检索增强提示用于 OOD 检测

Ruisong Han[a], Zongbo Han[a], Jiahao Zhang[a], Mingyue Cheng[b], Changqing Zhang[a]

[a] *College of Intelligence and Computing, Tianjin University, Tianjin, China*
[b] *State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, China*

Abstract

异常分布（Out-of-Distribution, OOD）检测对于机器学习模型在现实环境中的可靠部署至关重要，它能精确识别与训练数据分布不同的测试样本。现有的方法依赖于辅助的异常样本或分布内（In-Distribution, ID）数据来生成用于训练的异常信息，但由于异常样本的有限性及其与真正测试 OOD 样本的不匹配，它们往往无法提供充足的语义监督，导致次优表现。为此，我们提出了一种称为检索增强提示（Retrieval-Augmented Prompt, RAP）的新颖 OOD 检测方法。RAP 通过检索外部知识来增强预训练视觉语言模型的提示，为 OOD 检测提供更强的语义监督。在训练过程中，RAP 根据与外部文本知识的联合相似性检索出描述异常的词语，并使用这些词语来增强模型的 OOD 提示。在测试过程中，RAP 根据遇到的 OOD 样本实时动态更新 OOD 提示，使得模型能够快速适应测试环境。我们的广泛实验表明，RAP 在大规模 OOD 检测基准测试中达到了最先进的性能。例如，在 ImageNet-1k 数据集上的 1-shot OOD 检测中，RAP 将平均 FPR95 降低了 7.05 %，并将 AUROC 提升了 1.71 %，相较于以往的方法。此外，全面的消融研究验证了我们方法中每个模块的有效性以及基础动机。

Keywords: OOD detection, Retrieval-augmented prompt, Vision-language models

1. 介绍

OOD 检测使得模型能够在部署期间识别出与训练分布不同的样本，从而保护模型免受 OOD 数据的负面影响 [1, 2] 。这种能力在野外应用中尤其重要，因为测试样本通常与控制的训练环境明显不同。这种能力对诸如智能医疗 [3] 和

自动驾驶 [4] 等安全关键应用至关重要，因为由于 OOD 输入导致的不可靠决策可能会带来严重后果。

如图 1 所示，目前的 OOD 检测方法可以广泛分为四大类：事后方法、训练时正则化方法、使用视觉-语言预训练模型的方法和外部知识引导的方法。事后方法通过设计各种评分函数来区分分布内 (ID) 和 OOD 样本。虽然便利，但由于缺乏可以监督模型的 OOD 信息，这些方法在性能提升方面受到限制。因此，一些训练时正则化方法探索使用辅助异常值来协助模型训练。这些方法利用辅助异常值来采样或挖掘有价值的异常值，并设计各种损失函数进行学习，从而增强模型检测 OOD 样本的能力。然而，这些方法也面临在难以获得足够多样性异常值数据集的情况下效果不佳的缺点。其他仅基于 ID 数据的研究，建模 ID 样本的分布并合成异常值进行学习，在某种程度上解决了辅助异常值不可用的问题。最近，在大规模数据集上预训练的视觉-语言模型在各种下游任务中表现出了巨大潜力。在零样本或小样本设置下，基于这些模型的 OOD 检测方法可以通过测量图像与 ID/OOD 提示之间的相似性，达到甚至超过使用整个 ID 训练集或辅助异常值的原始方法的性能。外部知识能够通过提供超出训练数据 [5] 的附加上下文和事实信息来增强模型性能。为了进一步释放视觉-语言预训练模型在 OOD 检测任务中的潜力，一些方法，例如 NegLabel [6] ，探索了使用外部知识。通过利用外部知识与 ID 提示之间的相似性来搜索负面提示，这些方法在零样本设置下取得了良好的检测性能。尽管这些方法各不相同，但在有效使用监督信号和处理分布变化方面仍然存在共同的挑战。

总结来说，上述方法中的许多通过利用 ID 样本或辅助离群点获取监督信息，然后从这些样本中学习，以增强模型检测 OOD 样本的能力。虽然这些方法已经识别出离群点的监督信息，但在学习 OOD 检测器的策略上仍然存在一些问题：1. 由于有价值的离群点样本稀缺，以往基于梯度的学习方法难以有效利用这些样本进行学习，因为其数量有限，往往不能提供足够的监督信号。2. 用于训练模型的离群点样本与在实际测试环境中遇到的 OOD 样本之间存在分布差异，这可能导致获得的 OOD 检测器表现不佳。这表明，仅依赖有限且可能存在偏差的离群点数据进行基于梯度的学习可能不足以训练出具有鲁棒性的 OOD 检测器，因而促使我们探索更稳定和更丰富的监督来源。

为了解决上述问题，我们借助像 WordNet [7] 这样的大型语料库中的外部知识，这些语料库包含高度精炼和完整的信息。通过适当利用这些知识，我们可

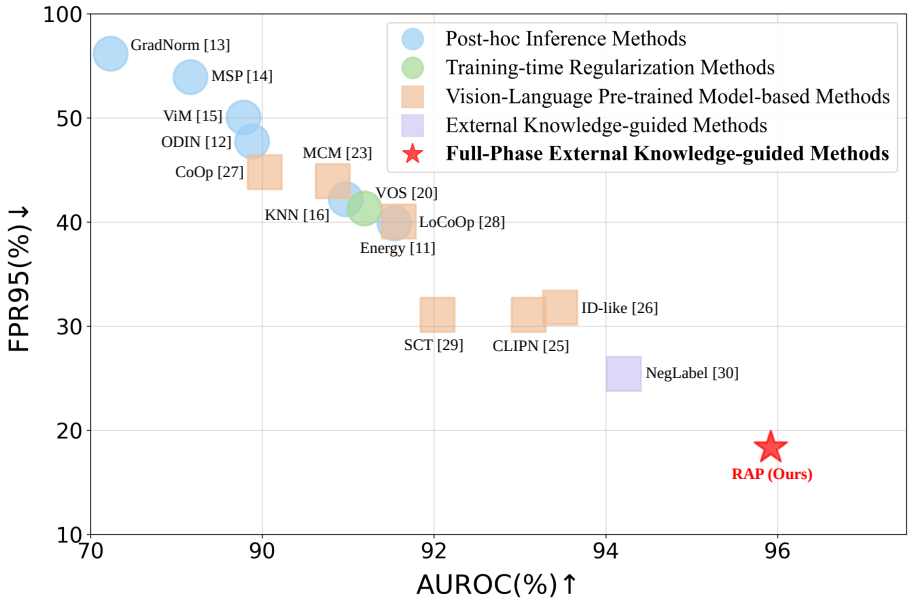

Figure 1: OOD 检测方法的演变。

以引入额外的语义监督。与使用有限异常监督的基于梯度的学习方法相比，结合用于提示检索增强的外部语义监督可以带来更好的 OOD 检测性能。我们提出了一种基于检索增强提示的 OOD 检测框架。通过利用来自知识库和图像的语义信息进行检索，我们可以高效地获取 OOD 提示以促进 OOD 检测。在训练阶段，我们首先使用少量的 ID 训练样本。通过随机裁剪训练图像并计算图像块与 ID 提示之间的相似性，我们获得有价值的异常图像表征。然后，我们根据联合相似性从外部知识中检索合适的文本作为 OOD 提示。联合相似性最大化准则使得 OOD 提示能够匹配前述有价值的异常图像表征，同时在细粒度图像和抽象文本语义中与 ID 表征保持距离，从而提高模型的检测性能。在测试阶段，为了减轻构造的异常与在实际测试环境中遇到的 OOD 样本之间差异的影响，我们通过动态感知其特征并细化 OOD 提示来连续适应测试样本。我们将我们的方法定义为全阶段外部知识引导法。贡献总结如下：

- 我们提出了一种基于检索增强提示的 OOD 检测框架，该框架利用外部知识的语义信息来增强模型的 OOD 提示。
- 在测试阶段，我们利用测试环境中遇到的潜在 OOD 样本来解决由训练期

间构建的异常值与实际 OOD 样本之间的差异导致的性能下降问题。

- 通过在多个 OOD 检测基准上的大量实验，我们的方法实现了最新的性能。我们还进行了深入分析来评估我们方法的效率。在各种基准上的消融研究也证实了在训练和测试阶段检索增强提示的有效性以及联合相似性的有效性。

自从 OOD 检测问题被提出以来，研究人员从不同的角度引入了各种算法，包括设计不同的 OOD 评分函数、结合辅助异常值监督、利用视觉-语言模型以及引入外部知识。具体来说，这些方法可以被分类为事后方法、训练时正则化方法、基于视觉-语言模型的方法和外部知识引导的方法。在本节中，我们详细介绍这四种类别。

