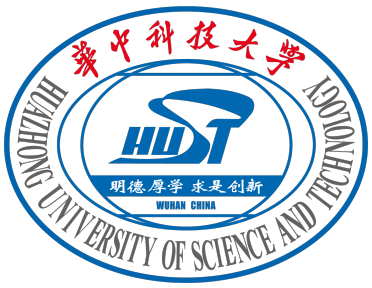

華中科技大學

学术规范与论文写作课程报告

选题（四）

分布外检测前沿进展：问题与方法

姓名：梁一凡

学号：U202115210

班级：人工智能本硕博 2101 班

2025 年 7 月 17 日

分布外检测前沿进展：问题与方法

姓名：梁一凡　　班级：人工智能本硕博 2101 班　　学号：U202115210

摘　要： 分布外检测旨在检测训练类别空间之外的测试样本，它是构建可靠机器学习系统的关键组成部分。现有的分布外检测综述主要聚焦于方法分类，通过对各种方法进行分类来研究该领域。然而，许多近期的工作集中于非传统的分布外检测场景，如测试时适应、多模态数据源以及其他新颖的情境。在本次综述中，我们首次从问题场景的角度独特地回顾了分布外检测的最新进展。依据训练过程是否完全可控，我们将分布外检测方法分为训练驱动型和训练无关型。此外，考虑到预训练模型的快速发展，基于大型预训练模型的分布外检测也被视作一个重要类别并单独进行讨论。再者，我们对评估场景、多种应用以及若干未来研究方向展开了讨论。我们相信，这种具有新分类法的综述将有助于新方法的提出以及更实际场景的拓展。

关键词： 分布外检测; 可信机器学习

Abstract

Out-of-distribution (OOD) detection aims to detect test samples outside the training category space, which is an essential componentin building reliable machine learning systems. Existing reviews on OOD detection primarily focus on method taxonomy, surveying the field by categorizing various approaches. However, many recent works concentrate on non-traditional OOD detection scenarios, such as test-time adaptation, multi-modal data sources and other novel contexts. In this survey, we uniquely review recent advances in OOD detection from the problem scenario perspective for the first time. According to whether the training process is completely controlled, we divide OOD detection methods into training-driven and training-agnostic. Besides, considering the rapid development of pre-trained models, large pre-trained model-based OOD detection is also regarded as an important category and discussed separately. Furthermore, we provide a discussion of the evaluation scenarios, a variety of applications, and several future research directions. We believe this survey with new taxonomy will benefit the proposal of new methods and the expansion of more practical scenarios.

Keywords: Out-of-distribution Detection, Trustworthy Machine Learning

1 引言

在封闭世界假设下，机器学习方法 [1] 取得了重大进展，其中测试数据与训练集来自相同的分布，被称为分布内数据。然而，在现实世界中，模型不可避免地会遇到不属于任何训练集类别的测试样本，通常被称为分布外数据。分布外检测 [2] 旨在识别并拒绝分布外样本，而非随意做出过度自信的预测 [3] ，同时保持对分布内数据的准确分类。具有卓越分布外检测能力的模型其能力更可靠，并在众多安全关键场景中具有重要应用。例如，在医疗诊断系统中，无法检测分布外样本的模型会误判未知疾病，从而导致严重误诊 [4] 。同样，自动驾驶算法 [5] 应检测未知场景，并依靠人工控制来避免因随意判断而引发的事故。

值得注意的是，近年来已有多项工作致力于对分布外检测进行综述和总结。Yang 等人 [6] 讨论了分布外检测及几个类似主题，并将现有工作分类为基于分类、基于密度、基于距离和基于重建的方法。Cui 和 Wang[7] 从方法学角度对分布外检测进行了综述，但采用了另一种分类标准，包括监督式、半监督式和无监督式方法。此外，Lang 等人 [8] 对自然语言处理中的分布外检测方法进行了回顾。然而，先前的工作过于侧重从方法角度进行讨论，缺乏从任务场景视角的深入探索。建立清晰的任务场景分类法有助于增强对该领域的全面理解，并协助从业者选择合适的方法。此外，鉴于近期新范式 (例如测试时学习范式 [9]) 和基于大型预训练模型的方法 [10] 的引入，迫切需要一项纳入最新技术的综述。

在本次综述中，我们首次以问题导向的分类法回顾了分布外检测的最新进展，如 Figure 1 所示。基于方法是否需要控制预训练过程，我们将分布外检测算法分为*训练驱动型*和*训练无关型*方法。考虑到当今大型预训练模型的快速发展，我们也将*基于大型预训练模型的*分布外检测作为单独的部分进行讨论。具体而言，训练驱动型方法通过设计训练阶段的优化过程来实现高检测能力。它们根据训练中是否使用分布外数据进一步分类和讨论。训练无关型方法基于训练良好的模型区分分布外数据和分布内数据，在实践中跳过耗时且昂贵的预训练过程。根据是否利用测试样本进一步提高分布外检测性能，我们将其分为事后方法和测试时方法。基于大型预训练模型的分布外检测方法聚焦于视觉语言模型或大型语言模型等模型，这些模型在大规模数据集上进行预训练，并在众多任务中表现出色。我们根据它们是否能访问少量示例 (包括零样本、小样本和全样本场景) 对其进行讨论。

本综述的其余部分组织如下。我们在第 2 节回顾分布外检测的相关工作。接下来，我们在第 3 节总结训练驱动的分布外检测方法，并在第 4 节介绍训练无关的分布外测方法。然后，在第 5 节中，我们介绍基于大型预训练模型的分布外检测。第 6 节对评估指标、实验协议和应用进行概述。随后，我们在第 7 节讨论有前景的趋势和开放挑战，以揭示未充分探索和潜在关键的途

径。最后，我们在第 8 节总结本综述。

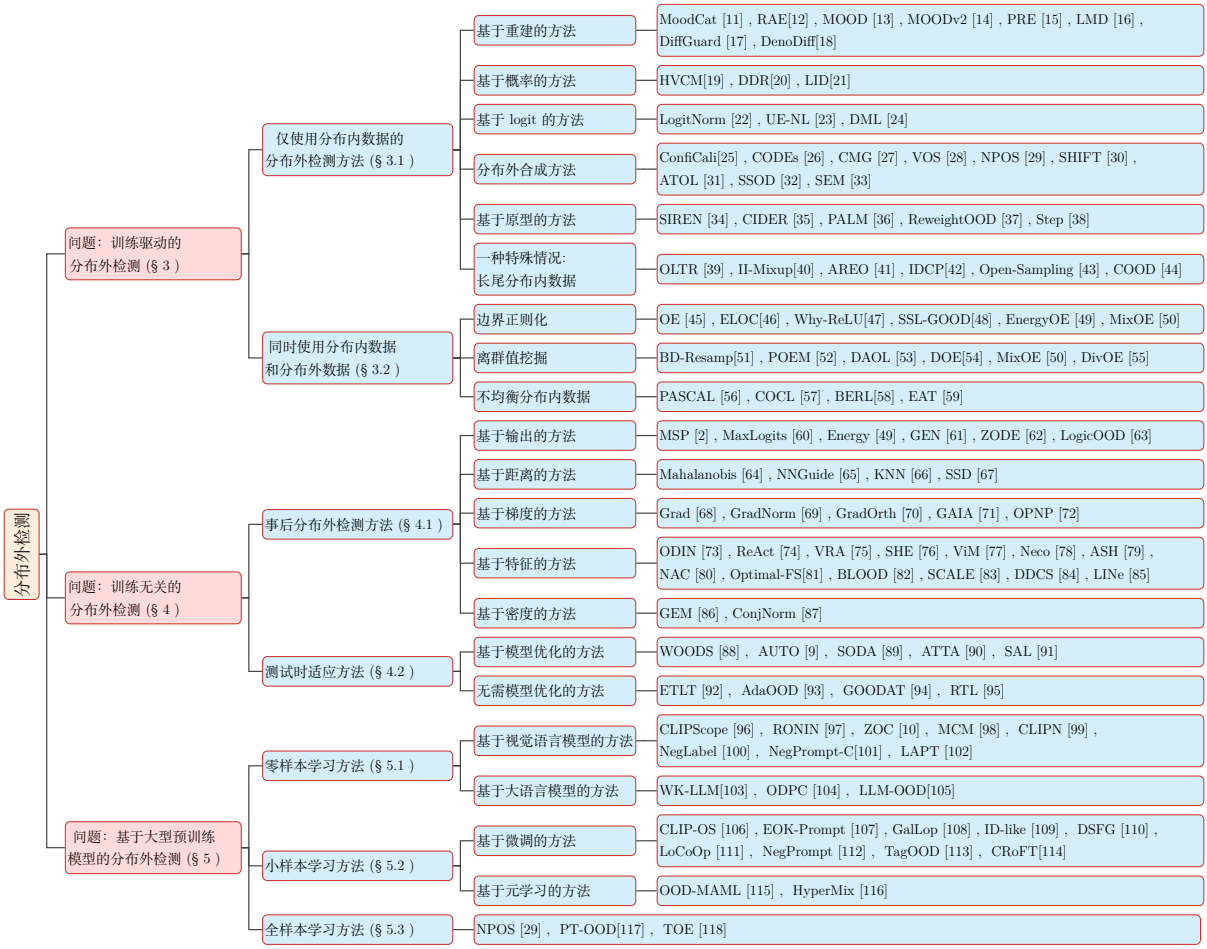

Figure 1: 分布外检测问题场景和解决方案的分类

2 相关工作

异常检测。异常检测涉及识别那些与数据集正常行为有显著偏差的数据点、事件或观测结果 [119] 。这些异常情况可能预示着诸如欺诈 [120] 、网络入侵 [121] 、系统故障 [122] 等重大事件。该过程需要运用统计学、机器学习或深度学习方法对正常行为进行建模并检测偏差。异常检测在网络安全、金融、医疗保健以及制造业等领域至关重要，在这些领域中，快速识别异常模式能够防止重大损失或提高运营效率。尽管异常检测和分布外检测二者均旨在识别意外的数据点，但异常检测侧重于单个分布（通常是训练集）内的偏差，而分布外检测则针对训练分布与新的、未曾见过的输入之间的差异。

新奇检测。新奇检测侧重于识别模型在训练期间未曾见过的新的或未知的数据点 [123] 。在系统需要适应不断变化的情况，或者标记新的、未曾出现过的场景至关重要的情形下，这一过程必不可少。与旨在找出偏离常规模式的异常检测不同，新奇检测的目的是发现全新的模式。其应用包括识别社交媒体中的新趋势 [124] 、在生物数据中发现新物种 [123] ，或者检测文档流

中的新主题 [125] 。从本质上讲，新奇检测处理的是熟悉情境中的意外情况，而分布外检测处理的则是来自不熟悉情境的数据。

开放集识别。开放集识别超越了传统的分类任务，它不仅对分布内进行分类，还能识别输入是否不属于任何分布内 [126] 。这对于现实世界中的应用来说至关重要，因为在现实应用中环境是动态的，系统会遇到在训练期间未曾出现过的实例。开放集识别在机器人技术 [127] 、自动驾驶汽车以及图像识别系统 [128] 等领域，在这些领域中遇到未知物体或场景是很常见的，并且系统必须能够妥善应对这些情况。开放集识别关注的是识别在训练期间未出现过的新类别，这通常是通过将同一数据集中的类别划分为基础类别和新类别来实现的，这意味着新类别和基础类别通常来自同一领域。相比之下，分布外检测侧重于识别任何与训练分布不同的数据，无论它是属于一个新类别还是完全来自不同的领域。从本质上讲，开放集识别是分布外检测的一个子集，特别针对的是对同一领域内的新类别类型进行分类。

离群值检测。离群值检测是识别与大多数数据存在显著差异的数据点的过程 [129] 。它与异常检测类似，但更侧重于单个数据点的识别，而非模式识别。离群值可能因测量的可变性、实验误差或真正的新变异而产生。其应用包括欺诈检测、故障检测以及从数据集中去除异常数据以提高模型精度 [130] 。与分布外检测相比，离群值检测更像是一种转导式场景，因为它本身就能接触到离群值。然而，传统的分布外检测只有在部署时才会遇到离群值。

零样本学习。零样本学习是机器学习中的一种范式，其目标是在训练阶段未曾见过某些对象示例的情况下对这些对象进行识别 [131, 132, 133] 。零样本学习中的根本挑战在于如何有效地将知识从已见过的类别迁移到未见过的类别 [131] 。这主要是通过属性来学习类别之间的语义关系，或者将已见过和未见过的类别都嵌入到一个共享语义空间中来解决的。虽然二者在推理过程中都涉及对未知事物的处理，但零样本学习试图对新类别进行分类 [134] ，而分布外检测则是标记那些对于训练数据分布而言属于异常或不熟悉的数据 [64] 。

选择性分类。选择性分类，也被称为拒绝选项分类，为模型提供了一种机制，使其在对自身预测不够有信心时可以选择不做决策 [135, 136] 。选择性分类涉及模型根据自身的置信水平来决定何时进行预测，实际上就是在不确定的时候选择不做预测 [137] 。另一方面，分布外检测是识别那些与训练分布不同的数据点，旨在将它们标记为模型所不熟悉的内容。虽然这两种方法都应对不确定性，但选择性分类处理的是针对分布内数据进行预测时的置信度问题 [138, 136] ，而分布外检测侧重于识别和处理那些未在训练集中出现的数据。

3 问题：训练驱动的分布外检测

在基于训练驱动的分布外检测问题中，研究人员会设计预训练流程，以获得具备出色分布外检测能力的模型。基于在训练期间是否能够获取分布外数据，我们进一步将此情形下的方法分为两类：仅使用分布内数据进行训练，以及同时使用分布内数据和分布外数据进行训练，如 Figure 2 所示。

