# 通过理由对齐实现忠实且一致的图神经网络解释

TIANXIANG ZHAO,美国宾夕法尼亚州立大学DONGSHENG LUO,美国佛罗里达国际大学XIANG ZHANG,美国宾夕法尼亚州立大学SUHANG WANG,美国宾夕法尼亚州立大学

arXiv:2301.02791v2 [cs.LG] 2 Sep 2023

近年来，揭示图神经网络（GNN）预测背后的理由日益受到关注。实例级图神经网络解释旨在找出目标图神经网络进行预测所依赖的关键输入元素，例如节点或边。尽管已有多种算法被提出，但其中大多数都将这项任务形式化为搜索能够保留原始预测的最小子图。然而，这一框架存在根深蒂固的归纳偏置：多个子图可能产生与原始图相同或相似的输出。因此，这些方法可能给出虚假解释，且无法提供一致性解释。将这些方法用于解释表现不佳的图神经网络，会进一步加剧这些问题。为解决这一问题，我们从因果视角对图神经网络的预测进行了理论分析。我们指出了产生虚假解释的两个典型原因：分布偏移等潜变量造成的混杂效应，以及不同于原始输入的因果因素。鉴于混杂效应和多样的因果理由都编码在内部表征中，我们提出了一种新的解释框架，并引入辅助对齐损失；理论上已证明，该损失本质上能够优化一个更忠实的解释目标。具体而言，针对这一对齐损失，我们探索了多种不同视角：基于锚点的对齐、基于高斯混合模型的分布对齐、基于互信息的对齐等。我们对这一新框架在解释忠实度和一致性方面的有效性，以及这些变体各自的优势，进行了全面研究。代码请参见以下链接：https://github.com/TianxiangZhao/GraphNNExplanation。

CCS 概念： • **计算方法** → **神经网络**；统计关系学习。

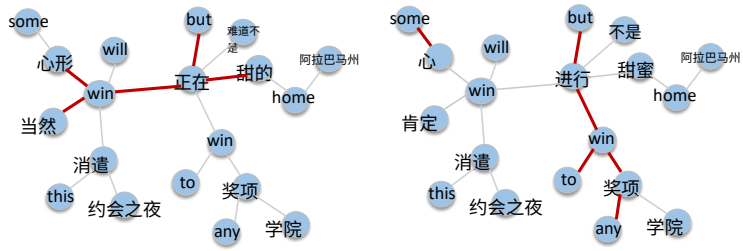
其他关键词和短语：图神经网络、可解释人工智能、机器学习 g

**ACM 引用格式：**
Tianxiang Zhao、DongshengLuo、Xiang Zhang 和 SuhangWang。2023年。通过理由对齐实现忠实且一致的图神经网络解释。*ACM*智能系统与技术汇刊1，1，第1篇文章（2023年1月），22页。
https://doi.org/10.1145/3616542

图结构数据在现实世界中无处不在，例如社交网络 [1–3],分子结构 [4, 5] 和知识图谱 [6, 7]。随着人们对从图中学习的兴趣日益浓厚，图神经网络（GNN）近年来受到越来越多的关注。通常，图神经网络采用消息传递机制，递归地传播并融合来自

作者地址：TianxiangZhao，宾夕法尼亚州立大学，美国宾夕法尼亚州斯泰特学院，tkz5084@psu.edu；Dongsheng Luo，佛罗里达国际大学，美国佛罗里达州迈阿密，dluo@fiu.edu；Xiang Zhang，宾夕法尼亚州立大学，美国宾夕法尼亚州斯泰特学院，xzz89@psu.edu；Suhang Wang，宾夕法尼亚州立大学，美国宾夕法尼亚州斯泰特学院，szw494@psu.edu。

允许为个人或课堂使用而免费制作本作品的全部或部分数字副本或纸质副本，前提是这些副本不得用于或分发以营利或商业利益为目的，且每份副本均须附上本声明及首页的完整引文。对于本作品中由作者以外的其他人拥有版权的部分，必须尊重其版权。注明出处后，可以摘录。若要以其他方式复制、再版、张贴到服务器或分发至列表，则须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2018年 版权所有归所有者/作者所有。出版权授予 ACM。2157-6904/2023/1-ART1 $15.00 https://doi.org/10.1145/3616542




