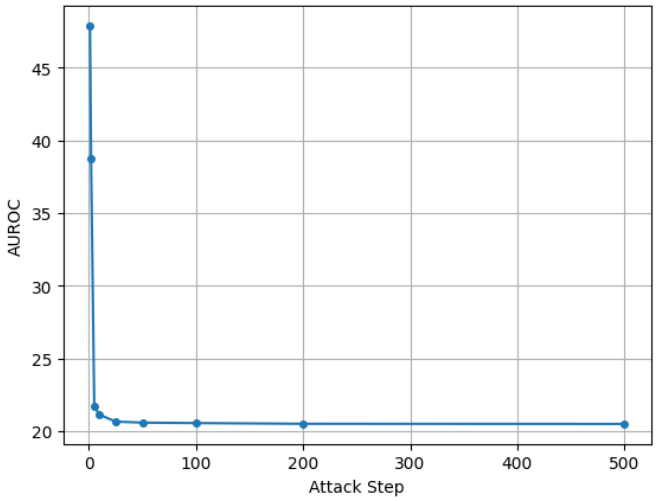


图 3.3　不同攻击步数下的 AUROC 性能曲线图

适应攻击时，只需要设置较为适中的攻击步数即可。攻击步数过大并不能带来更为强力的攻击性能，反而由于迭代次数过多增加时间成本；攻击步数过小会导致在扰动空间内没有得到充分的搜索，因而产生的对抗样本不够强力，模型鲁棒性得不到较好评估。由图可知，攻击步数取在 5 到 200 之间较为合适。

（3） 迭代步长分析实验

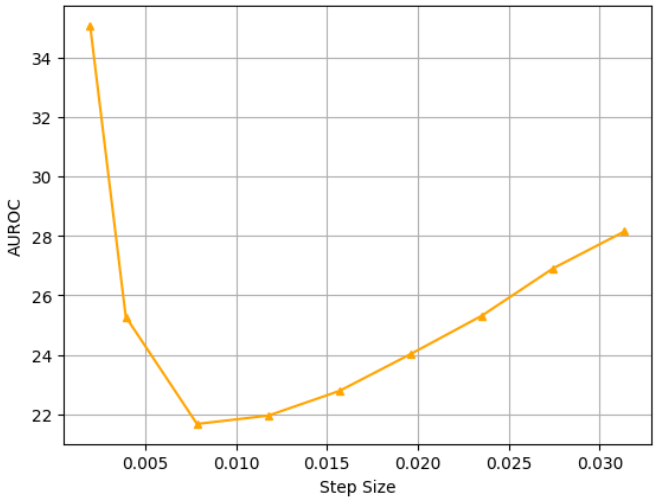


图 3.4　不同迭代步长下的 AUROC 性能曲线图

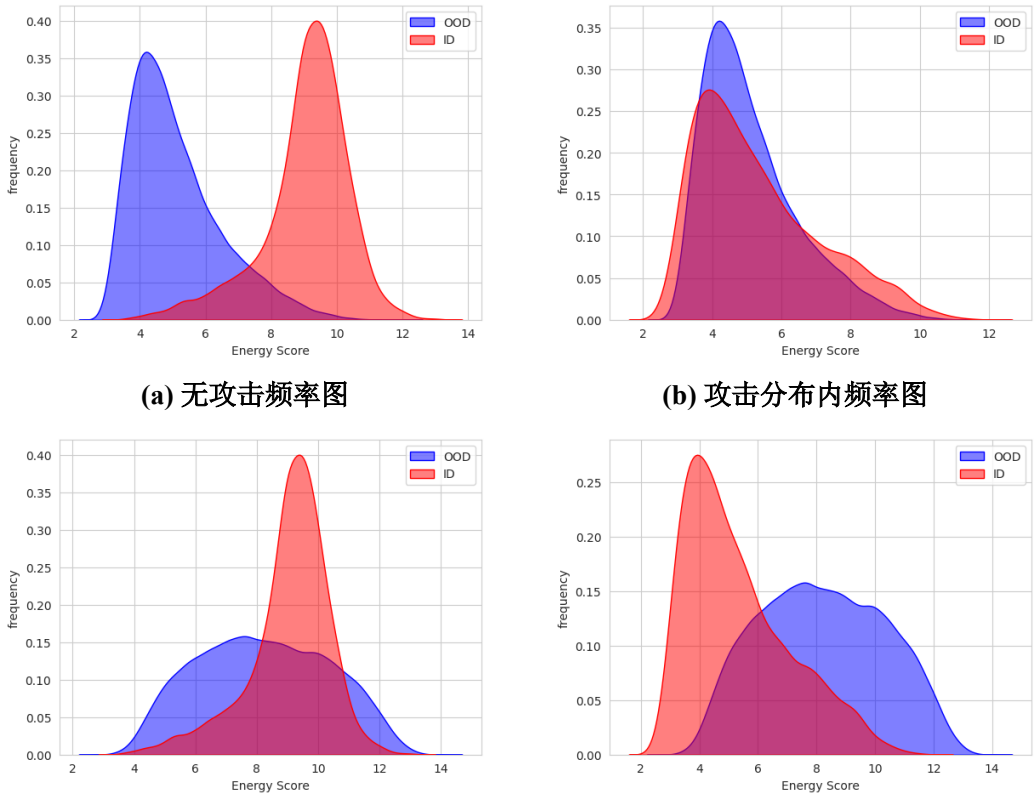
在不同的迭代步长下，以 MSP 方法为例展现迭代步长对性能产生的影响，结果如图3.4所示, 图中数据点依次表示迭代步长为 0.5/255、1/255、2/255、3/255、4/255、5/255、6/255、7/255 和 8/255 的情况。

分析图像可知，随着迭代步长的增加，AUROC 呈现先下降再上升的趋势。这是由于迭代步长较小时扰动还未收敛到最优点迭代就停止了，所以攻击性能较弱；而迭代步长较大时，会导致扰动在最优扰动附近来回振荡，导致错过最优扰动并最终获得一个较差的解。因此，在进行分数自适应攻击时，应设置一个合

适大小的迭代步长，不易过小或过大。由图中可知，设置为 2/255 较为合适。

3.4.4　可视化分析

本小节通过可视化分布外检测分数在不同攻击下的频率分布图来直观展示分数自适应攻击对检测分数的分布所带来的影响。将参数设置为攻击扰动边界 $\epsilon = 4/255$，攻击步数 $t = 5$，迭代步长 $\alpha = 2/255$。由于能量分数跨度较大，可视化效果较为清晰，所以选择 Energy 方法进行可视化，在 mnist 数据集分别获取无攻击、仅攻击分布内、仅攻击分布外和同时攻击分布内外的能量分数分布。可视化结果如图3.5所示。



(a) 无攻击频率图　　**(b) 攻击分布内频率图**

(c) 攻击分布外频率图　　**(d) 攻击分布内外频率图**

图 3.5　不同攻击情况下的频率图

观察图3.5可知，无攻击情况下分布内和分布外的能量分数具有较好的区分度。当攻击分布内样本时，分布内样本的能量分数总体下降，频率分布向减小的方向偏移；当攻击分布外样本时，分布外样本的能量分数总体上升，频率分布向增大的方向偏移。此时，分布内和分布外的检测分数有较大的重叠，区分度下降，因此检测性能下降。同时攻击分布内外时，分布内外的频率分布同时偏移，甚至产生了较大错位，使得此时对分布内外的样本都会产生大量错判。

3.5　本章小结

本章针对现有分布外检测的鲁棒性评估不够完善和强力的问题，分析了不同检测分数下生成的分布内对抗样本和分布外对抗样本的影响，提出了基于检测分数的分布外自适应攻击方法。该方法指出了在进行分布外检测的鲁棒性评估时，应该同时考虑对抗分布内和对抗分布外样本，并且根据不同的检测分数自适应调整优化目标。实验中利用该自适应攻击方法评估了现有分布外检测算法的鲁棒性并发现分布外性能产生大幅下降超 70%。本章揭示了现有分布外检测算法的脆弱性，并为分布外检测的鲁棒性提供一种强力的评估方法。

第 4 章　有额外数据场景基于有效点选取的鲁棒分布外检测方法

上一章揭示了分布外检测中的对抗样本对模型安全性所产生的威胁，本章在此基础上研究具有对抗鲁棒性的分布外检测方法。该方法针对在具有额外分布外数据的场景，通过有效点选取的策略选择优质的分布外数据辅助模型训练，使模型对分布外样本产生对抗鲁棒性。

4.1　引言

分布外检测的目标是在开放场景中识别不属于训练训练集分布的样本，对于开放世界中的深度学习安全性有重要意义。然而，现有大多数分布外检测方法只考虑干净样本的检测情况，忽视了在开放场景中的另一威胁——对抗攻击的存在，使得现有大多数分布外检测方法并不能满足安全关键系统的需求。

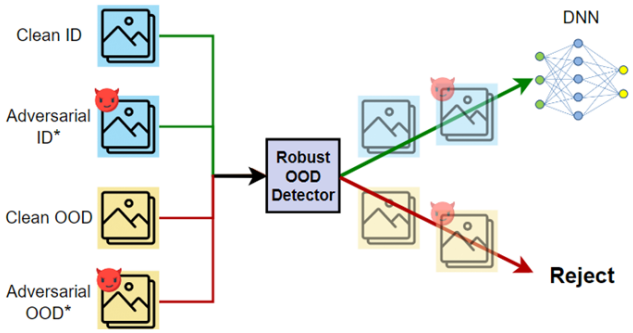


图 4.1　鲁棒分布外检测示意图

为了减少对抗攻击对分布外检测器的威胁，鲁棒分布外检测这个研究领域便诞生了，其示意图如图4.1所示。鲁棒分布外检测要求分布外检测器能够有效识别四种输入：干净分布内数据、对抗分布内数据、干净分布外数据、对抗分布外数据。其中干净分布内样本和干净分布外样本的识别是分布外检测的基本要求。在对抗性的设定下，还需要考虑对抗分布内数据和对抗分布外数据。对抗分布内数据能够欺骗检测器将其识别为分布外输入，频繁地引发系统报警影响系统正常运行；对抗分布外样本则是欺骗检测器将其识别为分布内输入，从而躲避警报系统产生更严重的安全后果。

为了使得分布外检测器能同时识别对抗分布内样本和对抗分布外样本，一些方法借用离群值暴露的思想，如图4.2所示。这些方法在训练期间引入一个额外的辅助分布外数据集，并利用额外分布外数据集产生分布外对抗样本，以此对模型进行正则化训练。如 ALOE[8]、ATOM[9] 和 ATD[10] 等在训练过程中都同时

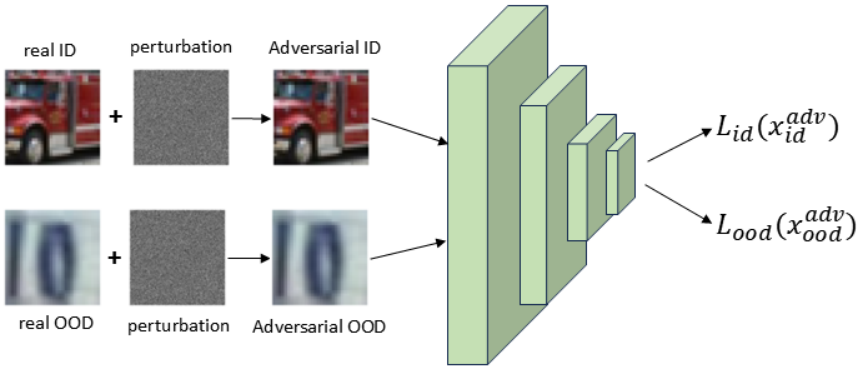


图 4.2　对抗离群值暴露方法示意图

使用分布内数据集和额外分布外数据集，获得更加紧密的鲁棒决策边界。

然而，有研究表明额外数据的选取对于基于离群值暴露的分布外检测方法至关重要[9,76]。所选取的额外分布外数据需要具有多样性并且与分布内的样本具有较高相似性，即需要近分布内样本。因为辅助数据需要表示分布外，若所选取的辅助分布外样本特征不够多样，则会导致该辅助样本并不能有效表示分布外的特征，且会使网络偏向于特定的分布外特征从而影响泛化能力。若所选取的辅助分布外样本与分布内样本相关性不够，则会引入大量无效的信息，使模型关注不重要的特征从而对模型训练过程产生误导。

针对上述问题，本章提出了基于有效点选取的鲁棒分布外检测方法 (Effective Points Select on robust out-of-distribution Detection, EPSD)。具体来说，该方法通过挑选辅助分布外数据集中的近分布内样本进行高效的训练以获得紧凑的决策边界，并采用对抗净化模块对对抗噪声进行去噪使模型对分布内和分布外输入具有鲁棒性。在有效点的选取过程中，该方法训练一个投影头将特征投影到低维嵌入空间，在该空间以类别原型对各个类别的分布建模。将辅助分布外数据映射到该空间后，利用马氏距离计算分布外数据与各个类别原型的距离，以该最小距离进行排序，将靠近分布内原型的点作为有效点。在下一轮训练中，优先使用有效点产生对抗样本进行训练。通过对辅助分布外数据进行有选择性的挑选，可以使模型学习到更有效的信息和更紧凑的边界。为了减少对抗扰动的影响，EPSD 采用神经微分方程并结合李雅普诺夫理论，在特征层面进行对抗扰动净化，提升了鲁棒分布外检测效果。

4.2　问题描述

在传统分布外检测中，分布外检测器 $G(x)$ 需要对每一个输入 x 给出判别，即判断输入是分布内样本 (ID) 还是分布外样本 (OOD)。而在鲁棒分布外检测中，由于加入了对抗性设定，输入不再单纯的是干净样本，还包括由干净样本产生的

对抗样本 x_{adv}。鲁棒分布外检测中，对抗样本分为对抗分布内样本和对抗分布外样本。对抗分布内样本的语义信息是分布内，但由于添加了对抗扰动，分布外检测器会将其错误的识别为分布外，用公式表示如下：

$$\begin{aligned} &\min_{\delta} \mathcal{L}\left(G(x_{adv}^{id}), OOD\right) \\ &\text{s.t. } x_{adv}^{id} = x^{id} + \delta,\ \|\delta\| \leqslant \epsilon \end{aligned} \tag{4.1}$$

其中 x^{id} 表示分布内输入，x_{adv}^{id} 表示对抗分布内样本，δ 表示对抗扰动，其大小被 ϵ 约束以保证人眼所见的语义信息不发生改变，经扰动后分布外检测器更容易将其判别为分布外样本。与此类似，对抗分布外样本 x_{adv}^{ood} 则导致检测器更容易将其判别为分布内样本。

为了使分布外检测器对分布外样本具有对抗鲁棒性，可通过收集额外的辅助分布外数据集 D_{ood}^{aux} 进行训练, 其中 D_{ood}^{aux} 的标签集合与分布内标签集合无重叠，即 $Y_{ood}^{aux} \cap Y_{id} = \varnothing$。在训练期间总的训练数据集包括原始分布内数据集和辅助分布外数据集，即 $D_{train} = D_{id} \cup D_{ood}^{aux}$。在训练期间，则可针对不同的输入类型产生不同的对抗样本，然后分别根据输入类型优化损失函数：

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D_{id}} \max_{\delta\sim B(x,\epsilon)} \mathcal{L}(G(x+\delta), \text{ID}) + \mathbb{E}_{(x,y)\sim D_{ood}^{aux}} \max_{\delta\sim B(x,\epsilon)} \mathcal{L}(G(x+\delta), \text{OOD}) \tag{4.2}$$