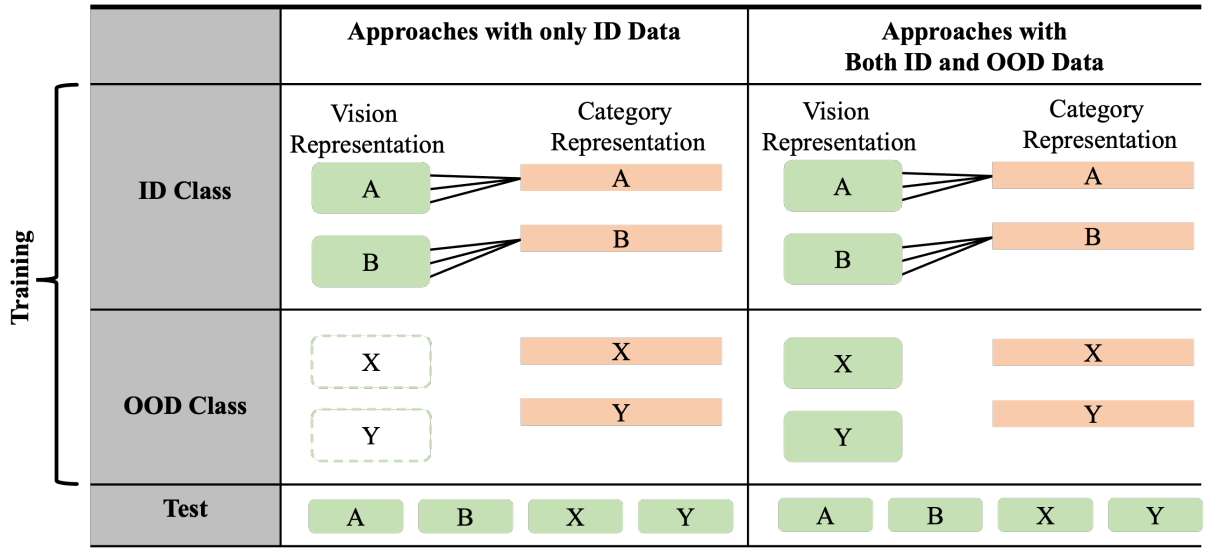

Figure 2: 训练驱动的分布外检测方法示意图。虚线边框表示它们在特定阶段未被使用。

3.1 仅使用分布内数据的分布外检测方法

概述。给定分布内数据，本节所涉及的方法会基于这些数据训练一个模型，并旨在利用该模型检测分布外测试样本，同时确保对分布内数据进行准确分类。仅使用分布内数据的方法特别侧重于从分布内数据中挖掘信息，而不会明确依赖来自现实世界中分布外数据的其他信息。我们进一步将这些方法细分为以下五类：基于重建的方法、基于概率的方法、基于 *logit* 的方法、分布外合成方法以及基于原型的方法。考虑到现实世界的需求，我们还深入探讨了一个特定场景：长尾分布内数据。

基于重建的方法。基于重建的方法论通过仔细研究样本在重建前后表征之间的差异，开辟了一条新的研究途径，该方法在很大程度上依赖于重建模型的性能。从根本上讲，重建任务的目标是在指定的监督信号引导下，恢复数据集内在的语义内容。这一前提基于这样一种观念：分布外数据本质上具有与分布内标签信息不一致的语义特征。通过对这种语义偏差程度进行量化评估，模型就能高精度地识别出分布外数据。此类方法的大致流程如 Figure 3 所示。

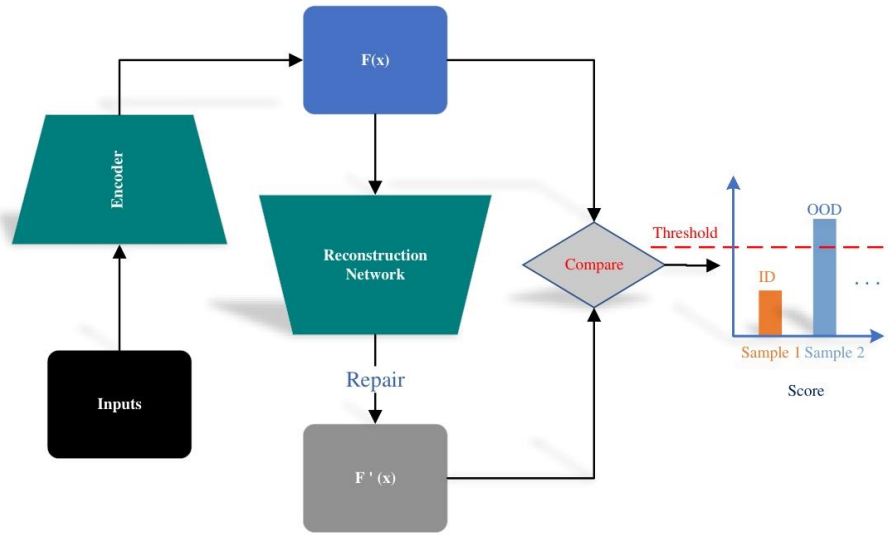

Figure 3: 基于重建的分布外检测方法示意图。通过神经网络提取图像的特征表示 $F(x)$ 后，将其输入重建网络 (如变分自编码器 (VAE) 或去噪扩散概率模型 (DDPM)) 以获得重建后的图像表示 $F'(x)$。通过比较原始图像和重建图像之间的差异，我们可以识别分布外数据，因为分布外样本在重建前后通常表现出更大的差异。

MoodCat [11] 会对输入图像的随机部分进行遮挡，并利用生成模型依据分类结果合成新图像，在合成过程中施加较强的约束条件。类似地，Zhou [12] 引入了一个辅助模块来提取特征向量的激活值，帮助模型对潜在的重建空间加以约束，以过滤潜在的分布外数据。在此之后，MOOD [13] 和 MOODv2 [14] 借助基于掩码图像建模的预训练任务，在学习数据内部的分布方面展现出显著优势。PRE [15] 引入了归一化流的概念，并结合基于典型性的惩罚机制来约束重建误差，从而能很好地辨别出分布外数据与分布内数据之间的差异。

近来，扩散模型在训练稳定性以及生成图像的质量方面都取得了显著进展。利用扩散模型来检测分布外数据已成为一个新的研究方向。Graham 等人 [18] 通过引入 DDPM 来解决这一问题，利用该模型对受噪声干扰的图像进行重建。该模型采用多维重建误差来识别分布外数据，其信息瓶颈可从外部进行调节。同样地，LMD[16] 利用扩散模型进行分布外验证。LMD 对数据进行扰乱，然后运用扩散模型重建与原始流形分离的图像，通过比较与原始流形的差异来区分分布外数据。DiffGuard [17] 直接采用预训练的扩散模型进行语义失配引导，旨在利用扩散模型放大重建的分布外图像与原始图像之间的差异。

基于概率的方法。基于概率方向的研究旨在建立概率模型来描述训练数据的分布情况。通过这些概率模型，开发出合适的评分函数，用以计算测试样本在分布内的得分，这些得分能够反映出这些样本是否属于分布内分布。在这一领域，Li 等人 [19] 在训练过程中使用多个高斯混合模型对每个分布内

类别进行建模，而在预测阶段，其结合马氏距离度量来评估异常类别的可能性。Huang 等人 [20] 通过引入两项正则化约束来解决这一问题。密度一致性正则化使解析密度与低维类别标签相匹配，对比分布正则化则有助于区分分布内样本和分布外样本之间的密度。此外，LID 方法 [21] 针对深度生成模型生成数据背景下的分布外检测悖论引入了一种新的检测准则。当数据被赋予较高概率且概率量不可忽略时，它通过估计生成模型所学习到的流形的局部固有维度 (LID) 来衡量数据是否应被归类为分布内数据。

基于 Logits 的方法。这些算法主要聚焦于神经网络的预测结果，尤其关注 Logits，它是神经网络输出层的输出结果。Logits 通常代表模型对每个类别的置信度或概率。LogitNorm [22] 提出了一种方法，即在训练期间使用 logits 归一化，对 logits 强制施加一个恒定的向量范数，以缓解模型过度自信的问题。同样地，源于贝叶斯网络的 UE-NL[23] 在对 logits 进行归一化的同时，还学习嵌入表示和不确定性得分。它在训练期间调整样本之间的学习强度，使学习过程更具稳健性。DML [24] 应对了可能对分布外检测性能产生阻碍的问题。借助经验性见解，DML 通过解耦 logits 并平衡各个组成部分来提升分布外检测性能，从而减轻各属性对结果的影响。它通过将场景最大范数作为一个参数来解耦 logits ，进而平衡各属性对结果的影响，以此提升分布外检测性能。

分布外合成。在分布外检测任务中，理论上认为在模型训练期间融入分布外数据的特征能够提升分布外检测性能。由于获取分布外样本的分布信息存在诸多挑战，一些方法会利用分布内数据来估计分布外数据的分布情况。这样做是为了模拟模型在现实世界中遇到分布外数据的场景。

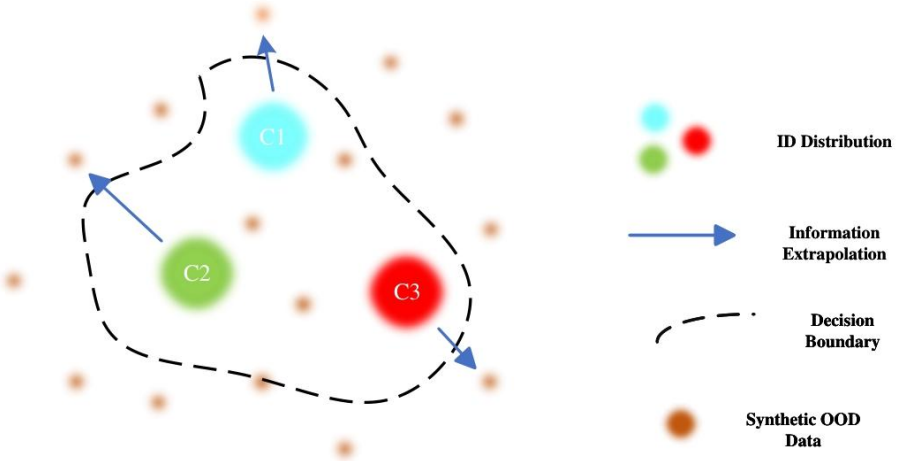

Figure 4: 分布外合成方法示意图。在缺乏真实世界分布外数据的情况下，利用从分布内数据的外推来识别合适的分布外数据，然后使用离群值暴露的方法训练模型，可增强分布外检测能力。图中的橙色点表示使用分布内数据合成的虚拟分布外数据。

Lee 等人 [25] 通过在原始损失函数中纳入两个补充项来添加额外的正则化项。第一项迫使分类器增加对分布外样本的不确定性，而第二项则鼓励 GAN 生成合适的分布外样本。具体而言，设计了一种联合训练方案，用于在分布内样本的低密度区域中选择合适的边界样本。同样地，VOS[28] 采用了离群值的自适应合成。它从特征空间中类别条件分布的低概率区域采样离群值，无需依赖外部数据。这使得模型能够合成分布外数据用于预测。值得注意的是，VOS 支持自适应离群值合成，使其能够轻松应用于任何分布内数据，而无需人工干预。注意到先前针对分布外数据的方法所提出的额外分布假设，一些研究从另一个角度来处理该问题。例如，NPOS[29] 认为，与其将离群值的特征空间建模为参数化的高斯分布，不对嵌入的分布内数据做任何分布假设反而会赋予模型强大的灵活性和通用性。它通过基于最近邻的非参数密度估计来选择边界点，这些边界点位于分布内数据和分布外数据之间。此外，SSOD[32] 引入了自监督采样来隐式地生成分布外数据。它直接从分布内数据图像的背景中采样自然的分布外信号，从而克服了在分布外合成阶段因偏差而产生的局限性。

与上述提及的分布外样本的隐式生成方法形成对比的是，一些方法另辟蹊径，直接利用从分布内样本中对分布外样本进行显式构建的方式来开展学习。为了获取分布外样本，CODEs[26] 首先通过对来自不同类别内部分布的样本进行切割和拼接来生成初始的种子分布外示例。然后，将这些示例输入到 Chamfer GAN 中进行分布转换，从而生成高质量的分布外样本。随后，CMG[27] 通过将混合的类别嵌入作为条件提供给 CVAE 来生成伪分布外数据，接着利用这些数据对分类器进行微调以用于分布外检测。SEM[33] 引入了全谱分布外检测的问题设定，在训练期间利用 Mixup 生成负样本。然后，它同时利用高级语义信息以及诸如风格等低级非语义特征来识别数据的来源。SHIFT[30] 提出基于训练样本直接合成分布外图像样本。它通过运用 CLIP 模型去除训练样本中的分布内对象区域来实现这一点。然后利用潜在扩散模型在考虑上下文背景的情况下用真实特征替换这些区域。这种方法由此建立了模型的拒绝能力。然而，确保生成数据的质量往往颇具挑战性。而且，生成数据本身的质量可能存在固有缺陷。ATOL[31] 引入了一项辅助任务，在该任务中辅助分布内数据和辅助分布外数据同时存在。在低维潜在空间中，手动为辅助分布内数据和辅助分布外数据寻找不同的区域，确保在输入空间中具有不重叠的特性。随后，生成器进行的数据生成会保证输入空间的不重叠属性。最后，通过将真实的分布内数据与辅助分布内数据进行对齐来保证可靠性，有效缓解与错误生成的分布外实例相关的问题。

基于原型的方法。在模型训练过程中，基于原型的分布外检测方法旨在利用原型对分布内数据进行建模，以学习分布内数据的常见分布特征。在测试阶段，模型会测量样本与类别层级原型之间的差异，从而确定样本的类别。这些方法的大致流程如 Figure 5 所示。