1.1. 经典 OOD 检测方法

经典的 OOD 检测方法主要指的是不依赖于视觉语言模型，并使用整个 ID 训练数据集进行训练或微调的方法。这些方法可以进一步分为后处理方法和训练时正则化方法。

后验方法旨在通过设计各种 OOD 评分函数来区分内部数据（ID）和外部数据（OOD），无需超出标准分类模型训练之外的额外训练。由于它们不修改训练过程或目标，因此相对简单和方便。在训练成本高的领域，这些方法特别有价值，因为不需要重新训练。最早的方法之一是最大软最大概率（MSP）[8]，它使用最大软最大值作为 OOD 评分，作为早期基准。ODIN [9] 通过对输入施加扰动和对 logits 进行温度缩放来增强 OOD 检测，提高 ID 和 OOD 数据之间软最大值的可分性。基于能量的 OOD 检测 [10] 利用样本能量与其在训练分布下的可能性之间的理论关系而得出的能量分数，进一步提高了 OOD 检测性能。GradNorm [11] 认为仅依赖输出空间或特征空间的评分函数常常忽略了来自梯度空间的信息。它提出使用通过反向传播获得的梯度范数，基于软最大值输出与均匀分布之间的 Kullback-Leibler（KL）散度作为 OOD 评分。基于 KNN 的 OOD 检测 [12] 采用非参数方法，使用测试数据表示与训练数据表示之间的最近邻距离作为 OOD 评分，克服了以前依赖高斯分布假设的工作的局限性。ASH [13] 通过截断和重新缩放模型输出激活来提高 OOD 评分的可分性。

ISH [14] 进一步解释和分析了 ASH 的有效机制，并提出了一种在训练过程中重新加权的方法以增强 OOD 检测性能。

训练时正则化方法可以根据是否利用辅助异常数据分为两类。第一类使用辅助异常数据集，将有价值的异常表示提取为监督信号。通过设计各种损失函数，这些方法正则化模型参数以增强 OODS 检测性能。ATOM [15] 指出，大多数辅助异常可能不会帮助，甚至可能损害 OOD 检测器决策边界的学习。它通过识别和学习更具挑战性和信息量的异常来解决这一问题，从而产生更强的 OOD 检测器。DivOE [16] 指出，现有的辅助异常数据集可能缺乏足够的特征多样性。它提出对异常进行对抗攻击，以生成具有更大特征多样性的噪声增强异常，从而提升 OOD 检测。DOS [17] 认为仅根据不确定性选择异常可能会阻止模型捕获异常分布的全面视图。为了解决这一问题，它引入了一种简单而有效的方法，将不确定性与多样性相结合，以识别用于模型正则化的信息丰富且多样的异常。DiverseMix [18] 通过对异常数据应用 Mixup [19] 进一步增加了异常表示的多样性。从理论上讲，它表明辅助异常值的更大多样性有利于 OOD 检测器的学习。虽然使用辅助异常值的方法是有效的，但由于异常值数据集的高成本或不可用性，它们往往面临挑战。为了解决这个问题，一些研究人员探讨了仅使用 ID 训练集生成虚拟异常值的方法。VOS [20] 使用多元高斯分布对训练数据表示进行建模，并通过从低概率区域采样来模拟异常值。NPOS [21] 通过消除高斯分布假设并采用非参数方法来模拟异常值，从而改进了 VOS，并进一步增强了模型的 OOD 检测能力。MUFAS [3] 结合合成对抗性 OOD 样本和不确定性估计训练用于细粒度医学图像 OOD 检测的检测器。MPCL [1] 提出在训练阶段使用多样化多原型对比学习，以增强模型的 OOD 检测能力。

1.2. 基于视觉-语言预训练模型的 OOD 检测方法

视觉-语言模型的方法可以根据是否融入外部知识分为两类。对于第一类，ZOC [22] 使用预训练的 CLIP [23] 模型为每个测试样本生成候选未知类别名称。通过将候选类别名称与已知类别名称结合，它计算一个置信分数以实现零样本 OOD 检测。MCM [24] 提出了一种简单而有效的零样本 OOD 检测方法，该方法利用视觉特征与文本概念之间的对齐，并使用 CLIP。它使用来自 CLIP 的最大分类概率输出作为 OOD 分数，在具有语义相似类别的复杂 OOD 任务中表现突出。CLIPN [25] 引入可学习的负提示和一个文本编码器，以利用图像中的

负面语义，显著提高了 OOD 检测性能。CoOp [26] 针对 CLIP 的下游适应问题，通过提出一种简单而有效的提示学习方法。具体来说，CoOp 将提示中的上下文词建模为可学习的向量，同时保持所有预训练模型参数不变，实现自动提示优化。它提供了两种方法：统一上下文和类别特定上下文，提高了视觉语言模型在下游任务（包括 OOD 检测）中的适应性。LoCoOp [27] 聚焦于使用视觉语言提示学习进行少样本 OOD 检测。与传统的提示学习方法如 CoOp 不同，后者可能会受到文本嵌入中无关信息的影响，LoCoOp 使用来自 CLIP 的局部特征应用 OOD 正则化。通过将这些局部特征视为 OOD 特征，并学习将其与 ID 类文本嵌入分离，LoCoOp 消除了无关干扰，增强了 OOD 检测能力。ID-like [28] 通过对少样本 ID 数据集进行随机裁剪生成近 ID 样本。它引入了提示多样性和提示对齐正则化，以提高 CLIP 的 OOD 检测性能。然而，尽管 ID-like 能捕捉到有价值的离群表示，但不足的离群监督信号以及这些表示与真实测验 OOD 样本之间的分布差异降低了提示学习的有效性。

在第二类中，NegLabel [6] 观察到大多数现有的 OOD 检测算法很少利用来自文本模态的额外信息。在零-shot 设置下，NegLabel 从 WordNet [7] 语料库中检索负促发词，并使用与现有 ID 促发词的余弦相似性，进一步增强视觉-语言模型的 OOD 检测能力。然而，NegLabel 引入的外部知识常常不够准确，导致检索到的知识与真正的 OOD 样本不匹配，并且不能快速适应测试环境中的真实 OOD 样本。与之前的方法不同，我们提出的方法使用基于联合相似性最大化原则的训练阶段的 OOD 促发词检索策略以及测试阶段动态 OOD 促发词检索更新策略，引入更准确的外部知识。这减少了模型的 OOD 促发词与真实测试 OOD 样本之间的差距，进一步增强了模型的 OOD 检测能力。

这一部分首先介绍模型的整体框架，然后描述训练和测试阶段 RAP 的算法过程，最后解释模型的推理过程。

在介绍总体模型框架之前，我们首先描述由所提出方法解决的问题设置。所提出的 RAP 基于训练时少样本提示学习和测试时提示学习设置。训练数据集 $\mathcal{D} = \{x_1, x_2, ...x_N\}$ 每个类仅包含极少样本（在后续实验中，我们使用每类一个样本和每类四个样本的设置），其中 N 是训练样本的总数，并且不使用辅助的 OOD 数据。此外，该算法使用测试数据集实时更新提示以适应测试环境。接下来，我们介绍 RAP 的整体框架。

如图 2 所示，所提方法由 ID 提示 $T^{ID} = \{t_1^{ID}, t_2^{ID}, ...t_K^{ID}\}$ 、OOD 提示 $T^{OOD} =$

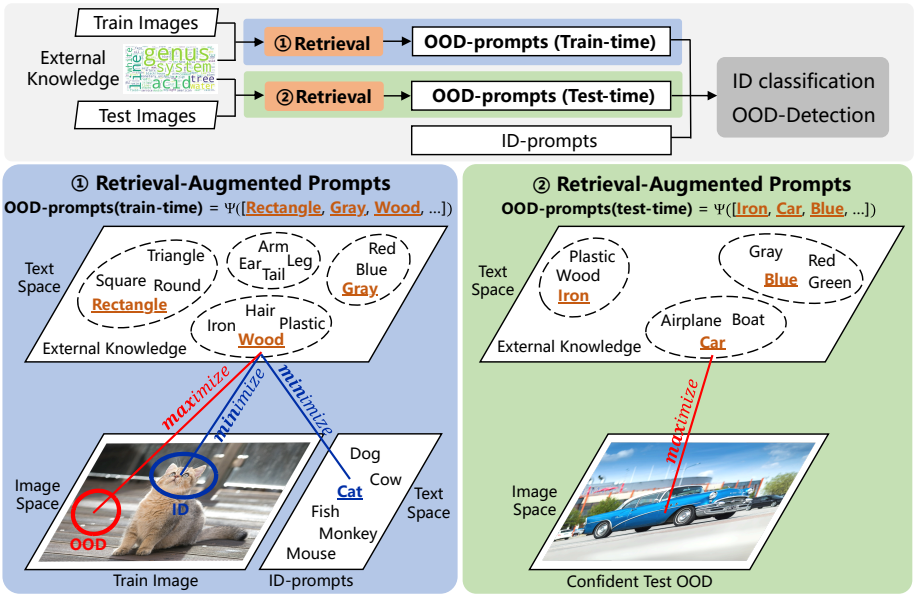

Figure 2: 提出的用于分布外检测的检索增强提示（RAP）概述。在训练阶段（图 2 的蓝色部分），基于词语表示与 ID、OOD 视觉表示以及 ID 提示表示的联合相似性来检索适合的 OOD 提示，而在测试阶段（图 2 的绿色部分），则基于词语表示与可信的测试 OOD 表示之间的相似性来检索 OOD 提示，以适应变化的测试环境。然后使用 ID 和 OOD 提示进行 ID 分类和 OOD 检测。