(a) 运行 1 (b) 运行 2

图 1. 领先基线 GNNExplainer 在来自 Graph-SST5 的同一输入图上得到的解释结果。红色边构成解释子结构。

图中邻居节点的信息。因此，学习到的节点表示既包含节点属性，也包含邻域信息，有助于完成各种下游任务，例如节点分类 [8–11], 图分类 [12], 和链接预测 [13–15]。

尽管图神经网络在各个领域都取得了成功，但与其他神经网络一样，图神经网络缺乏可解释性。理解图神经网络的内部工作机制可以带来诸多益处。首先，这有助于加深实践者对图神经网络模型特征的理解，例如模型是否按照预期运行，从而增强他们对图神经网络模型的信任。其次，这能提高模型的透明性，使其能够应用于对公平性、隐私和安全性挑战较为敏感的关键决策领域中的可信应用 [16–18]。因此，对图神经网络可解释性的研究正日益受到关注，已有许多相关工作付诸实施 [19–21]。

我们尤其关注事后实例级解释。给定一个训练后的图神经网络和一个输入图，该任务旨在发现能够解释图神经网络模型预测行为的子结构。现有研究已提出了一些解决方案 [19, 22, 23]。例如，为寻找预测所依赖的重要子结构，GNNExplainer通过扰动学习节点属性和边的重要性矩阵 [19]。能够保持原始预测结果的最小子结构被视为解释。在此思想基础上，PGExplainer训练图生成器，以利用解释中的全局信息，并在归纳设置下实现更快的推理 [20]。SubgraphX将解释约束为连通子图，并基于Shapley值执行蒙特卡洛树搜索 [24]。这些方法可以概括为一个标签保持框架，即候选解释由原始图的掩码版本构成，并被识别为最小判别性子结构。

然而，由于拓扑结构的复杂性以及候选子结构数量的组合爆炸，现有的标签保持方法不足以对图神经网络作出忠实且一致的解释。这些方法不稳定，容易将虚假相关性作为解释。图 1 展示了一个失败案例：在 Graph-SST5 [25] 上训练的图神经网络用于情感分类。每个节点代表一个词，每条边表示节点之间的句法依存关系。每个图根据句子的情感进行标注。图中，句子“Sweet home alabama isn’t going to win any academy awards, but this date-night diversion will definitely win some hearts”的标签为正面。在第一次运行中，GNNExplainer [19]将“definitely win some hearts”识别为解释；而在第二次运行中，它却将“win academy awards”识别为解释。GNNExplainer 得出的不同解释违背了一**致性**标准，即对于相同输入，解释方法在不同运行中应具有确定性和一致性 [26]。因此，现有方法给出的解释可能无法忠实反映待解释图神经网络的决策机制。

通过检查目标图神经网络的推理过程，我们发现，从因果视角可以理解不一致问题和伪解释。具体而言，现有解释方法可能因不同的因果因素而产生伪解释，也可能由于分布偏移的混杂效应而产生伪解释（识别出的子图可能超出分布范围）。这些失败案例源于一种特定的归纳偏置：预测标签足以表明哪些关键输入组件应被提取。这一潜在假设植根于现有研究采用的优化目标 [19, 20, 24]。然而，我们的分析表明，标签信息不足以滤除伪解释，因而会导致解释不一致且不忠实。

考虑图神经网络的推理，混杂效应和不同的因果关系都可以体现在内部表示空间中。基于这一观察，我们提出了一种新颖的目标，鼓励原始图与识别出的子图在内部嵌入空间中的嵌入对齐，从而获得更忠实且一致的图神经网络解释。具体而言，为了评估两个输入之间的语义相似度，并将对齐目标纳入解释中，我们设计并比较了多种设计方案，以衡量嵌入空间中的相似度。这些策略能够实现候选解释与原始输入之间的对齐，并且可以灵活地融入各种现有的图神经网络解释方法。特别是，除了直接使用欧几里得距离外，我们还提出了三种分布感知策略。第一种策略识别一组锚定嵌入，并利用它们作为参照的相对距离。第二种策略假设数据符合高斯混合模型，并通过样本落入各高斯中心的概率来捕捉分布。第三种策略训练深度神经网络来估计两个输入之间的互信息，该方法采用数据驱动的方式，几乎不依赖先验领域知识。进一步分析表明，所提出的方法实际上优化了一种新的解释框架，其设计更为忠实。我们的主要贡献如下：