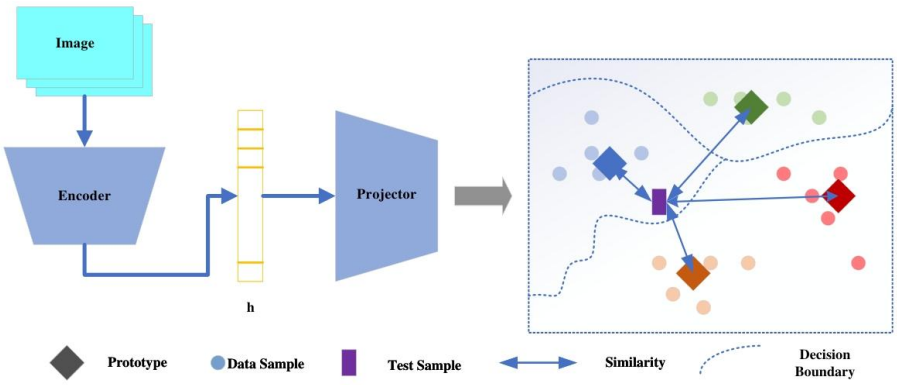

Figure 5: 基于原型的 OOD 检测方法示意图。基于原型学习的分布外检测算法大致由三种网络结构组成：1) 编码器。此结构负责通过神经网络传播样本数据以获得特征表示 h。2) 投影器。此类结构将编码在 h 中的图像特征投影到一个新的表示空间。3) 原型学习。在新的表示空间中，学习代表每个类别的原型，每个类学习其相应的原型中心。在测试期间，通过比较测试样本与每个分布内类别的原型之间的相似性来衡量异常数据的可能性。

SIREN [34] 首先使用 von Mises-Fisher(vMF) 模型对分布内数据的分布进行建模，这使得能够将每个类别表示为一个紧凑的簇，即一个类别原型。一般来说，vMF 分布建模公式可以表示为：$p_D(z;p_k,k) = Z_D(k)\exp(kp_k^T z)$，其中 p_k 是具有单位范数的第 k 个原型，$\kappa \geq 0$ 表示围绕均值的集中度，$Z_D(\kappa)$ 是归一化因子。在基于原型的方法中，嵌入向量 z 以如下归一化概率分配给类别 c：

$$p\left(y=c|z;(P_j,k_j)_{j=1}^{C}\right) = \frac{Z_D\left(k_c\right)\exp\left(k_c\mu_c^T z\right)}{\sum_{j=1}^{C} Z_d\left(k_j\right)\exp\left(k_j\mu_j^T z\right)}, \tag{1}$$

其中 $c \in \{1,2,\ldots,C\}$。此外，在损失函数中，它强制分布内样本的嵌入向量与类别原型之间保持对齐，以此约束每个分布内样本。这种参数化的分布外得分在训练后可直接获取，无需单独进行估计。CIDER[35] 以 SIREN 为基础，通过联合优化两项损失来增强数据的可判别性，促使不同类别原型之间的角度距离最大化以及同一类别中原型的内部紧凑性增强。模型训练期间的优化过程与各类别的原型相关。ReweightOOD[37] 认为，优化非类别数据会阻碍实现清晰的类别可分离性，而聚焦于较少的类别数据则难以实现更低的 MSE 得分。为解决这一问题，他们提出一种重加权优化策略来平衡不同损失的重要性。尽管 Step[38] 背后的思路有所不同，但在半监督任务的背景下，它本质上是通过对比学习过程生成未标记的分布内和分布外样本的聚类，这在概念上与原型学习类似。

然而，PALM [36] 注意到，使用单个原型对每个类别进行建模可能无法捕捉到数据内部信息的多样性。PALM 引入了一种混合原型策略来优化建模过程，利用多个原型的策略为每个类别学习富含信息的表示形式。通过同时学习类别层级的原型并对比类间原型，它在损失函数内优化了原型层级的类

内紧凑性以及类间可判别性。

一种特殊情况：长尾分布内数据。在实际场景中，由于收集难度以及出现频率等原因，分布内数据可能呈现长尾分布。这种不均衡性将极大地影响分布外检测的性能。人们提出了许多方法来应对分布内数据不均衡的挑战，并提升分布外检测能力。OLTR [39] 通过在头部和尾部嵌入之间关联视觉概念来解决尾部识别的鲁棒性问题。它借助视觉记忆动态调整嵌入范数来实现这一点。该方法包含两个用于增强长尾数据鲁棒性的组件：一个利用来自标准嵌入的直接特征，另一个则采用存储从直接特征派生的可区分质心的记忆特征，以此增强图像的直接特征。梅塔等人 [40] 提出了一种专门为处理中尾类设计的定制化复杂子集混合策略。这种方法通过将来自不同尾类的两个独立样本配对来创建混合样本，并计算混合样本与特定类别的原型之间的距离。该混合策略最终与原型学习相结合，以有效应对长尾场景带来的挑战。AREO [41] 引入了一种类似于置信度的、通过证据学习来量化样本不确定性的方法。整个训练过程由一种创新的多调度器学习机制管控，该机制会根据不同类别的重要性动态调整训练参数，以确保模型聚焦于少数类别的特征，从而平衡多数类和少数类之间的差异。总体而言，AREO 在训练期间识别出某些样本中较高的不确定性，并动态调整参数以减轻对这些样本的过拟合情况。

现有的分布外检测方法大多假定从分布外到分布内的概率呈均匀分布。在存在类别不均衡的情况下，Jiang 等人 [42] 提出了一种替代策略来增强当前的分布外检测方法。这涉及基于分布内的类别先验分布以及与预训练模型输出的 KL 散度来重新校准分布外得分。这种方法意味着有可能在先前优化方法的成果基础上进一步增强分布外检测的鲁棒性，强化整体性能。此外，Open-Sampling[43] 利用来自分布外数据集的噪声标签来重新平衡分布内训练数据集的类别先验分布。这些标签是从与原始类别先验分布互补的预定义分布中采样得到的。COOD [44] 使用监督模型将单个分布外度量组合成一个整体，类似于随机森林的概念。这种方法解决了单个分布外方法的局限性，还能克服与数据不均衡相关的问题。

3.2 同时使用分布内数据和分布外数据的分布外检测方法

概述。在一些已知的部署场景中，真实的分布外数据能够以较低成本轻松收集。基于这一假设的一些方法侧重于如何利用分布外数据来获得更好的检测性能。与涉及分布外数据合成的方法不同，在这些研究方向中，模型在训练阶段能够获取现实世界中的分布外数据。此类问题的主要关注点在于优化模型的决策边界，而非分布外数据本身。由于引入了真实的分布外信息，这两类 (分布内类别和分布外类别) 的边界将能被精准计算出来。

边界正则化。边界正则化这类方法属于传统的离群值暴露 (Outlier Exposure，OE) 方法。Hendrycks 等人 [45] 以及 Hein 等人 [47] 的核心思想是充分利用分布外数据来优化模型的决策边界，从而实现分布外检测。这一概念的支持者可以利用辅助异常数据集来增强分布外检测器，使其能够泛化并检测训练期间未遇到过的异常信息。从 Figure 6 中可以领会该方法的核心思想。

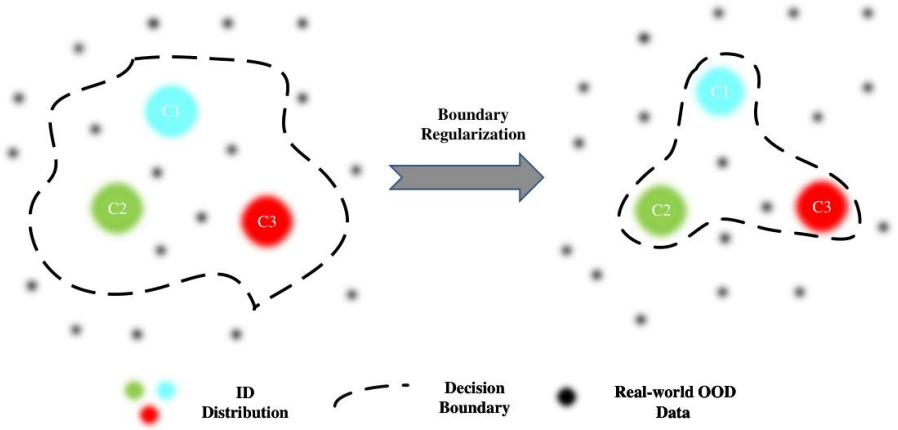

Figure 6: 边界正则化方法示意图。模型优化策略的关键在于模型分类的决策边界。边界正则化利用现有的已知分布内数据 (彩色点) 和一些分布外数据 (黑色点) 来优化潜在的决策边界，在不影响分布内数据分类的情况下，使分类边界与异常数据尽可能紧密地贴合。

具体而言，给定一个模型 f 和原始损失函数，模型训练过程旨在最小化目标：

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{in}}\left[\mathcal{L}(f(x),y)+\lambda\mathbb{E}_{x'\sim\mathcal{D}_{out}^{OE}}\left[\mathcal{L}_{OE}(U,f(x))\right]\right] \tag{2}$$

关于 f 的参数，其中 x' 表示辅助异常数据，U 表示均匀分布。L_{OE} 表示相对于 U 的交叉熵损失。其根本目的是迫使模型将分布外数据分布优化为均匀分布，这一原则在 OE 类型的方法中是通用的。L_{OE} 的具体设计可以取决于其他任务要求和所选的分布外分数。这种设计可以利用最大 softmax 概率基线 [2] 检测器来检测异常数据。与传统的 softmax 分数相比，EnergyOE [49] 在 OE 的基础上利用能量分数来更好地区分分布内和分布外样本，并且不太容易出现过度自信的问题。具体来说，其计算公式为：

$$E(x;f)=-T\cdot\log\sum_{i}^{K}e^{f_i(x)/T} \tag{3}$$

其中使用了温度系数 T，$f(x)$ 表示判别性神经分类器 $f(x):\mathbb{R}^D\to\mathbb{R}^K$，它将输入 $x\in\mathbb{R}^D$ 映射为 K 个实值 logits。

Mohseni 等人 [48] 使用自监督方法训练模型，利用伪标记对未标记的分布外样本优化目标函数，以泛化分布外检测能力。Vyas 等人 [46] 同样采用了

对分类器的自监督训练。与 OE 方法不同，其目的是找出分布外样本和分布内样本平均熵之间的差距。MixOE [50] 考虑到了细微的分布外样本在增强分布外检测泛化能力方面的有益作用。其主要思路是混合分布内和分布外数据样本，以拓宽分布外数据的泛化范围。使用这些离群值来训练模型，能够随着输入从分布内样本变为分布外样本，线性地降低预测置信度，从而显式地优化决策者的泛化能力。

离群值挖掘。传统的离群值暴露 (OE) 概念假定存在分布内输入数据 D_{in} 和分布外输入数据 D_{out}，它们各自独立且呈异质分布，来源于不同的数据源。然而，在当前的训练过程中，由于训练所用的分布外数据中可能存在噪声，这一前提无法完全得到保证。离群值挖掘与传统的异常暴露方法略有不同，尽管它同样利用现实世界中的分布外样本去解决问题，但它侧重于在现有的分布外数据中确定最优选择。其主要流程如 Figure 7 所示。

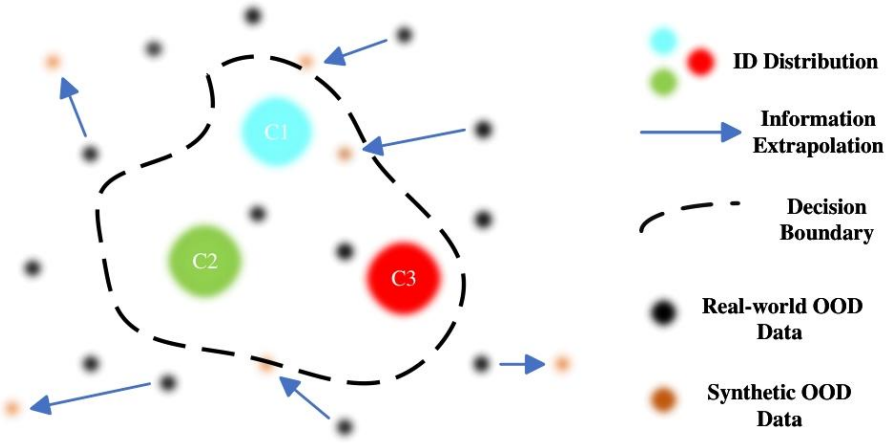

Figure 7: 离群值挖掘方法示意图。由于真实分布外数据是否存在的限制，像离群值挖掘这样的方法允许模型在训练期间访问真实分布外数据。图中的黑色点代表从其他数据集中采样的真实分布外数据，这些数据与分布内类别不重叠。利用此类分布外数据来推断更具代表性的分布外数据 (橙色点) 是解决此类问题的一种方法。

POEM [52] 所提出的方法侧重于挖掘具有更具代表性特征的异常情况。Li[51] 提出了一种数据重采样的方法，用于获取具有代表性的离群值数据来进行训练，会给难分负样本赋予更高的优先级得分重加权。利用对抗原理进行迭代优化来选择目标数据。POEM 采用后验采样，从一个庞大的辅助数据集中挖掘出边界得分较高的异常情况，有助于深入细致地理解分布外样本内部的复杂情况。DAOL [53] 以真实数据和辅助分布外数据之间的差异作为出发点来构建一个分布外数据集。利用 Wasserstein ball 对所有分布外数据的分布进行建模，选择该球内最具挑战性的分布外数据用于训练。

除了仅仅依赖原始数据之外，解决这一问题的另一个方向涉及通过信息外推利用真实的分布外数据来合成具有代表性的离群值数据。DivOE [55] 引入了一种新颖的学习目标，以缓解与辅助分布外数据集有限相关的挑战。它

通过在给定特定异常情况时，自适应地从替代分布外数据中推断并学习信息来实现这一点，具体方式是使生成的分布外数据与原始数据之间的差异最大化。这种自适应推断拓展到了更广泛的范围，解决了因辅助分布外数据集有限而带来的限制问题。此外，DOE [54] 引入了一种最小-最大学习策略，以便为合成模型识别最具挑战性的分布外数据。通过对模型进行扰动，数据被隐式地转换，然后模型继续从这种经过扰动的数据中学习，以提高其鲁棒性。