其中 $B(x, \epsilon)$ 表示以 x 为中心的一个 ϵ 有界区域内。最终获得的鲁棒分布外检测器将产生一个分布外检测分数 $S(x)$，利用该分数设置阈值则可判别对抗性输入 x_{adv} 属于分布内样本还是分布外样本。即：

$$G(x_{adv}) = \begin{cases} ID, & \text{如果 } S(x_{adv}) > \lambda \\ OOD, & \text{如果 } S(x_{adv}) < \lambda \end{cases} \tag{4.3}$$

在有额外辅助分布外数据集 D_{ood}^{aux} 的情况下，问题的关键在于如何有效的利用额外数据集中的分布外信息。若在训练过程中随机选择分布外数据进行训练，会导致引入大量的无用信息使得模型关注非有效特征，并且也会浪费大量的时间用于产生无用分布外对抗样本。因此，通过合理的策略选择有效的分布外数据进行训练是至关重要的。

4.3　算法设计

本章受基于聚类的离群点检测算法的启发，通过寻找辅助分布外样本在特征空间中的离群点以挑选更有效的辅助分布外数据进行模型训练。此外，为了增强模型的对抗鲁棒性，本章方法采用了经对抗训练的分类器作为特征提取器，并利用神经微分方程模块进行进一步的对抗净化。

接下来的小节中，本文将详细介绍本章所提出的具有额外数据集场景下的鲁棒分布外检测方法 EPSD。

4.3.1　模型框架

EPSD 方法框架如图4.3所示。训练框架包括三个阶段，第一个阶段利用鲁棒特征提取器进行特征提取；第二阶段进行辅助分布外数据集的有效点选取并训练神经 ODE 去噪模块；第三阶段利用全连接层用于标签映射，并获取分布外检测分数。

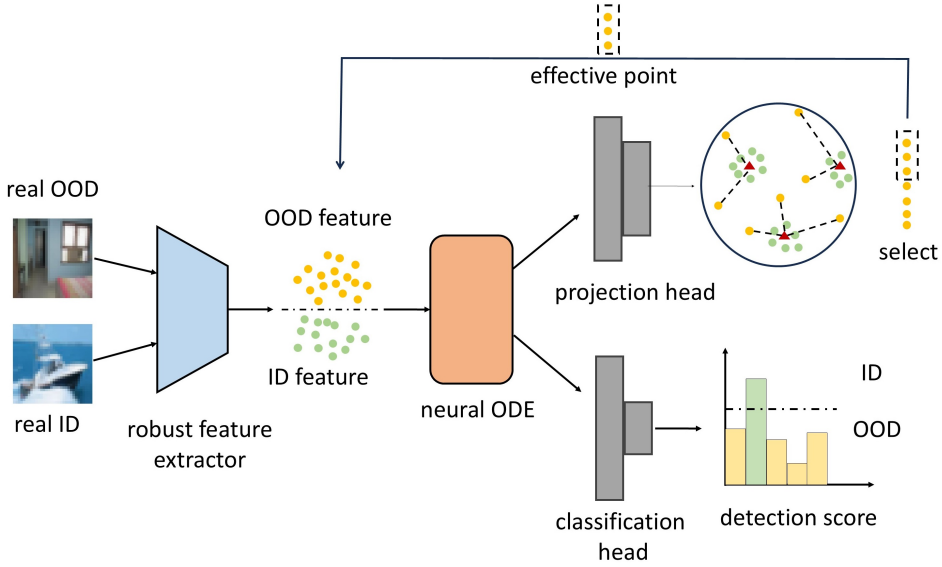


图 4.3　EPSD 方法框架图

在第一阶段中，输入图像会经过一个鲁棒特征提取器以获得鲁棒特征表示，将所有提取到的鲁棒特征保存下来以供后续模块训练使用，使用鲁棒特征提取器可进行初步的对抗去噪。其中，鲁棒特征提取器由一个经过对抗训练的鲁棒分类器获得，将该分类器的 softmax 层去掉后作为鲁棒特征提取器。

第二阶段的训练包括一个神经 ODE 模块和一个投影头。神经 ODE 模块用于对对抗噪声进行去噪，起到对抗净化的作用。投影头用于将神经 ODE 输出特征映射到低维嵌入空间，并在该空间对分布内样本进行聚类操作并计算各类别分布的原型表示以挑选辅助分布外数据的边界点。在每一轮训练中，计算辅助分布外数据距离各个类别原型的马氏距离，以最小马氏距离进行排序，并在下一轮训练中优先挑选具有最小马氏距离的辅助分布外样本进行训练。

第三阶段通过全连接层将输出映射到标签空间中并获取分布外检测分数。在分布外样本检测阶段，以最大 softmax 值作为分布外检测分数，设置阈值进行判定。最大 softmax 值高于阈值的样本判别为分布内，低于阈值的样本判别为分布外。

4.3.2　基于李雅普诺夫稳定性理论的神经 ODE 去噪器

近年来，基于神经 ODE 的对抗去噪器在对抗鲁棒性领域产生了良好的效果[59-61]。神经 ODE 去噪器在原网络结构的中间层嵌入了一个去噪器模块，在特征层面对输入进行对抗净化，具有良好的效果与较强的普适性，因此本文将其引入到分布外检测任务中以保证鲁棒性。神经 ODE 的核心思想是将神经网络视为一个动态系统[77-79]，其状态随时间连续变化，其将输入数据作为初始条件，通过求解一个由神经网络参数化的常微分方程，得到系统在不同时间点的状态，从而实现对数据的处理和建模。一个神经 ODE 由一个神经网络和 ODE 求解器组成，其输入 $z(0)$ 和输出 $z(T)$ 可由以下方程描述：

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = f_{\theta}(z(t), t) \tag{4.4}$$

其中 $\mathbf{z}:[0,\infty) \mapsto \mathbb{R}^n$ 表示神经 ODE 的状态 $f_{\theta}:\mathbb{R}^n \times [0,\infty) \mapsto \mathbb{R}^n$ 表示由权重 θ 参数化的神经网络，通常使用全连接或残差网络。现有研究[61]表明，可将微分方程简化为时不变的情况 $\frac{\mathrm{d}z(t)}{\mathrm{d}t} = f_{\theta}(\mathbf{z}(t))$，即假定神经 ODE 的行为与达到状态的特定时间无关，在此情况下可以简化问题分析并且达到较好的性能，本文也遵从这样的假定。

对于非线性系统 $\frac{\mathrm{d}z(t)}{\mathrm{d}t} = f_{\theta}(\mathbf{z}(t))$ 而言，在其双曲平衡点附近，根据 Hartman-Grobman 定理可将其行为通过线性系统近似。即可将非线性系统近似为：

$$\frac{\mathrm{d}\bar{\mathbf{z}}(t)}{\mathrm{d}t} = \nabla f_{\theta}(\mathbf{z}^*) \cdot \bar{\mathbf{z}}(t) \tag{4.5}$$

其中 z^* 表示系统的双曲平衡点。从而在其平衡点附近，可从线性系统的角度出发研究系统的稳定性。由李雅普诺夫稳定性理论，当且仅当方阵 A 的所有特征值具有负实部时，线性系统 $\frac{\mathrm{d}z(t)}{\mathrm{d}t} = A\mathbf{z}(t)$ 渐进稳定。即对李雅普诺夫稳定平衡点 $z(0)$ 施加一个小扰动变为 $\tilde{\mathbf{z}}(0)$ 后，当 $t \to \infty$ 时有 $\|\tilde{\mathbf{z}}(t) - \mathbf{z}(0)\|_2 \to 0$。

可将以上结论从对抗鲁棒性的角度进行分析。在神经网络对抗攻击的背景下，在李亚普诺夫稳定平衡点周围，如果神经 ODE 输入 $z(0)$ 周围的恶意扰动较小，那么若 T 足够大时输出 $z(T)$ 不会受到扰动的显著影响。因此，神经 ODE 层之后的后续网络层仍然可以表现良好，不会受到输入扰动的影响。李亚普诺夫稳定平衡点周围的扰动减弱现象可以看作噪声滤波器，对对抗噪声进行削弱和净化，从而起到抵御对抗攻击的作用。因此，对于一个神经 ODE 而言，若能通过训练策略使其满足雅可比矩阵 $\nabla f(\mathbf{z})$ 具有负实部，且使训练输入接近李雅普诺夫平衡点，则可对输入产生对抗净化的效果。

根据 Levy–Desplanques 定理，对于 $n \times n$ 的方阵 A, 若对任意的 $i, j \leqslant n$ 满足 $|\mathbf{A}_{ii}| > \sum_{j \neq i} |\mathbf{A}_{ij}|$ 且 A 的每个对角线元素 a_{ij} 都是负数，则 A 的所有特征值具有负实部。因此，为了使输入位于神经 ODE 的李雅普诺夫平衡点附近，网络的优

化目标为：

$$\begin{aligned}
&\min_{\theta} \mathbb{E}\ \mathcal{L}(\mathbf{z}(T), y) \\
&\text{s.t.}\ \ \mathbb{E}_{\nu}\ \|f_{\theta}(\mathbf{z}(0))\|_2 < \epsilon, f_{\theta} \in C^1(\mathbb{R}^n, \mathbb{R}^n), \\
&\qquad \mathbb{E}_{\nu}\ \left[\nabla f_{\theta}(\mathbf{z}(0))\right]_{ii} < 0, \forall i = 1, \cdots, n, \\
&\qquad \mathbb{E}_{\nu}\left[|[\nabla f_{\theta}(\mathbf{z}(0))]_{ii}| - \sum_{j \neq i} |[\nabla f_{\theta}(\mathbf{z}(0))]_{ij}|\right] > 0, \forall i = 1, \cdots, n,
\end{aligned} \tag{4.6}$$

其中，$z(T)$ 为神经 ODE 的输出，该输出经过后续网络后与标签计算损失函数。在约束项中，第一项约束使输入 $z(0)$ 接近于李雅普诺夫平衡点；第二项约束使雅可比矩阵所有对角线元素为负数；第三项约束使雅可比矩阵具有主对角优势，即 $|\mathbf{A}_{ii}| > \sum_{j \neq i} |\mathbf{A}_{ij}|$。

由于在网络训练过程中，直接优化上述目标较为困难，因此用以下的拉格朗日函数进行替代：

$$\begin{aligned}
\min_{\theta} \frac{1}{N} \sum_{k=0}^{N-1} &\left(\mathcal{L}\left(\mathbf{z}_k(T), y_k\right) + \alpha_1 \| f_{\theta}\left(\mathbf{z}_k(0)\right) \|_2 + \alpha_2 g_1 \left(\sum_{i=1}^{n} [\nabla f_{\theta}(\mathbf{z}_k(0))]_{ii} \right) \right. \\
&\left. + \alpha_3 g_2 \left(\sum_{i=1}^{n} (-|[\nabla f_{\theta}(\mathbf{z}_k(0))]_{ii}| + \sum_{j \neq i} |[\nabla f_{\theta}(\mathbf{z}_k(0))]_{ij}|) \right) \right.
\end{aligned} \tag{4.7}$$

其中 α_1,α_2,α_3 为超参数权重，取 $g_1(\cdot) = g_2(\cdot) = \exp(\cdot)$ 为单调递增有下界的函数，这有助于防止两个正则化项的无界效应。

在鲁棒分布外检测中，神经 ODE 的输入集同时包含分布内输入和辅助分布外输入，即 $Z = Z_{id} \cup Z_{ood}^{aux}$。因此，神经 ODE 将对分布内输入和分布外输入都具有对抗鲁棒性。进一步地，由于分布外输入的引入，损失 $\mathcal{L}$ 不再能够使用交叉熵损失。若将后续网络输出表示 $h(z)$, 则损失函数形式化如下：

$$\mathcal{L}(z, y) = \begin{cases} -\log \frac{e^{h_y(z)}}{\sum_{i=1}^{k} e^{h_i(z)}} & , \quad y \in Y_{id} \\ \mathcal{L}_{CE}(h(z), U_K) & , \quad y \in Y_{ood} \end{cases} \tag{4.8}$$

其中 $h_i(z)$ 表示最后的输出层在维度 i 上的值，U_K 表示 K 维均匀分布。该损失表示，若输入属于分布内，则计算其交叉熵分类损失；若输入属于分布外，则计算其与均值分布的 KL 散度作为损失。

4.3.3　基于原型学习的特征聚集投影

为了有效的利用额外数据集中的分布外信息，避免随机选取导致的无用信息引入，本框架利用特征聚集的策略，通过计算分布外样本在特征空间与各个类别的距离来进行选取。

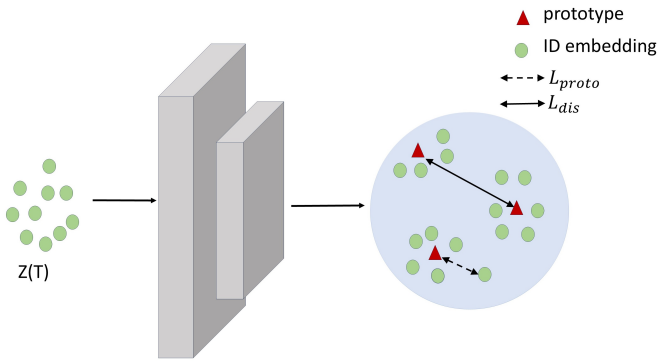


图 4.4　特征聚集示意图

特征聚集的思想来源于原型学习[28,80-81]。原型学习为每个类别计算类别原型作为该类别的代表，并利用该原型集聚该类样本的特征，从而获得类间更为紧凑的特征表示。为了在原型表示周围聚集特征，原型学习通常结合对比学习，使用以下损失：

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(Sim(\phi(z_i), p_{c(i)})\right)}{\sum_{j=1}^{C}\exp\left(Sim(\phi(z_i), p_j)\right)} \tag{4.9}$$

其中 p_j 表示类别 j 的原型表示，$\phi(z_i)$ 为投影空间的嵌入，通过相似度计算表征嵌入与原型之间的距离。对比学习可以使得各个类别的嵌入在该投影空间靠近与之对应的类别原型，并远离其他类别原型。

相似度计算可采用余弦相似度，对于向量 p_i,p_j，余弦相似度计算公式为：

$$\cos(p_i, p_j) = \frac{p_i \cdot p_j}{\|p_i\|\|p_j\|} \tag{4.10}$$

若在投影头输出后将嵌入和原型进行归一化，嵌入和对应原型的余弦相似度可用内积表示:$\cos(\phi(z_i), p_j) = \phi(z_i) \cdot p_j$。因此可将损失写为：

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\phi(z_i) \cdot p_{c(i)}/\tau\right)}{\sum_{j=1}^{C}\exp\left(\phi(z_i) \cdot p_j/\tau\right)} \tag{4.11}$$

一般来说，类别原型需要计算单个类所有嵌入的均值向量，但如果每轮训练都全部计算所有输入的均值会导致计算复杂度较大。因此一般采用指数移动平均的策略对类别原型进行更新：

$$\boldsymbol{p}_c := \alpha \boldsymbol{p}_c + (1-\alpha)\phi(z), \forall c \in \{1, 2, \cdots, C\} \tag{4.12}$$

为了进一步的增加各个类别之间的区分度，可增加类间离散损失以最大化各个类别之间的相似度，从而使得各个类别相互分离。类间离散损失为：

$$\mathcal{L}_{\text{dis}} = \frac{1}{C}\sum_{i=1}^{C}\log\frac{1}{C-1}\sum_{j=1}^{C}\mathbb{1}\{j \neq i\}e^{p_i \cdot p_j/\tau} \tag{4.13}$$

最终，将用于类内聚集的原型损失和类间离散损失相结合，并通过权重系数进行加权，以此优化投影头便可获得类内聚集和类间离散的投影表示。最终损失为：

$$\mathcal{L} = \mathcal{L}_{\text{proto}} + \lambda \cdot \mathcal{L}_{\text{dis}} \tag{4.14}$$

