基于能量的图神经网络分布外检测

Qitian Wu, Yiting Chen, Chenxiao Yang,Junchi Yan*上海交通大学计算机科学与工程系及人工智能教育部重点实验室{echo740, sjtucyt, chr26195, yanjunchi}@sjtu.edu.cn

arXiv:2302.02914v2 [cs.LG] 9 Mar 2023

摘要

图中的实例节点彼此相连，因此图上的学习已成为深度学习的核心问题之一。关系结构普遍存在，并会引发数据相互依赖，从而妨碍对现有方法进行简单调整；这些方法假设输入是独立同分布采样得到的。然而，当前模型大多专注于提升分布内数据的测试性能，在很大程度上忽视了分布外（OOD）测试样本带来的潜在风险：如果模型对这些样本过度自信地进行预测，可能会导致负面结果。本文研究了尚未得到充分探索的问题——图结构数据上的分布外检测，并提出一种可证明有效的分布外判别器。该判别器基于能量函数，直接从使用标准分类损失训练的图神经网络中提取。这为图上的图神经网络学习提供了一种简单、强大且高效的分布外检测模型，我们将其称为 GNNSafe。该模型还具有良好的理论性质，能够保证分布内样本与分布外样本的检测分数之间存在可区分的间隔。更重要的是，这一间隔还可通过一种无需学习的能量信念传播方案进一步扩大。为进行全面评估，我们提出了新的基准测试设置，用于评估模型检测合成和真实分布偏移下分布外数据的能力（跨领域图偏移和时序图偏移）。结果表明，GNNSafe的 AUROC 至多提升 17.0%，优于现有最先进方法；在这一尚未充分发展的领域中，它可以作为简单而强大的基线。代码可在以下地址获取：https://github.com/qitianwu/GraphOOD-GNNSafe。

1 引言

现实世界应用通常要求机器学习系统与开放世界交互，这打破了测试分布与训练分布相同这一常见假设。这促使学界投入越来越多的精力，研究如何提升模型的 泛化能力 (Muandet 等人，2013年) 和 可靠性(Liang 等，2018年)，以应对分布外（OOD）数据。然而，当前大多数方法都建立在这样一个假设之上：数据样本是独立生成的（例如图像识别，其中各实例之间没有交互）。这一前提阻碍了这些模型直接应用于图结构数据，因为其中的节点实例相互依赖（Zhao 等，2020；Ma 等人，2021年；Wu 等人，2022a年）。

分布外泛化。 为填补这一研究空白，近期越来越多关于图相关任务的研究不再局限于单一目标——分布内测试性能，而是更加关注模型如何 泛化，从而在分布外数据上表现良好（Koh 等人，2021年；Ye 等人，2021年）。一项开创性研究（Wu 等人，2022a年）对基于图的分布外泛化问题进行了形式化，并利用（因果）不变性原理，设计了一种面向图数据的新型域不变学习方法。与网格结构且独立生成的图像不同，图结构数据所涉及的分布偏移可能更加复杂，难以应对，因此往往需要图特定的技术创新。例如，Yang 等人（2022c）提出识别不变子结构，即具有因果效应的节点子集，以

*通讯作者：Junchi Yan，同时隶属于上海人工智能实验室。本研究部分由国家重点研发计划（2020AAA0107600）、国家自然科学基金委员会（62222607）和上海市科学技术委员会（22511105100）资助。

标签，在输入图中学习跨环境的稳定预测关系；而Yang 等人（2022b年）则借鉴热力学扩散在图上的类比，构建了一个原理严谨的知识蒸馏模型，用于几何知识迁移与泛化。

分布外检测。可信人工智能系统面临的一项关键挑战，是如何让深度学习模型具备足够的可靠性，使其能够认识到自身的未知之处，也就是检测出模型预期会对其置信度较低的分布外数据（Amodei 等，2016年；Liang 等，2018年）。在医学诊断（Kukar，2003）、自动驾驶（Dai & Van Gool，2018年）等安全关键型应用中，这一点尤为重要。文献将这一问题称为分布外检测，其目标是将分布外数据与分布内数据区分开来。尽管近期涌现了大量研究，探索各种有效的分布外检测方法（Hendrycks & Gimpel，2016年；Bevandi´c 等，2018年；DeVries & Taylor，2018年；Hein 等，2019年；Hsu 等，2020；Sun 等，2021年；Bitterwolf 等，2022年），但这些模型仍然默认输入是独立采样的，因此难以应用于图数据。

本文研究图上的分布外检测，其中模型需要特别处理由图引起的数据相互依赖。我们的方法以图神经网络（GNN）作为编码器，用于节点表示和预测，并纳入结构信息。为提升图神经网络模型应对分布外测试实例的可靠性，我们从采用标准学习目标训练的图神经网络分类器中识别出一种内在的分布外判别器。我们工作的关键见解是，标准图神经网络分类器本身就具备良好的能力，能够从未知测试观测中检测出分布外样本（即它们所不了解的内容），模型细节如下所述。

- 简洁性与通用性：我们模型中的分布外判别器基于能量函数，该函数通过简单变换直接从采用标准监督分类损失、在分布内数据上训练的图神经网络分类器的预测 logits 中提取。因此，我们的模型可以高效地部署于实践中，即无需训练图生成模型来进行密度估计，也无需额外的分布外判别器。此外，该模型形式通用，也就是说，基于能量的检测器不依赖图神经网络的架构，原则上能够提升任意现成图神经网络（或更广泛地说，图 Transformer）应对分布外数据的可靠性。

- 理论可靠性：尽管模型简单，但我们可以证明它能够有效区分分布内输入与分布外输入，并为它们生成可区分的分数。基于能量的信念传播方案是一种无需学习的方法，可进一步强化这种区分能力，在图拓扑上提升检测共识。我们还讨论了当训练数据包含额外的分布外观测数据作为异常值暴露时，如何恰当地纳入辅助正则化项，从而双重保证保留分布内学习能力并增强分布外可靠性。
- 实际效能：我们将模型应用于具有不同属性的多个节点分类数据集，并考察各种分布外类型。当使用纯分布内数据和标准交叉熵损失进行训练，并在推理时进行分布外检测测试时，我们的模型始终优于最先进方法，平均 AUROC 最高提升 12.9%；当使用辅助OOD暴露数据作为正则化进行训练，并在新的未见OOD数据上测试时，我们的模型优于强劲的竞争方法，平均 AUROC 最高提升 17.0%。

2 背景

图上的预测任务。 我们考虑一组索引为 $i \in \{1, 2, \cdots, N\} = \mathcal{I}$ 的实例，其生成过程涉及彼此之间的相互依赖关系，这种关系由观测图 $G = (V, E)$ 表示，其中 $V = \{i | 1 \leq i \leq N\}$ 表示包含所有实例的节点集，$E = \{e_{ij}\}$ 表示边集。观测到的边会生成邻接矩阵 $A = [a_{ij}]_{N \times N}$：如果存在连接节点 i 和 j 的边，则 $a_{ij} = 1$，否则为 0。此外，每个实例 i 都具有输入特征向量 $\mathbf{x}_i \in \mathbb{R}^D$ 和标签 $y_i \in \{1, \cdots, C\}$，其中 D 为输入维度，C 表示类别数。N 实例仅有部分标签，我们将 $\mathcal{I}_s$（分别将 $\mathcal{I}_u$）定义为有标签（分别为无标签）节点集，即 $\mathcal{I} = \mathcal{I}_s \cup \mathcal{I}_u$。图上的标准^（半）监督学习旨在使用 $Y = f(X, A)$ 训练节点级分类器 f，其中 $X = [\mathbf{x}_i]_{i \in \mathcal{I}}$ 和 ^ $Y = [\hat{\mathbf{y}}_i]_{i \in \mathcal{I}}$，以预测 $\mathcal{I}_u$ 中分布内实例的标签。

分布外检测。 除了在分布内测试节点（从与训练数据相同的分布中采样）上取得良好的预测性能外，我们还期望学习得到的分类器能够

检测与分布内实例具有不同数据生成分布的分布外（OOD）实例。更具体地说，分布外检测的目标是找到合适的决策函数 G （与分类器 f 相关联），使得对于任意给定的输入 $\mathbf{x}$：

$$G(\mathbf{x}, \mathcal{G}_{\mathbf{x}}; f) = \begin{cases} 1, & \mathbf{x} \text{ is an in-distribution instance}, \\ 0, & \mathbf{x} \text{ is an out-of-distribution instance}, \end{cases} \tag{1}$$

其中， $\mathcal{G}_{\mathbf{x}} = (\{\mathbf{x}_j\}_{j \in \mathcal{N}_{\mathbf{x}}}, A_{\mathbf{x}})$ 表示以节点 $\mathbf{x}$ 为中心的自我图， $\mathcal{N}_{\mathbf{x}}$ 是图中以 $\mathbf{x}$ 为中心的自我图（在一定跳数范围内）中的节点集合，而 $A_{\mathbf{x}}$ 是其对应的邻接矩阵。本文的问题设定考虑了数据生成过程中与中心节点相互依赖的邻近节点，并据此判断其是否属于分布外。在这种情况下，每个输入的分布外判定取决于图中的其他节点，这与计算机视觉领域的现有研究大不相同；后者假设输入独立同分布，且对每个输入独立作出判定。