不均衡问题。出于实际需求，针对训练期间分布内数据不均衡但仍需提供分布外数据信息的场景，相关研究正日益增多。PASCL [56] 背后的理念是，在训练阶段仅通过对比损失来分离尾部数据和分布外数据，帮助模型更好地区分这两者。为解决模型在分布外样本和尾部类别数据之间产生混淆的问题，COCL [57] 在训练过程中引入了一个可学习的尾部类别原型。该原型的作用是拉近尾部样本之间的距离，同时使其与分布外数据拉开距离，从而减轻模型对分布外样本的偏向性。Choi 等人 [58] 采取了不同的方法，他们认为影响分布外检测性能的因素可能与辅助分布外数据的跨类别分布不均衡有关。因此，他们提出了一种能量正则化损失，专门对多数类的辅助样本进行正则化，以解决分布外数据中的类别不均衡问题。EAT [59] 在模型训练过程中利用动态分配的虚拟标签来训练分布外数据，从而拓展了分类空间。

尽管像离群值暴露这类方法在研究领域取得了成功并受到了相当多的关注，但也存在一些质疑的声音，对在训练期间允许使用分布外数据这一做法的本质提出了疑问。而且，有人担忧在某些数据集中所观察到的出色分类性能未必能转化为在现实世界部署中的竞争力，这对分布外检测的初衷构成了挑战。

4 问题：训练无关的分布外检测

与训练驱动的分布外检测侧重于分类器性能不同，*训练无关的分布外检测*主要强调测试阶段的适应策略。根据其是否依赖测试数据之间的关联性，相关方法可分为两类：事后方法和测试时适应方法，如 Figure 8 所示。事后方法独立计算单个样本的结果，不受其他样本变化的影响。与之相反，测试时适应方法在测试样本之间存在关联性。

4.1 事后分布外检测方法

概述。给定一个经过良好训练的模型，在这种问题场景下，只需利用训练好的模型在测试期间计算出的中间结果，而无需修改模型的任何参数，就能完成分布外检测任务。事后 (Post-hoc) 方法因其轻量化的特性、较低的计算成本以及对模型和目标只需极少修改等优点而受到青睐。其主要目标是构建一个能准确反映分布内数据行为的有效评分函数。这些特性使得它们在实际场景中非常便于直接部署。

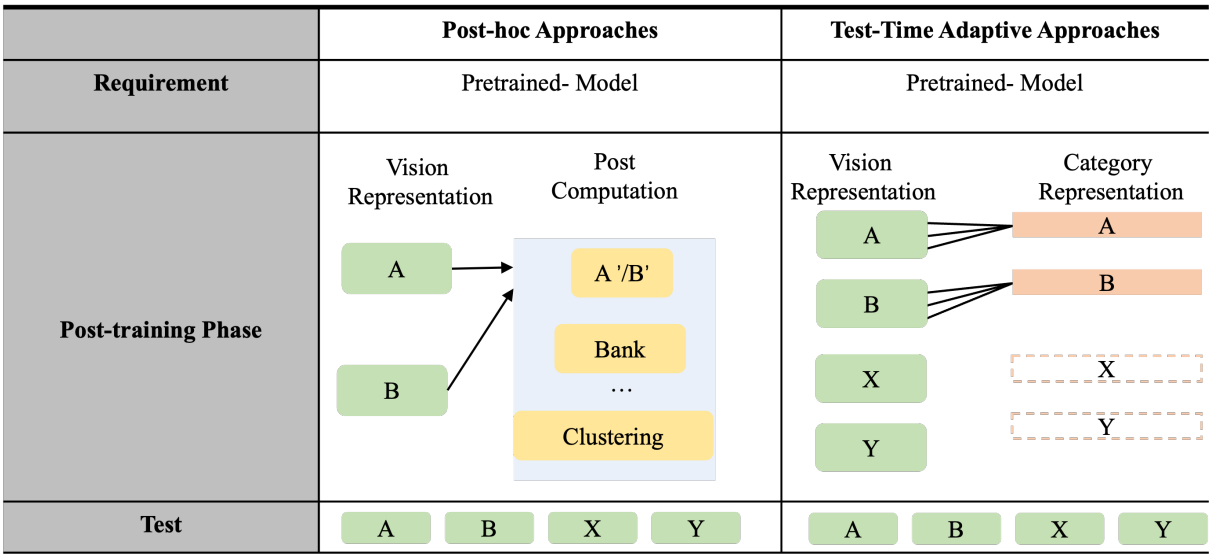

Figure 8: 训练无关的分布外检测方法示意图。这两种方法都需要访问预训练模型。事后方法在训练后阶段不涉及任何操作，而测试时适应方法需要根据测试期间遇到的样本进行调整。“A’ /B’” 表示原始特征发生了变形；“Bank” 表示存储了一些样本；“Clustering” 表示对分布内图像进行聚类。

事后方法可分为五类：基于输出的方法、基于距离的方法、基于梯度的方法、基于特征的方法以及基于密度的方法。针对这类问题的近期研究工作取得了一些新进展。Table 1 给出了此类方法所涉及关键要素的概述。

基于输出的方法。基于输出的算法主要旨在探究神经网络中间层输出的潜在表示，其中包括 logits 和类别分布等内容。MSP[2] 是首个采用最大 Softmax 值来验证分布外检测有效性的方法。对于分布外样本而言，它们的输出概率分布往往更趋近于均匀分布，这表明模型无法正确对其类别进行分类。与 MSP 方法不同，MaxLogits[60] 通过比较神经网络输出的 logits 向量中的最大 logits 来检测异常情况。logits 代表着模型对每个类别的置信度，它是神经网络在经过 Softmax 层之前的输出，尚未进行 Softmax 变换。与此同时，Energy[49] 引入了亥姆霍兹自由能，该能量在理论上与输入概率密度相契合，并且不太容易受到过度自信问题的影响。GEN[61] 引入了广义熵的概念，并直接利用布雷格曼散度来计算模型概率输出与均匀分布之间的统计距离，旨在识别分布外数据。此外，利用充足的先验知识可能是一个可行的解决方案。ZODE[62] 会同时对多个预训练模型的样本进行预测，以判断多个模型是否能够识别出分布外样本，并以此作为区分数据的依据。LogicOOD[63] 提出了一种新颖的方法，即利用一阶逻辑进行知识表示来执行分布外检测。这种推理系统运用先验知识来推断输入是否与关于训练分布的先验知识相符。它在可解释性方面对用户特别友好，因为它允许将样本的输出与知识库进行对比，以确定它们是否属于分布内数据。

Table 1: 分布外检测方法中关键组件的比较。

方法	**类型**	**空间**			
		特征	**logit**	**梯度**	**概率**
MSP [2]	基于输出的				✓
Maxlogits [60]			✓		
Energy [49]			✓		
GEN [61]		✓			✓
Mahalanobis [64]	Distance-Based	✓			
NNGuide [65]		✓			
KNN [66]		✓			
SSD [67]		✓			
Grad [68]	基于梯度的			✓	
GradNorm [69]		✓		✓	✓
GradOrth [70]		✓		✓	
GAIA [71]		✓		✓	
OPNP [72]		✓		✓	
ODIN [73]	基于特征的	✓		✓	✓
ReAct [74]		✓			
VRA [75]		✓			
SHE [76]		✓			
Vim [77]		✓	✓		
Neco [78]		✓	✓		
ASH [79]		✓			
NAC [80]		✓			
Optimal-FS [81]		✓	✓		
BLOOD [82]		✓			
SCALE [83]		✓			
ConjNorm [87]	基于密度的	✓	✓		
GEM [86]		✓	✓		✓

基于距离的方法。分布外检测研究中的另一种方法侧重于测量统计距离度量指标。马氏距离 (Mahalanobis)[64] 通常是通过计算特征向量与其均值之间的距离来得出的。具体而言，对于每个类别，我们会计算其特征向量的均值和协方差矩阵。在测试阶段，它会计算特征向量与每个类别均值之间的马氏距离。SSD[67] 本质上运用了马氏距离。在通过自监督表征学习对未标记的分布内数据进行训练后，它利用马氏距离作为一种统计度量，借助预训练模型进行分类。相比之下，马氏距离对数据有较强的分布假设，而 KNN[66] 则探究了非参数化最近邻距离在分布外检测方面的有效性。通过测量输入嵌入之间的 K 近邻距离以及训练集嵌入，设计一个阈值来判定数据是否属于分布内数据。NNGuide[65] 通过结合 KNN 的思路，在精细化方向上更进一步。它会依据样本与训练集中嵌入之间的最近邻距离，在传统的分布外得分之前分配权重。

基于梯度的方法。相关研究 [68] 表明，基于梯度的方法也有助于分布外检测，其方式是通过反向传播过程中反向传播的梯度来量化模型的不确定性。如果输入样本是分布内样本，模型针对这些样本的梯度往往相对较小且稳定。相反，对于分布外样本，其梯度通常更大或更不规则。GradNorm[69] 假定分布内数据的梯度幅值超过分布外数据的梯度幅值。利用这一观察结果，它采用梯度向量范数，通过 Softmax 输出以及均匀分布的反向传播来计算 KL 散度，以此检测分布外数据。与之不同的是，GradOrth[70] 采用了另一种视角，认为分布外数据的关键特征存在于低秩子空间中。因此，它将重点转向计算该子空间中的梯度投影范数来识别分布外数据。GAIA[71] 采用了梯度异常检查与聚合相结合的方式。它能让模型从属性角度解释不确定性，在无需先验知识的情况下引入通道平均异常和零压缩异常，以衡量数据分布变化的程度。OPNP[72] 方法发现，模型的分布外检测能力对接近零的参数和激活神经元高度敏感。因此，该方法利用对参数和神经元的修剪行为，去除那些会导致过拟合的部分，从而增强模型的泛化能力。

基于特征的方法。分布外检测研究的另一个方向涉及神经网络中间变量相对于最终预测的影响。受对抗样本的启发，ODIN[73] 通过实验发现，对输入图像特征添加微小扰动能够更有效地检测分布外数据。扰动公式如下：

$$\overline{x} = x + \varepsilon sign\left(\nabla_x log max_c p_c(x)\right), \tag{4}$$

其中参数 ε 是扰动幅度。对于给定的输入 x，计算其 logit 输出 $p_c(x)$。因此，ReAct[74] 聚焦于模型中间结果中的高激活值。这些激活值并不影响模型的分类，但截断高激活值能够显著提升分布外检测性能。VRA[75] 是 ReAct[74] 的扩展迭代版本，它基于 ReAct 仅截断高激活值这一前提构建。然而，VRA 认为这可能并非最优解，因此采用变分方法来寻求最优解。它利用分段函数模拟抑制或放大操作，帮助模型识别异常数据。SHE[76] 将倒数第二层的输出转换为一种存储模式，利用 Hopfield 能量 [139] 通过先存储后比较的模式

来进行分布外样本检测。它运用一种存储机制来计算每个类别的平均 logits，为后续的分布外检测任务准备这种简单的存储模式。受经验启发，DDCS[84] 在对神经网络中的每个通道进行校正后，会自适应地选择合适的通道用于数据分类。这些通道是基于类间相似度和方差来进行评估的，以此衡量它们对分布内数据的判别能力。LINe [85] 同样强调特征层面的神经元输出。它采用 Shapley value 剪枝方法，仅选择对预测贡献度高的神经元，同时屏蔽其余输入数据，从而减少无关输出的影响。

特征塑造正成为一个备受瞩目的研究重点领域。这种方法需要对模型前向传播阶段中的中间表示 (特别是中间特征) 进行细化。该方法的显著优势在于它不会干扰原始分类结果，并且兼具简单性和有效性。在此基础上，ViM[77] 同时考虑了特征、logits 和概率所起的作用，构建虚拟 logits 来聚合信息以辅助决策。通过对神经网络倒数第二层的特征进行分解，它识别出了与分类无关的零空间，但该空间在分布外检测中表现出色。其计算公式可表示如下：

$$-\alpha \left\| z^{P^{\perp}} \right\|^2 + logSumExpf(z), \tag{5}$$

其中 α 是一个由模型计算的缩放常数。这里 $z = z^P + z^{P^{\perp}}$，$z^{P^{\perp}}$ 是 z 到 $P^{\perp}$ 的投影。并且有 $Wz^{P^{\perp}} = 0$。此外，$LogSumExp$ 表示能量函数 [49] 的计算过程，$f(z)$ 表示模型的 logit 输出。这里的第一项代表虚拟 logits，而第二项代表能量函数的得分。随后，Neco[78] 揭示了当代神经网络中普遍存在的神经元坍缩现象，该现象影响了分布外检测性能。利用分布内数据和分布外数据特征之间的正交趋势来区分分布外数据。ASH[79] 是一种简单的动态激活修改方法，在模型训练后期，样本中的大部分激活要么被移除，要么被稍加调整。NAC[80] 引入了一种神经元激活覆盖率的度量方法，其前提是如果神经网络中的某个神经元很少被激活，那么这种状态可能意味着出现分布外数据的可能性更高。通过对这一统计特性进行量化，NAC 旨在区分分布内数据和分布外数据。基于现有的特征塑造相关文献，Zhao 等人 [81] 利用分段常数塑造函数将特征域划分为互不相交的区间，在每个区间内估计一个标量作为近似值。当区间宽度趋近于零时，就可以得到最大 logit 的近似值。BLOOD[82] 的目标是对神经网络中间层的表示进行平滑处理，以预测分布外数据。这项研究发现，与分布外数据相比，分布内数据在神经网络中间层表示中的变化更为平滑。利用这一特性，可以设计新的统计度量来甄别异常数据。此外，另一个研究方向聚焦于神经元激活剪枝，这是在 ASH 的研究基础上提出的一种新的激活塑造方案。SCALE[83] 强调缩放是评估样本的一个关键指标，并且同样发现分布外数据的剪枝率明显更低。因此，所提出的对中间层张量进行重塑的方法能够有效提升检测性能。

基于密度的方法。基于密度的分布外检测模型近期取得的进展表明，其性能有了大幅提升，这是基于这些模型能够准确捕捉并理解真实数据分布的

内在特征。例如，GEM[86] 将分布内数据的特征空间建模为类条件多元高斯分布。在此假设下，它设计了新的统计度量来验证模型的性能。利用基于高斯混合模型的建模方法，GEM Score 与真实的对数似然相契合，用于捕捉分布外的不确定性。然而，尽管 GEM 严格依赖高斯假设，但近期的研究成果 ConjNorm[87] 引入了一个基于布雷格曼散度的新颖框架，将对数据分布的考量范围扩展到涵盖指数分布族。通过重新定义数据的密度函数，该模型具有更广泛的应用范围。