4.3.4　基于马氏距离的有效点选取

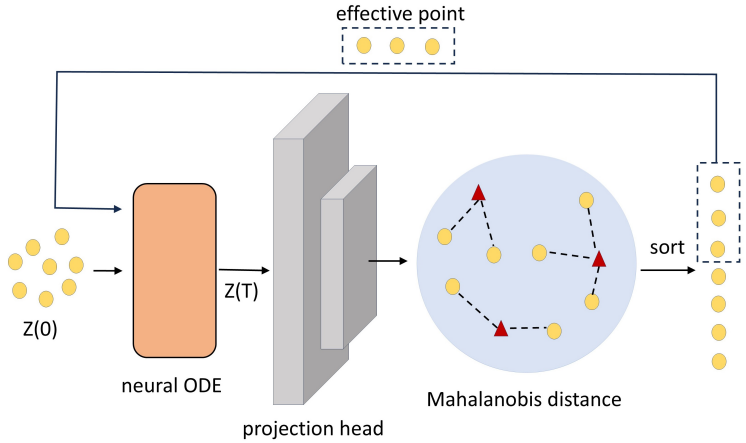


图 4.5　有效点选取示意图

利用投影头及所获得的类别原型，则可对分布外数据进行有效点的选取。具体来说，在每轮训练中，从辅助分布外数据集 D_{ood}^{aux} 中随机挑选 N 个数据点作为分布外候选集 S_N，再从 S_N 中选择与分布内最为相似的点作为有效点用于本轮训练。

为了衡量辅助分布外数据与分布内的相似性，在投影空间中利用距离作为指标进行衡量。若分布外数据与分布内的嵌入表示具有较小的距离，则认为该分布外数据与分布内数据更为相似，选取其用于训练将获得分布外和分布内之间更为紧凑的决策边界。在计算距离时，直接利用上一小节中的类别原型代表类别的分布，从而计算嵌入与类别原型间的距离以进行简化。

本章方法采用马氏距离作为嵌入间的距离衡量方法。马氏距离可以看作标准化后的欧式距离，它对于数据的尺度变换具有不变性，并且考虑到了数据维度之间的相关性，在分布外检测任务中有良好的效果[12]。马氏距离计算公式如下：

$$d_{maha}(\mathbf{x}) = \left(f(\mathbf{x}) - \hat{\mu}_c\right)^{\top} \hat{\Sigma}^{-1} \left(f(\mathbf{x}) - \hat{\mu}_c\right) \tag{4.15}$$

其中均值 μ 和协方差矩阵 Σ 通过以下公式计算：

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\mathbf{x}_i),\ \hat{\Sigma} = \frac{1}{N} \sum_{c} \sum_{i:y_i=c} \left(f(\mathbf{x}_i) - \hat{\mu}_c\right) \left(f(\mathbf{x}_i) - \hat{\mu}_c\right)^{\top} \tag{4.16}$$

对于候选集 S_N 中的任一样本，将其经过投影头进行降维投影得到嵌入表示 $\phi(z_i^S)$。类别原型可代表每个类别的均值向量，因此在投影空间中，通过类别

原型计算 p_c 计算马氏距离，并取最小马氏距离作为相似性指标：

$$M(z_i^S) = \min_c \left(\phi(z_i^S) - p_c\right)^\top \hat{\Sigma}^{-1} \left(\phi(z_i^S) - p_c\right) \tag{4.17}$$

按照 $M(z_i^S)$ 可对候选集 S_N 中的分布外样本进行从小到大的排序。直观上来说，分布外数据具有越小的 $M(z_i^S)$，则代表其越靠近分布内数据，选取其进行训练所获得的决策边界就越紧凑。若每轮训练使用 n 个辅助分布外数据，则可取排序后的 S_N 中的前 n 个样本作为有效点，以此获得最为有效的分布外数据点。

4.3.5　分布外评价分数

在式4.8中，对于分布内数据驱使其在对应的 one-hot 标签具有较高的 softmax 值，对于分布外数据则驱使其输出接近于均值分布。因此可采用最大 softmax 作为分布外评价分数。即对于最终层的每一个输出向量 $(h_1(z), h_2(z), \cdots, h_K(z))$, 其 MSP 分数为：

$$S(x) = \max_k \frac{\exp\left(h_k(z)\right)}{\sum_{j=1}^{K} \exp\left(h_j(z)\right)} \tag{4.18}$$

具有较大 $S(x)$ 的输入样本应该判断为分布内，反之为分布外。选取阈值 γ 作为判断标准，则有：

$$G(x) = \begin{cases} ID, & \text{如果 } S(x_{adv}) > \gamma \\ OOD, & \text{如果 } S(x_{adv}) < \gamma \end{cases} \tag{4.19}$$

4.4　实验与分析

本节介绍实验设计与相应的结果分析，包括实验所采用的基准数据集、评价指标、实验设置和对比实验，并进行了进一步的消融和可视化结果分析。

4.4.1　数据集与评价指标

在训练阶段，本章的分布内数据集使用鲁棒分布外检测领域通用的基准分布内数据集 CIFAR10 和 CIFAR100。与采用离群值暴露的方法一样，使用具有 8000 万张图像的 Tiny-Images 数据集作为额外的辅助分布外数据集。

在模型测试阶段，本章采用鲁棒分布外检测中常用的分布外基准数据集作为分布外数据，包括 LUSN、iSUN、SVHN 和 Places365。这些分布外数据集将通过随机裁剪和下采样将其调整至 32×32 大小，与 CIFAR 数据集大小保持一致。在实验中，如果分布内数据集是 CIFAR10，则将 CIFAR100 和这四个数据集共同作为分布外数据集；若分布内数据集是 CIFAR100，则将 CIFAR10 和这四个数据集共同作为分布外数据集。

鲁棒分布外检测是对抗性设定下的分布外检测，因此本章依然采用 AUROC 作为评价指标。

4.4.2　实验设置

鲁棒特征提取器所采取的网络架构为 WideResNet-70-16，相关网络模型和参数来源于对抗鲁棒性标准模型库 RobustBench[82]，选取其中具有优良性能的方法[83]，并去掉最后一层作为鲁棒特征提取器。对于神经 ODE 去噪器，本章方法与先前工作[61]保持一致，使用两层全连接作为参数化的微分方程，取超参数为 $\alpha_1 = 1$，$\alpha_2 = \alpha_3 = 0.05$，积分时间 $T = 5$，神经 ODE 的求解利用开源神经 ODE 求解库 torchdiffeq[77]，使用 5 阶 Runge-Kutta 方法求解。投影头使用两层全连接层，输出维度为 128，参数 $\lambda = 1$，$\alpha = 0.5$。输出模块使用两层全连接作标签映射，在 CIFAR10 数据集上输出维度为 10，在 CIFAR100 数据集上输出维度为 100。使用 SGD 作为优化器，初始学习率为 0.05，学习率使用余弦学习率衰减，共训练 50 个 epoch, 批次大小为 128。根据离群值暴露方法[19]的实验设定，辅助分布外数据应为分布内数据的两倍，则固定 $n = 100000$，选取候选点数 $N = 400000$。

测试期间，在对抗性的设定下使用 PGD 攻击来生成对抗样本。由于采用 MSP 作为分布外评价分数，根据第三章的分数自适应攻击方法，对于分布内样本应最小化其 MSP 分数，对于分布外样本应最大化其 MSP 分数。为了保证扰动强力，PGD 攻击参数设置为扰动边界 $\epsilon = 8/255$，攻击步数 $t = 100$，迭代步长 2/255。

4.4.3　基准数据集鲁棒分布外检测对比实验

本节分别以 CIFAR10 和 CIFAR100 作为分布内数据集，与多种方法进行对比以证明本章所提方法的有效性。在对比的方法中，AT 表示经过对抗训练获得的分类器并用 MSP 分数进行分布外检测，以展现直接将鲁棒分类器进行鲁棒分布外检测的性能，但未接触辅助分布外数据。OE 为标准离群值暴露方法，采用了辅助分布外数据但未考虑对抗样本的存在。ATOM、ALOE、ATD 在训练期间直接使用辅助分布外数据产生对抗分布外样本进行对抗训练；REDOE 使用 CLIP 模型中的额外信息生成分布外数据；AROS 使用了神经 ODE 模块但未采用辅助分布外数据集，为了公平比较在训练过程中利用随机选择的策略添加分布外数据，记为 AROS+。

（1） 分布内数据为 CIFAR10 的对比实验

表4.1展示了分布内数据集为 CIFAR10 时的对比实验结果。其中斜线左侧数据表示干净样本的 AUROC 指标，斜线右侧数据表示同时攻击分布内和分布外样本后的 AUROC 指标，AUROC 指标越大表示性能越好。每一行表示不同的鲁

表 4.1　CIFAR10 数据集上攻击前后 AUROC 变化表

方法	CIFAR100	SVHN	LSUN	iSUN	Texture	Places365	average
AT	60.9/31.1	73.3/28.2	68.2/31.9	69.8/34.6	65.5/33.2	52.3/26.4	65.3/30.9
OE	90.0/1.3	99.5/6.1	98.2/3.6	98.6/3.5	97.4/4.8	96.6/1.7	97.1/3.6
ATOM	94.2/2.5	89.2/5.8	99.1/1.9	99.5/3.4	98.1/6.3	98.7/6.5	97.1/4.2
ALOE	78.8/17.0	83.5/27.3	98.7/51.6	98.3/50.4	95.4/45.5	85.1/22.8	90.8/34.7
ATD	82.0/38.0	87.9/37.5	96.0/69.0	94.8/66.8	93.3/65.2	92.5/60.7	91.6/54.6
RODEO	95.6/38.7	83.0/36.9	99.0/86.0	97.7/79.6	95.4/69.7	96.2/71.1	95.5/62.5
AROS+	90.4/82.4	94.2/**88.7**	92.1/84.2	90.7/82.6	90.4/81.2	92.9/85.1	91.8/84.0
EPSD	89.2/**86.7**	90.3/87.8	90.0/**86.1**	85.2/**84.3**	89.3/**83.1**	90.3/**88.6**	89.1/**86.1**

棒分布外检测方法，最后一行为本章提出的方法。每一列表示不同的分布外数据集下的结果，分布外数据集包括近-分布外数据集 CIFAR100 和其余五个远-分布外数据集，最后一列为六个数据集结果的均值。

分析表中结果可知，在 CIFAR10 数据集作为分布内数据集时，EPSD 在除了 SVHN 外的其他数据集上都取得了最好的鲁棒分布外检测性能，并且六个数据集上的平均鲁棒分布外检测性能最优，性能比次优的方法提升 2.5%，证明了本章所提方法的优越性。EPSD 尤其在近-分布外场景下相比其他方法有更强的优越性，性能提升 5.2%。这是由于在选择辅助分布外样本时，EPSD 选取与分布内样本具有更高的相似性的有效点进行训练，获取了更为紧凑的决策边界，因此增加了近-分布外场景下的区分度。EPSD 的干净样本的 AUROC 指标相比于其他方法具有一定劣势，根据先前的研究[84-85]，这是因为模型的鲁棒性提升会导致一定程度的干净性能的下降，鲁棒性和干净样本性能之间需要根据不同情况进行权衡。

此外，在鲁棒分布外检测中 AT 和 OE 的性能较差，两者分别是常规对抗鲁棒性和常规分布外检测的方法。对于 AT 而言，由于模型缺乏分布外数据的信息与区分分布外特征的能力，使得对抗攻击能够方便利用分布外数据所包含的特征进行攻击[63]。对于 OE 而言，模型所接受的一直是干净数据，因此不具有利用鲁棒特征进行判断的能力，所以使得极易被对抗攻击所欺骗。因此，分布外数据与对抗攻击在鲁棒分布外检测中会互相影响，两者需要联合考虑。

（2） 分布内数据为 CIFAR100 的对比实验

表4.2展示了分布内数据为 CIFAR100 时的对比实验结果，表中数据格式依然为“攻击前/攻击后”。其中 CIFAR10 为近-分布外数据集，其余为远-分布外数据集。

表 4.2　CIFAR100 数据集上攻击前后 AUROC 变化表

方法	CIFAR10	SVHN	LSUN	iSUN	Texture	Places365	average
AT	55.8/42.7	65.2/55.0	60.6/49.6	59.7/50.3	54.3/49.4	57.4/48.3	58.0/49.0
OE	88.5/1.1	89.1/0.6	81.3/0.9	83.6/1.0	88.1/1.2	85.6/1.4	85.9/1.0
ATOM	87.5/3.7	92.8/5.1	96.6/3.2	96.4/3.1	93.9/5.3	94.8/4.7	93.8/4.0
ALOE	43.6/3.0	74.0/19.3	83.1/20.7	80.1/22.1	79.3/18.9	75.0/14.1	72.7/15.6
ATD	57.5/13.8	72.5/29.5	89.2/49.4	86.5/47.3	80.6/39.4	83.3/41.7	78.3/35.0
RODEO	61.5/30.7	76.9/33.5	98.1/**84.8**	95.1/**77.3**	91.2/66.9	93.0/68.3	86.7/57.6
AROS+	75.7/68.9	82.2/72.7	78.4/70.6	84.1/72.7	77.8/68.4	80.6/71.9	80.0/70.9
EPSD	76.1/**74.5**	79.5/**75.3**	81.2/76.4	79.4/73.1	76.3/**71.9**	78.9/**75.7**	78.4/**74.3**

分析表中数据可知，在 CIFAR100 数据集作为分布内数据集时，EPSD 也取得了优越的鲁棒分布外检测性能，近-分布外数据集上性能提升 8.1%，平均性能提升 4.8%。与表4.1对比可得出相同结论，本章方法在近-分布内场景下有较大的优势，但更强的鲁棒性也带来了一定程度的干净样本性能下降。

4.4.4　消融实验

本小节通过对模型结构中的核心组件和步骤进行消融实验，以展示本章方法中各个组件和步骤的重要性。具体来说，本小节对鲁棒特征提取、神经 ODE 去噪器和有效点选取步骤进行消融分析。对鲁棒特征提取的消融采用一个未经对抗训练的普通分类器去掉最后一层后替换掉鲁棒特征提取器；对神经 ODE 去噪器的消融采用一个不进行正则化的普通残差块代替；对有效点选取步骤的消融采用随机选择的策略进行代替。实验中以 CIFAR10 作为分布内数据集，CIFAR100 作为分布外数据集，攻击参数为扰动边界 $\epsilon = 8/255$，攻击步数 $t = 100$，迭代步长 2/255，以 AUROC 作为性能指标。消融结果如表4.3所示。

表 4.3　模块消融数据表

设置	鲁棒特征提取	神经 ODE 去噪	有效点选取	AUROC
A	-	✓	✓	60.1
B	✓	-	✓	26.7
C	✓	✓	-	81.4
D(ours)	✓	✓	✓	86.7

在设置 A 中，鲁棒特征提取器被替换为普通的特征提取器使得性能下降，这

是因为相比与鲁棒特征提取器，普通的特征提取器更容易受到对抗扰动的影响，从而使得施加扰动后特征产生较大偏移导致最终决策产生错误；在设置 B 中，基于李雅普诺夫理论的神经 ODE 模块被替换为不进行正则化训练的普通残差块，失去了对抗净化的作用，使得对抗扰动会较大地影响后续模块从而带来性能较大的下降；设置 C 将有效点选取替换为传统离群值暴露的随机选取，使得训练过程中带来了无用信息并导致决策边界不够紧凑，从而引发性能的下降。设置 D 为本章提出的方法，综合了所有模块获得了最优的性能。