基于能量的模型。近期研究（Xie 等人，2016年；Grathwohl 等人，2020）确立了输入独立同分布生成时，神经网络分类器与基于能量的模型（EBM）之间的等价性（Ranzato 等人，2007年）。设分类器为 $h(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^C$ ，它将一个输入实例映射为 C 维 logits；对 logits 应用 Softmax 函数即可得到预测的类别分布： $p(y|\mathbf{x}) = \frac{e^{h(\mathbf{x})[y]}}{\sum_{c=1}^{C} e^{h(\mathbf{x})[c]}}$ ，其中 $h(\mathbf{x})[c]$ 表示查询输出的第 c 个元素。基于能量的模型构建函数 $E(\mathbf{x}\ y) : \mathbb{R}^D \to \mathbb{R}$ ，将每个输入实例及任意给定的类别标签映射为一个称为能量的标量值，并导出一个玻尔兹曼分布： $p(y|\mathbf{x}) = \frac{e^{-E(\mathbf{x},y)}}{\sum_{y'} e^{-E(\mathbf{x},y')}} = \frac{e^{-E(\mathbf{x},y)}}{e^{-E(\mathbf{x})}}$ ，其中分母 $\sum_{y'} e^{-E(\mathbf{x},y')}$ 通常称为配分函数，它对 y 进行边缘化。通过将能量设为预测的 logits 值 $E(\mathbf{x}\ y) = -h(\mathbf{x})[y]$，即可确立 EBM 与判别式神经网络分类器之间的等价性。对于任意给定的输入，能量函数 $E(\mathbf{x})$ 可以

$$E(\mathbf{x}) = -\log \sum_{y'} e^{-E(\mathbf{x},y')}. \tag{2}$$

能量分数 $E(\mathbf{x})$ 凭借其良好的特性，可作为有效的分布外检测指标：对于分布内数据，得到的能量分数总体上低于分布外数据的能量分数（Liu 等，2020）。然而，据我们所知，目前关于基于能量的建模及其在分布外检测中的应用的研究，大多聚焦于独立同分布输入；能量模型对相互依赖数据的建模能力仍未得到充分探索。

3 提出的方法

我们介绍一种用于（半）监督节点分类的新型分布外检测模型，其中各实例相互依赖。我们将首先介绍该模型的基础版本，该版本采用基于能量的建模（第 3.1 节）和一种新颖的基于能量的信念传播方案（第 3.2 节）。之后，我们将使用辅助分布外训练数据（如果有）扩展该模型（第 3.3 节）。

3.1 具有数据依赖性的基于能量的分布外检测

为捕捉数据点之间的依赖关系，最常采用的模型类别遵循图神经网络（GNNs）的经典思想（Scarselli 等人，2008年），通过迭代聚合邻近节点的特征来更新中心节点的表示。这种操作通常称为消息传递，在现代图神经网络中发挥着核心作用，用于编码拓扑模式并吸收来自其他实例的全局信息。具体而言，假设 $\mathbf{z}_i^{(l)}$ 表示实例节点 i 在第 l层的表示，图卷积网络（GCN）（Kipf 和 Welling，2017）通过逐层归一化的特征传播来更新节点表示：

$$Z^{(l)} = \sigma\left(D^{-1/2}\tilde{A}D^{-1/2}Z^{(l-1)}W^{(l)}\right), \quad Z^{(l-1)} = [\mathbf{z}_i^{(l-1)}]_{i \in \mathcal{I}}, \quad Z^{(0)} = X, \tag{3}$$

其中， $\tilde{A}$ 是基于 A添加自环后的邻接矩阵， D 是其对应的对角度矩阵， $W^{(l)}$ 是第 l层的权重矩阵， σ 是非线性激活函数。

经过 L 层图卷积后，GCN 模型为每个节点输出一个 C 维向量 $\mathbf{z}_i^{(L)}$ 作为 logits，记作 $h_\theta(\mathbf{x}_i, \mathcal{G}_{\mathbf{x}_i}) = \mathbf{z}_i^{(L)}$ ，其中 θ 表示图神经网络的可训练参数

模型 h。根据图神经网络的更新规则，实例 $\mathbf{x}$ 的预测逻辑值表示为以实例 $\mathbf{x}$为中心的 L 阶自我图的函数，记作 $\mathcal{G}_{\mathbf{x}}$。利用 logits 和 Softmax 函数可得分类的类别分布：

$$p(y \mid \mathbf{x}, \mathcal{G}_{\mathbf{x}}) = \frac{e^{h_\theta(\mathbf{x}, \mathcal{G}_{\mathbf{x}})_{[y]}}}{\sum_{c=1}^{C} e^{h_\theta(\mathbf{x}, \mathcal{G}_{\mathbf{x}})_{[c]}}}. \tag{4}$$

将其与定义能量函数和概率密度之间关系的能量基模型相联系，我们可以得到由图神经网络模型诱导的能量形式，即 $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}},\ y; h_\theta) = -h_\theta(\mathbf{x}\ \mathcal{G}_{\mathbf{x}})[y]$。请注意，上述能量直接由图神经网络分类器的预测逻辑值获得（特别是，没有改变图神经网络的参数化方式）。此外，对 y 进行边缘化的自由能函数 $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}}; h_\theta)$ 可以用式 4 中的分母表示：

$$E(\mathbf{x}, \mathcal{G}_{\mathbf{x}}; h_\theta) = -\log \sum_{c=1}^{C} e^{h_\theta(\mathbf{x}, \mathcal{G}_{\mathbf{x}})_{[c]}}. \tag{5}$$

式 5 给出的非概率能量分数包含实例本身的信息，以及基于观测结构而可能存在依赖关系的其他实例的信息。

对于节点分类，图神经网络模型通常通过最小化带标签训练数据的负对数似然（即监督分类损失）进行训练

$$\mathcal{L}_{sup} = \mathbb{E}_{(\mathbf{x}, \mathcal{G}_{\mathbf{x}}, y) \sim \mathcal{D}_{in}} \left(-\log p(y \mid \mathbf{x}, \mathcal{G}_{\mathbf{x}}) \right) = \sum_{i \in \mathcal{I}_s} \left(-h_\theta(\mathbf{x}_i, \mathcal{G}_{\mathbf{x}_i})_{[y_i]} + \log \sum_{c=1}^{C} e^{h_\theta(\mathbf{x}_i, \mathcal{G}_{\mathbf{x}_i})_{[c]}} \right), \tag{6}$$

其中，$\mathcal{D}_{in}$ 表示有标签训练数据部分的采样分布。接下来我们将证明，对于任何使用式 6 训练的图神经网络模型，能量分数都可以作为一个定义良好的指标，用于判断输入数据是否来自分布内或分布外。

命题 1. 对 $\mathcal{L}_{sup}$ 进行梯度下降，总体上会降低任意分布内实例 $(\mathbf{x}\ \mathcal{G}_{\mathbf{x}}; h_\theta) \sim \mathcal{D}_{in}$的能量 $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}}; h_\theta)$。

该命题表明，由经过监督训练的图神经网络诱导出的能量函数 $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}}; h_\theta)$，保证了分布内数据的能量分数总体上会降低。这说明在推理时间使用能量分数检测分布外输入是合理的。

备注。上述结果与图神经网络架构无关，也就是说，基于能量的分布外判别器可以应用于任意现成的图神经网络。此外，除了在固定输入图上定义消息传递之外，近期提出的图 TransformerWu 等人 (2022b年;2023年) 已被证明是用于节点级预测的强大编码器。基于能量的模型也适用于这类类 Transformer 模型，我们将此类扩展留待未来工作。

3.2 基于能量的信念传播增强共识

由于半监督学习场景中的标注数据通常较为稀缺，直接从使用标注数据和监督损失训练的图神经网络模型中获得的能量分数，可能不足以实现良好的泛化。增强泛化能力的关键，在于如何利用训练集中的未标注数据，帮助更好地识别数据背后的几何结构（Belkin 等人，2006年；Chong 等人，2020）。受经典非参数半监督学习算法标签传播(Zhu 等人，2003）的启发，我们提出了基于能量的信念传播方法，在观测结构上于互连节点之间迭代传播估计能量。我们将$\mathbf{E}^{(0)} = [E(\mathbf{x}_i,\ \mathcal{G}_{\mathbf{x}_i};\ h_\theta)]_{i \in \mathcal{I}}$定义为图中各节点的初始能量分数向量，并考虑以下传播更新规则

$$\mathbf{E}^{(k)} = \alpha \mathbf{E}^{(k-1)} + (1-\alpha) D^{-1} A \mathbf{E}^{(k-1)}, \quad \mathbf{E}^{(k)} = [E_i^{(k)}]_{i \in \mathcal{I}}, \tag{7}$$

其中，$0 < \alpha < 1$是一个参数，用于控制节点自身及其他相连节点的能量所占的比重。经过 K步传播后，我们可以使用最终结果进行估计，即$\tilde{E}(\mathbf{x}_i, \mathcal{G}_{\mathbf{x}_i};\ h_\theta) = E_i^{(K)}$，而用于分布外检测的判定准则可以表示为

$$G(\mathbf{x}, \mathcal{G}_{\mathbf{x}}; h_\theta) = \begin{cases} 1, & \text{if} \quad \tilde{E}(\mathbf{x}, \mathcal{G}_{\mathbf{x}}; h_\theta) \le \tau, \\ 0, & \text{if} \quad \tilde{E}(\mathbf{x}, \mathcal{G}_{\mathbf{x}}; h_\theta) > \tau, \end{cases} \tag{8}$$

其中，τ 是阈值，返回0表示判定为分布外。设计能量信念传播的依据，是使其能够适应数据生成过程中存在实例级交互的物理机制。由于输入图能够反映流形上具有一定邻近性的数据样本之间的几何结构，因此节点的生成自然取决于其邻居，由此，相连节点往往从相似的分布中采样[1]。鉴于式5中的能量分数是分布外的有效指标，能量空间中的传播本质上模仿了这种数据生成的物理机制，从而有望增强检测置信度。命题2形式化地描述了在图拓扑上强化实例级分布外判定共识的整体趋势。