$\{t_1^{OOD}, t_2^{OOD}, ...t_C^{OOD}\}$ 以及 CLIP 模型 [23] 的视觉编码器 $\mathcal{I} : x \to \mathbb{R}^d$ 和文本编码器 $\mathcal{T} : t \to \mathbb{R}^d$ 组成。ID 样本的分类和 OOD 检测任务是基于 ID/OOD 提示和测试图像之间的余弦相似度进行的。这里，K 表示 ID 类别的数量，而 OOD 提示 $T^{OOD} = \{t_1^{OOD}, t_2^{OOD}, ...t_C^{OOD}\}$ 包括两个组成部分：训练中生成的 OOD 提示 $T_{train}^{OOD} = \{t_1^{OOD}, t_2^{OOD}, \ldots, t_{C_{train}}^{OOD}\}$ 和测试中生成的 OOD 提示 $T_{test}^{OOD} = \{t_1^{OOD}, t_2^{OOD}, \ldots, t_{C_{test}}^{OOD}\}$ 。C_{train} 和 C_{test} 分别表示在训练和测试阶段生成的 OOD 提示数量，C 是两者之和。这些提示通过从外部知识中检索适当的文本来获得，以实现准确的语义监督信号。在训练期间（在图 2 中用蓝色表示），根据其与 ID 和 OOD 视觉表示以及 ID 文本提示表示的相似性来检索单词，这些共同构成了联合相似性。检索到的词语作为语义监督信号，在训练期间增强 OOD 提示。在测试期间（在图 2 中用绿色表示），根据与被明确检测到的 OOD 样本的相似性，持续地检索词语并添加到 OOD 提示中。由于所提方法仅增强了 OOD 提示，因此不会影响 ID 分类的准确性。详细算法在算法 1 和算法 2 中描述。

1.3. 训练期间的检索增强提示

OOD 检测方法通常需要从 ID 数据或辅助 OOD 数据集中获得高质量的离群监控信号，以促进后续基于梯度的提示学习。因此，这项工作仅从有限的 ID 训练数据中构建有价值的离群表示。此外，由于离群表示在提供足够和有效的监督信号方面的困难，我们使用联合相似性从外部知识中检索语义监督信号。然后将检索到的语义监督信号用作 OOD 提示，以实现更有效的 OOD 检测。如算法 1 所示，该过程包括两个部分：(1) 从 ID 训练数据中构建有价值的离群表示，以及 (2) 在训练过程中使用检索增强的提示。

(1) 挖掘和学习困难的 OOD 表征可以使模型更准确地检测 OOD 样本 [29, 16, 17, 20, 21, 30, 28] ，因为这些具有挑战性的 OOD 表征通常与 ID 具有高度相关性，并且经常出现在 ID 样本附近，视觉上表现出 “看起来像 ID 但不是” 的特征 [28] 。我们假设局部采样 ID 训练图像可以产生既有价值的 ID 样本，也有价值的 “近似 ID” 样本（即离群点），这些样本被不足以特征化为 ID。因此，我们随机裁剪一小部分 ID 训练图像，并基于裁剪图像块和 ID 文本提示之间的余弦相似度获得有价值的离群点和 ID 表征 [28] 。由于 CLIP 模型 [23] 具有强大的零样本分类能力，我们直接使用其预定义的 ID 提示和预训练的视觉和文本编码器来获得有价值的 OOD 表征。

具体来说，我们首先获取训练图像的多个裁剪视觉表示。每张 ID 图像 x_i 都被随机裁剪 M 次以生成 $\mathcal{X}^{crop} = \{x_{1,1}^{\text{crop}}, x_{1,2}^{\text{crop}}, \ldots, x_{N,M}^{\text{crop}}\}$ ，然后使用 CLIP 视觉编码器 $\mathcal{I} : x \to \mathbb{R}^d$ 提取其表示以生成 $\mathcal{Z}^{crop} = \{z_{1,1}^{\text{crop}}, z_{1,2}^{\text{crop}}, \ldots, z_{N,M}^{\text{crop}}\}$ 。

接下来，我们使用 ID 提示及其对应的余弦相似度来选择有价值的 ID 和异常值表示。ID 提示 T^{ID} 是根据 ID 类别标签生成的模板。例如，给定标签 “狗”，模板将其转换为 “一张狗的照片”，从而创建相关的文本描述。该方法允许 CLIP 模型对图像和文本进行零样本分类对齐。ID 提示表示如下：

$$T^{ID} = \{\text{“a photo of a dog”}, \text{“a picture of a cat”}, \ldots\}. \tag{1}$$

使用 ID 提示 $T^{ID} = \{t_1^{ID}, t_2^{ID}, \ldots t_K^{ID}\}$ 和视觉表示 z ，余弦相似度计算如下：

$$\text{sim}(z, T^{ID}) = \frac{z \cdot T^{ID}}{\|z\| \times \|T^{ID}\|}. \tag{2}$$

我们选择与 ID 提示具有最高和最低余弦相似度的前 L 个视觉补丁，分别作为有价值的 ID 表示 $\mathcal{Z}_{train}^{ID}$ 和有价值的异常表示 $\mathcal{Z}_{train}^{OOD}$ 。形式上，对于每个图像的裁剪表示 $\mathcal{Z}_{j}^{crop}$ ，选择定义如下：

$$\mathcal{Z}_{train,j}^{ID} = \text{topk}(\text{sim}(\mathcal{Z}_{j}^{crop}, t^{ID}), \mathcal{Z}_{j}^{crop}, L), \tag{3}$$

$$\mathcal{Z}_{train,j}^{OOD} = \text{topk}(-\text{sim}(\mathcal{Z}_{j}^{crop}, t^{ID}), \mathcal{Z}_{j}^{crop}, L), \tag{4}$$

其中 $\text{sim}(\mathcal{Z}_{j}^{crop}, t^{ID})$ 表示相似度集，$\text{topk}(A, B, L)$ 表示根据集合 A 中的对应值从集合 B 中选择前 L 个元素。随后的有价值 ID 和异常值表示在训练阶段用于增强检索的提示学习。

(2) 现有的基于梯度的提示学习 OOD 检测方法由于在少量样本设置中缺乏有价值的异常样本而面临监督不足的挑战。这个问题使得这些方法难以提供足够准确和有效的监督信号，限制了它们的 OOD 检测能力。为了解决这个问题，如图 2 蓝色部分所示，RAP 从外部大规模结构化文本语料库 WordNet [7] 中检索单词作为额外的语义监督。这些外部语义监督信号直接在训练阶段用作 OOD 提示，用于后续的 OOD 检测，称为检索增强提示（训练时）。这些单词的选择是为了匹配有价值的异常表示，同时避免与 ID 表示在抽象的细粒度图像和文本语义层面上重叠，以确保准确性和有效性。

为了实现这个目标，我们的方法在检索过程中应用了一个联合相似性最大化原则。具体而言，对于 WordNet [7] 中的所有名词和形容词，我们使用 CLIP [23] 的文本编码器来提取词表示 $H^{W} = \{h_i^{W}\}_{i=1}^{N_W}$ ，并计算与 $\mathcal{Z}_{train}^{OOD}$ 、$\mathcal{Z}_{train}^{ID}$ 和 h^{ID} 的联合相似性 sim_{train} ，如下所示：

$$sim_{train} = \lambda_1 sim_1 + \lambda_2 sim_2 + \lambda_3 sim_3, \tag{5}$$

其中 sim_1 、sim_2 和 sim_3 分别表示词语表示 h_i^{W} 和 $\mathcal{Z}_{train}^{OOD}$ 、h_i^{W} 和 $\mathcal{Z}_{train}^{ID}$ 、以及 h_i^{W} 和 h^{ID} 之间的相似性。λ_1 、λ_2 和 λ_3 是平衡这些相似性（$\lambda_1 > 0$ 、$\lambda_2 < 0$ 、$\lambda_3 < 0$ ）的系数。为了获得 OOD 提示，我们首先选择具有最高联合相似性得分的前 P 个词。具体而言，选定的名词和形容词被纳入提示模板 $\Psi(\cdot)$ 中，如“the nice <n. >”和“这是一个 <adj. > 照片”，以构建用于训练的 OOD 提示，公式如下：

$$T_{train}^{OOD} = \Psi(\text{topk}(sim_{train}, H^W, P)). \tag{6}$$

最大化 sim_1 有助于通过语义监督获得与离群点一致的准确外部知识。这缓解了由于有价值的离群点表示的稀缺性导致的监督信号缺乏。同时，最小化 sim_2 和 sim_3 确保检索到的外部知识在细粒度图像和抽象文本语义两个层面上与 ID 样本表示保持距离，减少 ID 和 OOD 提示之间的语义重叠。具体而言，细粒度图像语义捕捉了详细的视觉信息，如物体形状、纹理和颜色，这有助于区分 OOD 和 ID 样本。另一方面，抽象文本语义关注高层次概念，包括物体类别和属性，支持更广泛的样本分类。通过在两个层面上将 ID 表示与外部知识进行比较，我们缓解了语义模糊性，提高了 OOD 检测的准确性和鲁棒性。