4.2 测试时适应分布外检测方法

概述。测试时适应方法使用在训练集上训练好的分类器，致力于利用测试数据 (可以是整个测试集，也可以是一系列未标记的小批次数据)，通过模型自适应来提升分布外检测性能。测试时适应方法基于这样一种理论见解 [140] ，即仅使用分布内样本且无任何额外知识的情况下，是不可能检测出分布外样本的。这些方法可分为两类。根据模型在测试期间是否被修改，可分为*基于模型优化的方法*和*无需模型优化的方法*。这两种方法都会经历一个训练后阶段，在此阶段，无论模型是否更新，经过训练的模型都可以进行自适应调整。

基于模型优化的方法。基于模型优化的方法会在训练后阶段利用未标记数据来对已训练的模型进行增强。

这类方法中的一系列方法 [88, 91] 主张利用未标记的分布内数据和分布外数据的组合 (被称为 “野生数据”，在现实场景中数量丰富且易于获取) 来进行分布外检测。WOODS[88] 的出发点是清理野生数据以获得可靠的分布外候选数据，然后利用这些数据进行模型正则化。此后，为了理解野生数据在分布外检测中所起的作用，SAL[91] 从可分离性和可学习性的角度解释了它们是如何助力分布外检测的。值得注意的是，如果野生分布外数据并非来自测试数据集，那么野生数据并不完全等同于测试分布外数据，这必然会导致将非预期的信息纳入模型当中。此外，WOODS 中额外的训练要求虽然被认为是必要的，但会产生高昂成本，通常是不受欢迎的。

另一类基于模型优化的方法从半监督学习技术 [141] 中获取灵感，旨在训练后阶段实现更高效、轻量化的训练过程。伪标签法 [142] 是一种对测试数据进行标注的简单却有效的方式，它能增强从测试阶段的未标记数据中进行学习的能力。Yang 等人 [9] 提出的名为 AUTO 的方法仅使用伪分布外数据来对模型进行优化。在 AUTO 方法中，伪分布内数据的作用是以语义一致的目标来减轻灾难性遗忘问题，从而维持分布内分类的准确性。与之不同的是，ATTA [90] 和 SODA[89] 同时利用伪分布内数据和伪分布外数据来优化已训练的模型。SODA 采用双损失方法来同时处理伪分布内数据和伪分布外数据，而 ATTA 则通过不同的加权技术对它们加以区分。

无需模型优化的方法。在某些对安全性要求敏感的场景中，修改原始的已训练模型是不可行的。因此，那些能够在无需更新模型的情况下实现测试时适应的方法 (被称为“无模型优化” (MOF) 技术) 正日益受到关注。这些方法通过记忆测试数据或者在原始模型之上添加额外模块的方式，来提高测试数据的利用率。

ETLT[92] 和 GOODAT[94] 都通过训练一个附加模块来调整分布外得分，而非改变原始模型本身，以此保持原始模型的完整性。ETLT 观察到特征表示与测试输入的分布外得分之间存在线性关联。换句话说，对于给定的图像，由其特征和分布外得分组成的数对 *(特征，分布外得分)* 呈现出线性相关性。此外，上述的分布内数据和分布外数据的数对是线性可分的。基于这一观察结果，ETLT 提议学习一个基于 *(特征，分布外得分)* 数对训练而成的线性回归模块。完整测试数据集并非总是能够获取得到，因此，Fan 等人 [92] 还提供了一个在线变体以确保更安全地部署。同样地，GOODAT 开发了一个名为图掩码器的附加模块，该模块专为图数据而设计。它整合了基于信息瓶颈提升的损失，并将其用作分布外评分的度量指标。与之相反，AdaOOD[93] 通过一种非参数的 K 近邻方法避免了任何额外的训练负担。AdaOOD 背后的核心原理是维护一个记忆库，这与 AUTO 所采用的方法类似。

在线与离线。大多数事后方法 [143, 92] 传统上侧重于离线场景，在这种场景下，分布外检测器在部署后保持静态和固定不变。相比之下，大多数测试时方法 [93, 90] 采用在线场景来动态获取决策边界，从而将每个时间步长上出现错误的分布外预测的风险降至最低。

更具挑战性的场景。在测试时分布外检测场景的背景下，一些学者提出了更具挑战性的设置，这些设置对模型的能力水平有着更高要求。MOL[144] 引入了一种更贴合现实的问题场景，即持续自适应分布外 (CAOOD) 检测，旨在应对现实世界中分布内和分布外分布不断变化的挑战。在 MOL 中，采用元学习方法来使模型能够快速适应在各种场景中遇到的复杂情况。

5 问题：基于大型预训练模型的分布外检测

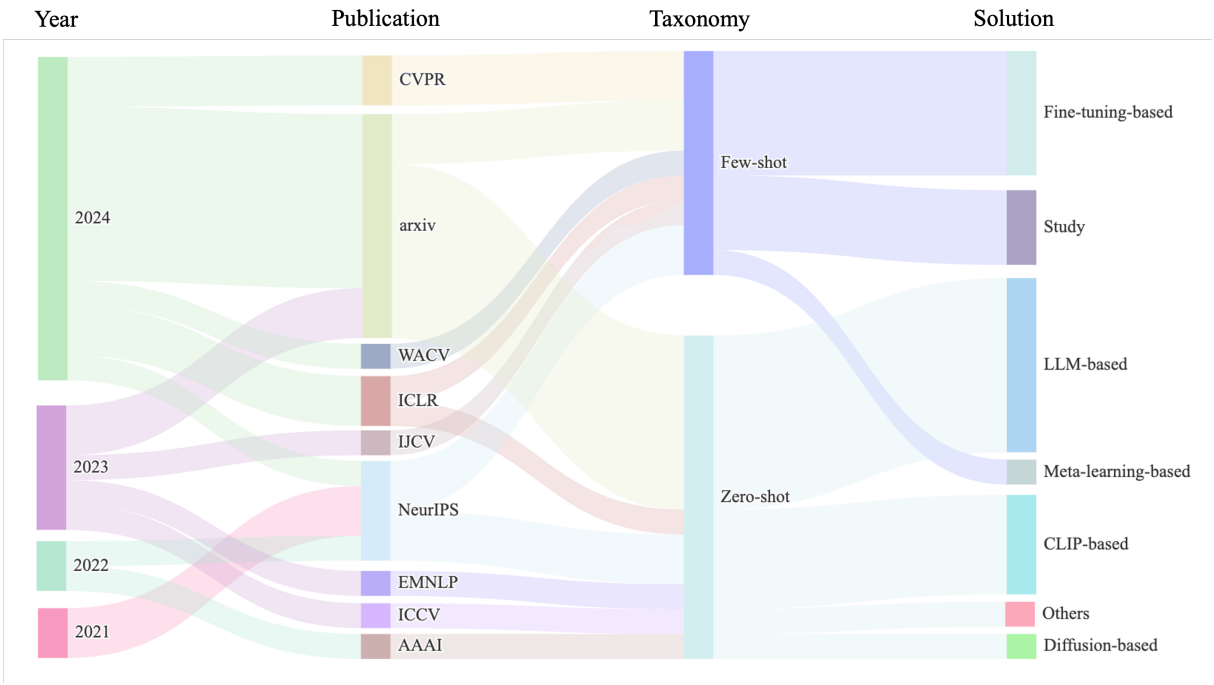

Figure 9: 基于大型预训练模型的分布外检测方法概述与趋势。

大型预训练模型在众多下游的分布内分类任务中展现出了卓越的性能，但它们在分布外检测任务中的潜力仍是一个尚未充分探索的领域。近期的研究 [145] 强调了更高的分布内分类准确率与更好的分布外检测性能之间存在关联。因此，基于大型预训练模型的分布外检测问题便自然而然地出现了。近年来，各种类型的大型预训练模型，包括单模态 (ViT[146] 、BERT[147] 、Diffusion[148])、视觉语言模型 (如 CLIP[149] 、多模态扩散模型 [150] 、ALIGN[151]) 以及大型语言模型 (如 GPT3[152])，越来越多地被用于分布外检测任务，如 Figure 9 所示。利用大型预训练模型强大的表征能力进一步放宽了分布外检测任务的限制，使得人们开始聚焦于更具挑战性和更贴合现实的场景，这已然成为一个新的研究热点。鉴于大型预训练模型所接触到的分布内样本数量，基于大型预训练模型的分布外检测可分为零样本学习 (Zero-shot)、小样本学习 (Few-shot) 和全样本学习 (Full-shot) 的分布外检测，如 Figure 10 所示。Table 2 总结了几种相关竞争方法的性能评估情况，以便了解该领域分布外检测的性能水平。

5.1 零样本检测方法

概述。给定大型预训练模型和分布内的类名，我们承担与分布外检测相同的任务，即精准地检测出分布外数据以避免对其进行预测，并对分布内数据进行准确分类。请注意，我们无需基于视觉语言模型的方法。现有的利用

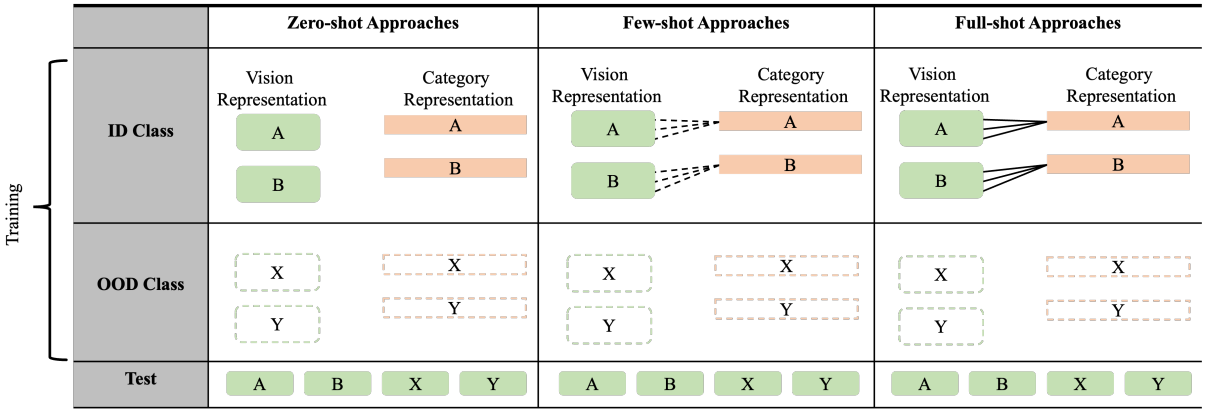

Figure 10: 基于大型预训练模型的分布外检测方法示意图。在训练阶段，零样本方法仅需要分布内类别的类别标签。小样本方法需要每个分布内类别的一部分图像以及类别标签 (用虚线表示)。全样本方法使用每个分布内类别的类别标签和所有图像。这些方法均不使用分布外类别的标签或图像。

Table 2: 使用 ImageNet-1K 数据集作为分布外数据集，以及 iNaturalist、SUN、Places 和 Textures 作为分布外数据集的一些竞争方法的性能评估。最佳结果以粗体显示。

场景	方法	iNaturalist		SUN		Places		Textures		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
零样本	ZOC [10]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
	MCM [98]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
	CLIPN [99]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
	NegLabel [153]	**99.49**	**1.91**	95.49	**20.53**	91.64	35.59	90.22	43.56	94.21	25.40
小样本	CoOp [154]	93.77	29.81	93.29	40.83	90.58	40.11	89.47	45.00	91.78	51.68
	LoCoOp [111]	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
	NegPrompt [112]	98.73	6.32	**95.55**	22.89	**93.34**	**27.60**	**91.60**	**35.21**	**94.81**	**23.01**

视觉语言模型进行零样本学习 (Zero-shot) 分布外检测的研究，依据将视觉语言模型用作基础框架的不同方式，可大致分为两类主要方法。**基于扩散模型的方法**。传统的生成式方法，例如扩散模型 [148] ，能够有效地捕捉分布内分布。随后，这些方法通过评估给定测试输入源自分布外来源的可能性，来识别分布外样本。在近期关于 RONIN[97] 的一项研究中，扩散模型已被应用于实现分布外目标检测，该研究利用扩散模型生成分布内图像修复内容，同时运用 CLIP 模型来计算相似度得分。**基于 CLIP 的方法**。鉴于 CLIP 在关联文本和图像方面表现出色，大量研究人员已尝试利用 CLIP 进行零样本学习分布外检测。通常，在零样本学习分布外检测设置中，会在预训练模型上添加附加组件，使其能更好地适应分布外检测任务。ZOC[10] 通过在一个大型图像字幕数据集 [155] 上训练图像描述生成器来实现零样本学习分布外检测，使模型能够生成候选的未见过的标签。需要注意的是，ZOC 仅仅将 CLIP 当作特征提取器，并未赋予 CLIP 自身分布外检测能力。后续的 CLIPN[99] 研究通过一个“无提示”编码器赋予 CLIP 说“不”的能力。其他方法则减少了额外训练的需求，侧重于提升事后零样本学习分布外检测的性能。基于 CLIP 的零样本学习分布外检测的一个简单基准是将归一化的文本 - 图像相

似度用作分布外得分。Ming 等人 [98] 进一步用 MCM 得分替代该得分，并对这一修改给出了全面的理论解释。此外，NegLabel[153] 、LAPT[102] 和 CLIPScope[96] 通过预定语料库丰富文本信息。具体而言，NegLabel 从语料库中选取与分布内标签不同的否定词 (标签)，以提升分布外检测性能。此外，CLIPScope 利用这些标签通过应用贝叶斯规则来修正原始得分。与之不同的是，LAPT 利用文本到图像生成模型或图像检索模型来获取与给定的分布内和否定类别标签相对应的图像，随后进行提示微调。值得注意的是，ZOC、CLIPN、NegLabel 和 CLIPScope 各自设计了独特的分布外得分，以便整合到各自的模型中。相反，MCM 得分因其在不同模型中的通用性而脱颖而出。MCM 得分的计算方式如下：

$$S_{MCM}(x';Y_{in},T,\tau)=\max_i \frac{e^{s_i(x')/\tau}}{\sum_{j=1}^{K} e^{s_j(x')/\tau}} \tag{6}$$