命题2。 对于任意给定的实例$\mathbf{x}_i$，如果其一跳邻居节点的平均能量分数（即 $\frac{\sum_j a_{ij} E_j^{(k-1)}}{\sum_j a_{ij}}$）小于（分别地，大于）其自身的能量（即 $E_i^{(k-1)}$），那么根据式 7 更新后得到的能量将满足 $E_i^{(k)} < E_i^{(k-1)}$（分别地，$E_i^{(k)} > E_i^{(k-1)}$）。

命题2表明，传播方案会将能量推向邻居节点中的多数，从而有助于放大分布内样本与分布外样本之间的能量差距。这一结果颇具启发性，说明能量模型可以在推理时通过简单的传播方案得到增强，而无需额外的训练成本。我们将在实验中通过实证比较验证其有效性。

3.3 带界限保证的能量正则化学习

到目前为止所介绍的模型并不假设训练过程中使用了任何分布外数据。作为模型的扩展，我们进一步考虑分布外检测已有研究中的另一种设置，例如（Hendrycks 等，2019b；Liu 等，2020；Bitterwolf 等，2022年），使用额外的分布外训练数据（通常称为分布外暴露）。在这种情况下，我们可以通过正则化损失对能量差距施加硬约束 $\mathcal{L}_{reg}$，以界定分布内数据（即 $\mathcal{I}_s$ 中带标签的实例）以及分布外数据（即来自不同分布的另一组辅助训练实例）$\mathcal{I}_o$ 的能量。这样，我们可以将最终训练目标定义为 $\mathcal{L}_{sup} + \lambda \mathcal{L}_{reg}$，其中 λ 是一个权衡权重。正则化项 $\mathcal{L}_{reg}$有很大的设计灵活性；这里我们将其具体设定为绝对能量的界限约束（Liu 等，2020）： $\mathcal{L}_{reg} =$

$$\frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \left(\text{ReLU} \left(\tilde{E} \left(\mathbf{x}_i, \mathcal{G}_{\mathbf{x}_i}; h_\theta \right) - t_{in} \right) \right)^2 + \frac{1}{|\mathcal{I}_o|} \sum_{j \in \mathcal{I}_o} \left(\text{ReLU} \left(t_{out} - \tilde{E} \left(\mathbf{x}_j, \mathcal{G}_{\mathbf{x}_j}; h_\theta \right) \right) \right)^2 . \quad (9)$$

上述损失函数会促使 $[t_{in}, t_{out}]$ 内的能量值对于分布内（相应地，分布外）样本更低（相应地，更高），这正是我们所期望的。

然而，令人担忧的是，引入额外的损失项可能会干扰原有的监督学习目标 $\mathcal{L}_{sup}$。换言之，能量有界正则化的目标可能与监督学习的目标不一致，前者可能会将全局最优点改变到某个位置，使图神经网络模型无法对带标签实例进行最优拟合。幸运的是，接下来的命题为模型中的能量正则化提供了理论依据。

命题 3。 对于任意能量正则化 $\mathcal{L}_{reg}$，若图神经网络模型 h_{θ^*} 最小化$\mathcal{L}_{reg}$ 后，对于分布内（相应地，分布外）样本，能量分数的上界（相应地，下界）为 t_{in}（相应地，t_{out}），其中 $t_{in} < t_{out}$ 为两个间隔参数，则相应的 softmax 类别分布 $p(y|\mathbf{x}, \mathcal{G}_\mathbf{x}; h_{\theta^*})$ 也会使 $\mathcal{L}_{sup}$ 最小化。

这一命题具有重要意义，表明我们采用的 $\mathcal{L}_{reg}$ 具有最优能量分数，能够可靠地检测分布外输入，并且与图神经网络模型的监督训练保持一致。由此，我们得到了一种新的训练目标，它能够严格保证训练集中分布内实例与分布外实例之间所需的区分，同时不会损害模型对分布内数据的拟合。

诚然，能量正则化学习的一个缺点是，训练期间需要接触额外的分布外数据。不过，正如我们将在实验中展示的那样，在不少实际场景中，能量正则化并非必需。

[1]即使数据分布相似，相连节点也可能具有不同的标签或特征。此处的数据分布是关于节点特征、标签和图结构的联合分布。

4 实验

模型在测试集上应对分布外数据的能力，可通过其区分分布外测试样本与分布内样本的能力来衡量。同时，我们不希望模型在分布内测试实例上的性能下降。理想的模型应能在不牺牲分布内准确率的情况下实现分布外检测。

4.1 设置

实现细节 我们将使用纯监督损失训练的模型称为 GNNSAFE，将使用额外能量正则化训练的版本称为 GNNSAFE++。我们基本上将传播层数 K 设置为 2，将权重 α 设置为 0.5。为进行公平比较，所有模型均使用层深为 2、隐藏维度为 64 的 GCN 模型作为骨干编码器。

数据集与划分 我们的实验基于五个真实世界数据集，这些数据集被广泛用作节点分类基准数据集：`Cora`、`Amazon-Photo`、`Coauthor-CS`、`Twitch-Explicit`和 `ogbn-Arxiv`。遵循近期关于图上的分布外泛化的工作（Wu 等人，2022a年），我们通常考虑两种在实验中创建分布外数据的方式，这些方式可以反映涉及在图中相互连接的分布外实例的真实世界场景。1）多图场景： 分布外实例来自不同的图（或子图），该图与训练集中的任何节点都不相连。2）单图场景： 分布外测试实例与训练实例位于同一图中，但这些分布外实例在训练期间未被见过。具体而言，

- 对于`Twitch`，我们使用 DE 子图作为分布内数据，并将其他五个子图作为分布外数据。ENGB 子图用作训练图神经网络的分布外暴露数据SAFE++。这些子图的规模、边密度和度分布各不相同，因此可以视为来自不同分布的样本Wu 等人（2022a年）。
- 对于作为单个图的 Arxiv，我们将节点划分为两组：2015年之前发表的论文作为分布内数据，2017年之后发表的论文作为分布外数据。2015年和2016年发表的论文在训练期间用作分布外暴露数据。
- 对于没有明确领域信息的 Cora、Amazon 和 Coauthor，我们通过三种方式合成分布外数据。结构操纵：使用原始图作为分布内数据，并采用随机块模型随机生成图作为分布外数据。特征插值：通过随机插值为分布外数据创建节点特征，并将原始图作为分布内数据。标签留出：使用具有部分类别的节点作为分布内数据，并将其余类别留出作为分布外数据。对于每个数据集中的分布内数据，仅将部分节点的标签用于监督训练，另一部分节点用于验证。其余节点则作为分布内测试数据。数据集和划分的详细信息见附录 B.1。

指标我们遵循常见做法，使用 AUROC、AUPR 和 FPR95 作为分布外检测的评估指标。分布内性能通过测试节点上的准确率衡量。附录 B.3 中提供了关于这些评估指标的更多信息。

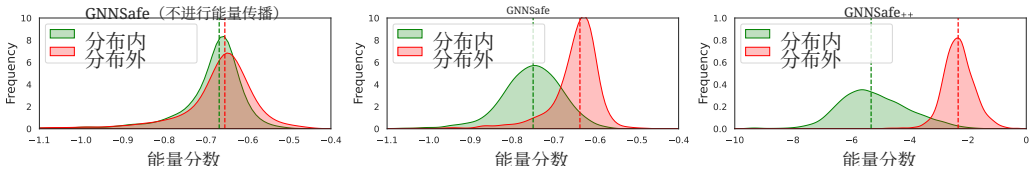
竞争方法 我们针对分布外检测比较了两类模型。第一类基线模型主要针对计算机视觉中的分布外检测，其输入（例如图像）被假定为独立同分布采样：MSP（Hendrycks & Gimpel，2016年）、ODIN（Liang 等，2018年）、马氏距离（Lee 等，2018年）、异常值暴露（OE）（Hendrycks 等，2019b）（训练时使用分布外暴露）、能量和 Energy Fine-Tune（Liu 等，2020）（训练时同样使用分布外暴露）。我们将这些模型的卷积神经网络骨干替换为本模型使用的 GCN 编码器，以处理图结构数据。第二类基线模型专为处理图机器学习中的分布外数据而设计，我们将其与两种最先进方法进行比较：基于图的核狄利克雷 GCN 方法 GKDE（Zhao 等，2020）和图后验网络 GPN（Stadler 等，2021）。这些模型的更多实现细节见附录 B.2。

4.2 对比结果

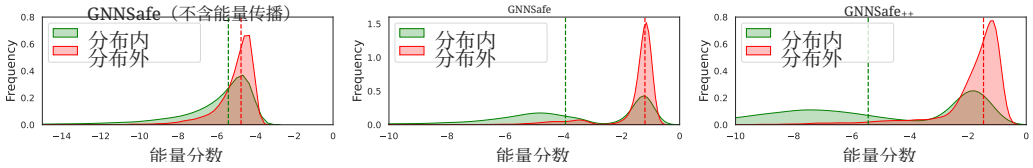
GNNSAFE++ 与竞争方法相比，表现如何？ 表 1 和表 2 报告了 GNNSAFE++ 与竞争模型的实验结果。我们发现，在所有实验中，GNNSAFE++ 始终以显著优势胜过所有竞争方法

表 1：Twitch 上的分布外检测结果（以**AUROC** (↑) / **AUPR** (↑) / **FPR95** (↓)衡量）；在该数据集中，不同子图中的节点属于分布外数据。Arxiv 数据集上的结果则将 2017 年后发表的论文视为分布外数据。此处还报告了分布内测试准确率，用于校准。附录 C 展示了各分布外数据集（即子图或年份）的详细结果。使用 24GB GPU 时，GPN 在 Arxiv 数据集上出现了内存不足问题。