4.4.5　超参数分析实验

本小节通过控制变量实验分析本章方法中的超参数对于方法性能产生的影响。分析的超参数包括: 与神经 ODE 模块相关的积分时间 T; 与有效点选取有关的候选点数量 N；与攻击强度有关的扰动边界 ϵ。实验均以 CIFAR10 作为分布内数据集，以 CIFAR100 作为分布外数据集，在同时攻击分布内和分布外数据的情况下以 AUROC 作为指标评价性能。

（1） 积分时间分析实验

在神经 ODE 不同的积分时间下分析积分时间对鲁棒分布外检测性能的影响，结果如图4.6所示。

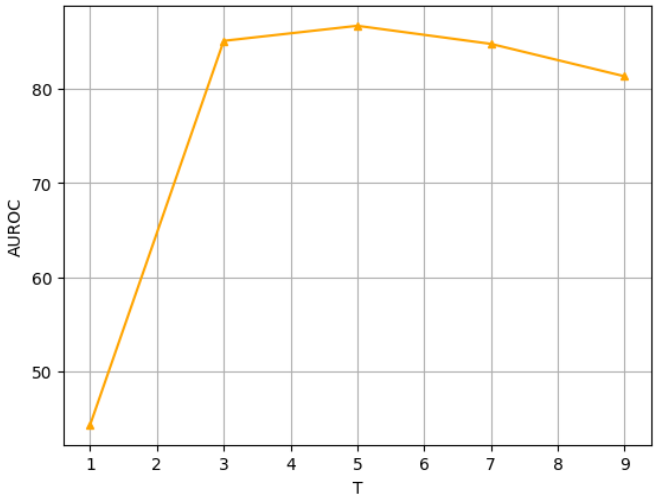


图 4.6　不同积分时间的 AUROC 性能曲线图

分析图可知，积分时间过小会导致模型的鲁棒分布外检测严重下降。积分时间小于 5 时，随着积分时间越来越大模型的鲁棒分布外检测性能会逐渐上升；在积分时间大于 5 时，继续增大积分时间性能又产生略微的下降。这是因为在李雅普诺夫稳定性理论的保证下，对输入施加扰动后经过神经 ODE 模块会逐渐收敛至其平衡点。所以，过小的积分时间导致扰动点还未收敛至平衡点附近，从而还未对扰动进行净化，因此后续模块会受到扰动较大的影响导致鲁棒分布外检测

性能低。而当积分时间逐渐增大时，越来越多的扰动点有足够的时间收敛到平衡点，因此鲁棒分布外检测性能逐渐上升，所以前半段呈现上升趋势。而后半段的略微下降趋势可能是由于神经 ODE 是通过数值积分求解，数值积分会产生误差，当积分时间过大时，累计误差影响了神经 ODE 的求解准确度从而使得系统偏离理论上的平衡点，造成性能的下降。

（2） 候选点数量分析实验

每轮训练中，有效点从候选点中选取获得，改变从辅助分布外数据集中随机选择的候选点数量以分析其对鲁棒分布外检测性能的影响，结果如图4.7所示。

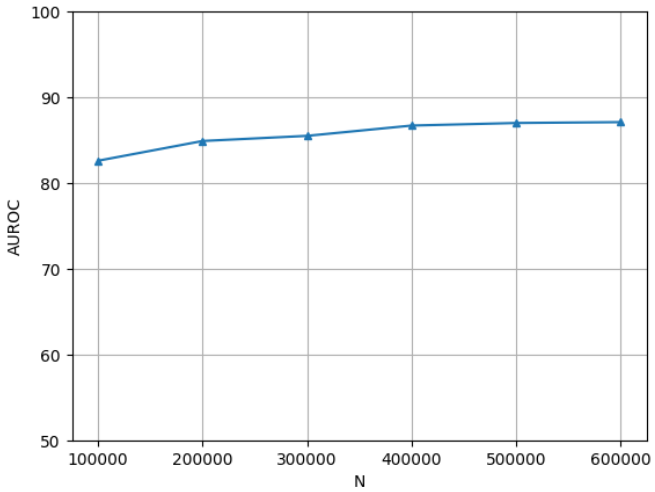


图 4.7　不同候选点数量的 AUROC 性能曲线图

分析图像可知，随着每轮随机选取的候选点数的增加，AUROC 指标总体呈现上升趋势，但后续趋于平稳。这是由于随着候选点数的增加，从候选点中选取有效点时就更容易选取到丰富的与分布内样本相似的有效点，从而使得鲁棒分布外检测性能上升。当候选点数量较少时，据此选取的有效点中就容易包含无用信息，所以此时性能较低。当候选点数量到达一定规模时，带来的具有新特征的有效点变得有限，所以鲁棒分布外检测性能提升幅度变小并趋于平稳。但总体来说，只要候选点数较大，模型性能对于候选点数并不敏感。

（3） 扰动边界分析实验

扰动边界影响着攻击强度的大小，通过修改不同的扰动边界以检测本章方法在不同攻击强度下的性能，结果如图4.8所示。

分析图像可知，随着攻击强度的增加 AUROC 逐渐下降，但在较大的扰动边界下依然能保持较高的鲁棒分布外检测能力，在扰动边界为 8/255 时性能仅下降 2.8%。EPSD 的性能未出现图3.2中大幅下降的情况，表现出了较好的鲁棒性。

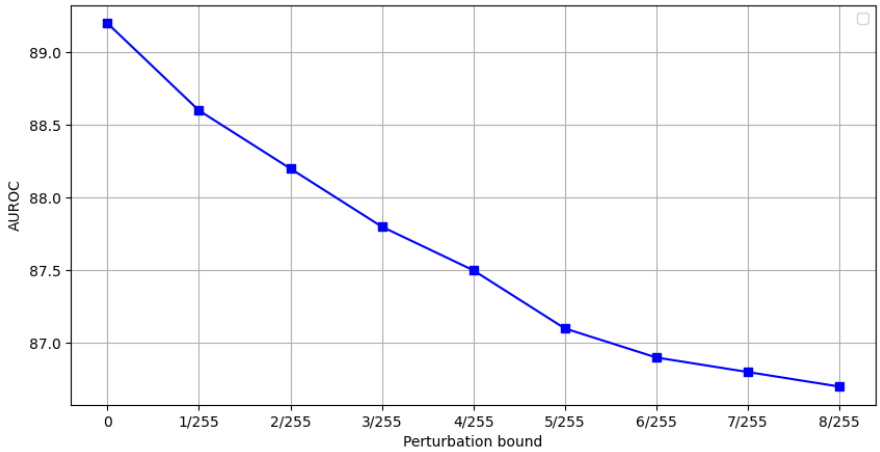


图 4.8　不同扰动边界下的 AUROC 性能曲线图

4.4.6　可视化分析

本小节通过可视化分析来直观展现本章方法的特性，其中包括攻击前后的分布外检测分数频率分布图和特征聚集前后的 UMAP 可视化图。

（1） 分布外检测分数可视化

为了展现对抗攻击的影响，分别在 CIFAR10 和 CIFAR100 作为分布内样本时获取攻击前后的分布外检测分数频率分布。实验参数与对比实验中保持一致。可视化结果如图4.9所示。

分析图像可知，分布内和分布外样本在攻击前后的频率分布都产生了轻微的偏移，但总体偏移幅度较小，因此区分度不会有较大的改变，模型的 AUROC 指标在攻击后也就不产生较大幅度的下降，进一步说明了 EPSD 良好的鲁棒分布外检测性能。

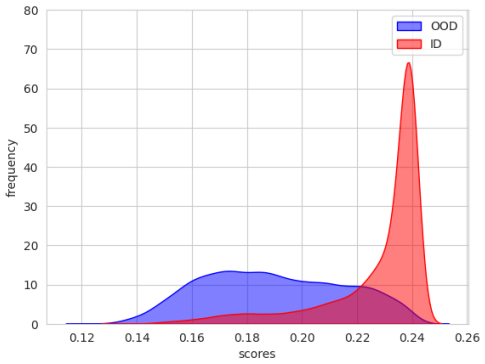
（2） UMAP 可视化

UMAP 可以将高维特征空间中的特征点降维至低维空间中以进行可视化分析。为了展现特征聚集的作用，将特征聚集前后分布内数据特征点的 UMAP 图可视化，结果如图4.10所示，其中不同颜色代表了不同的类别。

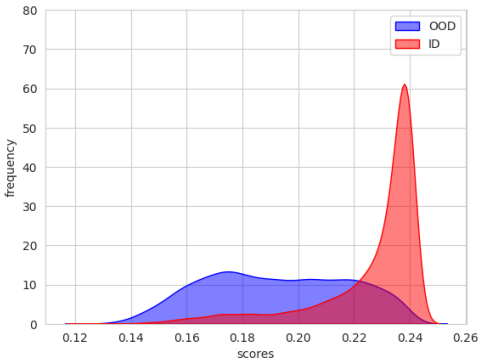
观察图4.10发现，特征聚集前，类别间出现了较大的重叠部分，区分性较差。特征聚集后，类内具有更好的紧凑性，类间具有更好的分散性，体现了原型损失和类间离散损失的作用。

为直观体现特征聚集对于有效点选取的作用，将特征聚集前后分布内和分布外样本特征点的 UMAP 图进行可视化，结果如图4.11a所示。其中橘黄色代表分布内特征点，灰色代表分布外特征点。

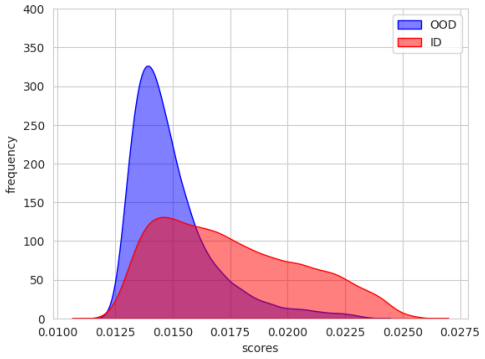
观察图4.11a发现，特征聚集前，分布内和分布外数据特征点之间有较大程度的重叠，且分布内数据较为发散，使得此时挑选与分布内数据相近的特征点较为困难。在图4.11b中，经过特征聚集后，分布内数据更加紧密，分布内外数据重



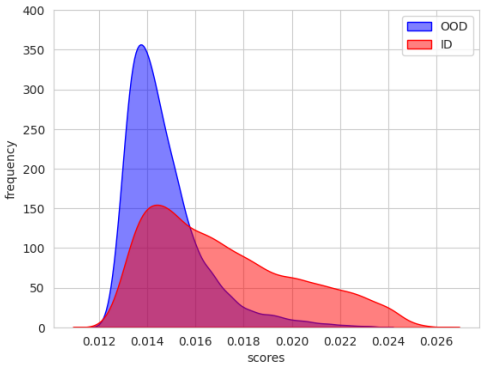
(a) CIFAR10 无攻击频率图



(b) CIFAR10 有攻击频率图

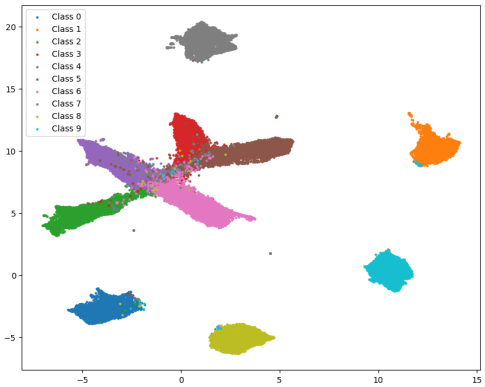


(c) CIFAR100 无攻击频率图

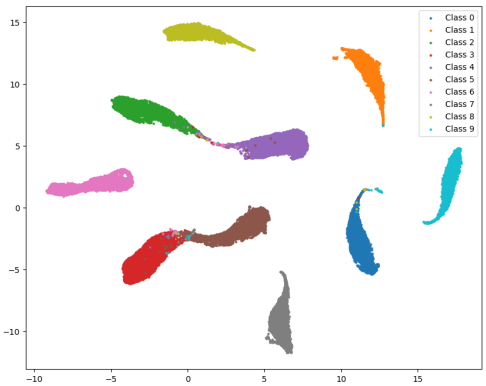


(d) CIFAR100 有攻击频率图

图 4.9　不同数据集攻击前后的频率图



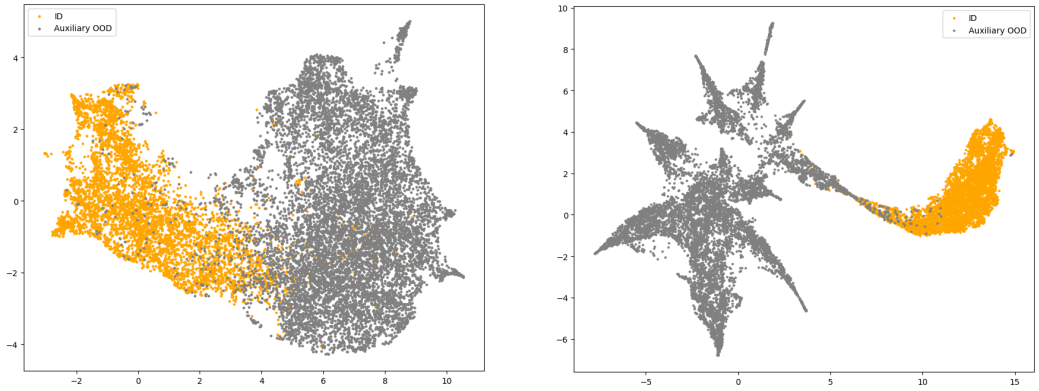
(a) 特征聚集前各类别 UMAP 可视化图



(b) 特征聚集后各类别 UMAP 可视化图

图 4.10　特征聚集前后各类别 UMAP 可视化图

叠部分更小，此时可以较为方便的在特征空间中利用距离衡量分布外数据与分布内数据的相似性，甚至可以简单的通过计算与分布内类别中心的距离来衡量相似性。因此，特征聚集有利于使用距离进行有效点的挑选。



(a) 特征聚集前分布内外 UMAP 可视化图　　**(b) 特征聚集后分布内外 UMAP 可视化图**

图 4.11　特征聚集前后分布内外数据 UMAP 可视化图

4.5　本章小结

本章在有额外分布外数据集的场景下，针对现有工作使用额外数据时会引入大量无用信息的问题，提出了基于有效点选取的鲁棒分布外检测方法 EPSD。EPSD 利用特征聚集和马氏距离对额外数据集进行有效的选取，同时为了保证模型鲁棒性结合了基于李雅普诺夫理论的神经 ODE 去噪器。实验表明，EPSD 在 CIFAR10 数据集上鲁棒分布外检测平均性能比现有方法提升 2.5%，CIFAR100 数据集上平均性能比现有方法提升 4.8%，面对攻击时性能下降不超过 5%，为有额外数据集场景下的鲁棒分布外检测提供了有效的方法。

第 5 章　无额外数据场景基于虚拟离群点生成的鲁棒分布外检测方法

上一章在具有额外数据集的场景下研究鲁棒分布外检测方法，本章考虑更为严苛的场景，即无额外数据集场景下的鲁棒分布外检测。在无额外数据集的情况下，该方法仅利用分布内数据集，在鲁棒特征空间中通过对抗性传播和非参数化采样合成虚拟离群点，从而弥补了分布外数据的缺失。

5.1　引言

在鲁棒分布外检测领域，大多数方法在训练期间采用额外的分布外数据集，通过结合离群值暴露和对抗训练的思想使得分布外检测器对分布外输入也具有鲁棒性。上一章中，本文也使用了额外辅助分布外数据，通过有效点选取和神经 ODE 去噪在对抗性设定下进行分布外检测。