其中 x' 是测试输入图像；$y_{in}=\{y_1,y_2,\ldots,y_K\}$ 表示标签集；y_i 表示标签集中的第 i 个标签；$T(t_i)$ 对应于文本提示 t 的概念向量；$s_i(x')$ 是图像特征 $I(x')$ 与概念向量 $T(t_i)$ 之间的余弦相似度得分，计算为 $s_i(x')=\frac{I(x')\cdot T(t_i)}{\|I(x')\|\|T(t_i)\|}$；$e^{s_i(x')/\tau}$ 是由温度参数 τ 归一化的指数化相似度得分；K 是标签集中的标签数量；τ 是 softmax 函数的温度参数；S_{MCM} 一些研究已经探索了零样本学习 (Zero-shot) 分布外检测的变体或更具挑战性的场景，例如检测分布内对象 [156] 或处理带有噪声的标签 [157] ，这些内容将在第 7节中进行讨论。

基于大型语言模型的方法。随着大型语言模型的蓬勃发展，其在分布外检测领域的应用带来了新的机遇。大型语言模型的优势在于它们拥有广博的世界知识，能够提供有关分布内标签的详尽信息。然而，在利用大型语言模型执行分布外检测任务时，会产生虚假或误导性信息 (即所谓的“幻觉”现象) 的问题，这构成了一项重大挑战。现有的方法 [104, 103] 基于这样一种观察：两类图像可能具有非常相似的视觉特征，但在语义空间中却具有不同的属性。Dai 等人 [103] 引入了一种基于一致性的、结合目标检测器的校准方法，以减轻“幻觉”现象，并利用大型语言模型的世界知识来描述分布内。而 Huang 等人 [104] 提出了基于文本和图像模态中的分布内标签及分布外样本生成对等类别的方法，即 ODPC。需要注意的是，ODPC 需要分布内图像来训练 MLP，这意味着它并非一种零样本学习方法。关于分布外得分，Dai 等人 [103] 采用了 MSP，然而 ODPC 使用的是基于 K 近邻的得分。后来，有一项针对大型语言模型的研究 [105] 旨在探讨大型语言模型进行分布外检测时的三个问题：大型语言模型对近分布外和远分布外情况的倾向、不同微调方法的影响以及分布外得分的选择。Liu 等人 [105] 发现大型语言模型天生就是远分布外检测工具，生成式微调优于判别式微调，并且由于大型语言模型的嵌入空间呈现各向异性，所以余弦距离足以作为分布外得分。

备注。如上文所述，此处的“零样本学习”是指不接触分布内图像，仅

能获取分布内标签。然而，Fort 等人 [158] 认为，在部署场景中，分布外的标签要么是现成可用的。他们提出了一种基准方法的变体，将分布外类别的名称作为候选标签纳入其中，从而提升了性能。不过，在现实生活场景中，能够获取分布外标签的情况很少见，而且这种做法似乎不太合理。

5.2 小样本检测方法

概述。给定大型预训练模型以及少量分布内数据，我们可以利用这些分布内数据对模型进行调整，随后检测分布外测试数据。零样本学习 (Zero-shot) 分布外检测不需要任何训练图像，这使其适用于对安全性要求较高的场景。然而，它可能会面临与分布内下游数据的领域差异相关的挑战，这会限制零样本学习方法的性能。因此，在分布外检测中采用了许多小样本学习方法，而且这些小样本学习方法的有效性往往优于零样本学习分布外检测的有效性。

研究情况。利用若干分布内样本对大型预训练模型进行调整的直接方式就是微调。明和李 [159] 以及董等人 [110] 在视觉语言模型背景下就微调对分布外检测的影响开展了研究，尤其侧重于对 CLIP 的研究。明和李 [159] 更关注参数高效微调 (PEFT) 方法的作用以及分布外得分。明等人 [98] 提出的 MCM 得分作为一种创新举措，连同基于提示学习来检测分布外实例的方法，因其有效性而受到认可。同样地，Dong 等人 [110] 对诸如参数高效微调 (PEFT) 和传统方法等各种微调方法进行了广泛比较，发现参数高效微调 (PEFT) 在检测分布外实例方面表现更优，这与 [160] 的结论相呼应。在此之后，Kim 等人 [161] 认为 Finetune Like You Pretrain(FLYP) 这种微调方法因其在分类任务上的良好表现值得关注。具体而言，FLYP 旨在模仿像 CLIP 那样的对比语言-图像预训练。在比较了 FLYP 和 PEFT 方法在零样本学习分布外检测方面的性能之后，金等人 [161] 发现 FLYP 比 PEFT 能产生更好的分布外检测性能。需要注意的是，虽然 Fort 等人 [158] 讨论了小样本学习分布外检测，但他们的方法依赖的是少量离群样本而非分布内样本，因此不适合纳入本节内容之中。

基于微调的方法。利用有限的分布内数据对 CLIP 进行微调的一种常用方法涉及提示学习 [162, 163] ，在这种微调中，重点在于提示的上下文词语，保持预训练参数不变。为了区分分布内数据和分布外数据，LoCoOp[111] 和 ID-like Prompting Learning[109] 这两种方法都是从近分布内特征的角度来增强对上下文向量的学习。LoCoOp 利用熵最大化的方式，将与分布内无关的局部特征 (比如分布内图像中的背景) 与分布内文本嵌入拉开距离。同样地，ID-like Prompting Learning 会生成类分布内数据——位于分布内样本附近的离群值——来优化上下文向量。此外，它还引入了多样性损失，以增强所采样的分布外数据之间的差异性。然而，近期的 EOK-Prompt[107] 和 GalLop[108] 这两项工作指出，仅利用图像中的众多特征对一个提示进行优

化会浪费宝贵的信息。相反，EOK 提出了一种正交方法来优化新的局部提示，以便更好地利用有限的图像数据，而 GalLop 则分别利用全局特征和局部特征来优化全局提示和局部提示。此外，基于上述研究，DSFG[110] 认为，应用于 CLIP 的传统微调方法可能会不经意间导致关键的分布外知识丢失。为解决这一问题，DSFG 采取了一种策略，即将原始特征与经过微调修改后的特征合并，以此保留广泛的分布外知识，随后再训练一个分类器。DSFG 的即插即用功能使其能与所有基于微调的方法无缝兼容，增强了其实用性和价值。上述这些小样本学习微调方法不可避免地会在有限次样本上出现过拟合问题。Chen 等人 [164] 提出了一种无需训练的小样本学习分布外检测方法，Dual-Adapter[164] ，该方法通过提取正负特征构建了两种类型的适配器，以辅助进行分布外检测。目前所提到的小样本学习方法都缺乏对分布外的监督，因此 CLIP-OS[106] 试图寻找这样的监督，并且也取得了令人瞩目的成果。

基于元学习的方法。元学习旨在设计一种能够快速适应新挑战的学习方法 [165, 166] 。OOD-MAML[115] 对模型无关元学习 (MAML) 进行了调整，以用于小样本学习的分布外检测。它生成分布外样本，并将这些样本与分布内数据一同纳入经过调整的 N 路 K 次学习任务中，该任务会被划分为 N 个子任务，每个子任务侧重于 K 次学习的分布外检测。测试数据是否为分布外数据的判定是基于这些快速且简单的 N 个子任务的结果来做出的。与之不同的是，HyperMix[116] 主张采用基于超网络的方法来增强样本扩充，且无需额外的离群值。这是因为未包含在特定元训练任务中的类别可以充当分布外样本。

5.3 全样本检测方法

概述。这种设置通常不如前两种 (零样本学习和小样本学习) 设置贴合现实情况。不过，我们将它们单独列出，以确保对现有各类方法进行全面的回顾。给定完整的分布内数据及其相应标签，视觉语言模型通过微调能够显著增强分布外检测效果。此外，还引入了一项名为 “PT-OOD” 检测的新任务。

基于微调的方法。在能够获取完整数据集的情况下，就可以利用更多的数据对大型预训练模型进行微调，或者利用这些数据更好地模拟分布内分布，从而便于区分分布外数据。NPOS[29] 提出了一种非参数离群值合成技术，通过使用完整的分布内数据对 CLIP 进行微调，以此来区分分布内数据和分布外数据。与之不同的是，TOE[118] 在微调过程中同样使用交叉熵损失来约束模型，它以 OE 的理念为基础，着重关注 CLIP 框架内的文本离群值，以此来控制模型的识别能力，这与直接使用分布外图像的做法有着显著差异。

预训练-分布外 (PT-OOD) 检测。“PT-OOD” 样本是指在预训练数据中存在重叠情况的分布外样本。在对各种预训练方法 (有监督的、自监督的) 对预训练-分布外检测的影响进行调查和阐释之后，宫井等人 [117] 观察到特

征空间中较低的线性可分性会显著降低预训练-分布外检测的性能。他们建议针对每个实例使用具有区分性的特征，以此来区分分布内样本和分布外样本。

6 评估与应用

6.1 评估指标

在视觉领域的绝大多数分布外检测任务中，通常会用到以下评估指标：

受试者工作特征曲线下面积 (AUROC)。该指标用于量化分类器给分布内样本打出的分数高于分布外样本分数的可能性。较高的受试者工作特征曲线下面积值意味着模型性能更优，表明其区分分布内实例和分布外实例的能力更强。因此，该值越高越好。

精确率-召回率曲线下面积 (AUPR)。当把分布内视作正类时，该指标就与之相关，并且在类别分布不均衡的情况下尤为有价值。它用于评估精确率和召回率之间的平衡，精确率-召回率曲线下面积值越高，表明模型性能越优。因此，该值越高越好。

95% 真阳性率下的假阳性率 (FPR@95)。该指标描述的是在真阳性率达到 95% 这一节点处的假阳性率情况。它主要衡量的是被错误地识别为分布内的分布外样本所占的比例，由此可以了解模型在高敏感度阈值下出现误报的倾向。在保持对分布内样本高敏感度的同时，95% 真阳性率下更低的假阳性率意味着模型在正确标记分布外样本方面具有更强的特异性。因此，该指标的值越低越好。

这些指标从不同方面为分布外检测性能提供了全面的见解，比如区分分布内样本和分布外样本的能力、应对类别不平衡情况的能力、对分布内样本的敏感度，以及在特定真阳性率阈值下控制假阳性率的情况。

6.2 实验方案

在传统的分布外检测实验方案中，测试数据被严格划分为分布内数据或分布外数据。然而，随着该领域的发展，如今在分布外数据和分布内数据之间有了更细致的区分，这也导致了评估过程出现了变化。

随后，依据与分布内数据的协变量偏移程度，分布外数据被进一步细分为近分布外数据和远分布外数据。这种分类方式与将分布外检测任务划分为近分布外检测和远分布外检测相对应。显然，近分布外检测任务更具挑战性，不过，许多方法 [158, 98, 109] 已在这一领域展现出了卓越的性能。

近期，Yang 等人 [33] 、Bai 等人 [167] 提出，我们应当考虑分布内数据中发生协变量偏移的情况，而这在之前是未被纳入考量的。这对于防止模型泛化能力的丧失至关重要。前面提到的这类样本被称作协变量偏移分布内数据。因此，一种新的实验方案已被探索出来，名为全谱分布外检测。在测试

阶段，期望模型能够识别近分布外和远分布外实例。此外，它还应当拒绝对分布外数据进行预测，并准确地对分布内数据和协变量偏移分布内数据进行预测。

6.3 应用

计算机视觉。分布外检测方面的大部分工作都集中在计算机视觉领域，我们将众多与视觉相关的任务列举如下：

- **图像分类**。本所讨论的大多数任务都聚焦于图像分类领域内的分布外检测。在此类任务场景中，常用的分布内数据集包括 MNIST[168] 、CIFAR-10[169] 、CIFAR-100[169] 以及 ImageNet-1K[170] 。相应地构建了各种各样的分布外数据集，用以评估不同的方法 [171] ，其中最常使用的数据集有 iNaturalist[172] 、SUN[173] 、Places[174] 以及 Textures[175] 。此外，全谱分布外检测通常依据 Yang 等人 [33] 提出的三个基准 (即数字、物体和新冠数据集) 来进行评估。
- **语义分割**。近期的一些研究 [90] 已开始深入探究密集分布外检测任务，该任务也被称作异常分割。用于评估的数据集包括 Cityscapes 数据集 [176] 、Road Anomaly 数据集 [177] ，以及近期开发的面向语义分割的 SOOD-ImageNet 数据集 [178] 。
- **目标检测**。与分布外检测相关的方法在目标检测领域的应用还处于相对初期的阶段，只有少数研究在探索这一领域 [34] 。通常使用诸如 PASCAL-VOC[179] 和 Berkeley DeepDrive-100K[180] 等数据集来进行评估。
- **自动驾驶**。长期以来，自动驾驶一直是分布外检测的一项关键的实际应用领域。近期，Mao 等人 [181] 利用 CARLA[182] 系统来模拟并评估分布外检测在自动驾驶场景中的性能表现。
- **医学图像分析** 在医学图像分析领域，分布外检测至关重要。依据医学图像的具体类别，分布外检测会采用包括 CIFAR-10 以及 KvasirCapsul[183] 在内的各类数据集。

分布外检测在诸如人类行为识别 [184] 和太阳图像分析 [185, 186] 等多个领域有着重要应用。

自然语言处理。在众多自然语言处理应用的各类任务中，也对分布外检测进行了探索。最常见的两种应用如下：

- **意图检测**。意图检测是分布外检测在自然语言处理中的一项重要应用。用于评估的数据集包括 CLINC150[187] 、Banking[188] 、StackOverflow[189] 等。

- **文本分类**。在文本分类的分布外检测应用中，诸如 News Category[190] 和 SST-2[191] 等数据集通常被用来构建分布内/分布外类别对，并评估模型的检测能力。