为了进一步防止局部最优，我们计算每个词表示 h_i^W 与 sim_1 及 sim_2 的所有 ID/异常视觉表示 z 之间的平均相似度如下：

$$sim_1 = \frac{1}{N \times S} \sum_{j=1}^{N} \sum_{l=1}^{L} sim(h_i^W, z_{j,l}^{OOD}), \tag{7}$$

$$sim_2 = \frac{1}{N \times S} \sum_{j=1}^{N} \sum_{l=1}^{L} sim(h_i^W, z_{j,l}^{ID}). \tag{8}$$

为了增强检索的鲁棒性 [6] ，我们使用词表示与 ID 提示表示之间的 η 百分位相似度来计算 sim_3 ，如下所示：

$$sim_3 = \text{percentile}_{\eta}(\{sim(h_i^W, h_j^{ID})\}_{j=1}^{K}). \tag{9}$$

1.4. 测试期间的检索增强提示

现有的大多数 OOD 检测方法仅在训练阶段优化提示，这使得它们难以快速有效地适应测试环境。为了解决这一问题，我们提出了一种方法，可以在测试时基于测试样本快速更新 OOD 提示。如图 2 绿色部分所示，我们首先根据测试样本的 ID 分数获得置信且有价值的 OOD 表示。然后，从外部知识中检索适当的词语，并将其添加到现有的 OOD 提示中，以实现对测试环境的在线适配。具体而言，如算法 2 所示，该过程包括两个主要组成部分：（1）从测试数据中获取 OOD 样本；（2）在测试期间增强检索的提示。

Algorithm 1 训练期间的检索增强提示

Input: Training data: $\mathcal{D} = \{x_1, x_2, ...x_N\}$.

Initialization: Pretrained CLIP model with image encoder $\mathcal{I} : x \to \mathbb{R}^d$ and text encoder $\mathcal{T} : t \to \mathbb{R}^d$.

/* Constructing Valuable Outlier representations from ID training data */

for $i = 1, 2, \ldots, N$ **do**

  Perform random cropping on each image x_i to obtain $\mathcal{X}^{crop}$.

  Use $\mathcal{I}$ to extract visual representations $\mathcal{Z}^{crop}$ from $\mathcal{X}^{crop}$.

  Generate ID prompts using Equation 1.

  Obtain valuable ID representations using Equation 3.

  Obtain valuable outlier representations using Equation 4.

end for

Collect all representations as $\mathcal{Z}^{ID}_{train}$ and $\mathcal{Z}^{OOD}_{train}$.

Compute sim_1 , sim_2 , sim_3 using Equations 7, 8, 9.

Calculate sim_{train} using Equation 5.

/* RAP During Training */

Retrieve and augment OOD prompts from WordNet using Equation 6.

Output: ID prompts T^{ID} and OOD prompts T^{OOD}_{train} .

(1) 在测试阶段，由于真正的 OOD 样本是可用的，我们不从 ID 测试样本中构造 OOD 图像表示。相反，我们从测试数据中选择自信和有价值的异常值作为监督信号，这基于由模型预测的 OOD 检测分数。具体来说，分数在 u_1 和 u_2 之间的测试样本被识别为有价值的 OOD 样本。这个中间分数阈值确保模型能够准确检测这些样本，这些样本也是相对有价值的，可能是真正的 OOD 样本，从而为模型提供更有效的监督信号，如下方方程式所示：

$$\mathcal{Z}^{testOOD}_{j} = \left\{ z_j \mid u_1 \leq S(z_j) \leq u_2 \right\}. \tag{10}$$

(2) 为了及时适应不断变化的测试环境，RAP 通过从有价值的测试 OOD 样本中学习，持续使用外部知识增强 OOD 提示。具体而言，如图 2 绿色部分所示，计算单词表示 h^W_i 和有价值的测试 OOD 图像表示 $z^{testOOD}_j$ 之间的相似性

$sim_{test} = sim(h_i^W, z_j^{testOOD})$ 。选择与测试 OOD 样本相似度最高的前 Q 个单词作为新的 OOD 提示 $T_{test,new}^{OOD}$ ，并将其添加到现有的测试时 OOD 提示 T_{test}^{OOD} 中。选定的新 OOD 提示 $T_{test,new}^{OOD}$ 由以下公式定义：

$$T_{test,new}^{OOD} = \Psi(\text{topk}(sim_{test}, H^W, Q)). \tag{11}$$

测试时的 OOD 提示会不断更新，以确保适应不断变化的测试环境。

1.5. ID 分类和 OOD 检测

该过程由两个主要部分组成：（1）ID 分类；（2）OOD 检测。

(1) ID 分类结果是基于视觉表示 z 与每个 ID 提示表示 t^{ID} 之间的余弦相似度获得的，如下公式所定义：

$$\hat{y} = \underset{y_i \in \mathcal{Y}}{\arg\max}\{sim(z, t_i^{ID})\}. \tag{12}$$

(2) 通过计算视觉表示与 ID 和 OOD 提示表示之间的余弦相似度，使用以下公式获得 ID 分数 $S(x)$ ：

$$S(x) = \frac{\sum_{i=1}^{K} \varphi(z, h_i^{ID})}{\sum_{i=1}^{K} \varphi(z, h_i^{ID}) + \sum_{i=1}^{C} \varphi(z, h_i^{OOD})}, \tag{13}$$

其中 $\varphi(z,h) = e^{\frac{sim(z,h)}{\tau}}$ ，而 τ 是来自 CLIP 的温度参数 [23] 。所提出的方法构建了一个 OOD 检测器 $F(x)$ 来执行 OOD 检测任务，其基于 OOD 检测分数 $S(x)$ ，如下式所述：

$$F(x) = \begin{cases} \text{ID}, & S(x) \geq \gamma \\ \text{OOD}, & S(x) < \gamma \end{cases}, \tag{14}$$

其中 γ 是判断样本是 ID 还是 OOD 的阈值。

在连续检索增强过程中增加 OOD 提示可能导致与 ID 样本 [6] 的语义重叠，从而降低 OOD 检测性能。为了缓解这一问题，所提出的方法遵循 NegLabel [6] 中使用的提示集合策略。OOD 提示被随机分成 N_g 组，其中每组 OOD 提示都与所有 ID 提示组合以计算 ID 分数。使用所有组的平均分作为最终的 OOD 检测分数，定义如下：

$$S_g(x) = \frac{1}{N_g} \sum_{j=1}^{N_g} S_j(x). \tag{15}$$

Algorithm 2 测试期间的检索增强提示

Input: Test data x_{test} , trained prompts T^{ID} and T^{OOD}_{train} .
/* Obtain Valuable OOD Representations */
Compute OOD detection scores using Equation 15.
Identify valuable OOD samples using Equation 10.
/* RAP During Testing */
if valuable OOD samples exist **then**
    Generate new OOD prompts $T^{OOD}_{test,new}$ using Equation 11.
    Update OOD prompts: $T^{OOD}_{test} \leftarrow T^{OOD}_{test} \cup T^{OOD}_{test,new}$.
end if
/* Inference */
Perform ID classification using Equation 12.
Perform OOD detection using Equation 14.
Output: ID predictions and OOD detection results.

2. 实验

本节详细描述了实验设置、结果、分析和各种消融研究。

2.1. 实验设计

RAP 中的实验包括以下几个部分：（1）主要实验，其中将所提出的方法与其他基准方法在两个大规模的 OOD 检测基准数据集上进行性能比较。（2）在训练和测试阶段应用提示检索增强的有效性。（3）在训练阶段，提示检索增强的联合相似性最大化原则的有效性。（4）检索提示数量对检测性能的影响。

2.2. 数据集描述

根据 MCM [24] 、NegLabel [6] 和 ID-like [28] 的设置，所提出的方法在标准和具有挑战性的 OOD 检测基准上进行了全面评估。在标准基准中，

ImageNet-1k [11] 被用作 ID 数据集，而 iNaturalist [31] 、Places [32] 、SUN [33] 和 Textures [34] 被用作 OOD 数据集。在具有挑战性的基准中，ImageNet-10 [24] 、ImageNet-20 [24] 和 ImageNet-100 [24] 交替用作 ID 和 OOD 数据集，确保 ID 和 OOD 数据之间没有语义重叠。以下是这些数据集的详细信息：

(1) iNaturalist [31] 是一个细粒度数据集，包含超过 5,000 种植物和动物的 859,000 张图像。每张图像的最大边长被调整为 800 像素。从 ImageNet-1k [11] 中不存在的 110 种植物中手动选择，总共随机抽样了这些类别的 10,000 张图像，用作标准基准测试的 OOD 数据集。