但是，这类方法对额外分布外数据集都具有一定的要求[76,86]，需要额外分布外数据集具有多样性、相关性、规模性和语义不重叠性。多样性指的是分布外数据需覆盖潜在异常类型的广泛分布，避免模型过拟合到特定模式。相关性指分布外数据应该与分布内数据存在一定的语义或结构相似性，由此才能获得更为紧凑的决策边界。如动物图像作为分布内数据时，植物图像比数字图像更适合作为辅助分布外数据。规模性指辅助分布外数据集应有较大的样本数量。一般来说，辅助分布外数据的样本数量应该是分布内数据集的两倍及以上[9,19]。语义不重叠性指分布外数据的语义信息要确保不属于分布内类别以防止检测器产生语义混淆。

然而在一些情况下，满足上述条件的额外辅助分布外数据集是难以获得的。由于多样性和规模性的要求，需要收集数量足够多且语义足够丰富的分布外数据，由此可能产生较大的成本。再加上语义不重叠性的要求，收集分布外数据后需要进行数据清洗以剔除与分布内数据具有语义重复的样本，由此会进一步产生更大的成本。在一些特殊的领域，如将分布外检测用于罕见病检测时，由于此时分布外数据属于长尾分布中的罕见类别，收集成本极高且难以形成规模。并且由于相关性要求，若分布外数据与实际应用场景差距过大时 (如将自然图像模拟工业缺陷)，也会导致训练无效，因此也不能随便收集。AROS[65] 未使用分布外数据，其将特征空间假设为类条件高斯分布，并在低似然区域采样获得虚拟分布外数据，但该方法对特征空间具有较强的分布假设，并且虚拟分布外数据未考虑对抗性的因素。

为了在上述额外分布外数据集难以获得的场景下进行鲁棒的分布外检测，本章提出了无额外数据集场景下基于虚拟离群点生成的鲁棒分布外检测方法(Adversarial Propagation Non-parametric Outlier Synthesis, APNOS)。具体来说，该方法通过对抗性传播和非参数化采样来生成虚拟分布外嵌入，使用生成的虚拟分布外嵌入作为分布外输入的表示以对网络进行优化。为了增加模型鲁棒性，本章沿用了上一章中提到的神经 ODE 模块进行对抗扰动净化。本章方法还将模型的输出层替换为 logit 归一化层，使网络更加关注 logit 向量的方向以进一步的增强分布外检测能力。

5.2　问题描述

对于无额外鲁棒分布外数据的场景下，没有额外的辅助分布外数据集 D_{ood}^{aux} 可以利用，输入数据集仅包含分布内数据 D_{id}。在这样的情况下，鲁棒分布外检测器在测试阶段依然需要能够识别对抗分布外样本和对抗分布内样本。将对抗分布外样本的产生公式重写如下：

$$\begin{aligned} &\min_{\delta} \mathcal{L}\left(G(x_{adv}^{ood}), ID\right) \\ &\text{s.t. } x_{adv}^{ood} = x^{ood} + \delta,\ \|\delta\| \leqslant \epsilon \end{aligned} \tag{5.1}$$

其中 x^{ood} 表示分布外输入,δ 表示对抗扰动，其大小被 ϵ 约束。对抗分布内样本的产生在式4.1中已经给出。

鲁棒分布外检测器产生检测分数 $S(x)$，并基于此分数设置阈值 λ 判别输入属于分布内样本还是分布外样本。在对抗性的设定下，检测器需要满足：

$$G(x_{adv}) = \begin{cases} ID, & \text{如果 } S(x_{adv}) > \lambda \\ OOD, & \text{如果 } S(x_{adv}) < \lambda \end{cases} \tag{5.2}$$

在有额外辅助分布外样本的情况下，鲁棒分布外检测器可按式4.2优化，重写如下：

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D_{id}} \max_{\delta\sim B(x,\epsilon)} \mathcal{L}(G(x+\delta), \text{ID}) + \mathbb{E}_{(x,y)\sim D_{ood}^{aux}} \max_{\delta\sim B(x,\epsilon)} \mathcal{L}(G(x+\delta), \text{OOD}) \tag{5.3}$$

为了弥补 D_{ood}^{aux} 的缺失，可通过合成虚拟分布外样本的方法进行替代。即利用分布内样本，通过一定的机制获得虚拟的分布外样本集：

$$D_{ood}^{virtual} = \{x_i^{virtual} | x_i^{virtual} = g_\theta(D_{id}), i = 1, 2, \cdots\} \tag{5.4}$$

其中 g_θ 为由 θ 参数化的虚拟分布外样本产生器，利用分布内数据集 D_{id} 产生虚拟分布外数据 $x^{virtual}$。利用虚拟分布外数据集替代辅助分布外数据集，鲁棒分布

外检测器优化目标可写为：

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D_{id}} \max_{\delta\sim B(x,\epsilon)} \mathcal{L}(G(x+\delta),\mathrm{ID}) + \mathbb{E}_{(x,y)\sim D_{ood}^{virtual}} \max_{\delta\sim B(x,\epsilon)} \mathcal{L}(G(x+\delta),\mathrm{OOD}) \quad (5.5)$$

基于以上的分析，在无额外分布外数据的场景下进行鲁棒的分布外检测，问题的关键在于如何设计虚拟分布外样本生成器，生成良好的虚拟分布外样本以弥补分布外数据的缺失。若产生的虚拟分布外样本质量较差，与分布内样本有重叠或者距离分布内样本过远，会导致分布外检测器产生混淆或造成无效的训练。因此，通过合理的策略利用分布内数据合成有效的虚拟分布外数据是至关重要的。

5.3　算法设计

本章受基于距离的离群点检测方法的启发，通过离群点检测获取分布内数据集的边界样本，在该边界样本周围的低似然区域采样获得虚拟分布外样本辅助模型进行训练。本章沿用了上一章中所描述的神经 ODE 模块进行对抗净化以获得更强的对抗鲁棒性，但为了进一步增强分布外检测能力，本章还将 logit 归一化层与融入神经 ODE 去噪。

接下来的小节中，将详细介绍本章所提出的无额外数据集场景下的鲁棒分布外检测方法。

5.3.1　模型框架

模型整体框架如图5.1所示。训练框架分为三个阶段，第一个阶段采用鲁棒特征提取器进行特征提取，并利用对抗性传播使分布内数据向分布外偏移；第二个阶段通过非参数化的离群点检测获取分布内特征边界点，并在边界点周围的低似然区域采样获得虚拟分布外样本；第三个阶段训练神经 ODE 去噪模块并添加 logit 归一化层增强分布外检测能力。

在第一个阶段中，利用对抗攻击算法对分布内样本产生对抗性扰动，使分布内样本产生对抗性传播。再将对抗分布内样本经过鲁棒特征提取器以获得特征表示，鲁棒特征提取器由经过对抗训练的鲁棒分类器去掉最后一层后获得。其中，对抗性传播的目的是使得分布内样本向分布外的空间偏移，并能获得需要关注的不鲁棒分布内边界点，为采样做准备。

第二个阶段进行虚拟分布外数据的采样合成。在采样前需要获得分布内样本的边界点，因此利用非参数化的离群点检测方法获取分布内数据离群点，并以此作为分布内边界点。其中，非参数化的检测方法相较于其他的参数化方法具有更强的适用性。获取边界点后以该边界点作为中心在其周围低似然区域进行采

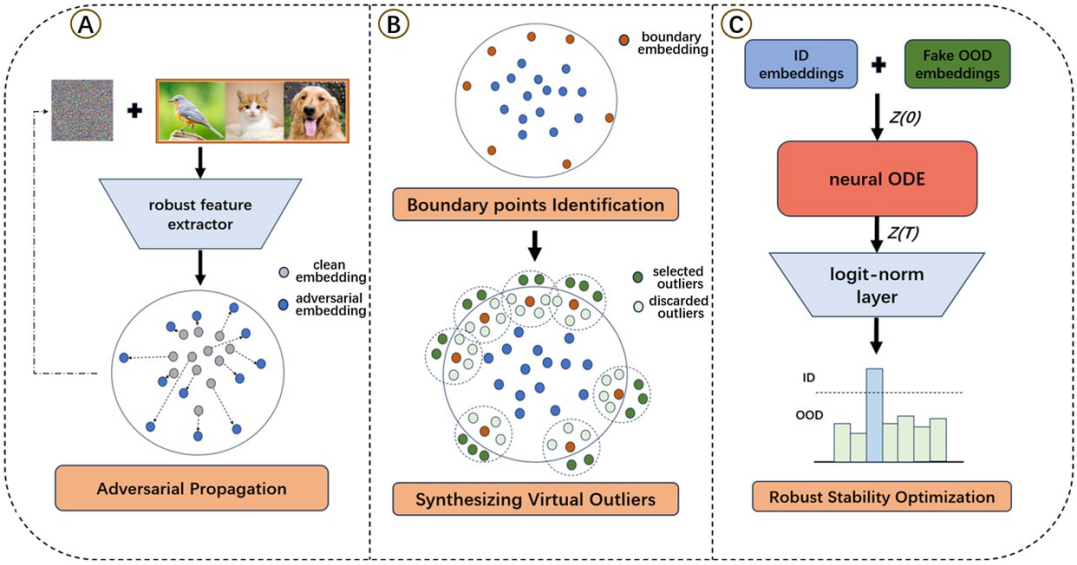


图 5.1　APNOS 方法框架图

样。采样后为了避免混淆剔除靠近分布内的点，将其余采样点收集起来作为合成的虚拟分布外数据用作后续训练。

第三个阶段训练神经 ODE 去噪器和 logit 归一化模块。神经 ODE 去噪器用于对噪声去噪，起到对抗净化的作用以增强模型鲁棒性。logit 归一化将 logit 向量归一化到超球面空间，使得模型更加关注 logit 输出的方向而不是大小，有助于缓解网络过度自信的问题。

5.3.2　对抗虚拟分布外数据点合成

针对辅助分布外数据的缺失，已有一些方法提出了通过生成虚拟分布外样本的策略以替代分布外数据。有的方法[30,64]利用 GAN 等生成模型生成虚拟图像作为代替，然而一些方法[10,33]指出在图像层面生成虚拟分布外样本是困难且没有必要的，以此提出了在特征层面生成虚拟嵌入的方法。本章也在特征层面进行虚拟分布外数据的合成，关键步骤如下：

（1） 对抗性传播

为了在特征空间进行合成，APNOS 利用了鲁棒特征提取器将输入映射到鲁棒特征空间。该鲁棒特征提取器利用分布内数据通过对抗训练再剔除网络结构最后一层获得。

在特征映射时，APNOS 还通过对抗传播的策略使得分布内样本在特征空间产生偏移，使其向分布外的空间移动从而更加靠近分布外。具体来说，对抗性传播利用 PGD 攻击对输入进行多次迭代，用公式表示如下：

$$x_{adv}^{t+1} = \mathrm{Proj}\left\{ x_{adv}^{t} + \alpha \cdot \mathrm{sign}(\nabla_{x_{adv}^{t}} S(x_{adv}^{t})) \right\} \tag{5.6}$$

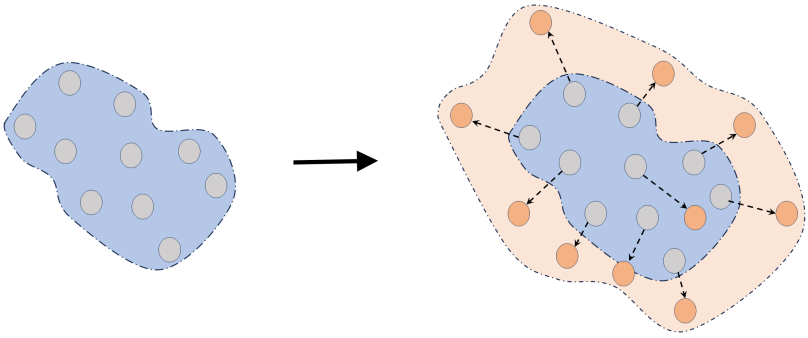


图 5.2　对抗传播示意图

其中，$S(x)$ 采用 MSP 分数以使得在迭代过程中分布内输入向分布外的空间中靠近。MSP 分数即最大 softmax 值，公式表示为：

$$S(x)=\max_{k}\frac{\exp\left(f_k(x)\right)}{\sum_{j=1}^{K}\exp\left(f_j(x)\right)} \tag{5.7}$$

其中 $f(x)$ 为鲁棒特征提取器输出向量。

对抗传播示意图如图5.2所示，其中灰色点为干净样本，蓝色区域为由干净样本所得到的分布内区域，橙色点为对抗传播后的嵌入点，橙色区域为对抗传播后分布内区域增加的部分。橙色点在语义上依然是分布内样本，如果不经过对抗性传播，仅利用干净样本获得的决策边界 (蓝色区域) 会导致部分橙色点被错误地判断为分布外。经过对抗性传播后，一些不鲁棒的分布内嵌入会在特征空间中发生显著的移位成为边界点，以对抗传播后的点构造决策边界，将减少对抗分布内输入的误判。

（2） 边界点鉴别

在鲁棒分布外检测领域，AROS 方法假设特征空间满足类条件高斯分布，并在高斯分布的低似然区域采样获取虚拟分布外嵌入。然而，类条件高斯分布是一个强力的假设，并不能随时满足。因此本章采用一种非参数化的方法，避免了高斯分布假设，具有更强的适用性。

将分布内数据映射到鲁棒特征空间后，需要获取分布内数据的边界点以进行虚拟分布外数据的采样，在本章方法中使用 k-近邻距离来鉴别离群值。具体来说，收集到分布内数据的特征嵌入集合 $Z_{id}=\{z_1,z_2,\cdots,z_N\}$ 后，对于任意属于类别 c 的嵌入 z^c，可计算其 k-近邻距离为:

$$d_{kNN}(z^c,Z_{id})=\left\|z^c-z^c_{(k)}\right\|_2 \tag{5.8}$$

其中 $z^c_{(k)}$ 为 z^c 在嵌入集合 Z_{id} 中的第 k 个相邻点。如果一个特征嵌入具有较大的 d_{kNN}，那么说明其在类别 c 中属于偏离集群的点，可被作为边界点。使用 k-近邻距离来鉴别边界点不依赖特征空间中的高斯分布假设，因此具有更强的适用性。

为了防止被选取的边界点集中于某一区域，可采用多轮选取的策略。每轮选取中，对于每个类别，先从分布内嵌入集合 Z_{id} 中随机挑选该类的 N 个候选点，计算所有候选点在该类的嵌入集合 Z_{id}^{c} 中 k-近邻距离并从大到小排序，选取最大的 n 个作为边界点。经过多轮选取将所有边界点收集起来形成边界点集 $\boldsymbol{B}$。

（3） 虚拟分布外数据点采样

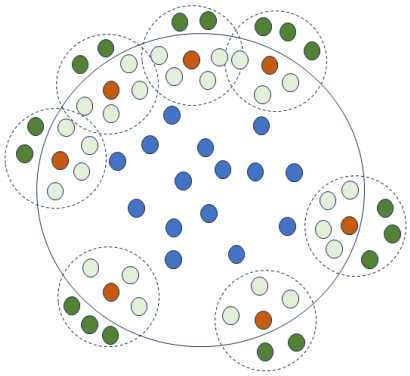


图 5.3　虚拟分布外数据点采样示意图

获取边界点集合后，便可以利用边界点在其周围的低似然区域采样获取虚拟分布外数据点。虚拟分布外数据点采样示意图如图5.3所示，其中橙色点为边界点，深绿色点为被选取的采样点，浅绿色点为被抛弃的采样点，虚线圆形表示采样高斯核。具体来说，对于任意一个边界点 $z' \in \boldsymbol{B}$，将其作为一个高斯核的中心，利用该高斯核在其周围采样：

$$\upsilon \sim \mathcal{N}(z', \sigma^2 \boldsymbol{I}) \tag{5.9}$$

其中 υ 表示以 z' 为中心的虚拟离群值，σ 调节高斯核方差。对于每个边界点 z' 而言，可在其周围采样获取离群值集合 $V = (\upsilon_1, \upsilon_2, ..., \upsilon_m)$。