超越计算机视觉和自然语言处理领域。除了上述两种数据模态之外，分布外检测在各类数据中仍有着诸多重要应用。

- **音频数据**。在音频分布外检测中，MSCW(Micro-EN)[192] 和 Vocalsound[193] 通常被用作分布内数据集，并且它们也互为对方的分布外数据集。
- **图数据**。近期的研究已经针对图数据中的分布外检测提出了各种各样的方法。现有的图级别分布外检测基准总共包含了来自 TU[194] 和 OGB 数据集 [195] 的 10 对数据集，这些数据集已得到了广泛应用。
- **强化学习**。目前，将分布外检测与强化学习 [196] 相结合以增强模型鲁棒性的趋势日益明显。穆罕默德和瓦尔登内格罗-托罗 [197] 提供了一个基准，并探索了构建可生成分布外数据的定制强化学习环境的方法。

7 新兴趋势与开放挑战

尽管分布外检测取得了快速进展，但仍存在众多新兴趋势和尚未充分探索的挑战。在本节中，我们从三个不同的视角探讨新兴趋势和开放挑战：方法、场景和应用。

7.1 更好的分布外检测方法

元学习适应。面对测试时分布外检测中的快速适应挑战，元学习算法提供了一种“学习如何学习”的范式，能有效地使模型适应新的测试数据，这或许是一种解决方案。此外，在处理训练驱动型分布外检测中潜在离群点所固有且数量巨大的样本空间问题时，改进的采样方法可能为高效利用离群点提供途径 [52] 。

理论驱动的分数设计。在传统的单模态体系中，许多事后分数是为分布外检测精心设计的，有效地降低了训练成本。然而，随着该领域进入多模态领域，对理论驱动的分数设计的需求日益增加。MCM[98] 是一个显著的例子，但它仅仅是 Softmax 的扩展，并未深入探究文本与图像之间的关系。因此，需要更先进的分数设计。

Table 3: 数据集总结。CARLA 系统是一个专为评估自动驾驶领域中分布外检测而设计的模拟平台，因此其整行都用"-" 填充。其他"-" 符号代表根据使用情况而变化的数字。

任务	数据集名称	数据类型	类别数	样本数	论　　文
图像分类	CIFAR-10	图像	10	60,000	[99, 10]
	CIFAR-100	图像	100	60,000	[99, 10]
	MNIST	图像	10	70,000	[2, 81]
	ImageNet-1K	图像	1,000	1,431,167	[98, 99]
	iNaturalist	图像	5,089	675,170	[98, 99]
	SUN	图像	397	108,754	[98, 99]
	Places	图像	> 205	$>$ 2,500,000	[98, 99]
	Textures	图像	47	5,640	[98, 99]
语义分割	Cityscapes	图像	-	25,000	[90]
	Road Anomaly Dataset	图像	-	100	[90]
物体检测	PASCAL VOC	图像	20	2,913	[97, 34]
	BBD100K	视频、图像	可变	100,000	[97, 34]
自动驾驶	CARLA System	-	-	-	[181]
医学影像处理	Kvasir-Capsul	图像	-	4,741,621	[198]
文本分类	News Category	文本	-	210,000	[199]
	SST-2	文本	-	215,154	[199]
意图检测	CLINC150	文本	150	22,500	[200]
	Banking	文本	77	13,083	[200]
	StackOverflow	文本	20	20,000	[200]
音频	MSCW	音频	-	$>$ 23400000	[201]
	Vocalsound	音频	-	21,024	[201]
图数据	TU	图数据	-	可变	[94]
	OGB	图数据	-	可变	[94]

7.2 分布外检测的更实际场景

在当前趋势下，由于现有不切实际限制的局限性，对出现更实际场景的需求日益增长。

快速测试时适应。像 CAOOD[144] 这样的测试时场景的出现，显著推进了分布外检测在现实世界中的应用，有望提高可靠性和适应性。

多模态检测。探索多模态的分布外检测能够增强我们对数据动态的把握，并提升模型在不同感官输入情况下的效能。此外，将大语言模型整合到多模态的分布外检测中的即将到来的举措有望变革该领域，它将先进的语言分析与其他数据类型相结合，为复杂应用打造更有效且更通用的检测系统。

开放词汇场景。现有的基于大型预训练模型的方法假定可以获取分布内类别。然而，在开放词汇场景中，这一假定并不成立。Li 等人 [112] 通过学习可迁移的负向提示来解决这一问题。它使用分布内标签的一个子集进行提示学习，并且这些学习到的提示能够应用于其他未见过的标签。开放词汇环境下的分布外检测进一步放宽了对事先获取分布内知识的要求，为开发更具通用性的模型铺平了道路。

噪声设置。虽然先前的研究大多聚焦于标准的干净标签设置，但很少有研究探索其另一面：噪声标签场景。Humblot 等人 [202] 对实际场景进行了调研，并检验了不同噪声水平对分布外检测的影响。随后，他们给出了在噪声标签环境下改进分布外检测的几个关键要点。

7.3 分布外检测的新应用

其他模态。尽管分布外检测在多个不同领域已取得长足进展，但其在语音和生理信号分析方面的潜力在很大程度上仍未得到充分挖掘。尤其在与情绪相关的生理信号中，由于受试者情绪状态的多变性，分布外问题十分普遍。因此，这是一个很有前景、值得进一步探索的领域。当前的一些方法 [203] 也提供了可供参考的基准。

人机协同应用。由 Vishwakarma 等人 [204] 提出的一个关键的未来发展方向，是将人类的洞察力融入检测过程。这种人机协同的方法，尤其在高风险决策中，对于提高分布外检测系统的准确性和响应能力、将人类直觉与算法精度相结合而言，将是至关重要的。

网络图像抓取。为了实现从互联网上自动抓取图像的流程，Miyai 等人 [156] 提出了一项零样本分布内检测任务，这一概念源于零样本分布外检测。如果一幅图像包含分布内对象，那么它将被归类为分布内图像；只有当它不包含任何分布内对象时，才会被视作分布外图像。这为分布外检测的应用提供了一个新颖的视角，值得进一步探索。

8 结论

分布外检测对于可信机器学习至关重要。在本文中，我们全面综述了分布外检测的最新进展，首次聚焦于问题场景视角：训练驱动型、训练无关型以及基于大型预训练模型的分布外检测。我们还总结了广泛使用的评估指标、实验协议和多样化的应用。我们相信，对现有论文的新颖分类以及对新兴趋势的广泛讨论将有助于更好地理解当前研究现状，协助研究人员选择合适的方法，并激发新的研究热点。

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. ICLR*, 2017.

[3] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. CVPR*, 2015.

[4] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Proc. IPMI*, 2017.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. CVPR*, 2012.

[6] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[7] Peng Cui and Jinjia Wang. Out-of-distribution (ood) detection based on deep learning: A review. *Electronics*, 11(21):3500, 2022.

[8] Hao Lang, Yinhe Zheng, Yixuan Li, Jian Sun, Fei Huang, and Yongbin Li. A survey on out-of-distribution detection in nlp. *Transactions on Machine Learning Research*, 2023.

[9] Puning Yang, Jian Liang, Jie Cao, and Ran He. Auto: Adaptive outlier optimization for online test-time ood detection. *arXiv preprint arXiv:2303.12267*, 2023.

[10] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proc. AAAI*, 2022.

[11] Yijun Yang, Ruiyuan Gao, and Qiang Xu. Out-of-distribution detection with semantic mismatch under masking. In *Proc. ECCV*, 2022.

[12] Yibo Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proc. CVPR*, 2022.

[13] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proc. CVPR*, 2023.

[14] Jingyao Li, Pengguang Chen, Shaozuo Yu, Shu Liu, and Jiaya Jia. Moodv2: Masked image modeling for out-of-distribution detection. *arXiv preprint arXiv:2401.02611*, 2024.

[15] Genki Osada, Tsubasa Takahashi, Budrul Ahsan, and Takashi Nishide. Out-of-distribution detection with reconstruction error and typicality-based penalty. In *Proc. WACV*, 2023.

[16] Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *Proc. ICML*, 2023.

[17] Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *Proc. ICCV*, 2023.

[18] Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proc. CVPR*, 2023.

[19] Jinglun Li, Xinyu Zhou, Pinxue Guo, Yixuan Sun, Yiwen Huang, Weifeng Ge, and Wenqiang Zhang. Hierarchical visual categories modeling: A joint representation learning and density estimation framework for out-of-distribution detection. In *Proc. ICCV*, 2023.

[20] Wenjian Huang, Hao Wang, Jiahao Xia, Chengyan Wang, and Jianguo Zhang. Density-driven regularization for out-of-distribution detection. In *Proc. NeurIPS*, 2022.

[21] Hamidreza Kamkari, Brendan Leigh Ross, Jesse C Cresswell, Anthony L Caterini, Rahul G Krishnan, and Gabriel Loaiza-Ganem. A geometric explanation of the likelihood ood detection paradox. In *Proc. ICML*, 2024.

[22] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proc. ICML*, 2022.

[23] Mouxiao Huang and Yu Qiao. Uncertainty-estimation with normalized logits for out-of-distribution detection. In *Proc. CAICE*, 2023.

[24] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proc. CVPR*, 2023.

[25] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proc. ICLR*, 2018.

[26] Keke Tang, Dingruibo Miao, Weilong Peng, Jianpeng Wu, Yawen Shi, Zhaoquan Gu, Zhihong Tian, and Wenping Wang. Codes: Chamfer out-of-distribution examples against overconfidence issue. In *Proc. ICCV*, 2021.

[27] Mengyu Wang, Yijia Shao, Haowei Lin, Wenpeng Hu, and Bing Liu. Cmg: A class-mixed generation approach to out-of-distribution detection. In *Proc. ECML & PKDD*, 2022.

[28] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *Proc. ICLR*, 2022.

[29] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *Proc. ICLR*, 2023.

[30] Gitaek Kwon, Jaeyoung Kim, Hongjun Choi, Byungmoo Yoon, Sungchul Choi, and Kyu-Hwan Jung. Improving out-of-distribution detection performance using synthetic outlier exposure generated by visual foundation models. In *Proc. BMVC*, 2023.

[31] Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. In *Proc. NeurIPS*, 2023.

[32] Sen Pei. Image background serves as good proxy for out-of-distribution data. In *Proc. ICLR*, 2024.

[33] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *IJCV*, pages 1–16, 2023.

[34] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Proc. NeurIPS*, 2022.

[35] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *Proc. ICLR*, 2023.

[36] Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. In *Proc. ICLR*, 2024.

[37] Sudarshan Regmi, Bibek Panthi, Yifei Ming, Prashnna K Gyawali, Danail Stoyanov, and Binod Bhattarai. Reweightood: Loss reweighting for distance-based ood detection. In *Proc. CVPR*, pages 131–141, 2024.

[38] Zhi Zhou, Lan-Zhe Guo, Zhanzhan Cheng, Yu-Feng Li, and Shiliang Pu. Step: Out-of-distribution detection in the presence of limited in-distribution labeled data. In *Proc. NeurIPS*, volume 34, pages 29168–29180, 2021.

[39] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proc. CVPR*, 2019.

[40] Deval Mehta, Yaniv Gal, Adrian Bowling, Paul Bonnington, and Zongyuan Ge. Out-of-distribution detection for long-tailed and fine-grained skin lesion images. In *Proc. MICCAI*, 2022.

[41] Hitesh Sapkota and Qi Yu. Adaptive robust evidential optimization for open set detection from imbalanced data. In *Proc. ICLR*, 2023.

[42] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Detecting out-of-distribution data through in-distribution class prior. In *Proc. ICML*, 2023.

[43] Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *Proc. ICML*, 2022.

[44] Laurens E Hogeweg, Rajesh Gangireddy, Django Brunink, Vincent J Kalkman, Ludo Cornelissen, and Jacob W Kamminga. Cood: Combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification. In *Proc. CVPR*, pages 3971–3980, 2024.

[45] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proc. ICLR*, 2019.

[46] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proc. ECCV*, 2018.

[47] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proc. CVPR*, 2019.

[48] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proc. AAAI*, 2020.

[49] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proc. NeurIPS*, 2020.

[50] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proc. WACV*, 2023.

[51] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proc. CVPR*, 2020.

[52] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *Proc. ICML*, 2022.

[53] Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. In *Proc. NeurIPS*, 2023.

[54] Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. In *Proc. ICLR*, 2023.

[55] Jianing Zhu, Geng Yu, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. In *Proc. NeurIPS*, 2023.

[56] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *Proc. ICML*, 2022.

[57] Wenjun Miao, Guansong Pang, Tianqi Li, Xiao Bai, and Jin Zheng. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. In *Proc. AAAI*, 2024.

[58] Hyunjun Choi, Hawook Jeong, and Jin Young Choi. Balanced energy regularization loss for out-of-distribution detection. In *Proc. CVPR*, 2023.

[59] Tong Wei, Bo-Lin Wang, and Min-Ling Zhang. Eat: Towards long-tailed out-of-distribution detection. In *Proc. AAAI*, 2024.

[60] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proc. ICML*, 2022.

[61] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proc. CVPR*, 2023.

[62] Feng Xue, Zi He, Yuan Zhang, Chuanlong Xie, Zhenguo Li, and Falong Tan. Enhancing the power of ood detection via sample-aware model selection. In *Proc. CVPR*, pages 17148–17157, 2024.

[63] Konstantin Kirchheim, Tim Gonschorek, and Frank Ortmeier. Out-of-distribution detection with logical reasoning. In *Proc. WACV*, pages 2122–2131, 2024.

[64] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proc. NeurIPS*, 2018.

[65] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proc. ICCV*, 2023.

[66] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proc. ICML*, 2022.

[67] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *Proc. ICLR*, 2021.

[68] Jinsol Lee and Ghassan AlRegib. Gradients as a measure of uncertainty in neural networks. In *Proc. ICIP*, 2020.

[69] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Proc. NeurIPS*, 2021.

[70] Sima Behpour, Thang Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Proc. NeurIPS*, 2023.

[71] Jinggang Chen, Junjie Li, Xiaoyang Qu, Jianzong Wang, Jiguang Wan, and Jing Xiao. Gaia: Delving into gradient-based attribution abnormality for out-of-distribution detection. In *Proc. NeurIPS*, 2023.