模型	OOD Expo	Twitch				Arxiv			
		AUROC	AUPR	FPR	分布内准确率	AUROC	AUPR	FPR	分布内准确率
MSP	No	33.59	49.14	97.45	68.72	63.91	**75.85**	**90.59**	53.78
ODIN	No	**58.16**	**72.12**	93.96	70.79	55.07	68.85	100.0	51.39
马氏距离	No	55.68	66.42	**90.13**	70.51	56.92	69.63	94.24	51.59
能量	No	51.24	60.81	91.61	70.40	**64.20**	75.78	90.80	53.36
GKDE	No	46.48	62.11	95.62	67.44	58.32	72.62	93.84	50.76
GPN	No	51.73	66.36	95.51	68.09	-	-	-	-
GNNSAFE	No	**66.82**	**70.97**	**76.24**	70.40	**71.06**	**80.44**	**87.01**	53.39
OE	Yes	55.72	70.18	95.07	70.73	69.80	80.15	85.16	52.39
能量微调	Yes	**84.50**	**88.04**	**61.29**	70.52	**71.56**	**80.47**	**80.59**	53.26
GNNSAFE++	Yes	**95.36**	**97.12**	**33.57**	70.18	**74.77**	**83.21**	**77.43**	53.50



(a) Twitch 上的能量分布，其中不同子图中的节点属于分布外实例



（b）Arxiv 上的能量分布，其中未来出现的节点被视为分布外实例

图 1：比较原始能量模型（见第 3.1 节）、采用基于能量的传播（第 3.2 节提出）以及进一步采用能量正则化训练（第 3.3 节提出）所取得的基于能量的分布外检测性能。这些消融结果表明，所提出的能量传播和正则化训练都有助于扩大分布内输入与分布外输入之间的能量间隔。

在这些情况下，Twitch（Cora-Structure）的平均 AUROC 提高了 12.8%（17.0%），平均 FPR95 降低了 44.8%（21.0%），同时保持了相当的分布内准确率。这表明，我们的基于能量的模型在图上的分布外检测方面更具优势。此外，即使不使用能量正则化，GNNSAFE 仍能取得极具竞争力的分布外检测结果，在 Cora-Structure 上将平均 AUROC 提高了 12.9%。尤其值得一提的是，它在所有数据集上的表现都优于异常值暴露（OE），并且在大多数情况下，在 Cora、Amazon 和 Coauthor 上的表现超过能量微调（Energy FT），尽管这两种竞争方法都使用额外的分布外暴露进行训练。这体现了我们所提出模型的一项关键优势：无需额外消耗计算资源，就能增强图神经网络应对分布外数据的可靠性。为定量展示其效率优势，我们在附录 C 的表 8 中比较了训练和推理时间。可以看到，GNNSAFE 在训练和推理过程中都比最先进模型 GPN 快得多，且效率与其他竞争方法相当。总之，这些结果表明，GNNSAFE 在实际图相关应用中检测分布外异常具有良好潜力，因为在这些应用中，效果和效率都是关键考量因素。

能量传播如何提升推理时分布外检测？我们通过比较 Energy 和 GNNSAFE 的性能，研究所提出的基于能量的信念传播的有效性（在我们的设定中，Energy 模型可以看作是从 GNNSAFE 中移除传播模块后的消融版本）。表 1 和表 2 的结果一致表明，传播方案确实显著提升了性能。此外，图 1 对比了 GNNSAFE（中间列）和 Energy（左列）生成的能量分数。图中显示，传播方案确实有助于区分分布内输入和分布外输入的分数。

表 2：在 Cora、Amazon、Coauthor 数据集上，采用三种分布外类型（**AUROC** (↑) 衡量分布外检测性能：**S**结构操纵、**F**特征插值、**L**标签留出）。AUPR、FPR95 和分布内准确率的其他结果见附录 C。

模型	OOD Expo	Cora			Amazon			Coauthor		
		S	F	L	S	F	L	S	F	L
MSP	No	70.90	85.39	91.36	98.27	97.31	**93.97**	95.30	97.05	94.88
ODIN	No	49.92	49.88	49.80	93.24	81.15	65.97	52.14	51.54	51.44
马氏距离	No	46.68	49.93	67.62	71.69	76.50	73.25	80.46	93.23	85.36
能量	No	71.73	**86.15**	**91.40**	**98.51**	**97.87**	93.81	**96.18**	**97.88**	**95.87**
GKDE	No	68.61	82.79	57.23	76.39	58.96	65.58	65.87	80.69	61.15
GPN	No	**77.47**	85.88	90.34	97.17	87.91	92.72	34.67	72.56	83.65
GNNSAFE	No	**87.52**	**93.44**	**92.80**	**99.58**	**98.55**	**97.35**	**99.60**	**99.64**	**97.23**
OE	Yes	67.98	81.83	89.47	**99.60**	98.39	95.39	97.86	99.04	96.04
能量微调	Yes	**75.88**	**88.15**	**91.36**	98.83	**98.55**	**97.35**	**98.84**	**99.43**	**96.23**
GNNSAFE++	Yes	**90.62**	**95.56**	**92.75**	**99.82**	**99.64**	**97.51**	**99.99**	**99.97**	**97.89**

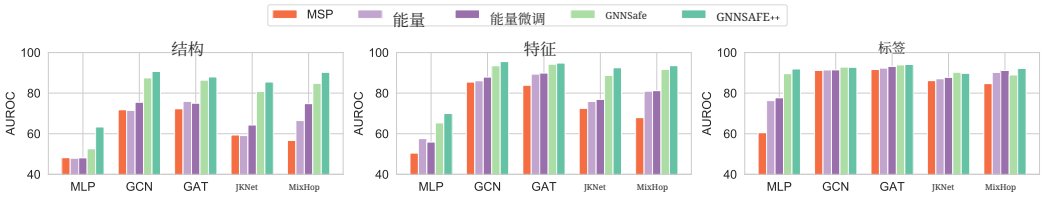


图 2：在具有三种分布外类型的 Cora 上，不同编码器骨干网络下 MSP、能量、EnergyFT、GNNSAFE 和 GNNSAFE++ 的性能比较。

这些结果也验证了我们在第 3.2 节中的假设：结构化传播可以利用拓扑信息，通过数据相互依赖提升分布外检测性能。

能量正则化训练如何帮助分布外检测? 接下来，我们进一步研究所引入的能量正则化训练的有效性。与 GNNSAFE 相比，我们发现 GNNSAFE++ 确实在表 1 和表 2 的各项实验中表现出更好的相对性能；图 1 中的可视化结果也进一步验证了这一点。在图 1 中，我们比较了未使用和使用正则化损失训练的模型的能量分数（分别对应中间列和右列），可以看到 GNNSAFE++ 确实能够在分布内数据与分布外数据之间产生更大的间隔。此外，我们注意到，表 2 中取得的性能提升并没有表 1 中那么显著。具体而言，在 Amazon 和 Coauthor 上，GNNSAFE 与 GNNSAFE++ 的 AUROC 相当。可能的原因是，多图数据集 Twitch 和大型时序数据集 Arxiv 中的分布外数据是由上下文信息引入的，分布偏移可能并不显著，因此更难进行区分，而能量正则化能够显著促进分布外检测。另一方面，对于分布外数据与分布内数据之间分布偏移足够明显的情况（如其他三个数据集），在辅助分布外训练数据上进行能量正则化可能并非必要。

4.3 超参数分析

接下来，我们分析使用不同编码器骨干网络和超参数的影响，以进一步阐明所提出模型的实际效能和用法。

不同编码器骨干网络会影响检测结果吗? 图 2 对比了在使用不同编码器骨干网络时，MSP、能量、能量微调、GNNSafe 和 GNNSafe++ 的性能，其中编码器骨干网络包括在 Cora 上针对三种分布外类型使用的 MLP、图注意力网络（Velickovic 等人，2018年）、JKNet（Xu 等人，2018年）和 MixHop（Abu-El-Haija 等人，2019年）。我们发现，不同骨干网络确实会影响检测性能；尤其是，由于图神经网络骨干网络在编码拓扑特征方面具有更强的表达能力，其 AUROC 显著高于 MLP。不过，我们观察到，五种模型的相对性能总体上与表 2 中使用图卷积网络作为骨干网络时的结果一致，并且几乎在所有情况下，两个 GNNSAFE 模型都优于其他竞争方法。在右图中，即使采用 MLP 骨干网络，两个 GNNSAFE 模型的表现也能与采用图神经网络骨干网络的对应模型相媲美，这也验证了我们的能量传播方案能够有效利用图结构中的拓扑信息。

传播深度和自环强度会如何影响推理时的性能? 图3(a)展示了 GNNSAFE++ 在传播步数 K 从

1 增至 64、自环强度 α 在 0.1 到 0.9 之间变化时的结果。如图所示，随着传播步数增加，AUROC 分数先上升后下降。这是因为适量的传播步数确实有助于利用结构依赖来提升分布外检测性能，但传播深度过大可能会过度平滑估计能量，使不同节点的检测分数难以区分。对于 α也观察到了类似趋势：取适中的值（例如 0.5 或 0.7）时，性能更好。这可能是因为 α 过大会抵消邻居信息的作用，而 α 过小则会过度关注其他节点。

正则化权重和间隔如何影响训练? 我们在图 3(b) 和 (c) 中讨论了 GNNSAFE++ 在 t_{in}、 t_{out} 和 λ 的配置下的性能变化。我们发现，GNNSAFE++ 的性能很大程度上取决于 λ，它控制着能量正则化的强度；而且其最优值在不同数据集之间也有所不同，尽管我们仅在较粗略的搜索范围 $\{0.01, 0.1, 1.0\}$内对其进行调优。此外，合理设置间隔 t_{in} 和 t_{out}可以获得稳定的良好性能，而间隔过大则会导致性能下降。