在 V 中的点一部分靠近分布内区域，一部分远离分布内区域。靠近分布内区域的点可能会与真实分布内样本点产生混淆，因此应该将其剔除。保留从采样点集 V 中选择与分布内点集 Z_{id} 具有较高 k-近邻距离的 q 个点作为被选虚拟离群点, 将其他点进行剔除。将所有被选点视为合成的虚拟分布外数据点集 $Z_{ood}^{virtual}$。

5.3.3　神经 ODE 去噪器训练

上一章中使用到了基于李雅普诺夫稳定性理论的神经 ODE 去噪器进行对抗净化，本章继续沿用该模块以增强模型对抗鲁棒性。其中，神经 ODE 模块在李

雅普诺夫理论下需要进行以下优化：

$$\begin{aligned}\min_{\theta}\frac{1}{N}\sum_{k=0}^{N-1}\Bigg(&\mathcal{L}\left(\mathbf{z}_k(T),y_k\right)+\alpha_1\|f_{\boldsymbol{\theta}}\left(\mathbf{z}_k(0)\right)\|_2+\alpha_2 g_1\left(\sum_{i=1}^{n}[\nabla f_{\boldsymbol{\theta}}(\mathbf{z}_k(0))]_{ii}\right)\\&+\alpha_3 g_2\left(\sum_{i=1}^{n}(-|[\nabla f_{\boldsymbol{\theta}}(\mathbf{z}_k(0))]_{ii}|+\sum_{j\neq i}|[\nabla f_{\boldsymbol{\theta}}(\mathbf{z}_k(0))]_{ij}|)\right)\end{aligned}\tag{5.10}$$

其中 $z(0)$ 为神经 ODE 的输入，$z(\boldsymbol{T})$ 为其输出。

在具有额外数据集的情况下，神经 ODE 的输入为分布内数据特征点集与辅助分布外数据特征点集的并集 $Z_{train}=Z_{id}\cup Z_{ood}^{aux}$。在无额外数据集场景下，经过虚拟分布外数据点的合成之后，可用合成点集替代缺失的辅助分布外点集，即此时神经 ODE 输入为 $Z_{train}=Z_{id}\cup Z_{ood}^{virtual}$。

神经 ODE 模块之后通常连接全连接层以映射到标签空间，全连接层的输出称为 logit，记作 $h(z)$。logit 向量可以分解为两个部分:

$$h(z)=\|h(z)\|_2\cdot\hat{h}\left(z\right)\tag{5.11}$$

其中 $\|h(z)\|_2$ 为 logit 向量的 l_2 范数，$\hat{h}\left(z\right)$ 为其方向向量。

常规的交叉熵损失优化以下训练目标：

$$\mathcal{L}_{CE}(h(z),y)=-\log p(y|\boldsymbol{x})=-\log\frac{e^{\|\boldsymbol{h}(\boldsymbol{z})\|_2\cdot\hat{h}_y(z)}}{\sum_{i=1}^{k}e^{\|\boldsymbol{h}(\boldsymbol{z})\|_2\cdot\hat{h}_i(z)}}\tag{5.12}$$

可以发现训练损失取决于大小 $\|h(z)\|_2$ 和方向 $\hat{h}\left(z\right)$。固定方向向量分析大小对损失的影响。当 $y=\arg\max_i(f_i)$ 时，可以看到为了进一步增加 $-\log p(y|\boldsymbol{x})$，模型将会促使产生更大的 $\|h(z)\|_2$。这意味着，对于那些已经正确分类的训练样本，训练损失的优化将进一步增加网络输出 logit 的大小以产生更高的 softmax 置信度分数，从而获得更小的损失。这将鼓励模型对输入都倾向于产生越来越大的 logit 向量，从而产生过度自信的 softmax 分数导致分布内和分布外难以区分。

因此为了进一步提升分布外检测能力，本章方法在最后添加了 logit 归一化操作以减轻模型的过度自信问题。归一化操作使得计算损失时 $h(z)$ 被替换为其方向向量 $\hat{h}\left(z\right)$，这促使 logit 向量的方向与其 one-hot 标签对齐，而不关注其大小，有助于减轻网络的过度自信问题，进一步增强分布外检测能力。对于虚拟分布外样本，其理想输出是均匀分布，因此我们使用 KL 散度计算其余均匀分布之间的偏差。总的来说，式5.10中的第一项损失 $\mathcal{L}$ 形式化如下：

$$\mathcal{L}(z,y)=\begin{cases}-\log\frac{\mathrm{e}^{h_y(z)/\|h(z)\|_2}}{\sum_{i=1}^{k}\mathrm{e}^{h_i(z)/\|h(z)\|_2}} & , \quad y\in Y_{in}\\ \mathcal{L}_{CE}(\frac{h(z)}{\|h(z)\|_2},U_K) & , \quad y\in Y_{ood}\end{cases}\tag{5.13}$$

其中 $\boldsymbol{U}_K$ 为具有 $\boldsymbol{K}$ 个维度的均匀分布。损失函数根据输入是分布内还是分布外采用不同的形式。

5.3.4　分布外评价分数

在式5.13中，为了减轻模型的过度自信对输出 logit 进行了归一化操作。本章依然采用 MSP 作为分布外检测分数，在 logit 归一化后其 MSP 分数为：

$$S(x)=\max_{k}\frac{\exp\left(h_k(z)/\left\|h(z)\right\|_2\right)}{\sum_{j=1}^{K}\exp\left(h_j(z)/\left\|h(z)\right\|_2\right)} \tag{5.14}$$

具有较大 $S(x)$ 的输入样本应该判断为分布内，反之为分布外。选取阈值 γ 作为判断标准，则有：

$$G(x)=\begin{cases} ID, & \text{如果 } S(x_{adv})>\gamma \\ OOD, & \text{如果 } S(x_{adv})<\gamma \end{cases} \tag{5.15}$$

5.4　实验与分析

本节介绍实验设置与相应的结果分析，包括实验所采用的基准数据集、评价指标、实验设置和对比实验，并进行了进一步的消融和可视化结果分析。

5.4.1　数据集与评价指标

本章所使用的数据集与上一章保持一致。对于分布内数据集，使用鲁棒分布外检测领域常用的两个基准数据集 CIFAR10 和 CIFAR100 作为分布内数据集。对于分布外数据集，依然使用 LUSN、iSUN、SVHN 和 Places365 作为测试阶段的分布外数据集。与上一章不同的是，本章训练阶段没有使用辅助分布外数据集 Tiny-images，因此将其作为测试阶段数据集。由于 Tiny-images 与 CIFAR 数据集具有高相似性，因此可将其作为近-分布外的基准数据集。

本章依然是评价鲁棒分布外检测器区分对抗分布内样本和对抗分布外样本的能力，所以依然采用 AUROC 作为评价指标。

5.4.2　实验设置

对于实验中的参数设置，本章依然使用了鲁棒特征提取器，相关模型与参数与第 4 章一致。实验中生成与分布内数据点数一致的虚拟离群点，取 $k=300$，$N=800$，$n=100$，$m=600$，$q=2$，共进行 25 轮。神经 ODE 去噪器的超参数依然与上一章保持一致，取超参数为 $\alpha_1=1$，$\alpha_2=\alpha_3=0.05$，积分时间 $T=5$。输出模块使用两层全连接并最后添加归一化操作，在 CIFAR10 数据集上输出维度为 10，在 CIFAR100 数据集上输出维度为 100。使用 SGD 作为优化器，初始学习率为 0.05，学习率使用余弦学习率衰减，共训练 50 个 epoch, 批次大小为 128。

测试期间，在对抗性的设定下使用 PGD 攻击来生成对抗样本。由于采用 logit 归一化后的 MSP 分数作为评价分数，根据第三章的分数自适应攻击方法，对于

分布内样本应最小化其 MSP 分数，对于分布外样本应最大化其 MSP 分数。为了保证扰动强力，PGD 攻击参数设置为扰动边界 $\epsilon = 8/255$，攻击步数 $t = 100$，迭代步长 $\alpha = 2/255$。

5.4.3　基准数据集鲁棒分布外检测对比实验

本小节分别以 CIFAR10 和 CIFAR100 作为分布内数据集，与多种无额外分布外数据的方法进行对比以证明本章所提方法的有效性。在对比的方法中，AT 表示经过对抗训练的分类器并以 MSP 作为检测分数；CSI、VOS 和 CIDER 均是干净样本下不使用额外数据的分布外检测方法，其中 CSI 作为数据增强的代表，VOS 作为合成分布外数据的代表，CIDER 作为特征投影的代表；OSAD 是开集识别中考虑对抗样本的方法，可迁移至鲁棒分布外检测中；AROS 是鲁棒分布外检测中唯一不使用额外分布外数据的方法。

（1） 分布内数据为 CIFAR10 的对比实验

表5.1展示了分布内数据为 CIFAR10 时的对比实验结果。与上一章采用同样的数据格式，即斜线左侧为干净样本的 AUROC 指标，斜线右侧为对抗样本的 AUROC 指标。每一列表示在不同的分布外数据集下的结果，其中 CIFAR100 和 Tiny-images 为近-分布外数据集，其余为远-分布外数据集。

表 5.1　CIFAR10 数据集上攻击前后 AUROC 变化表

方法	CIFAR100	Tiny-images	SVHN	LSUN	iSUN	Places365	average
AT	60.9/31.1	61.7/32.9	73.3/28.2	68.2/31.9	69.8/34.6	52.3/26.4	64.0/30.9
CSI	92.2/2.1	91.2/1.6	97.4/1.6	97.7/0.0	95.4/3.5	93.6/0.1	94.9/1.6
VOS	86.2/3.5	88.6/3.0	88.7/1.6	94.3/2.1	92.9/1.1	89.7/1.8	89.6/1.9
CIDER	90.8/1.1	92.1/0.8	99.7/1.9	99.0/0.6	96.6/2.6	94.0/2.2	95.7/1.5
OSAD	79.9/18.1	81.9/19.3	88.4/26.5	86.4/20.7	84.0/20.3	83.3/21.2	84.4/20.9
AROS	88.2/80.4	89.7/81.1	93.0/87.1	90.6/82.6	88.9/81.7	90.8/83.9	90.0/82.8
APNOS	88.6/**85.2**	88.0/**84.3**	91.8/**89.3**	88.1/**85.7**	86.7/**83.9**	90.3/**87.6**	88.4/**85.9**

分析表中数据可知，当 CIFAR-10 作为分布内数据集时，无论是在近-分布外数据集下还是远-分布外数据集下，APNOS 都取得了最好的鲁棒分布外检测性能，体现了其性能优越性。其中，在近-分布外数据集下鲁棒分布外检测性能提升 5.0%，平均鲁棒分布外检测性能提升 3.7%，攻击前后性能仅下降 2.8%。AROS 与 APNOS 在鲁棒分布外检测性能上较为突出，这是由于两者都使用了神经 ODE 去噪器模块，体现了神经 ODE 模块在鲁棒分布外检测任务中的优越性能。由于 APNOS 使用了更为有效的虚拟离群点生成方法，并使用 logit 归一化操作增强分

布外检测能力，使得 APNOS 比 AROS 的性能有了进一步的提升。但是 APNOS 的干净样本性能相比其他方法有略微的劣势，这是由于鲁棒性提升会导致一定程度的干净性能的下降[84-85]，鲁棒性能与干净性能之间存在一定的权衡关系。

（2） 分布内数据为 CIFAR100 的对比试验

表5.2展示了分布内数据为 CIFAR100 时的对比实验结果，其中近-分布内数据集为 CIFAR-10 和 Tiny-images。

表 5.2　CIFAR100 数据集上攻击前后 AUROC 变化表

方法	CIFAR10	Tiny-images	SVHN	LSUN	iSUN	Places365	average
AT	55.8/42.7	55.0/41.5	65.2/55.0	60.6/49.6	59.7/50.3	57.4/48.3	58.0/48.3
CSI	53.2/0.9	55.6/1.3	90.5/4.7	63.4/2.2	81.4/3.1	73.6/0.3	70.7/2.1
VOS	78.7/1.2	81.3/1.6	86.6/0.6	92.9/4.0	70.2/3.9	80.2/1.4	81.9/2.1
CIDER	73.3/4.4	72.8/4.2	95.1/5.1	96.3/1.0	82.9/3.4	73.4/2.9	83.6/3.3
OSAD	50.3/10.3	48.3/9.9	61.8/13.9	55.6/10.4	54.8/10.6	55.7/12.1	54.3/10.9
AROS	74.3/67.3	76.5/69.8	81.5/70.8	77.0/69.5	72.8/68.5	77.0/69.4	76.4/69.2
APNOS	75.6/**73.6**	76.4/**72.5**	76.8/**74.7**	78.1/**74.5**	75.3/**72.2**	78.2/**76.5**	76.9/**73.8**

分析表中数据可知，当 CIFAR-100 作为分布内数据集时，APNOS 依然在各种分布外数据集下都取得了最好的结果。在 CIFAR100 实验中，APNOS 相比于 AROS 在近-分布外数据集下性能提升了 6.7%，平均性能提升 6.6%，比 CIFAR10 实验中 (5.0% 与 3.7%) 有更大的提升幅度，攻击前后性能仅下降 4.0%。且 CIFAR100 实验中的干净性能与 AROS 持平，与表5.1中 CIFAR10 的结果相比，表现出 APNOS 在更复杂数据集下的优越性能。

5.4.4　消融实验

本小节通过对模型结构中的核心组件和步骤进行消融实验，以展示本章方法中各个组件和步骤的重要性。具体来说，本小节对鲁棒特征提取、神经 ODE 去噪器、虚拟离群点生成和 logit 归一化进行消融分析。对鲁棒特征提取的消融采用普通分类器去掉最后一层后代替；对神经 ODE 的消融采用无正则化的普通残差块代替；对虚拟离群点生成的消融采用不产生虚拟离群点来代替；对 logit 归一化的消融采用原始 logit 向量代替。实验中以 CIFAR10 作为分布内数据集，CIFAR100 作为分布外数据集，攻击参数为扰动边界 $\epsilon = 8/255$，攻击步数 $t = 100$，迭代步长 $\alpha = 2/255$，以 AUROC 作为性能指标。消融结果如表5.3所示。

设置 A 和 B 对鲁棒特征提取器和神经 ODE 去噪器进行消融，这两个模块起到对抗去噪的作用，消融后性能产生了大幅度的下降，体现了鲁棒去噪的重要

表 5.3　模块消融数据表

设置	鲁棒特征提取	神经 ODE 去噪	虚拟离群点生成	logit 归一化	AUROC
A	-	✓	✓	✓	57.7
B	✓	-	✓	✓	24.9
C	✓	✓	-	✓	78.4
D	✓	✓	✓	-	81.7
E(ours)	✓	✓	✓	✓	85.9

性；设置 C 取消了虚拟离群点生成步骤，使得训练过程中缺乏分布外信息，模型缺失对分布外输入的平衡性训练，导致针对分布外样本的鲁棒分布外检测能力下降；设置 D 取消了 logit 归一化操作，使得模型更加容易产生过度自信现象，因此性能产生下降；E 为本章提出的方法，综合了所有模块获得了最好的性能。