(2) SUN [33] 包含 397 个类别和超过 13 万张图像，所有图像都大于 200×200 像素。由于 SUN [33] 与 ImageNet-1k 存在一些类别重叠，因此选择了 50 个独特的自然相关类别（例如，森林、冰山），并随机抽取 10,000 张图像作为标准基准的 OOD 数据集。

(3) Places365 [32] 是一个和 SUN [33] 概念上相似的场景数据集，图像经过调整大小以确保最小尺寸为 512 像素。从中手动选择了 50 个在 ImageNet-1k 中不存在的类别，并从这些类别中选取了 10,000 张图像作为标准基准测试的 O0D 数据集。

(4) Textures [34] 包含 5,640 张纹理图像，尺寸范围从 300×300 到 640×640 像素。整个数据集用作标准基准测试的 OOD 数据集。

(5) ImageNet-1k [11] 涵盖了各种各样的真实世界对象类别。与常用的 CIFAR 数据集 [35] 相比，ImageNet-1k 包含了多达十倍以上的标签，其图像分辨率显著高于 CIFAR（ 32×32 ）或 MNIST（ 28×28 ）。ImageNet-1k 被用作标准基准的 ID 数据集。

(6) ImageNet-10、ImageNet-20 和 ImageNet-100 [24] 是 ImageNet-1k 的子集，其中的类别展示了某些语义上的相似性，使得它们更难以区分。例如，将狗（ID）与狼（OOD）区分开来。这些数据集在具有挑战性的基准测试中交替用作 ID 和 OOD 数据集。

2.3. 评价指标

所提出的方法使用 OOD 检测任务中的两个常用指标进行评估：(1) FPR95：当真正例率为 95 % 时，测量的假正例率。较低的 FPR95 表明模型在区分 ID

和 OOD 样本方面表现更好。(2) AUROC：代表接收者操作特征（ROC）曲线下的面积。较高的 AUROC 值（接近 1）表示模型性能更好。

2.4. 基线方法

所提出的方法与各种 OOD 检测方法进行了比较，包括事后方法（完全监督）、训练时正则化方法（完全监督）、基于视觉语言预训练模型的方法（零样本或少样本），以及外部知识引导的方法（零样本）。

事后方法：MSP [8] 使用最大 softmax 输出概率作为 OOD 得分。ODIN [9] 通过应用输入扰动和温度缩放改进了 MSP。Energy [10] 使用能量函数计算 OOD 得分。GradNorm [11] 使用梯度的向量范数作为 OOD 得分。ViM [36] 结合特征残差范数和主空间来计算 OOD 得分。KNN [12] 执行基于非参数最近邻的 OOD 检测。

训练时正则化方法：VOS [20] 使用类条件高斯分布合成虚拟 OOD 样本。NPOS [21] 通过在不假设 ID 嵌入分布的情况下生成合成 OOD 数据来扩展 VOS。

基于视觉-语言预训练模型的方法：Mahalanobis [37] 和 Energy [10] 最初被用作用于 OOD 检测的后处理方法，其中 Mahalanobis 距离和能量函数被用作 OOD 检测分数。按照 NegLabel [6] 的方法，我们将它们应用于视觉-语言模型中的零样本 OOD 检测。ZOC [22] 通过生成候选未知类别名称来扩展 CLIP [23] 。MCM [24] 使用来自 CLIP 的最大分类概率。CLIPN [25] 介绍了可学习的负提示和文本编码器用于 OOD 检测。CoOp [26] 使用常规提示学习用于 OOD 检测。LoCoOp [27] 通过利用 CLIP 中的局部特征增强 OOD 检测。ID-like [28] 通过使用随机裁剪生成近 ID 样本并应用提示学习。SCT [38] 引入了 OOD 正则化影响的自适应调节，以增强基于提示的 OOD 检测性能，特别是在有限的 ID 数据条件下。

外部知识引导的方法：NegLabel [6] 根据余弦相似度从 WordNet [7] 中检索负提示，以进一步增强 OOD 检测。

所有方法，包括 RAP，都采用 CLIP/ViT-B/16 作为预训练模型。对于训练集中的每个类别，我们随机选择 1 个和 4 个样本作为小样本训练数据集。在训练阶段，当从 ID 训练数据中构建异常值表示时，每个样本会进行 M 次随机裁剪，并选择与预定义提示最相似的 L 个样本，以及与预定义提示最不相似的

L 个样本，分别作为有价值的 ID 和异常值数据。当 M 太大时，过多的裁剪会导致计算成本增加和资源浪费。相反，当 M 太小时，很难对训练图像进行足够的裁剪以获得有价值的 ID 和异常值表示。同样，如果 L 太小，有价值表示的数量可能不足，会对 OOD 检测性能产生负面影响。另一方面，如果 L 太大，所选择的 ID 和异常值表示可能包含过多的噪声和无关信息。因此，参考类 ID [28] ，我们为 M 和 L 选择了适中的值，在我们的实验中将 M 设置为 256，将 L 设置为 32。

Table 1: 不同数据集的超参数设置

ID	OOD	Hyper Parameters				
		λ_1	λ_2	λ_3	u_1	u_2
ImageNet-1K	iNaturalist	0.2	-0.2	-1	0.5	0.6
	SUN	0.2	-0.2	-1	0.2	0.3
	Places	0.2	-0.2	-1	0.4	0.5
	Textures	0.2	-0.2	-1	0.4	0.6
ImageNet-10	ImageNet-20	0.05	-0.005	-1	0.0	0.5
ImageNet-20	ImageNet-10	0.1	-0.02	-1	0.0	0.2
ImageNet-100	ImageNet-10	0.1	-0.01	-1	0.0	0.2

在 RAP 的训练阶段，用于检索增强的提示数量 P 设置为 10,000。η 百分位值设置为 5。τ 设置为 0.01。在测试阶段，每个测试样本用于检索增强的提示数量 Q 设置为 4。提示集成中的组数设置为 100。联合相似度中的权重系数 λ_1 、λ_2 和 λ_3 以及 u_1 和 u_2 的值在表格 1 中显示。

表 2 展示了主要的实验结果。值得注意的是，我们提出的方法在 1-shot 设置中表现出强劲的性能，显著优于需要完整训练数据的方法。具体而言，使用完整训练数据的方法平均 AUROC 最高达到 91.54 %，FPR95 最低为 37.93 %，而基于视觉-语言预训练模型在 few-shot 或 zero-shot 设置中的方法则实现了 94.21 % 的最高平均 AUROC 和 25.40 % 的最低 FPR95。相比之下，我们的方法在 1-shot 设置中每个类别仅使用一个样本，即实现了平均 AUROC 95.92 % 和平均 FPR95 18.35 %，明显超越了使用完整训练数据的方法以及现有使用

Table 2: 标准基准上与多个基线的 OOD 检测性能比较。所有数值均为百分比。↑ 表示数值越大越好，↓ 表示数值越小越好。加粗的 红色 表示最好，加粗的 蓝色 表示第二好。

Methods	OOD Dataset								Average	
	iNaturalist		SUN		Places		Textures			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
	Full/Sub Data Fine-tune									
MSP [8]	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ODIN [9]	94.65	30.22	87.17	54.04	85.54	55.06	87.85	51.67	88.80	47.75
Energy [10]	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm [11]	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
ViM [36]	93.16	32.19	87.19	54.01	83.75	60.67	87.18	53.94	87.82	50.20
KNN [12]	94.52	29.17	92.67	35.62	91.02	39.61	85.67	64.35	90.97	42.19
VOS [20]	94.62	28.99	92.57	36.88	91.23	38.39	86.33	61.02	91.19	41.32
NPOS [21]	96.19	16.58	90.44	43.77	89.44	45.27	88.80	46.12	91.22	37.93
	Zero-shot									
Mahalanobis [37]	55.89	99.33	59.94	99.41	65.96	98.54	64.23	98.46	61.50	98.94
Energy [10]	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
ZOC [22]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
MCM [24]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
CLIPN [25]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
NegLabel [6]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
	One-shot									
CoOp [26]	91.26	43.38	91.95	38.53	89.09	46.68	87.83	50.64	90.03	44.80
LoCoOp [27]	92.49	38.49	93.67	33.27	91.07	39.23	89.13	49.25	91.59	40.06
ID-like [28]	97.35	14.57	91.08	44.02	91.08	41.74	94.38	26.77	93.47	31.78
SCT [38]	95.70	19.16	94.58	23.52	91.23	32.81	86.66	48.87	92.04	31.09
RAP(Ours)	99.55	1.48	96.35	17.55	93.68	27.13	94.09	27.25	95.92	18.35
	Four-shot									
CoOp [26]	92.60	35.36	92.27	37.06	89.15	45.38	89.68	43.74	90.93	40.39
LoCoOp [27]	93.93	29.45	93.24	33.06	91.32	41.13	90.54	44.15	92.26	36.95
ID-like [28]	98.19	8.98	91.64	42.03	90.57	44.00	94.32	25.27	93.68	30.07
SCT [38]	97.03	13.88	94.85	22.13	92.09	30.20	87.96	45.53	92.98	27.93
RAP(Ours)	99.53	1.65	96.34	17.40	93.54	27.98	94.19	26.63	95.90	18.42

zero-shot 或 few-shot ID 训练数据的方法。这表明 RAP 有效利用了视觉-语言预训练模型在 OOD 检测中的潜力。像 ID-like [28] 和 NegLabel [6] 这样的方法由于其引入的监督信号或外部知识的准确性和不足性而表现不佳。相比之下，我们的方法通过联合相似性最大化原则提升了提示准确性，并在测试阶段基于真实的 OOD 样本更新提示，将 FPR95 降低了 7.05 %，同时提升 AUROC 1.71 %。这些结果表明，我们提出的 RAP 通过引入结构化的外部知识文本词汇缓解了当前方法中缺乏有效监督信号的问题，为提示学习提供了更准确和有效的文本语义监督。