5 相关工作

神经网络的分布外检测近期文献对检测模型应具有较低置信度的分布外样本进行了广泛研究。从不同的技术角度出发，已有研究提出采用预训练模型的置信度分数（Hendrycks&Gimpel，2016年；Hendrycks 等，2019a）、在辅助分布外训练数据上添加正则化（Bevandi´c 等，2018年；Liu 等，2020；Mohseni 等人，2020），或训练生成模型以检测分布外样本（Ren 等人，2019年）。然而，这些方法大多假设实例（例如图像）是独立同分布生成的，忽略了广泛存在的涉及相互依赖数据的问题场景。

图上的分布外检测 相比之下，针对包含相互依赖节点的图数据进行分布外检测，仍是一个研究不足的领域。近期，Li 等人(2022年)提出了一种灵活的生成式框架，纳入节点特征、标签和图结构，并推导出用于分布外检测的后验分布。此外，Bazhenov 等人 (2022年)从不确定性估计的角度比较了不同的分布外检测方案。这些研究聚焦于图分类，其中每个实例本身就是一张图，也可以视为独立同分布样本。对于实例间相互依赖不可忽略的节点分类，Zhao 等(2020)和 Stadler 等 (2021年)提出了能够检测图中分布外节点的贝叶斯图神经网络模型。与这些方法不同，我们模型中的分布外判别器直接从采用监督分类损失训练的标准图神经网络的预测逻辑值中提取，因此形式简单且通用，不受图神经网络架构的限制。此外，我们的模型适用于现成的判别式图神经网络分类器，无需训练复杂的图生成模型来进行密度估计，也无需额外的分布外判别器。

分布外泛化 在分布偏移学习这一总体框架下，另一个与我们研究方向正交的相关问题是提升模型在分布外测试数据上的预测性能，即分布外泛化（Muandet 等人，2013年）。图上的分布外数据因不同场景和数据集中的具体分布偏移而类型多样，包括但不限于实验中使用的跨域图偏移和时序图偏移，见Wu 等人 （2022a年）、节点分类中的子群属性偏移（Ma 等人，2021年；Bi 等人，2023年）、图规模偏移（Bevilacqua 等人，2021年）（其特殊情形包括分子性质预测中的分子大小/骨架偏移（Yang 等人，2022c）、推荐系统中的时间上下文偏移（Yang 等人，2022a年）、表格预测任务中的特征空间偏移（Wu 等人，2021年）等。作为图上分布外检测的一项探索性工作，我们在实验中聚焦于跨域偏移和时间偏移这两种真实场景。未来一个颇具前景的方向是将分布外检测的评估扩展到图上的其他分布偏移。

6 结论

本文证明，使用监督损失训练的图神经网络本身就可以成为有效的分布外判别器，用于检测模型应避免进行预测的分布外测试数据。我们的模型简单、通用，并有充分的理论分析作为依据。我们还通过在真实世界数据集上针对各种分布外数据类型开展的大量实验，验证了其实际效能。我们希望这项工作能够助力图神经网络应用于重视安全性的真实场景。

参考文献

Sami Abu-El-Haija、Bryan Perozzi、Amol Kapoor、Nazanin Alipourfard、Kristina Lerman、Hrayr Harutyunyan、Greg Ver Steeg 和 Aram Galstyan。MixHop：通过稀疏化邻域混合实现高阶图卷积架构。发表于国际机器学习会议，第 21–29 页，2019年。

Dario Amodei、Chris Olah、Jacob Steinhardt、Paul Christiano、John Schulman 和 Dan Man´e。人工智能安全中的具体问题。*arXiv*预印本 *arXiv:1606.06565*，2016年。

Gleb Bazhenov、Sergei Ivanov、Maxim Panov、Alexey Zaytsev 和 Evgeny Burnaev。从不确定性估计的角度探索图分类中的分布外检测。*arXiv* 预印本 *arXiv:2206.10691*，2022年。

Mikhail Belkin、Partha Niyogi 和 Vikas Sindhwani。流形正则化：一种基于几何的框架，用于从有标签和无标签样本中学习。《机器学习研究杂志》，7（11），2006年。

˘ Petra Bevandi´c、Ivan Kre˘so、Marin Or˘si´c 和 Sini˘sa Segvi´c。用于语义分割的判别式分布外检测。*arXiv*预印本 *arXiv:1808.07703*，2018年。

Beatrice Bevilacqua、Yangze Zhou 和 Bruno Ribeiro。《用于图分类外推的尺寸不变图表示》。发表于国际机器学习会议，第 837–851 页。PMLR，2021年。

Wendong Bi、Bingbing Xu、Xiaoqian Sun、Li Xu、Huawei Shen 和 Xueqi Cheng。《预测图上的沉默的大多数：知识可迁移图神经网络》。发表于万维网大会，2023。

Julian Bitterwolf、Alexander Meinke、Maximilian Augustin 和 Matthias Hein。《剖析分布外检测：许多基于分布外训练数据的方法估计的是相同核心量的组合》。发表于国际机器学习会议，第 2041–2074 页。PMLR，2022年。

Yanwen Chong、Yun Ding、Qing Yan 和 Shaoming Pan。《基于图的半监督学习：综述》。《style id='1'>《神经计算》，408:216–230，2020。

Dengxin Dai 和 Luc Van Gool。暗光模型适应：从白天到夜间的语义图像分割。载于*2018*年 第*21*届国际智能交通系统会议（*ITSC*），第3819–3824页。IEEE，2018年。

Terrance DeVries 和 Graham W Taylor。为神经网络中的分布外检测学习置信度。*Arxiv* 预印本 *Arxiv:1802.04865*，2018年。

Will Grathwohl、Kuan-Chieh Wang、J¨orn-Henrik Jacobsen、David Duvenaud、Mohammad Norouzi 和 Kevin Swersky。你的分类器其实是一个基于能量的模型，你应该像对待基于能量的模型一样对待它。载于国际学习表征会议，2020。

Matthias Hein、Maksym Andriushchenko 和 Julian Bitterwolf。为什么 ReLU 网络会在远离训练数据的地方产生高置信度预测，以及如何缓解这一问题。载于*IEEE/CVF*计算机视觉与模式识别会议，第41–50页，2019年。

Dan Hendrycks 和 Kevin Gimpel。《神经网络中误分类和分布外样本检测的基线方法》。*arXiv* 预印本 *arXiv:1610.02136*，2016年。Dan Hendrycks、Steven Basart、Mantas Mazeika、Mohammadreza Mostajabi、Jacob Steinhardt 和 Dawn Song。《异常分割基准》。2019年a。Dan Hendrycks、Mantas Mazeika 和 Thomas Dietterich。《利用异常值暴露进行深度异常检测》。发表于国际学习表征会议，2019年b。

Yen-Chang Hsu、Yilin Shen、Hongxia Jin 和 Zsolt Kira。广义 ODIN：无需利用分布外数据进行训练即可检测分布外图像。载于*IEEE/CVF*计算机视觉与模式识别会议，第 10951–10960 页，2020 年。

Weihua Hu、Matthias Fey、Marinka Zitnik、Yuxiao Dong、Hongyu Ren、Bowen Liu、Michele Catasta 和 Jure Leskovec。开放图基准：用于机器学习的图数据集。神经信息处理系统进展，33：22118–22133，2020 年。

Thomas N. Kipf 和 Max Welling。基于图卷积网络的半监督分类。发表于国际学习表征会议 *(ICLR)*，2017年。

Pang Wei Koh、Shiori Sagawa、Henrik Marklund、Sang Michael Xie、Marvin Zhang、Akshay Balsub- ramani、Weihua Hu、Michihiro Yasunaga、Richard Lanas Phillips、Irena Gao、Tony Lee、Etienne David、Ian Stavness、Wei Guo、Berton Earnshaw、Imran Haque、Sara M. Beery、Jure Leskovec、Anshul Kundaje、Emma Pierson、Sergey Levine、Chelsea Finn 和 Percy Liang。WILDS：真实环境中的分布偏移基准。发表于国际机器学习会议 *(ICML)*，第 5637–5664 页，2021年。

Matjaˇz Kukar。医学诊断的传导式可靠性估计。医学人工智能，29(1-2):81–106，2003年。

Kimin Lee、Kibok Lee、Honglak Lee 和 Jinwoo Shin。《用于检测分布外样本和对抗攻击的简单统一框架》。载于 神经信息处理系统进展，2018年。

Zenan Li、Qitian Wu、Fan Nie 和 Junchi Yan。《GraphDE：用于图上去偏学习和分布外检测的生成式框架》。载于 神经信息处理系统进展，2022。

Shiyu Liang、Yixuan Li 和 Rayadurgam Srikant。《增强神经网络中分布外图像检测的可靠性》。载于 国际学习表征会议，2018年。

Weitang Liu、Xiaoyun Wang、John Owens 和 Yixuan Li。《基于能量的分布外检测》。神经信息处理系统进展，33：21464–21475，2020。

Jiaqi Ma、Junwei Deng 和 Qiaozhu Mei。《图神经网络的子群泛化与公平性》。神经信息处理系统进展，34：1048–1061，2021年。

Julian McAuley、Christopher Targett、Qinfeng Shi 和 Anton Van Den Hengel。关于风格和替代品的基于图像的推荐。载于第 *38* 届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，第 43–52 页，2015年。

Sina Mohseni、Mandar Pitale、JBS Yadawa 和 Zhangyang Wang。用于可泛化的分布外检测的自监督学习。载于*AAAI* 人工智能会议，第 5216–5223 页，2020。