[72] Chao Chen, Zhihang Fu, Kai Liu, Ze Chen, Mingyuan Tao, and Jieping Ye. Optimal parameter and neuron pruning for out-of-distribution detection. In *Proc. NeurIPS*, volume 36, 2024.

[73] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proc. ICLR*, 2018.

[74] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Proc. NeurIPS*, 2021.

[75] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. Vra: Variational rectified activation for out-of-distribution detection. In *Proc. NeurIPS*, 2023.

[76] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *Proc. ICLR*, 2022.

[77] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proc. CVPR*, 2022.

[78] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. In *Proc. ICLR*, 2024.

[79] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *Proc. ICLR*, 2023.

[80] Yibing Liu, Chris Xing Tian, Haoliang Li, Lei Ma, and Shiqi Wang. Neuron activation coverage: Rethinking out-of-distribution detection and generalization. In *Proc. ICLR*, 2024.

[81] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Towards optimal feature-shaping methods for out-of-distribution detection. In *Proc. ICLR*, 2024.

[82] Fran Jelenić, Josip Jukić, Martin Tutek, Mate Puljiz, and Jan Šnajder. Out-of-distribution detection by leveraging between-layer transformation smoothness. In *Proc. ICLR*, 2024.

[83] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *Proc. ICLR*, 2024.

[84] Yue Yuan, Rundong He, Yicong Dong, Zhongyi Han, and Yilong Yin. Discriminability-driven channel selection for out-of-distribution detection. In *Proc. CVPR*, pages 26171–26180, 2024.

[85] Seong Tae Kim Yong Hyun Ahn, Gyeong-Moon Park. Line: Out-of-distribution detection by leveraging important neurons. In *Proc. CVPR*, 2023.

[86] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proc. AAAI*, 2022.

[87] Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. In *Proc. ICLR*, 2024.

[88] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *Proc. ICML*, 2022.

[89] Andrew Geng, Kangwook Lee, and Yixuan Li. Soda: Stream out-of-distribution adaptation. 2023.

[90] Zhitong Gao, Shipeng Yan, and Xuming He. Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. In *Proc. NeurIPS*, 2023.

[91] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *Proc. ICLR*, 2024.

[92] Ke Fan, Yikai Wang, Qian Yu, Da Li, and Yanwei Fu. A simple test-time method for out-of-distribution detection. *arXiv preprint arXiv:2207.08210*, 2022.

[93] YiFan Zhang, Xue Wang, Tian Zhou, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Model-free test time adaptation for out-of-distribution detection. *arXiv preprint arXiv:2311.16420*, 2023.

[94] Luzhi Wang, Dongxiao He, He Zhang, Yixin Liu, Wenjie Wang, Shirui Pan, Di Jin, and Tat-Seng Chua. Goodat: Towards test-time graph out-of-distribution detection. *arXiv preprint arXiv:2401.06176*, 2024.

[95] Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, and Xingqun Jiang.

Test-time linear out-of-distribution detection. In *Proc. CVPR*, pages 23752–23761, June 2024.

[96] Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad Khorrami. Clipscope: Enhancing zero-shot ood detection with bayesian scoring. *arXiv preprint arXiv:2405.14737*, 2024.

[97] Quang-Huy Nguyen, Jin Peng Zhou, Zhenzhen Liu, Khanh-Huyen Bui, Kilian Q Weinberger, and Dung D Le. Zero-shot object-level ood detection with context-aware inpainting. *arXiv preprint arXiv:2402.03292*, 2024.

[98] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Proc. NeurIPS*, 2022.

[99] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proc. ICCV*, 2023.

[100] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. In *Proc. ICLR*, 2023.

[101] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *Proc. ICLR*, 2024.

[102] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *ECCV*, 2024.

[103] Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. Exploring large language models for multi-modal out-of-distribution detection. In *Proc. EMNLP*, 2023.

[104] K Huang, G Song, Hanwen Su, and Jiyan Wang. Out-of-distribution detection using peer-class generated by large language model. *arXiv preprint arXiv:2403.13324*, 2024.

[105] Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are large language models at out-of-distribution detection? In *Proc. COLING*, 2024.

[106] Hao Sun, Rundong He, Zhongyi Han, Zhicong Lin, Yongshun Gong, and Yilong Yin. Clip-driven outliers synthesis for few-shot ood detection. *arXiv preprint arXiv:2404.00323*, 2024.

[107] Fanhu Zeng, Zhen Cheng, Fei Zhu, and Xu-Yao Zhang. Enhancing outlier knowledge for few-shot out-of-distribution detection with extensible local prompts. *arXiv preprint arXiv:2409.04796*, 2024.

[108] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. *arXiv preprint arXiv:2407.01400*, 2024.

[109] Yichen Bai, Zongbo Han, Changqing Zhang, Bing Cao, Xiaoheng Jiang, and Qinghua Hu. Id-like prompt learning for few-shot out-of-distribution detection. In *Proc. CVPR*, 2024.

[110] Jiuqing Dong, Yongbin Gao, Heng Zhou, Jun Cen, Yifan Yao, Sook Yoon, and Park Dong Sun. Towards few-shot out-of-distribution detection. *arXiv preprint arXiv:2311.12076*, 2023.

[111] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Proc. NeurIPS*, 2023.

[112] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proc. CVPR*, 2024.

[113] Jinglun Li, Xinyu Zhou, Kaixun Jiang, Lingyi Hong, Pinxue Guo, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Tagood: A novel approach to out-of-distribution detection via vision-language representations and class center learning. *arXiv preprint arXiv:2408.15566*, 2024.

[114] Lin Zhu, Yifeng Yang, Qinying Gu, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. Croft: Robust fine-tuning with concurrent optimization for ood generalization and open-set ood detection. In *Proc. ICML*, 2024.

[115] Taewon Jeong and Heeyoung Kim. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. In *Proc. NeurIPS*, volume 33, pages 3907–3916, 2020.

[116] Nikhil Mehta, Kevin J Liang, Jing Huang, Fu-Jen Chu, Li Yin, and Tal Hassner. Hypermix: Out-of-distribution detection and classification in few-shot settings. In *Proc. WACV*, 2024.

[117] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Can pre-trained networks detect familiar out-of-distribution data? *arXiv preprint arXiv:2310.00847*, 2023.

[118] Sangha Park, Jisoo Mok, Dahuin Jung, Saehyung Lee, and Sungroh Yoon. On the powerfulness of textual outlier exposure for visual ood detection. In *Proc. NeurIPS*, volume 36, pages 51675–51687, 2023.

[119] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[120] Tahereh Pourhabibi, Kok-Leong Ong, Booi H Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303, 2020.

[121] Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *Proc. SDM*, pages 25–36, 2003.

[122] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proc. CCS*, pages 1285–1298, 2017.

[123] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal processing*, 99:215–249, 2014.

[124] Quanzhi Li, Armineh Nourbakhsh, Sameena Shah, and Xiaomo Liu. Real-time novel event detection from social media. In *Proc. ICDE*, pages 1129–1139, 2017.

[125] Cem Aksoy, Fazli Can, and Seyit Kocberber. Novelty detection for topic tracking. *Journal of the american society for information science and technology*, 63(4):777–795, 2012.

[126] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proc. CVPR*, pages 1563–1572, 2016.

[127] Stephen Marsland, Ulrich Nehmzow, and Jonathan Shapiro. On-line novelty detection for autonomous mobile robots. *Robotics and Autonomous Systems*, 51(2-3):191–206, 2005.

[128] Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Daniela Rus. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training debiasing. In *Proc. IROS*, pages 568–575, 2018.

[129] Karanjit Singh and Shuchita Upadhyaya. Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, 9(1):307, 2012.

[130] Amruta D Pawar, Prakash N Kalavadekar, and Swapnali N Tambe. A survey on outlier detection techniques for credit card fraud detection. *IOSR Journal of Computer Engineering*, 16(2):44–48, 2014.

[131] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *Proc. ICML*, pages 2152–2161, 2015.

[132] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[133] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.

[134] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[135] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Proc. NeurIPS*, volume 30, 2017.

[136] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[137] Yair Wiener and Ran El-Yaniv. Agnostic selective classification. In *Proc. NeurIPS*, volume 24, 2011.

[138] Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

[139] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. In *Proc. CVPR*, volume 29, 2016.

[140] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Proc. NeurIPS*, 2022.

[141] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

[142] Eric Arazo, Diego Ortego, Paul Albert, Noel E O' Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proc. IJCNN*, 2020.

[143] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proc. ICLR*, 2018.

[144] Xinheng Wu, Jie Lu, Zhen Fang, and Guangquan Zhang. Meta ood learning for continuously adaptive ood detection. In *Proc. ICCV*, 2023.

[145] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *Proc. ICLR*, 2022.

[146] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021.

[147] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, 2019.

[148] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, volume 33, pages 6840–6851, 2020.

[149] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763, 2021.

[150] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. NeurIPS*, volume 35, pages 36479–36494, 2022.

[151] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*, pages 4904–4916, 2021.

[152] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. NeurIPS*, volume 33, pages 1877–1901, 2020.

[153] Anonymous. Negative label guided OOD detection with pretrained vision-language models. In *Proc. ICLR*, 2024.

[154] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[155] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755, 2014.

[156] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*, 2023.

[157] Choubo Ding and Guansong Pang. Zero-shot out-of-distribution detection with outlier label exposure. *arXiv preprint arXiv:2406.01170*, 2024.

[158] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Proc. NeurIPS*, 2021.

[159] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision*, 132(2):596–609, 2024.

[160] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *IJCV*, 2023.

[161] Jeonghyeon Kim, Jihyo Kim, and Sangheum Hwang. Comparison of out-of-distribution detection performance of clip-based fine-tuning methods. In *Proc. ICEIC*, pages 1–4, 2024.

[162] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proc. CVPR*, pages 16816–16825, 2022.

[163] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proc. CVPR*, pages 11175–11185, 2023.

[164] Xinyi Chen, Yaohui Li, and Haoxing Chen. Dual-adapter: Training-free dual adaptation for few-shot out-of-distribution detection. *arXiv preprint arXiv:2405.16146*, 2024.

[165] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95, 2002.

[166] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[167] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *Proc. ICML*, 2023.

[168] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[169] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[170] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[171] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. In *Proc. NeurIPS*, 2022.

[172] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proc. CVPR*, pages 8769–8778, 2018.

[173] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, pages 3485–3492, 2010.

[174] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[175] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proc. CVPR*, 2014.

[176] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. CVPR*, 2016.

[177] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proc. ICCV*, pages 2152–2161, 2019.

[178] Alberto Bacchin, Davide Allegro, Stefano Ghidoni, and Emanuele Menegatti. Sood-imagenet: a large-scale dataset for semantic out-of-distribution image classification and semantic segmentation. *arXiv preprint arXiv:2409.01109*, 2024.

[179] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[180] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, Trevor Darrell, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[181] Zhenjiang Mao, Dong-You Jhong, Ao Wang, and Ivan Ruchkin. Language-enhanced latent representations for out-of-distribution detection in autonomous driving. In *Proc. ICRA*, 2024.

[182] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Proc. CoRL*, pages 1–16, 2017.

[183] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142, 2021.

[184] Minho Sim, Young-Jun Lee, Dongkun Lee, Jongwhoa Lee, and Ho-Jin Choi. A simple debiasing framework for out-of-distribution detection in human action recognition. *Proc. ECAI*, 2023.

[185] Marius Giger and André Csillaghy. Unsupervised anomaly detection with variational autoencoders applied to full-disk solar images. *Space Weather*, 22(2):e2023SW003516, 2024.

[186] Suman Das, Michael Yuhas, Rachel Koh, and Arvind Easwaran. Interpretable latent space for meteorological out-of-distribution detection via weak supervision. *ACM Transactions on Cyber-Physical Systems*, 2024.

[187] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach,

Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proc. EMNLP-IJCNLP*, 2019.

[188] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proc. NLP4ConvAI*, pages 38–45, 2020.

[189] Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short text clustering via convolutional neural networks. In *Proc. NAACL*, 2015.

[190] Rishabh Misra. News category dataset. *arXiv preprint arXiv:2209.11429*, 2022.

[191] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*, pages 1631–1642, 2013.

[192] Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al. Multilingual spoken words corpus. In *Proc. NeurIPS*, 2021.

[193] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *Proc. ICASSP*, pages 151–155, 2022.

[194] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *Proc. ICML Workshops*, 2020.

[195] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Proc. NeurIPS*, 2020.

[196] Stefanos Leonardos, Daiki Matsunaga, Jongmin Lee, Jaeseok Yoon, Pieter Abeel, and Kee-Eung Kim. Addressing out-of-distribution joint actions in offline multi-agent rl via alternating stationary distribution correction estimation. In *Proc. NeurIPS*, 2023.

[197] Aaqib Parvez Mohammed and Matias Valdenegro-Toro. Benchmark for out-of-distribution detection in deep reinforcement learning. In *Proc. NeurIPS*, 2021.

[198] Qiaozhi Tan, Long Bai, Guankun Wang, Mobarakol Islam, and Hongliang Ren. Endoood: Uncertainty-aware out-of-distribution detection in capsule endoscopy diagnosis. In *Proc. ISBI*, 2024.

[199] Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. In *Proc. EMNLP*, 2021.

[200] Li-Ming Zhan, Haowen Liang, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. A closer look at few-shot out-of-distribution intent detection. In *Proc. COLING*, 2022.

[201] Zaharah Bukhsh and Aaqib Saeed. On out-of-distribution detection for audio with deep nearest neighbors. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[202] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? *arXiv preprint arXiv:2404.01775*, 2024.

[203] Hao Dong, Yue Zhao, Eleni Chatzi, and Olga Fink. Multiood: Scaling out-of-distribution detection for multiple modalities. *arXiv preprint arXiv:2405.17419*, 2024.

[204] Harit Vishwakarma, Heguang Lin, and Ramya Korlakai Vinayak. Taming false positives in out-of-distribution detection with human feedback. In *Proc. AISTATS*, 2024.