5.4.5　超参数分析

本小节通过控制变量实验分析本章方法中超参数对于方法性能产生的影响。分析的超参数包括：k-近邻距离中的距离基数 k；采样高斯核的方差 σ^2；每轮获取的边界点数 n。实验均以 CIFAR10 作为分布内数据集，以 CIFAR100 作为分布外数据集，在同时攻击分布内和分布外数据的情况下以 AUROC 作为评价指标。

（1） 距离基数分析实验

在计算 k-近邻距离时，通过选取不同的距离基数分析其对方法性能的影响，结果如图5.4所示。

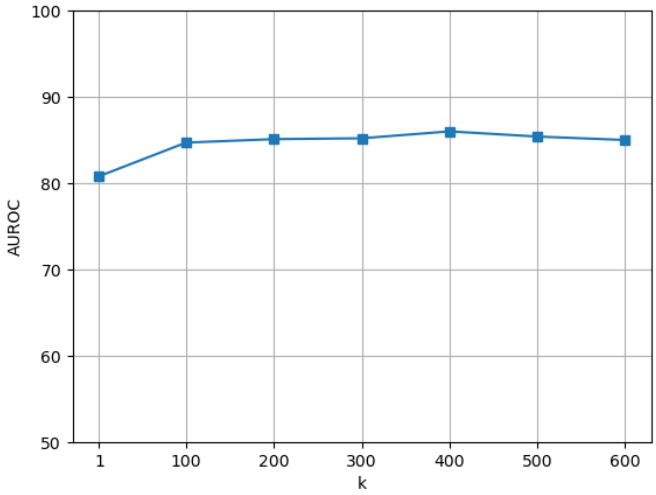


图 5.4　不同距离基数的 AUROC 性能曲线图

观察图可知，在较大的合适范围内不同 k 的取值只会引起性能小幅度的波动，但过小的 k 值会引起性能一定程度的下降。这是因为 k 过小时容易将非边界

的点识别为边界点，从而导致在非边界区域采样。因此在合适的范围内选取 k 值不会对性能产生较大影响，但要避免较为极端的取值。

（2） 采样方差分析实验

进行虚拟离群点合成时，需要在边界点处利用方差为 σ^2 的高斯核进行采样，不同的高斯核方差影响采样点的离散程度。通过选取不同的采样方差分析其对性能的影响，结果如图5.5所示。

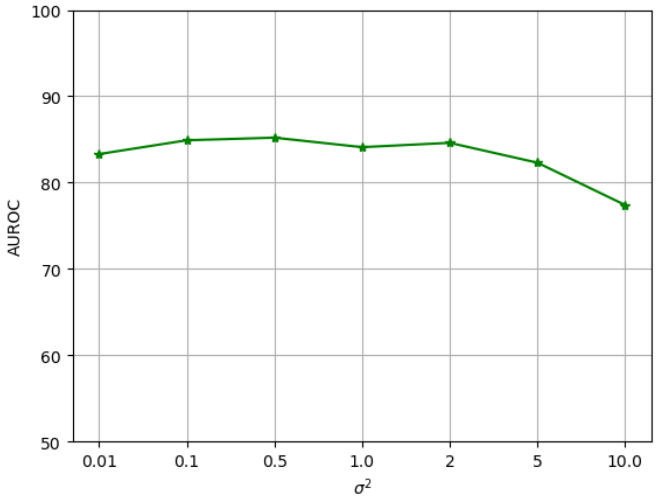


图 5.5　不同采样方差的 AUROC 性能曲线图

分析图像可知，在采样方差取 0.01 至 2 之间时，采样方差的变化只会引起性能的小幅波动，而在方差为 10 时，性能产生了较大幅度的下降。这是由于将采样点作为虚拟分布外样本时需要避免与分布内产生重叠混淆，当采样方差过大时采样点会较远偏离当前边界，甚至与其他类别的特征产生混淆，从而导致性能的下降。

（3） 边界点数分析实验

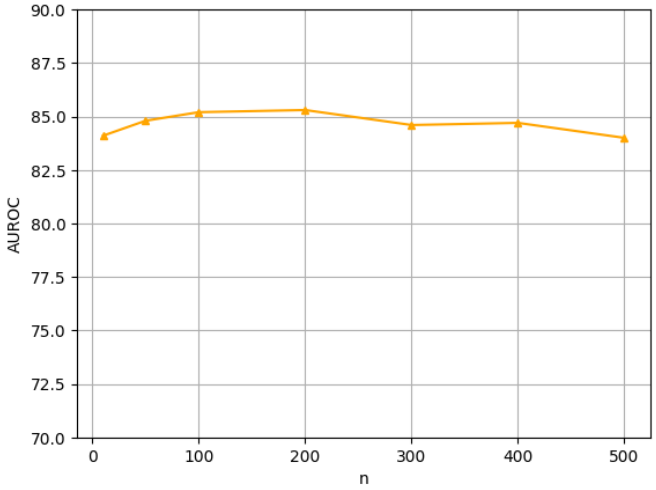


图 5.6　不同边界点数的 AUROC 性能曲线图

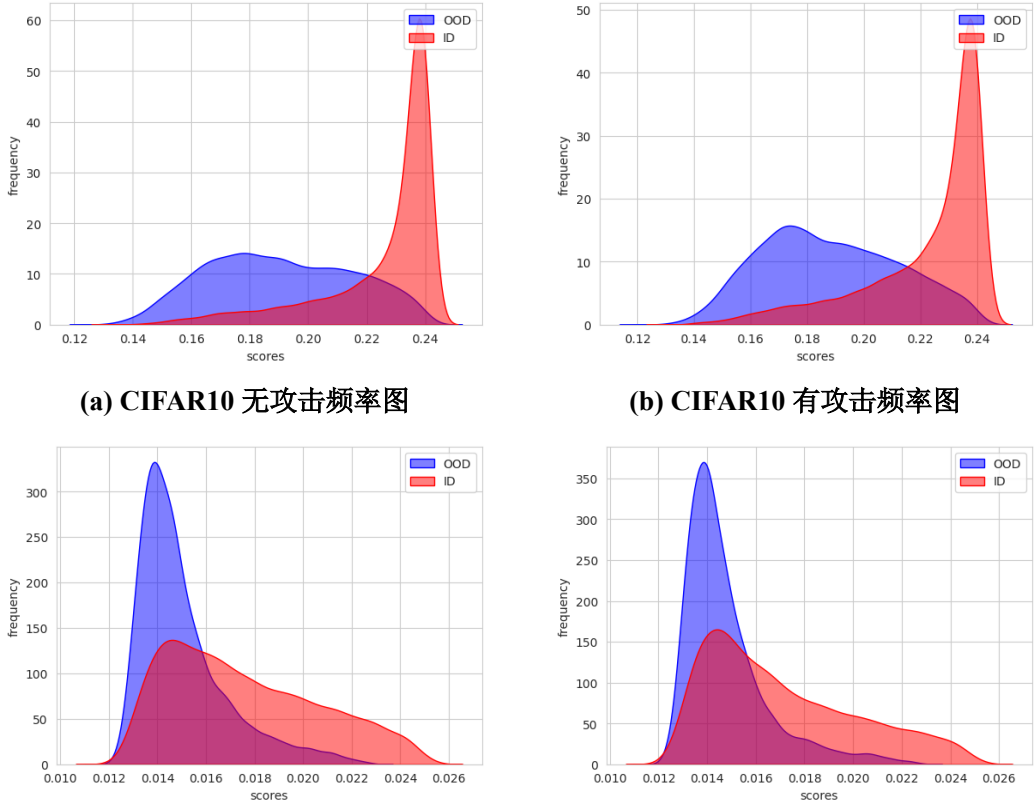
在合成虚拟点前，需要先选取分布内样本的边界点，选取的策略是每轮从 $N = 800$ 个候选点中选取具有较高 k-近邻距离的 n 个作为边界点。通过改变不同的边界点数分析其对性能的影响，结果如图5.6所示。

分析图像可知，随着 n 从小到大的变化，性能曲线整体呈现先增后减的趋势。由于 N 固定，当 n 过小时多轮边界点的选取容易选取到相同的点，因此产生的虚拟离群点就不够多样，容易聚集在某些边界点周围。当 n 过大时，每轮会挑选过多的边界点，因此边界点的选取就越近似于随机选择，不能足够的代表边界。但曲线整体波动较小，性能对 n 并不敏感，因此只需要避免选择过大或过小的极端取值即可。

5.4.6　可视化分析

本小节通过可视化分析来直观展现本章方法的特性，其中包括攻击前后的分布外检测频率分布图和虚拟离群点的 UMAP 可视化图。

（1） 分布外检测分数可视化



(a) CIFAR10 无攻击频率图　　**(b) CIFAR10 有攻击频率图**

(c) CIFAR100 无攻击频率图　　**(d) CIFAR100 有攻击频率图**

图 5.7　不同数据集攻击前后的频率图

为直观展现对抗攻击的影响及本章方法的有效性，分别在 CIFAR10 和 CIFAR100 作为分布内样本时获取攻击前后的分布外检测分数频率分布图。实验参

数与对比试验中一致。检测分数可视化结果如图5.7所示。

分析图像可知，在进行攻击之后，检测分数仅产生了小幅度的偏移，说明本章方法拥有较好的抵抗对抗分布外攻击的能力，在面对对抗攻击时仍然能保持较好的分布外检测能力，进一步说明了本章方法的有效性。

（2） UMAP 可视化

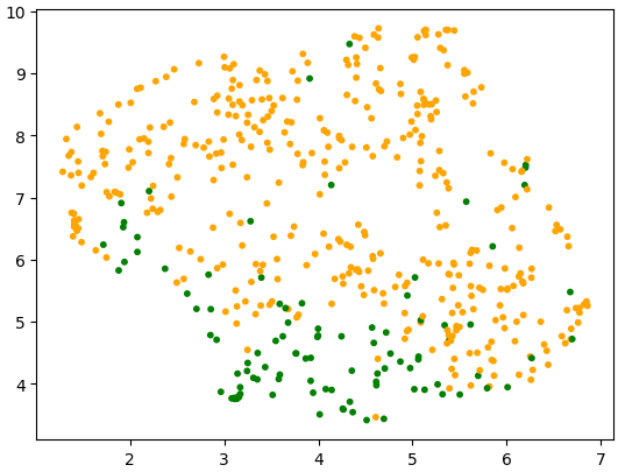


图 5.8　边界点检测 UMAP 可视化图

为直观展现边界点鉴别和虚拟分布外数据点采样的效果，将特征点进行 UMAP 降维可视化。选取一个轮次中的边界点鉴别过程，将其 UMAP 可视化结果展示如图5.8所示，其中绿色点表示被本轮次被鉴别为边界的点，橘黄色为本轮次中的其他点。

观察图5.8可知，通过计算 k-近邻距离并选取具有较大 k-近邻距离的点可获得特征空间中位于边界区域的点，由此可以利用边界点进行虚拟离群值的生成。每一轮次只能获取一部分特定区域的边界点，因此为了能够获取多样的边界点，需要采用对分布内数据随机采样和多轮获取的策略。

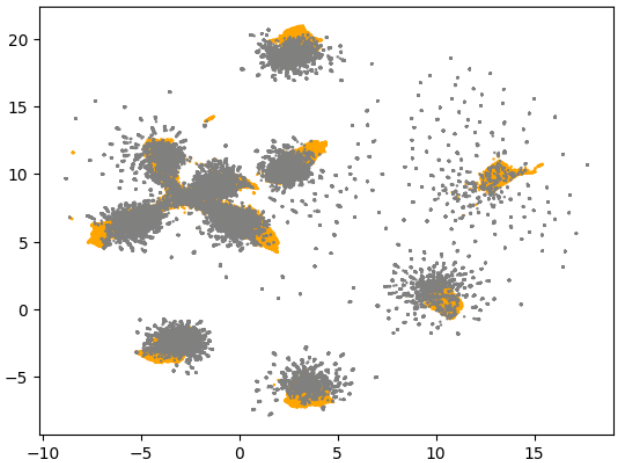


图 5.9　虚拟离群点 UMAP 可视化图

将所有分布内数据点和生成的虚拟离群值可视化如图5.9所示，其中橘黄色

为分布内数据点，灰色为生成的虚拟离群值。观察图像可知，生成的虚拟离群点大多位于分布内数据点周围，有利于获得更加紧凑的决策边界。再加上对抗性传播，生成的虚拟离群点位于不鲁棒区域，有利于使得模型增加鲁棒性。

5.5　本章小结

本章针对现有工作生成虚拟分布外数据时未考虑鲁棒性因素并且具有较强假设依赖的问题，提出了基于虚拟离群点生成的鲁棒分布外检测方法 APNOS。APNOS 通过对抗性传播和非参数化的虚拟特征采样进行虚拟离群点的生成，同时为了保证模型鲁棒性及削弱过度自信问题将 logit 归一化融入神经 ODE 去噪器。实验表明，APNOS 在 CIFAR10 数据集上鲁棒分布外检测平均性能比现有方法提升 3.7%，在 CIFAR100 数据集上提升 6.6%，且面对攻击时性能下降不超过 4%，为无额外数据集下的鲁棒分布外检测提供了有效的方法。

第 6 章　总结与展望

6.1　本文工作总结

在开放世界场景中，深度学习模型的安全性会受到分布外输入和恶意攻击的共同威胁。鲁棒分布外检测作为一项在对抗性设定下进行分布外检测的任务，对于模型的安全性有着重要的作用。本文对鲁棒分布外检测进行了系统的研究，提出了一种分布外检测的鲁棒性评估方法，并在两种场景下设计鲁棒分布外检测算法。具体包括：

(1) 提出了一种基于检测分数的分布外检测自适应攻击方法。该方法针对现有分布外检测的鲁棒性评估不够完善和强力的问题，同时考虑对抗分布内和对抗分布外样本，根据不同的检测分数自适应调整优化目标。利用该自适应攻击方法攻击现有分布外检测算法发现，现有分布外检测分布外在被攻击后性能大幅下降超 70%，揭示了现有分布外检测算法的脆弱性。实验还分析了不同的超参数对于攻击强度的影响，并给出了适当的参数设置范围，为分布外检测的鲁棒性评估提供了参考。

(2) 在有额外分布外数据集的场景下提出了基于有效点选取的鲁棒分布外检测方法 EPSD。EPSD 针对现有工作使用额外数据时会引入大量无用信息的问题，通过原型学习进行特征聚集，并在聚集后使用最近邻马氏距离对额外数据集进行有效点的选取。为了保证模型鲁棒性，EPSD 还结合了基于李雅普诺夫理论的神经 ODE 去噪器，将有效点和分布内数据点规范到平衡点附近以进行对抗净化。实验表明，EPSD 在 CIFAR10 数据集上鲁棒分布外检测性能比现有方法提升 2.5%，在 CIFAR100 数据集上比现有方法提升 4.8%，且面对攻击时性能下降不超过 5%，为有额外数据集下的鲁棒分布外检测提供了有效的方法。

(3) 在无额外分布外数据集的场景下提出了基于虚拟离群点生成的鲁棒分布外检测方法 APNOS。APNOS 针对现有工作生成虚拟分布外数据时未考虑鲁棒性因素并且具有较强假设依赖的问题，通过对抗性传播将分布内特征驱使到不鲁棒区域，利用非参数化的方法进行边界点鉴别及虚拟特征采样以获得虚拟离群点。为了保证模型鲁棒性及削弱过度自信问题，APNOS 将 logit 归一化融入神经 ODE 去噪器，使模型只需要关注输出向量的方向。实验表明，APNOS 在 CIFAR10 数据集上鲁棒分布外检测性能比现有方法提升 3.7%，在 CIFAR100 数据集上比现有方法提升 6.6%，且面对攻击时性能下降不超过 4%，为无额外数据集下的鲁棒分布外检测提供了有效的方法。

6.2　未来工作展望

本文工作虽对鲁棒分布外检测进行了系统性的研究，但仍然存在能够继续深入研究的地方，具体体现在：

(1) 针对分布外检测的黑盒攻击方法。本文提出的基于分数的自适应攻击方法需要知道模型参数及检测分数，是一种白盒攻击方法，可用作鲁棒性评估。但在更加现实的场景下，黑盒攻击更符合现实系统部署时的应用场景，模型在部署时遭遇的往往是黑盒攻击。因此，在未来的工作中，可更加深入研究针对分布外检测的黑盒攻击方法，以加深该领域的现实意义。

(2) 针对更复杂模型的鲁棒分布外检测研究。本文所有算法均针对于图像分类模型，这是因为图像分类任务是许多其他视觉任务如目标检测、语义分割、图像分析等的基础。因此，基于在图像分类模型下的研究，可将鲁棒分布外检测延伸到其他更加复杂的模型下，以扩展该领域的应用广度。

(3) 大规模数据集下的鲁棒分布外检测研究。本文实验中所使用的数据集都是较小规模的数据集，这是因为在和对抗性相关的领域，对抗样本的生成会耗费巨大的时间成本。因此，在未来的研究中，若得益于计算效率的提升，可在大规模数据集下进行鲁棒分布外检测的研究。

参考文献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[2] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. 2021: 1-12.