Krikamol Muandet、David Balduzzi 和 Bernhard Sch¨olkopf。通过不变特征表示实现领域泛化。载于国际机器学习会议 *(ICML)*，第 10–18 页，2013年。

Marc'Aurelio Ranzato、Y-Lan Boureau、Sumit Chopra 和 Yann LeCun。用于无监督学习的统一基于能量的框架。载于人工智能与统计学，第 371–379 页。PMLR，2007年。

Jie Ren、Peter J Liu、Emily Fertig、Jasper Snoek、Ryan Poplin、Mark Depristo、Joshua Dillon 和 Balaji Lakshminarayanan。《用于分布外检测的似然比》。神经信息处理系统进展，32，2019年。

Benedek Rozemberczki 和 Rik Sarkar。《Twitch 游戏玩家：用于评估邻近性保持和基于结构角色的节点嵌入的数据集》，2021年。

Franco Scarselli、Marco Gori、Ah Chung Tsoi、Markus Hagenbuchner 和 Gabriele Monfardini。图神经网络模型。*IEEE* 神经网络汇刊，20(1):61–80，2008年。Prithviraj Sen、Galileo Namata、Mustafa Bilgic、Lise Getoor、Brian Galligher 和 Tina Eliassi-Rad。网络数据中的集体分类。《人工智能杂志》，29(3):93–93，2008年。

Arnab Sinha、Zhihong Shen、Yang Song、Hao Ma、Darrin Eide、Bo-June Hsu 和 Kuansan Wang。微软学术服务（MAS）及其应用概述。载于第*24*届万维网国际会议论文集，第243–246页，2015年。

Maximilian Stadler、Bertrand Charpentier、Simon Geisler、Daniel Z¨ugner 和 Stephan G¨unnemann。图后验网络：用于节点分类的贝叶斯预测不确定性。神经信息处理系统进展，34：18033–18048，2021年。

Yiyou Sun、Chuan Guo 和 Yixuan Li。ReAct：利用校正激活进行分布外检测。神经信息处理系统进展，34：144–157，2021年。

Petar Velickovic、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Li` o 和 Yoshua Bengio。图注意力网络。载于国际学习表征会议（*ICLR*），2018年。

Qitian Wu、Chenxiao Yang 和 Junchi Yan。迈向开放世界特征外推：一种归纳式图学习方法。载于神经信息处理系统进展，2021年。

Qitian Wu、Hengrui Zhang、Junchi Yan 和 David Wipf。处理图上的分布偏移：一种不变性视角。载于国际学习表征会议，2022a年。

Qitian Wu、Wentao Zhao、Zenan Li、David Wipf 和 Junchi Yan。Nodeformer：一种用于节点分类的可扩展图结构学习 Transformer。载于神经信息处理系统进展，2022b年。

Qitian Wu、Chenxiao Yang、Wentao Zhao、Yixuan He、David Wipf 和 Junchi Yan。DIFFormer：由能量约束扩散诱导的可扩展（图）Transformer。载于国际学习表征会议，2023年。

Jianwen Xie、Yang Lu、Song-Chun Zhu 和 Ying Nian Wu。生成式卷积网络理论。载于国际机器学习会议，第2635–2644页，2016年。

Keyulu Xu、Chengtao Li、Yonglong Tian、Tomohiro Sonobe、Ken-ichi Kawarabayashi 和 Stefanie Jegelka。利用跳跃知识网络进行图上的表征学习。载于*ICML*，2018年。

Chenxiao Yang、Qitian Wu、Qingsong Wen、Zhiqiang Zhou、Liang Sun 和 Junchi Yan。面向分布外序列事件预测：一种因果处理方法。载于神经信息处理系统进展，2022a年。

Chenxiao Yang、Qitian Wu 和 Junchi Yan。几何知识蒸馏：图神经网络的拓扑压缩。载于神经信息处理系统进展，2022b年。

Nianzu Yang、Kaipeng Zeng、Qitian Wu、Xiaosong Jia 和 Junchi Yan。学习分布外分子表示的子结构不变性。载于神经信息处理系统进展，2022c。

Nanyang Ye、Kaican Li、Lanqing Hong、Haoyue Bai、Yiting Chen、Fengwei Zhou 和 Zhenguo Li。OOD-Bench：分布外泛化数据集与算法的基准评测和理解。*CoRR*，abs/2106.03721，2021年。

Xujiang Zhao、Feng Chen、Shu Hu 和 Jin-Hee Cho。《图数据上的不确定性感知半监督学习》。神经信息处理系统进展，33：12827–12836，2020。

Xiaojin Zhu、Zoubin Ghahramani 和 John D. Lafferty。使用高斯场和调和函数进行半监督学习。载于国际机器学习会议 *(ICML)*，第 912–919 页，2003。

A 技术结果的证明

A.1 命题1的证明

证明。 证明过程相当直接，遵循了Liu 等 (2020) 第 3.1 节中的类似推理。关于 $\mathcal{L}_{nll}$ 相对于 θ 的梯度为

$$\begin{aligned}\frac{\partial \mathcal{L}_{nll}}{\partial \theta} &= \sum_{i\in\mathcal{I}_{tr}}\left(-\frac{\partial h_\theta(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})_{[y_i]}}{\partial\theta}+\sum_{c=1}^{C}\frac{\partial h_\theta(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})_{[c]}}{\partial\theta}\frac{e^{h_\theta(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})_{[c]}}}{\sum_{c'=1}^{C}e^{h_\theta(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})_{[c']}}}\right)\\ &= \sum_{i\in\mathcal{I}_{tr}}\left(\frac{\partial E(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i},y_i)}{\partial\theta}-\sum_{c=1}^{C}\frac{\partial E(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i},c)}{\partial\theta}\frac{e^{h_\theta(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})_{[c]}}}{\sum_{c'=1}^{C}e^{h_\theta(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})_{[c']}}}\right)\\ &= (1-p(y=y_i\mid\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i}))\frac{\partial E(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i},y_i;h_\theta)}{\partial\theta}-\sum_{c\neq y_i}p(y=c\mid\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i})\frac{\partial E(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i},c;h_\theta)}{\partial\theta}.\end{aligned} \tag{10}$$

从上式可以看出，整体上，最小化 $\mathcal{L}_{nll}$ 的一阶梯度的训练过程会降低能量分数 $E(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i},y_i;h_\theta)$，并提高能量 $E(\mathbf{x}_i,\mathcal{G}_{\mathbf{x}_i},c;h_\theta)$。一般而言，对于任何分布内实例 $(\mathbf{x}\ \mathcal{G}_{\mathbf{x}},y)\sim\mathcal{D}_{in}$，训练后的模型产生的能量值往往会更低。 □

A.2 命题2的证明

证明。 如果 $\frac{\sum_j a_{ij}E_j^{(k-1)}}{\sum_j a_{ij}} < E_i^{(k-1)}$，我们可以将公式 7 中的更新重写为每个实例的标量形式:

$$\begin{aligned}E_i^{(k)} &= \alpha E_i^{(k-1)}+(1-\alpha)\frac{\sum_j a_{ij}E_j^{(k-1)}}{\sum_j a_{ij}}\\ &> \alpha E_i^{(k-1)}+(1-\alpha)E_i^{(k-1)}\\ &= E_i^{(k-1)}.\end{aligned} \tag{11}$$

对于情况 $\frac{\sum_j a_{ij}E_j^{(k-1)}}{\sum_j a_{ij}} > E_i^{(k-1)}$，可以类似地证明相反的结论。 □

A.3 命题 3 的证明

证明。 我们将 $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}};h_{\theta^*})$ 定义为模型 h_{θ^*} 通过最小化正则化损失 $\mathcal{L}_{reg}$得到的能量。而 $h_{\theta^\dagger}$ 表示产生最优预测 softmax 分布且能最小化监督分类损失的模型 $\mathcal{L}_{sup}$，即:

$$\frac{e^{h_{\theta^\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})_{[y]}}}{\sum_{c=1}^{C}e^{h_{\theta^\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})_{[c]}}} = \mathrm{argmin}_{p(y|\mathbf{x},\mathcal{G}_{\mathbf{x}})}\,\mathbb{E}_{(\mathbf{x},\mathcal{G}_{\mathbf{x}},y)\in\mathcal{D}_{in}}[-\log p(y|\mathbf{x},\mathcal{G}_x)]. \tag{12}$$

注意， $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}};h_\theta) = -\log\sum_{c=1}^{C}e^{h_\theta(\mathbf{x}\ \mathcal{G}_{\mathbf{x}})_c}$。因此我们有

$$\begin{aligned}E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^*}) &= E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^*})-E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^\dagger})-\log\sum_{c=1}^{C}e^{h_{\theta^\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})}\\ &= -\log\left(e^{-E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^*})+E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^\dagger})}\cdot\sum_{c=1}^{C}e^{h_{\theta^\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})}\right)\\ &= -\log\sum_{c=1}^{C}e^{h_{\theta^\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})-E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^*})+E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta^\dagger})}.\end{aligned} \tag{13}$$

上述关系表明，能量 $E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}};h_{\theta^*})$ 实际上等价于由预测逻辑值 $h_{\theta^\dagger}(\mathbf{x}\ \mathcal{G}_{\mathbf{x}})-E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}};\ h_{\theta^*})+E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}};\ h_{\theta^\dagger})$ 所诱导的能量。因此，我们可以将

$E(\mathbf{x}\ \mathcal{G}_{\mathbf{x}};\ h_{\theta*})$ 的预测 softmax 分布记为

$$
\begin{aligned}
p(y|\mathbf{x},\mathcal{G}_{\mathbf{x}}) &= \frac{e^{h_{\theta\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})_{[y]}-E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta*})+E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta\dagger})}}{\sum_{c=1}^{C} e^{h_{\theta\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})_{[c]}-E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta*})+E(\mathbf{x},\mathcal{G}_{\mathbf{x}};h_{\theta\dagger})}} \\
&= \frac{e^{h_{\theta\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})_{[y]}}}{\sum_{c=1}^{C} e^{h_{\theta\dagger}(\mathbf{x},\mathcal{G}_{\mathbf{x}})_{[c]}}}.
\end{aligned}
\tag{14}
$$