如图 5 和图 6 所示（放大查看细节），我们对标准基准上 RAP 的有效性进行了案例可视化分析。图 5 中的所有图像均来自 ID 数据集 ImageNet-1k [11] ，而图 6 中的所有图像均来自四个 OOD 数据集：iNaturalist [31] 、Places [32] 、SUN [33] 和 Textures [34] ，每个数据集选择了三张图片。每个子图由两部分组成：左侧显示了原始图像、其文件名、数据集名称、OOD 检测评分，以及一个指示样本是否为真实 ID/OOD 样本的标志（flag=1 表示真实 ID 实例，flag=0 表示真实 OOD 实例）。右侧展示了与 RAP 使用的 ID（蓝色）和负（绿色）标签相关联的前 5 个 softmax 正常化的亲和度。ID（蓝色）和负（绿色）标签与模板结合形成 RAP 的 ID 和 OOD 提示。RAP 引入的外部知识可以有效捕捉测试集中真实 OOD 样本的特征，同时与 ID 样本保持低相似度。例如，图 6 的第一行第一列的测试图像被负（绿色）标签“死荨麻”准确描述，而第一行第二列的图像与标签“曼陀罗”匹配。这些可视化分析证明了 RAP 在训练期间使用联合相似度最大化原则检索 OOD 提示的有效性，同时在测试时基于潜在的真实 OOD 样本更新 OOD 提示。

图 3 比较了我们的方法和类似 ID 的 [28] 在 ImageNet-1k 基准上使用 PLACES 作为 OOD 数据集的得分分布（可放大查看细节）。x 轴表示测试样本的 OOD 检测得分，y 轴显示这些得分的概率分布。蓝色和红色曲线分别代表 ID 和 OOD 样本的得分分布。与类似 ID 的 [28] 相比，我们的方法在 ID 和 OOD 得分分布之间明显重叠较少。具体来说，类似 ID 的 [28] 保留了大量得分接近 0 的 ID 样本和得分接近 1 的 OOD 样本，表明得分分布严重重叠。这使得难以找到合适的阈值来区分 ID 和 OOD 数据。相比之下，与类似 ID 的 [28] 相比，我们的方法几乎没有得分接近 0 的 ID 样本，并且得分接近 1 的 OOD 样本显著减少，显示出更优越的判别能力。

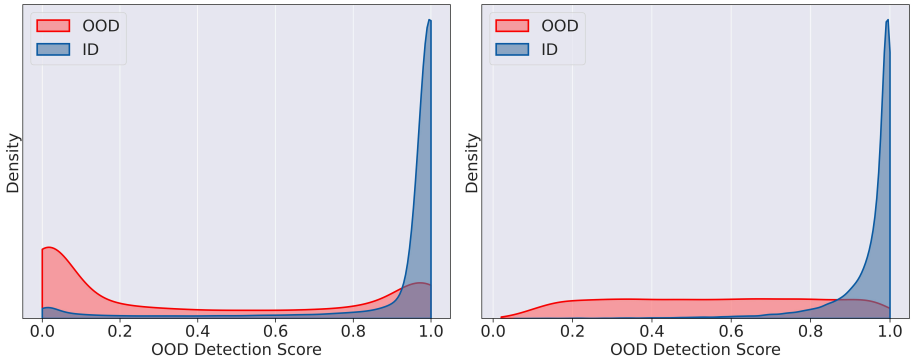

Figure 3: 得到的 ID 和 OOD 分数的密度，其中 ID 样式（左侧）和提出的方法 RAP（右侧）。

从表 2 中我们观察到，在标准基准测试中，我们的方法在 1-shot 和 4-shot 设置下的表现相似。这表明只用每类一个样本学习 OOD 表示，再结合我们方法引入的额外准确语义监督信号，已经足够实现强大的性能。然而，在更具挑战性的基准测试中，4-shot 设置表现优于 1-shot 设置，这表明更多的训练样本对于具有更高语义复杂度的数据集是有益的，能够提供更全面的监督信息。

此外，与需要数小时训练的基于梯度的提示学习方法相比，我们提出的方法在训练阶段仅需约 15 分钟。在测试阶段，当不需要更新提示时，RAP 每张图像仅引入额外的 $O((P+Q)D)$ FLOPs（其中 P 和 Q 分别为训练和测试的 OOD 提示数量，D 为嵌入维度），导致网络时延不到 1 %。当需要更新提示时，RAP 引入额外的 $O(WD)$ FLOPs，其中 W 为 WordNet 词汇数量。整个测试过程每张图像仅需几毫秒，这突显了我们的方法在实时推理和快速产品迭代场景中的高效性。

为了进一步验证我们方法在 ID 和 OOD 类之间具有较高语义相似度的困难基准上的性能，我们遵循 NegLabel 和 MCM 的实验设置，交替使用 ImageNet-10、ImageNet-20 和 ImageNet-100 作为 ID 和 OOD 数据集来评估 OOD 检测性能。结果如表 3 所示，其中方法的下标表示训练中每类使用的样本数量。例如，ID-like $_1$ [28] 表示每类一个样本，ID-like $_4$ 表示每类四个样本。

在具有挑战性的基准测试上的结果进一步凸显了现有基于梯度的少样本方法的局限性。例如，ID-like [28] 由于缺乏有价值的离群点表示以及这些表示与真实测试 OOD 样本之间的分布差异，在处理困难样本时显得力不从心，导致过

Table 3: 与多个基线在具有挑战性基准上的 OOD 检测性能比较。所有值均为百分比。方法名称中的数字下标表示每类使用的样本数。例如，RAP $_1$ 对应于 1-shot 设置。粗体 红色 表示最好。

ID Dataset	OOD Dataset	Method	AUROC↑	FPR95↓
ImageNet-10	ImageNet-20	NegLabel [6]	98.58	5.40
		ID-like $_1$ [28]	92.52	46.30
		RAP $_1$ (Ours)	98.98	3.20
		ID-like $_4$ [28]	94.22	34.90
		RAP $_4$ (Ours)	99.14	2.40
ImageNet-20	ImageNet-10	NegLabel [6]	97.33	14.00
		ID-like $_1$ [28]	92.86	42.80
		RAP $_1$ (Ours)	98.23	7.20
		ID-like $_4$ [28]	84.43	73.20
		RAP $_4$ (Ours)	98.31	6.40
ImageNet-100	ImageNet-10	NegLabel [6]	85.40	64.00
		ID-like $_1$ [28]	81.76	82.20
		RAP $_1$ (Ours)	88.40	53.60
		ID-like $_4$ [28]	84.75	69.80
		RAP $_4$ (Ours)	88.60	51.20

Table 4: 关于是否在标准基准测试的训练/测试阶段应用提示检索增强的消融研究。

Methods	OOD Dataset									
	iNaturalist		SUN		Places		Textures		Average	
	AUROCâ† ‘	FPR95â† “	AUROCâ† ‘	FPR95â† “	AUROCâ† ‘	FPR95â† “	AUROCâ† ‘	FPR95â† “	AUROCâ† ‘	FPR95â† “
train	99.48	1.78	95.69	20.16	92.06	34.01	91.63	36.19	94.71	23.03
train & test	99.55	1.48	96.35	17.55	96.38	27.13	94.09	27.25	95.92	18.35

拟合和有限的 OOD 检测性能。相反，我们的方法引入了来自外部知识的额外准确且充分的文本监督信号进行提示学习，从而实现了卓越的性能。

为了验证 RAP 在训练和测试阶段的有效性，我们在标准和具有挑战性的基准上进行了消融实验。具体来说，在训练阶段，我们像往常一样执行提示检索增强，而在测试阶段，我们没有应用 RAP（测试时），仅使用训练期间检索到的 ID 提示和 OOD 提示进行 OOD 检测。

如表 4 和表 5 所示，令人振奋的是，即使在测试阶段没有使用 RAP，所提出的方法在标准基准上的平均 AUROC 达到 94.71 %，平均 FPR95 达到 23.03 %。相比之下，类似 ID 的 [28] ，每类使用一个样本，仅实现了平均 AUROC 93.49 % 和平均 FPR95 31.77 %，这明显低于我们所提方法的性能。