[3] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 427-436.

[4] HENDRYCKS D, GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[C]//International Conference on Learning Representations. 2017: 1-12.

[5] LIANG S, LI Y, SRIKANT R. Enhancing the reliability of out-of-distribution image detection in neural networks[C]//International Conference on Learning Representations. 2018: 1-11.

[6] YANG J, ZHOU K, LI Y, et al. Generalized out-of-distribution detection: A survey[J]. International Journal of Computer Vision, 2024, 132(12): 5635-5662.

[7] SALEHI M, MIRZAEI H, HENDRYCKS D, et al. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges[J]. Transactions on Machine Learning Research, 2022(234).

[8] CHEN J, LI Y, WU X, et al. Robust out-of-distribution detection for neural networks[C]//The AAAI-22 Workshop on Adversarial Machine Learning and Beyond. 2022: 1-10.

[9] CHEN J, LI Y, WU X, et al. Atom: Robustifying out-of-distribution detection using outlier mining[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2021: 430-445.

[10] AZIZMALAYERI M, SOLTANI MOAKHAR A, ZAREI A, et al. Your out-of-distribution detection method is not robust![J]. Advances in Neural Information Processing Systems, 2022, 35: 4887-4901.

[11] LIU W, WANG X, OWENS J, et al. Energy-based out-of-distribution detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21464-21475.

[12] LEE K, LEE K, LEE H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks[J]. Advances in Neural Information Processing Systems, 2018, 31: 7167-7177.

[13] SUN Y, MING Y, ZHU X, et al. Out-of-distribution detection with deep nearest neighbors[C]//

International Conference on Machine Learning. 2022: 20827-20840.

[14] WANG H, LI Z, FENG L, et al. Vim: Out-of-distribution with virtual-logit matching[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4921-4930.

[15] HUANG R, GENG A, LI Y. On the importance of gradients for detecting distributional shifts in the wild[J]. Advances in Neural Information Processing Systems, 2021, 34: 677-689.

[16] SUN Y, GUO C, LI Y. React: Out-of-distribution detection with rectified activations[J]. Advances in Neural Information Processing Systems, 2021, 34: 144-157.

[17] DJURISIC A, BOZANIC N, ASHOK A, et al. Extremely simple activation shaping for out-of-distribution detection[C]//International Conference on Learning Representations. 2023: 1-12.

[18] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. 2021: 8748-8763.

[19] HENDRYCKS D, MAZEIKA M, DIETTERICH T G. Deep anomaly detection with outlier exposure[C]//International Conference on Learning Representations. 2019: 1-11.

[20] YU Q, AIZAWA K. Unsupervised out-of-distribution detection by maximum classifier discrepancy[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9518-9526.

[21] YANG J, WANG H, FENG L, et al. Semantically coherent out-of-distribution detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8301-8309.

[22] ZHANG J, INKAWHICH N, LINDERMAN R, et al. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 5531-5540.

[23] MING Y, CAI Z, GU J, et al. Delving into out-of-distribution detection with vision-language representations[J]. Advances in Neural Information Processing Systems, 2022, 35: 35087-35102.

[24] WANG H, LI Y, YAO H, et al. Clipn for zero-shot ood detection: Teaching clip to say no[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 1802-1812.

[25] NIE J, ZHANG Y, FANG Z, et al. Out-of-distribution detection with negative prompts[C]// International Conference on Learning Representations. 2024: 1-12.

[26] HENDRYCKS D, MAZEIKA M, KADAVATH S, et al. Using self-supervised learning can improve model robustness and uncertainty[J]. Advances in Neural Information Processing Systems, 2019, 32: 15637-15648.

[27] TACK J, MO S, JEONG J, et al. Csi: Novelty detection via contrastive learning on distribu-

tionally shifted instances[J]. Advances in Neural Information Processing Systems, 2020, 33: 11839-11852.

[28] MING Y, SUN Y, DIA O, et al. How to exploit hyperspherical embeddings for out-of-distribution detection?[C]//International Conference on Learning Representations. 2023: 1-13.

[29] LU H, GONG D, WANG S, et al. Learning with mixture of prototypes for out-of-distribution detection[C]//International Conference on Learning Representations. 2024: 1-15.

[30] LEE K, LEE H, LEE K, et al. Training confidence-calibrated classifiers for detecting out-of-distribution samples[C]//International Conference on Learning Representations. 2018: 1-10.

[31] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.

[32] CHEN G, PENG P, WANG X, et al. Adversarial reciprocal points learning for open set recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8065-8081.

[33] KONG S, RAMANAN D. Opengan: Open-set recognition via open data generation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 813-822.

[34] DU X, WANG Z, CAI M, et al. VOS: learning what you don't know by virtual outlier synthesis [C]//International Conference on Learning Representations. 2022: 1-14.

[35] TAO L, DU X, ZHU J, et al. Non-parametric outlier synthesis[C]//International Conference on Learning Representations. 2023: 1-14.

[36] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C]//International Conference on Learning Representations. 2014: 1-10.

[37] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. 2015: 1-11.

[38] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world [C]//International Conference on Learning Representations. 2017: 1-11.

[39] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//International Conference on Learning Representations. 2018: 1-18.

[40] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.

[41] CARLINI N, WAGNER D A. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy. 2017: 39-57.

[42] CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]//International Conference on Machine Learning: Vol. 119. 2020:

2206-2216.

[43] CHEN P Y, ZHANG H, SHARMA Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.

[44] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]//International Conference on Learning Representations. 2018: 1-12.

[45] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C]//International Conference on Machine Learning. 2018: 2137-2146.

[46] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765-1773.

[47] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.

[48] WANG Z, GUO H, ZHANG Z, et al. Feature importance-aware transferable adversarial attacks [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7639-7648.

[49] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[C]//International Conference on Learning Representations. 2018: 1-14.

[50] HUANG B, WANG Y, WANG W. Model-agnostic adversarial detection by random perturbations[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019: 4689-4696.

[51] CINTAS C, SPEAKMAN S, AKINWANDE V, et al. Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020: 876-882.

[52] HUANG R, XU B, SCHUURMANS D, et al. Learning with a strong adversary[A]. 2016. arXiv: 1511.03034.

[53] CAI Q, LIU C, SONG D. Curriculum adversarial training[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018: 3740-3747.

[54] SHAFAHI A, NAJIBI M, GHIASI M A, et al. Adversarial training for free![J]. Advances in Neural Information Processing Systems, 2019, 32: 3353-3364.

[55] WONG E, RICE L, KOLTER J Z. Fast is better than free: Revisiting adversarial training[C]// International Conference on Learning Representations. 2020: 1-13.

[56] SONG Y, KIM T, NOWOZIN S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples[C]//International Conference on Learning Representations. 2018: 1-14.

[57] SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[C]//International Conference on Learning Representations. 2018: 1-12.

[58] LIAO F, LIANG M, DONG Y, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1778-1787.

[59] YAN H, DU J, TAN V Y F, et al. On robustness of neural ordinary differential equations[C]// International Conference on Learning Representations. 2020: 1-12.

[60] LI X, XIN Z, LIU W. Defending against adversarial attacks via neural dynamic system[J]. Advances in Neural Information Processing Systems, 2022, 35: 6372-6383.

[61] KANG Q, SONG Y, DING Q, et al. Stable neural ode with lyapunov-stable equilibrium points for defending against adversarial attacks[J]. Advances in Neural Information Processing Systems, 2021, 34: 14925-14937.

[62] HEIN M, ANDRIUSHCHENKO M, BITTERWOLF J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 41-50.

[63] SEHWAG V, BHAGOJI A N, SONG L, et al. Analyzing the robustness of open-world machine learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. 2019: 105-116.

[64] MIRZAEI H, JAFARI M, DEHBASHI H R, et al. Rodeo: Robust outlier detection via exposing adaptive out-of-distribution samples[C]//International Conference on Machine Learning. 2024: 35744-35778.

[65] MIRZAEI H, MATHIS M W. Adversarially robust out-of-distribution detection using lyapunov-stabilized embeddings[C]//International Conference on Learning Representations. 2025: 1-18.

[66] ZHANG H, YU Y, JIAO J, et al. Theoretically principled trade-off between robustness and accuracy[C]//International Conference on Machine Learning. 2019: 7472-7482.

[67] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization[C]// International Conference on Learning Representations. 2018: 1-12.

[68] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. Technical Report, Citeseer, 2009.

[69] TORRALBA A, FERGUS R, FREEMAN W T. 80 million tiny images: A large data set for nonparametric object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958-1970.

[70] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]//NIPS Workshop on Deep Learning and Unsupervised Feature Learning: Vol. 2011. 2011: 1-5.

[71] YU F, SEFF A, ZHANG Y, et al. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop[A]. 2016. arXiv: 1506.03365.

[72] XU P, EHINGER K A, ZHANG Y, et al. Turkergaze: Crowdsourcing saliency with webcam based eye tracking[A]. 2015. arXiv: 1504.06755.

[73] CIMPOI M, MAJI S, KOKKINOS I, et al. Describing textures in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3606-3613.

[74] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1452-1464.

[75] YANG J, WANG P, ZOU D, et al. Openood: Benchmarking generalized out-of-distribution detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 32598-32611.

[76] SHAFAEI A, SCHMIDT M, LITTLE J J. A less biased evaluation of out-of-distribution sample detectors[C]//In Proceedings of the British Machine Vision Conference. 2019: 23-36.

[77] CHEN R T, RUBANOVA Y, BETTENCOURT J, et al. Neural ordinary differential equations [J]. Advances in Neural Information Processing Systems, 2018: 6572-6583.

[78] DUPONT E, DOUCET A, TEH Y W. Augmented neural odes[J]. Advances in Neural Information Processing Systems, 2019: 3134-3144.

[79] GRATHWOHL W, CHEN R T, BETTENCOURT J, et al. Ffjord: Free-form continuous dynamics for scalable reversible generative models[C]//International Conference on Learning Representations. 2019: 1-10.

[80] YANG H M, ZHANG X Y, YIN F, et al. Robust classification with convolutional prototype learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3474-3482.

[81] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[82] CROCE F, ANDRIUSHCHENKO M, SEHWAG V, et al. Robustbench: a standardized adversarial robustness benchmark[A]. 2021. arXiv: 2010.09670.

[83] WANG Z, PANG T, DU C, et al. Better diffusion models further improve adversarial training [C]//International Conference on Machine Learning. 2023: 36246-36263.

[84] TSIPRAS D, SANTURKAR S, ENGSTROM L, et al. Robustness may be at odds with accuracy[C]//International Conference on Learning Representations. 2019: 1-12.

[85] MOAYERI M, BANIHASHEM K, FEIZI S. Explicit tradeoffs between adversarial and natural distributional robustness[J]. Advances in Neural Information Processing Systems, 2022, 35: 38761-38774.

[86] HSU Y C, SHEN Y, JIN H, et al. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10951-10960.

致谢

行文至此，意味着三年的硕士生涯即将画上句号，也意味着我即将离开这个生活了七年的学校，意味着求学生涯将告一段落。我从贵州的大山里出发，这一路遇到过很多坎坷。这途中，我的包袱越来越重，但得益于许多人的关心和帮助，得益于他们在背后支撑着我，在前面拉我一把，使得我即将踏入更为广阔的天地。今天，我想用平淡的话语，在此一一感谢。

感谢我的导师赵云波教授。还记得当初保研时给赵老师发邮件，他当晚就约了我在中区一起散步，我顿时就被赵老师的真诚和亲和力所打动，也决定跟着赵老师求学。赵老师是一位有独到科研思维的老师，这使得我的思维模式的塑造与方法论的建立得到很大的提升。赵老师的教导往往鞭辟入里，直击要害，这教会了我分析问题的核心，教会了我在看待问题时不随波逐流，有自己的见解。可以说，赵老师对我的影响是超过学术范畴的。

感谢我的父亲、我的母亲。我很庆幸拥有这样的父母，他们总是支持我、尊重我，他们会给我提意见但不替我做决定，他们总是在我看不到的地方为我付出，让我一直被爱所包围。我的父亲是一位极其负责的父亲，也是一位没有爹味的父亲，是我最知心的朋友，我不怕他但很听他的话，直到现在我还经常和他打一个多小时的电话谈天论地。我的母亲是一位比较“笨”的母亲，我总是埋怨她每隔一段时间没听到我的电话就坐立难安，我总是教她要懂得自己享受，不要克扣自己，她的世界总是围绕着我转，也不懂得自转一下。

感谢一路走来的朋友。我的朋友都很纯粹和简单。杨小龙，高中时期最好的朋友和榜样，一起打球一起讨论题目，我考上科大有他不小的功劳；夏厚，大学的铁哥们，一起旅游唱 k，一起交谈人生理想；李冰，饭友和天才 (我知道你喜欢听)，教我搞了不少作业；张雯、余碧桢、黄康杰和陈龙鑫，是一群“酒囊饭袋们”；阳阳，我的世界中一颗璀璨的星。

最后，感谢自己。感谢你的努力，感谢你不轻言放弃，一路磕磕绊绊走到这里；感谢你永怀初心，一直做一个善良和正直的人；感谢你更加认清现实和自己，接受了自己的普通，并能更加鼓起勇气往前走。祝你未来一切顺利，活出自我，前途似锦！

在读期间取得的科研成果

已发表论文:

1. Zhongyue Wang, Yun-Bo Zhao. Robust Out-of-Distribution Detection Based on Effective Points Select[C]//2025 6th International Conference on Computer Engineering and Application (ICCEA). IEEE, 2025: 1066-1070.
2. Zhongyue Wang, Dong Li. Robust out-of-distribution detection based on non-parametric adversarial virtual outlier generation[C]//Fourth International Conference on Electronics Technology and Artificial Intelligence (ETAI). SPIE, 2025, 13692: 1297-1304.

已公开专利:

1. 赵云波，王中月，谢祖浩，梁秀华. 一种基于检测分数的分布外检测自适应攻击方法 [P]. 安徽省：CN202510856981.8
2. 赵云波，王中月，谢祖浩，梁秀华. 一种基于有效点选取的分布外检测方法 [P]. 安徽省：CN202510858561.3
3. 赵云波，王中月，谢祖浩，梁秀华. 一种基于对抗虚拟离群点的鲁棒分布外检测方法 [P]. 安徽省：CN202510858564.7

已登记软件著作权:

1. 赵云波，王中月. 分布外数据鲁棒性验证平台 V1.0. 2025SR0697304