上述预测分布恰好最小化了 $\mathcal{L}_{sup}$，其定义见公式 12。因此，我们证明了，能够最小化 $\mathcal{L}_{reg}$ 的最优能量会内在地诱导出能够最小化 $\mathcal{L}_{sup}$ 的预测 softmax 分布。另一方面，基于我们的假设，即最优能量函数能够很好地区分分布内输入和分布外输入的分数，我们由此得到了正则化损失 $\mathcal{L}_{reg}$ 的合适选择，使其最优能量满足第 3.3 节开头提到的两项要求。

□

B 实验细节

我们补充实验细节以确保结果可复现。我们的实现基于 Ubuntu 16.04.6、Cuda 10.2、Pytorch 1.9.0 和 Pytorch Geometric 2.0.3。大多数实验使用配备 11GB 显存的 NVIDIA 2080Ti 运行；对于模型需要更大 GPU 显存的情况，则使用配备 24GB 显存的 NVIDIA 3090 进行实验。

B.1 数据集信息

我们实验中使用的数据集均公开可用，是评估图学习模型的常用基准。对于 ogbn-Arxiv，我们使用 OGB 软件包[2]提供的预处理数据集和数据加载器。对于其他数据集，我们使用 Pytorch Geometric 软件包[3]提供的数据加载器。

`Cora` （Sen 等人，2008年）是一个引文网络，其中每个节点表示一篇已发表的论文，每条边表示引用关系。目标是预测每篇论文的主题，并将其作为标签类别。该数据集包含 2,708 个节点、5,429 条边、1,433 个特征和 7 个类别。我们按照第 4.1 节所述，采用三种不同方法合成分布外数据。对于分布内数据的训练/验证/测试划分，我们遵循 Kipf 和 Welling（2017）提出的半监督学习设置，并采用其提供的划分。

`Amazon-Photo` （McAuley 等人，2015年）是 Amazon 上的商品共同购买网络，其中每个节点表示一种产品，每条边表示相连的两种产品经常被共同购买。节点标签表示产品的类别。该数据集包含 7,650 个节点、238,162 条边、745 个特征和 8 个类别。与 `Cora`类似，我们采用三种方法合成分布外数据。对于分布内数据，我们遵循通行做法（Kipf 和 Welling，2017年），并采用 1:1:8 的随机划分用于训练/验证/测试。

`Coauthor-CS` (Sinha 等人，2015年) 是一个计算机科学领域的合著者网络。节点表示作者；如果两位作者共同撰写过一篇论文，则他们之间会以边相连。该数据集的目标是根据作者论文中的关键词（作为节点特征），预测作者各自的研究领域。该数据集包含 18,333 个节点、163,788 条边、6,805 个特征和 15 个类别。我们还采用前述三种方式构建分布外数据；对于分布内数据，则遵循常见做法，按 1:1:8 的比例随机划分为训练集、验证集和测试集。

`Twitch-Explicit` (Rozemberczki&Sarkar，2021年) 是一个多图数据集，其中每个子图对应一个地区的社交网络。该数据集中的节点代表 Twitch 上的游戏玩家，边表示两个用户互相关注。节点特征是 Twitch 用户所玩游戏的嵌入，标签类别表示用户是否直播成人内容。子图中的节点数量为 1,912 至 9,498，边数量为 31,299 至 153,138，共享特征维度为 2,545。如第 4.1 节所述，我们使用子图 DE 作为分布内

[2]https://github.com/snap-stanford/ogb [3] https://pytorch-geometric.readthedocs.io/en/latest/modules/datasets.html

数据，并使用 1:1:8 的随机划分进行训练/验证/测试。此外，子图 ENGB 用作分布外暴露，ES、FR、RU 用作分布外测试数据。

ogbn-Arxiv (Hu 等人，2020) 是一个规模相对较大的图数据集，记录了 1960 年至 2020 年间的引文信息。每个节点代表一篇论文，其学科领域作为待预测的标签。边表示引用关系，每个节点都有一个 128 维特征向量，该向量由其标题和摘要的词嵌入得到。如第 4.1 节所述，我们使用时间信息来划分分布内数据和分布外数据。对于分布内部分，我们还采用 1:1:8 的随机划分，用于训练/验证/测试。请注意，这里对分布内数据和分布外数据的划分使我们无法采用 Hu 等人（2020）提出的原始划分方式；该方式利用时间信息来划分训练节点和测试节点。在我们的设置中，随机划分还保证了（分布内数据中的）训练、验证和测试节点来自同一分布，且该分布不同于分布外数据。

B.2 实现细节

架构如第 4.1 节所述，我们将基本编码器骨干网络设为图卷积网络，同时也考虑多层感知机、图注意力网络、JKNet 和 MixHop 等其他模型，以作进一步讨论。我们总结了实验中使用的所有编码器骨干网络的架构信息如下。

- 图卷积网络：2 个 GCNConv 层，隐藏维度为 64，采用 ReLU 激活、自环和批归一化。
- 多层感知机：2 个全连接层，隐藏维度为 64，采用 ReLU 激活。
- 图注意力网络：2 个 GATConv 层，隐藏维度为 64，采用 ELU 激活，注意力头数为 $[2, 1]$，并进行批归一化。
- JKNet：2 个 GCNConv 层，隐藏维度为 64，采用 ReLU 激活、自环和批归一化。跳跃知识模块采用最大池化。
- MixHop：采用 2 个 MixHop 层，跳数为 2，隐藏维度为 64，使用 ReLU 激活和批归一化。

超参数 如第 4.1 节所述，我们将以下超参数设为默认值：传播步数 $K = 2$ 和传播自环权重 $\alpha = 0.5$。对于 GNNSAFE，我们仅使用验证集，在 $\{0.1, 0.01, 0.001\}$ 范围内对学习率进行网格搜索。对于 GNNSAFE++，我们使用验证集进一步搜索间隔超参数$t_{in} \in \{-9, -7, -5\}$，$t_{out} \in \{-1, -2, -3, -4\}$ 和正则化权重 $\lambda \in \{0.01, 0.1, 1.0\}$。

训练详情 模型 GNNSAFE 使用带标签训练数据上的监督分类损失进行训练，而 GNNSAFE++ 则使用分布外训练数据上的监督损失与正则化损失之和进行训练。我们使用验证集上的监督分类损失（即交叉熵）作为轮次选择的指标，以报告测试结果。具体而言，每次运行中，我们以 200 个轮次为固定训练预算，并报告模型在验证数据上分类损失最低的轮次所取得的测试性能。

竞争方法 对于竞争模型，我们采用原始论文提供的实现，并针对我们的设置进行了一些调整。如第 4.1 节所述，我们为基线方法 MSP、ODIN、马氏距离、能量和能量微调使用相同的编码器骨干网络图卷积网络（隐藏维度均为 64，层数均为 2）。对于 GKDE 和 GPN，我们使用其公开实现，参照论文中报告的超参数设置，并针对不同数据集进行微调。

B.3 评估指标

我们进一步详细说明实验中使用的评估指标。遵循分布外检测领域的惯例，我们使用 AUROC、AUPR 和 FPR95 来评估测试集中分布外数据的检测性能。这些指标独立于阈值τ。具体而言，AUROC 指受试者工作特征（ROC）曲线下面积。ROC 曲线展示了在

$0 \sim 1$ 中不同阈值下真正率与假阳性率之间的权衡。直观而言，AUROC 可以理解为：对于任意一对正例（即分布内样本）和负例（即分布外样本），正例的估计分数高于负例的概率。然而，AUROC 并非总是理想的指标，尤其是在正类和负类不平衡时。在这种情况下，AUPR 会针对不同的正类/负类基准率进行调整。AUPR 指 PR 曲线下面积，该曲线展示了 $0 \sim 1$ 中不同阈值下精确率与召回率之间的权衡。此外，FPR95 也是一种广泛使用的指标，指真正率（TPR）为 95% 时的假阳性率。FPR95 可以解释为：当 TPR 高达 95% 时，分布外样本被误分类为分布内样本的概率。

C 补充实验结果

本节补充了更多实验结果。具体而言，我们在表 3 和表 4 中报告了各个分布外数据集上的详细分布外检测性能（Twitch 的不同子图以及 Arxiv 的不同年份），作为正文中表 1 的补充。此外，我们在表 5、表 6 和表 7 中报告了 Cora、Amazon 和 Coauthor 上的 AUROC/AUPR/FPR95 以及分布内测试准确率，作为正文中表 2 的补充。在表 8 中，我们比较了模型与竞争方法的每轮训练时间和推理时间。此外，我们还在表 8 中比较了 GNNSAFE++ 与竞争方法的每轮训练时间和推理时间。另外，我们在图 3 中展示了更多超参数分析结果，包括 α、 K、 t_{in}、 t_{out} 和 λ。