在具有挑战性的基准测试中，改进更加显著。例如，当将 ImageNet-10 用作 ID 数据集而将 ImageNet-20 用作 OOD 数据集时，我们的方法实现了平均 AUROC 为 98.60 % 和平均 FPR95 为 4.60 %，显著优于 ID-like 的 AUROC 为 92.52 % 和 FPR95 为 46.30 %。这突显了利用外部知识生成更有效提示的优势。现有的基于梯度的提示学习方法由于缺乏有价值的 OOD 表示而难以提供足够的监督信号。相比之下，我们的方法利用联合相似性最大化原则从外部来源检索适当的知识，提供准确的语义监督。这些检索的文本直接用作 OOD 提示，以进一步促进提示检索增强，从而在 OOD 检测中获得显著的性能提升。

此外，在测试阶段应用 RAP 进一步提高了 OOD 检测效果。由于测试集中的 OOD 样本通常与训练期间使用的有价值的 OOD 表示不同，现有方法在有效区分 ID 和 OOD 数据方面面临困难。我们的方法通过根据 OOD 检测分数选择有价值的 OOD 样本，并从外部知识中检索匹配的词汇以更新 OOD 提示来缓解此问题。这种自适应检索策略增强了检测性能，展示了所提出方法在测试场景中的鲁棒性。

Table 5: 关于在具有挑战性的基准测试中，训练/测试阶段是否应用提示检索增强的消融研究。

ID Dataset	OOD Dataset	Method	AUROC↑	FPR95↓
ImageNet-10	ImageNet-20	train	98.60	4.60
		train & test	98.98	3.20
ImageNet-20	ImageNet-10	train	98.05	8.80
		train & test	98.23	7.20
ImageNet-100	ImageNet-10	train	87.66	57.60
		train & test	88.40	53.60

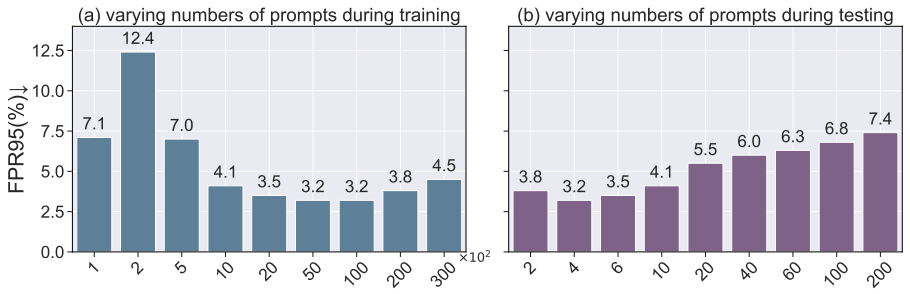

Figure 4: 消融研究了在训练和测试阶段提示数量变化对检测性能的影响。

为了评估每个相似性在联合相似性最大化原则中的贡献，我们进行了一个消融研究，通过每次去掉一项相似性（sim_1 、sim_2 或 sim_3 ）并使用剩余的两项进行 RAP。标准和挑战性基准上的结果如表 6 所示。具体而言，在标准基准上，去掉 sim_1 、sim_2 和 sim_3 分别导致 FPR95 上升至 21.27 %、23.38 % 和 19.97 %，AUROC 下降至 94.95 %、94.66 % 和 95.55 %。相比之下，使用所有三项相似性则导致 FPR95 为 18.35 %，AUROC 为 95.92 %。在具有挑战性的基准上也观察到了类似趋势，实现了 FPR95 为 21.40 %，AUROC 为 95.20 %，这凸显了每个相似性的的重要性。

sim_1 确保 OOD 提示在细粒度图像层面上更符合 OOD 表示，减少误报。细粒度图像特征，如纹理、边缘和背景细节，有助于准确识别 OOD 样本。sim_2 和

Table 6: 在标准和具有挑战性的基准测试中训练阶段的提示检索增广过程中，对联合相似性中每个相似性的消融研究。

Joint Similarity	Standard Benchmark		Challenging Benchmark	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
w.o. sim_1	94.95	21.27	93.20	32.40
w.o. sim_2	94.66	23.38	95.23	21.70
w.o. sim_3	95.55	19.97	94.73	25.87
w. $sim_{1,2,3}$	95.92	18.35	95.20	21.33

sim_3 在抽象语义和细粒度图像层面上确保 OOD 提示与 ID 表示之间的充分分离，尽量减少分布重叠。每种相似性都有助于从不同的角度优化检测策略，使联合相似性最大化原则成为准确文本监督的有效方法。

为了分析 RAP 中使用的提示数量对检测性能的影响，我们评估了模型在训练和测试阶段使用不同数量提示的表现。结果如图 4 所示。当检索的提示数量不足时，模型缺乏必要的语义监督信号，导致检测性能不佳。这在复杂的 OOD 样本中特别明显，由于模型的判别能力下降，FPR95 显著增加。

然而，当检索到过多提示时，模型获取了过多的信息，其中包括与 OOD 样本相关性低的无关词语。这在 OOD 提示中引入了语义噪音，增加了 OOD 和 ID 提示之间的重叠，从而降低了检测性能。因此，选择适当数量的提示对于有效的 OOD 检测至关重要。检索到的提示数量的平衡可确保准确的 OOD 语义监督表示，同时最大限度地减少 ID 和 OOD 之间的重叠，从而最大化检测性能。

在这项工作中，我们提出了一种新颖的检索增强提示学习框架（RAP）用于 OOD 检测。通过利用外部知识来源丰富的文本信息，我们检索语义相关的文本以增强 OOD 提示，从而引入有效的额外语义监督。为了解决当前 OOD 检测方法中的两个基本挑战——可用监督的缺乏以及由训练和测试数据分布偏移引起的性能下降——我们设计了训练期间的检索增强提示和测试期间的检索增强

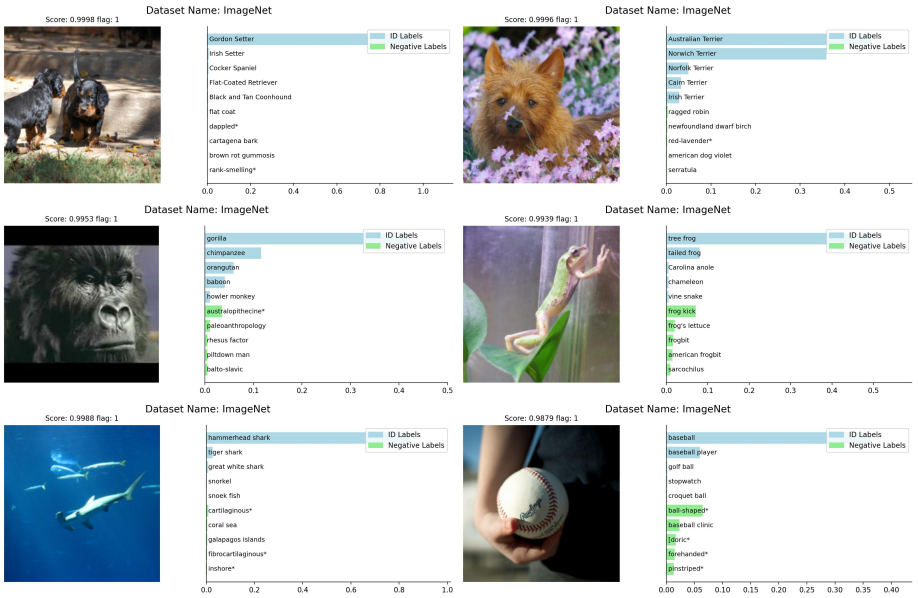

Figure 5: 身份图片的案例可视化。每个子图左侧显示原始图像，以及其文件名、数据集、OOD 检测得分和表示真实 ID 的标志（标志 =1）。右侧显示 RAP 使用的对 ID 标签（蓝色）和负标签（绿色）的 softmax 归一化相关性。

提示。这些组件有助于弥合训练和推理场景之间的差距，并提高模型对分布不一致的鲁棒性。在多个基准测试中进行的大量实验以及对真实数据集的案例研究，证明了我们的方法在 OOD 检测中的卓越性能。我们进一步提供了对 RAP 效率的全面分析，并进行了消融研究，以验证每个组件的有效性，包括用于检索的联合相似性策略。该领域的研究人员可以在真实的开放世界测试环境中部署我们的方法，以实现快速有效的 OOD 检测。此外，我们的框架为其他研究人员提供了一个灵活的基础，可以结合更准确和多样化的外部知识来源，以进一步提升检测性能。虽然我们的方法达到了强大的 OOD 检测性能，但目前它仅依赖于从大规模文本语料库中检索到的外部知识。这种依赖可能会限制进一步的性能提升，尤其是在更具挑战性的 OOD 场景中。作为未来的工作，我们计划探索整合嵌入在大型语言模型中的丰富和结构化知识，我们相信这可以进一步增强现有视觉-语言模型在 OOD 检测中的语义理解和泛化能力。

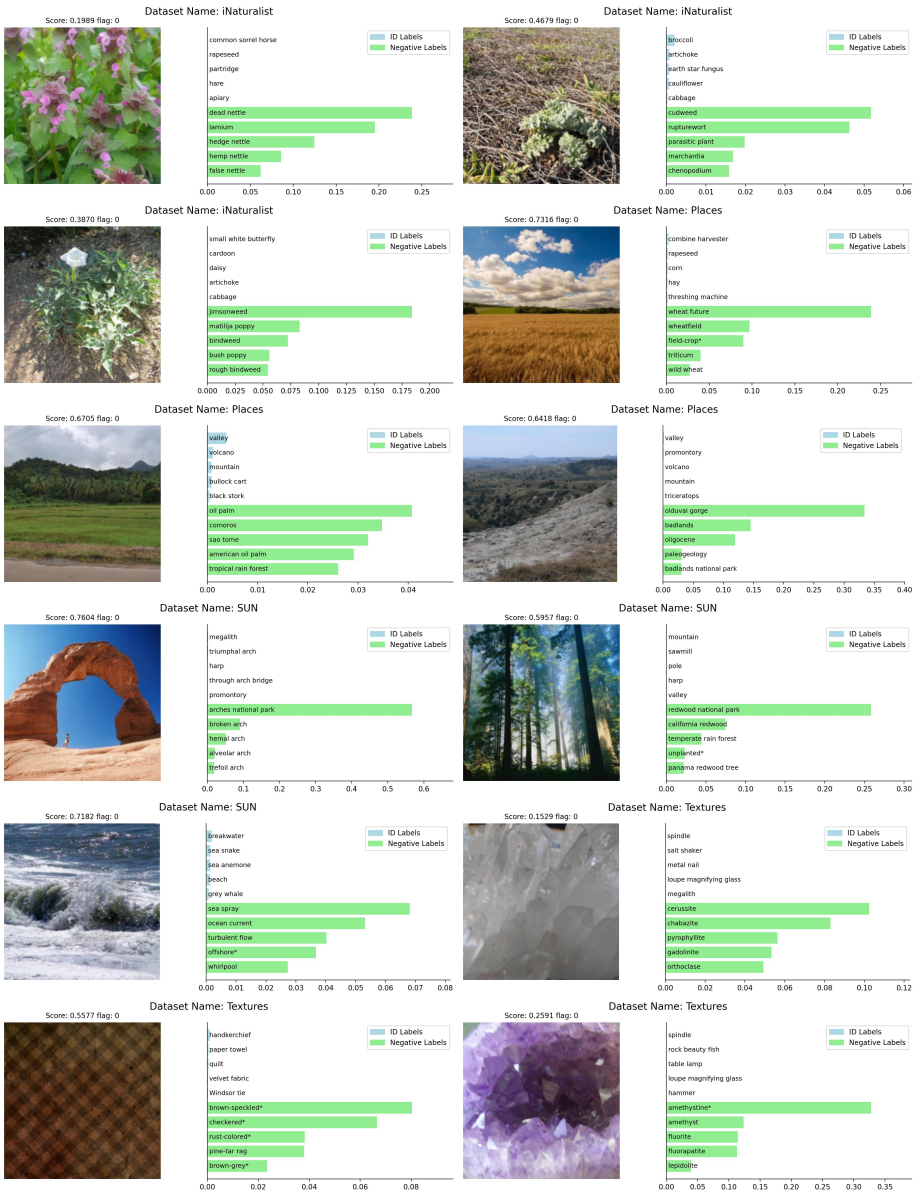

Figure 6: OOD 图像的案例可视化。每个子图左侧显示原始图像，连同其文件名、数据集、OOD 检测分数和标记为真实 OOD（标记 =0）的标识。右侧展示了 RAP 使用的对 ID 标签（蓝色）和负标签（绿色）进行 softmax 归一化的亲和力。

References

[1] Y. Jia, J. Li, G. Zhao, S. Liu, W. Sun, L. Lin, G. Li, Enhancing out-of-distribution detection via diversified multi-prototype contrastive learning, Pattern Recognition 161 (2025) 111214.

[2] Y. Zhang, J. Hu, D. Wen, W. Deng, Unsupervised evaluation for out-of-distribution detection, Pattern Recognition 160 (2025) 111212.

[3] J. Wei, G. Wang, S. Zhang, Fine-grained medical image out-of-distribution detection through multi-view feature uncertainty and adversarial sample generation, Pattern Recognition 162 (2025) 111401.

[4] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, et al., Planning-oriented autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17853–17862.

[5] M. Cheng, Y. Luo, J. Ouyang, Q. Liu, H. Liu, L. Li, S. Yu, B. Zhang, J. Cao, J. Ma, et al., A survey on knowledge-oriented retrieval-augmented generation, arXiv preprint arXiv:2503.10677 (2025).

[6] X. Jiang, F. Liu, Z. Fang, H. Chen, T. Liu, F. Zheng, B. Han, Negative label guided ood detection with pretrained vision-language models, in: International Conference on Learning Representations, 2024.

[7] C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 1998.
URL `https://mitpress.mit.edu/9780262561167/`

[8] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations, 2017.

[9] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, in: International Conference on Learning Representations, 2018.

[10] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Advances in Neural Information Processing Systems 33 (2020) 21464–21475.

[11] R. Huang, A. Geng, Y. Li, On the importance of gradients for detecting distributional shifts in the wild, Advances in Neural Information Processing Systems 34 (2021) 677–689.

[12] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: International Conference on Machine Learning, PMLR, 2022, pp. 20827–20840.

[13] A. Djurisic, N. Bozanic, A. Ashok, R. Liu, Extremely simple activation shaping for out-of-distribution detection, in: International Conference on Learning Representations, 2023.

[14] K. Xu, R. Chen, G. Franchi, A. Yao, Scaling for training time and post-hoc out-of-distribution detection enhancement, in: International Conference on Learning Representations, 2024.

[15] J. Chen, Y. Li, X. Wu, Y. Liang, S. Jha, Atom: Robustifying out-of-distribution detection using outlier mining, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21, Springer, 2021, pp. 430–445.

[16] J. Zhu, Y. Geng, J. Yao, T. Liu, G. Niu, M. Sugiyama, B. Han, Diversified outlier exposure for out-of-distribution detection via informative extrapolation, Advances in Neural Information Processing Systems 36 (2023) 22702–22734.

[17] W. Jiang, H. Cheng, M. Chen, C. Wang, H. Wei, Dos: Diverse outlier sampling for out-of-distribution detection, in: International Conference on Learning Representations, 2024.

[18] H. Yao, Z. Han, H. Fu, X. Peng, Q. Hu, C. Zhang, Out-of-distribution detection with diversification (provably), in: Advances in Neural Information Processing Systems, 2024.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.

[20] X. Du, X. Wang, G. Gozum, Y. Li, Unknown-aware object detection: Learning what you don't know from videos in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13678–13688.

[21] L. Tao, X. Du, X. Zhu, Y. Li, Non-parametric outlier synthesis, in: International Conference on Learning Representations, 2023.

[22] S. Esmaeilpour, B. Liu, E. Robertson, L. Shu, Zero-shot out-of-distribution detection based on the pre-trained model clip, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 36, 2022, pp. 6568–6576.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[24] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, Y. Li, Delving into out-of-distribution detection with vision-language representations, Advances in Neural Information Processing Systems 35 (2022) 35087–35102.

[25] H. Wang, Y. Li, H. Yao, X. Li, Clipn for zero-shot ood detection: Teaching clip to say no, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1802–1812.

[26] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, International Journal of Computer Vision 130 (9) (2022) 2337–2348.

[27] A. Miyai, Q. Yu, G. Irie, K. Aizawa, Locoop: Few-shot out-of-distribution detection via prompt learning, Advances in Neural Information Processing Systems 36 (2023).

[28] Y. Bai, Z. Han, B. Cao, X. Jiang, Q. Hu, C. Zhang, Id-like prompt learning for few-shot out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17480–17489.

[29] Y. Ming, Y. Fan, Y. Li, Poem: Out-of-distribution detection with posterior sampling, in: International Conference on Machine Learning, PMLR, 2022, pp. 15650–15665.

[30] X. Du, Y. Sun, J. Zhu, Y. Li, Dream the impossible: Outlier imagination with diffusion models, Advances in Neural Information Processing Systems 36 (2023).

[31] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The inaturalist species classification and detection dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8769–8778.

[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2017) 1452–1464.

[33] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3485–3492.

[34] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014, pp. 3606–3613.

[35] A. Krizhevsky, Learning multiple layers of features from tiny images, Master's thesis, University of Tront (2009).

[36] H. Wang, Z. Li, L. Feng, W. Zhang, Vim: Out-of-distribution with virtual-logit matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4921–4930.

[37] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in Neural Information Processing Systems 31 (2018).

[38] G. Yu, J. Zhu, J. Yao, B. Han, Self-calibrated tuning of vision-language models for out-of-distribution detection, Advances in Neural Information Processing Systems (2024).