表 3：在 Twitch 数据集的分布外子图 ES、FR 和 RU 上，通过 AUROC(↑)/AUPR(↑)/FPR95(↓) 衡量的分布外检测性能。

模型	OOD Expo	Twitch-ES			Twitch-FR			Twitch-RU		
		AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
MSP	No	37.72	53.08	98.09	21.82	38.27	99.25	41.23	56.06	95.01
ODIN	No	83.83	80.43	33.28	59.82	64.63	92.57	58.67	72.58	93.98
马氏距离	No	45.66	58.82	95.48	40.40	46.69	95.54	55.68	66.42	90.13
能量	No	38.80	54.26	95.70	57.21	61.48	91.57	57.72	66.68	87.57
GKDE	No	48.70	61.05	95.37	49.19	52.94	95.04	46.48	62.11	95.62
GPN	No	53.00	64.24	95.05	51.25	55.37	93.92	50.89	65.14	99.93
OE	Yes	55.97	69.49	94.94	45.66	54.03	95.48	55.72	70.18	95.07
能量微调	Yes	80.73	87.56	76.76	79.66	81.20	76.39	93.12	95.36	30.72
GNNSAFE	No	49.07	57.62	93.98	63.49	66.25	90.80	87.90	89.05	43.95
GNNSAFE++	Yes	94.54	97.17	44.06	93.45	95.44	51.06	98.10	98.74	5.59

表4：在 OOD Expo 上，分别针对 2018年、2019年和 2020 年发表的论文所对应的分布外数据集，使用 AUROC(↑)/AUPR(↑)/FPR95(↓) 衡量的分布外检测性能。

模型	OOD Expo	arXiv-2018			arXiv-2019			arXiv-2020		
		AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
MSP	No	61.66	70.63	91.67	63.07	66.00	90.82	67.00	90.92	89.28
ODIN	No	53.49	63.06	100.0	53.95	56.07	100.0	55.78	87.41	100.0
马氏距离	No	57.08	65.09	93.69	56.76	57.85	94.01	56.92	85.95	95.01
能量	No	61.75	70.41	91.74	63.16	65.78	90.96	67.70	91.15	89.69
GKDE	No	56.29	66.78	94.31	57.87	62.34	93.97	60.79	88.74	93.31
GPN	No	-	-	-	-	-	-	-	-	-
OE	Yes	67.72	75.74	86.67	69.33	72.15	85.52	72.35	92.57	83.28
能量微调	Yes	69.58	76.31	82.10	70.58	72.03	81.30	74.53	93.08	78.36
GNNSAFE	No	66.47	74.99	89.44	68.36	71.57	88.02	78.35	94.76	83.57
GNNSAFE++	Yes	70.4	78.62	81.47	72.16	75.43	79.33	81.75	95.57	71.50

Table 5：使用 AUROC(↑)/AUPR(↑)/FPR95(↓) 衡量的分布外检测性能：o n Cora，包含三种分布外类型（结构操纵、特征插值、标签留出）。

模型	Cora-Structure				Cora-Feature				Cora-标签			
	AUROC	AUPR	FPR95	分布内准确率	AUROC	AUPR	FPR95	分布内准确率	AUROC	AUPR	FPR95	分布内准确率
MSP	70.90	45.73	87.30	75.50	85.39	73.70	64.88	75.30	91.36	78.03	34.99	88.92
ODIN	49.92	27.01	100.0	74.90	49.88	26.96	100.0	75.00	49.80	24.27	100.0	88.92
马氏距离	46.68	29.03	98.19	74.90	49.93	31.95	99.93	74.90	67.62	42.31	90.77	88.92
能量	71.73	46.08	88.74	76.00	86.15	74.42	65.81	76.10	91.40	78.14	41.08	88.92
GKDE	68.61	44.26	84.34	73.70	82.79	66.52	68.24	74.80	57.23	27.50	88.95	89.87
GPN	77.47	53.26	76.22	76.50	85.88	73.79	56.17	77.00	90.34	77.40	37.42	91.46
OE	67.98	46.93	95.31	71.80	81.83	70.84	83.79	73.30	89.47	77.01	46.55	87.97
能量微调	75.88	49.18	67.73	75.50	88.15	75.99	47.53	75.30	91.36	78.49	37.83	90.51
GNNSAFE	87.52	77.46	73.15	75.80	93.44	88.19	38.92	76.40	92.80	82.21	30.83	88.92
GNNSAFE++	90.62	81.88	53.51	76.10	95.56	90.27	27.73	76.80	92.75	82.64	34.08	91.46

表 6：在 Amazon 上，采用三种分布外类型（结构操纵、特征插值、标签留出）测得的分布外检测性能：AUROC(↑)/AUPR(↑)/FPR95(↓)。

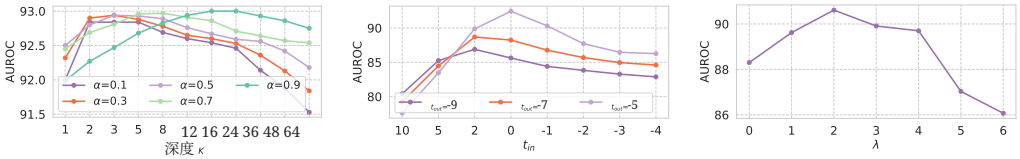
模型	Amazon-结构				Amazon-特征				Amazon-标签			
	AUROC	AUPR	FPR95	分布内准确率	AUROC	AUPR	FPR95	分布内准确率	AUROC	AUPR	FPR95	分布内准确率
MSP	98.27	98.54	6.13	92.84	97.31	95.16	8.72	92.89	93.97	91.32	26.65	95.76
ODIN	93.24	95.26	65.44	92.84	81.15	78.47	100.0	92.71	65.97	57.80	90.23	96.08
马氏距离	71.69	79.01	99.91	92.79	76.50	71.14	76.12	92.86	73.25	66.89	74.30	95.76
能量	98.51	98.72	4.97	92.86	97.87	95.64	6.00	92.96	93.81	91.13	28.48	95.72
GKDE	76.39	81.58	99.25	87.57	58.96	66.76	99.28	86.18	65.58	65.20	96.87	89.37
GPN	97.17	96.39	11.65	88.51	87.91	84.77	49.11	90.05	92.72	90.34	37.16	90.07
OE	99.60	99.61	0.51	92.61	98.39	96.24	4.34	92.30	95.39	92.53	17.72	95.72
能量微调	98.83	99.14	1.31	92.79	98.68	96.82	2.84	92.52	96.61	94.92	13.78	94.83
GNNSAFE	99.58	99.76	0.00	92.53	98.55	98.99	0.31	92.81	97.35	97.12	6.59	95.76
GNNSAFE++	99.82	99.89	0.00	92.22	99.64	99.68	0.13	92.39	97.51	97.07	6.18	95.84

表 7：在 Coauthor 上针对三种分布外类型（结构操纵、特征插值、标签留出）的分布外检测性能，以 AUROC(↑)/AUPR(↑)/FPR95(↓) 衡量。

模型	Coauthor-结构				Coauthor-特征				Coauthor-标签			
	AUROC	AUPR	FPR95	分布内准确率	AUROC	AUPR	FPR95	分布内准确率	AUROC	AUPR	FPR95	分布内准确率
MSP	95.30	94.37	24.75	92.47	97.05	96.93	15.55	92.45	94.88	97.99	23.81	95.18
ODIN	52.14	48.83	99.92	92.34	51.54	45.50	100.0	92.39	51.44	74.79	100.0	95.15
马氏距离	80.46	76.65	70.75	92.33	93.23	90.88	28.10	92.34	85.36	93.61	45.41	95.19
能量	96.18	95.25	18.02	92.75	97.88	97.69	9.75	92.75	95.87	98.34	18.69	95.20
GKDE	65.87	72.65	99.48	88.62	80.69	86.47	96.57	84.72	61.15	81.39	94.60	89.05
GPN	34.67	40.21	99.57	89.45	81.77	80.56	74.46	87.05	93.24	97.55	34.78	91.68
OE	97.86	96.81	9.23	92.60	99.04	98.80	4.44	92.64	96.04	98.50	18.17	95.10
能量微调	98.84	97.78	3.97	92.61	99.43	99.25	2.25	92.50	96.23	98.51	17.07	95.20
GNNSAFE	99.60	99.69	0.26	92.73	99.64	99.66	0.51	92.73	97.23	98.98	12.06	95.21
GNNSAFE++	99.99	99.99	0.02	92.92	99.97	99.95	0.09	92.87	97.89	99.24	9.43	95.24

Tabl图 8：五个数据集上所有模型的每轮训练时间和推理时间比较 s.

模型	Cora		Amazon		Coauthor		Twitch		Arxiv	
	TR（秒）	分布内（秒）	训练（秒）	分布内（秒）	训练（秒）	分布内（秒）	训练（秒）	分布内（秒）	训练（秒）	分布内（秒）
MSP	0.008	0.013	0.011	0.023	0.060	0.166	0.008	0.040	0.043	0.173
ODIN	0.007	0.025	0.009	0.053	0.055	0.311	0.008	0.081	0.018	0.010
马氏距离	0.008	0.251	0.013	0.306	0.073	1.081	0.009	0.006	0.024	0.034
能量	0.013	0.016	0.016	0.022	0.114	0.160	0.011	0.048	0.063	0.192
GKDE	0.007	0.015	0.008	0.025	0.069	0.236	0.006	0.043	0.030	0.171
GPN	10.02	40.30	6.480	25.86	28.58	114.44	2.850	13.88	-	-
OE	0.014	0.015	0.018	0.024	0.118	0.165	0.012	0.044	0.077	0.252
能量微调	0.013	0.018	0.018	0.024	0.131	0.171	0.016	0.052	0.069	0.218
GNNSAFE	0.015	0.015	0.015	0.023	0.115	0.164	0.011	0.049	0.062	0.219
GNNSAFE++	0.015	0.019	0.015	0.027	0.141	0.189	0.016	0.052	0.075	0.239



(a) (b) (c)

图3：（a）传播步数 K 和自环强度 α的超参数分析。（b）间隔超参数 t_{in} 和 t_{out}的影响。（c）GNNSAFE++ 能量正则化的权重超参数 λ 的影响。我们在 Cora 上采用结构操纵进行实验。