- 我们指出，现有图神经网络解释模型识别出的理由存在忠实性和一致性问题。这些问题源于其保留标签框架中的归纳偏置，该框架仅将预测作为引导信息；
- 我们提出一种有效且易于应用的对策，即对齐中间嵌入。我们实现了一组采用不同对齐策略的变体，可灵活地融入各种图神经网络解释模型。我们还进行了理论分析，以理解并验证所提出的框架。
- 在真实世界数据集和合成数据集上进行的大量实验表明，我们的框架有助于各种图神经网络解释模型实现更忠实且一致的解释。

## 1 相关工作

本节回顾相关工作，包括图神经网络以及图神经网络的可解释性。

### 1.1 图神经网络

近年来，图神经网络（GNN）随着对关系数据结构学习需求的不断增长而迅速发展 [1, 27–30]。通常，现有的图神经网络可分为两类：基于图信号处理理论的基于谱的方法 [31–33]，以及依赖邻域聚合的基于空间的方法 [34–36]。尽管存在差异，大多数图神经网络变体都可以概括为消息传递框架，该框架由每一层中的模式提取和交互建模组成 [37]。具体而言，图神经网络基于节点表示构建消息，再通过

各种消息传递机制传播这些消息，以改进节点表示，并将其用于下游任务 [10, 13, 28, 38]。相关探索包括解耦传播过程 [39–41] 或利用外部原型 [42, 43]。研究人员还探讨了表达能力 [44, 45] 以及不同核所引入的潜在偏差 [46, 47] ，以设计更有效的图神经网络。尽管图神经网络在网络表示学习方面取得了成功，但它们仍是不可解释的黑盒模型。即使已知所采用的消息传递机制和参数，要理解其行为仍然颇具挑战。此外，传统深度神经网络中的实例是独立同分布的，而图神经网络则联合考虑节点特征和图拓扑，这使得可解释性问题更难处理。

### 1.2 图神经网络解释方法

近年来，已有一些工作致力于解释图神经网络模型，并为其预测提供解释 [48]。根据粒度，现有方法大致可分为两类：（1）实例级解释 [19], , 通过识别重要子结构，为每个实例的预测提供解释；（2）模型级解释 [49, 50],, 旨在理解目标图神经网络所捕获的全局决策规则。从方法论角度来看，现有方法可分为：（1）自解释图神经网络 [50, 51],, 图神经网络可以同时给出预测及其解释；（2）事后解释 [19, 20, 24], , 采用另一种模型或策略来解释目标图神经网络。由于事后解释与模型无关，即可应用于各种图神经网络，因此本文聚焦于事后实例级解释 [19], , 即给定一个训练后的图神经网络模型，识别每个输入实例中的关键子结构，以解释其预测。 [52]中有一份全面的综述。

文献中探索了多种事后实例级解释策略，包括利用梯度信号的方法 [49, 53], 、基于扰动预测的方法 [19,20, 24, 54], 以及基于分解的方法 [49, 55]。在这些方法中，基于扰动预测的方法最为流行。其基本思路是学习一个扰动掩码，过滤掉不重要的连接，并识别出能够保留原始预测的主导性子结构 [25]。识别出的重要子结构被用作该预测的解释。例如，GNNExplainer [19] 分别在节点属性和图结构上使用两个软掩码矩阵，并在最大化互信息（MMI）框架下以端到端方式学习这两个矩阵。PGExplainer [20] 通过引入图生成器来利用全局信息，从而对其进行扩展。它可应用于归纳设置，并避免针对每个待解释实例从头重新学习这一繁重任务。SubgraphX [24] 要求解释采用子图而非边集合的形式，并使用蒙特卡洛树搜索（MCTS）寻找能够保留预测结果的连通子图，保留程度通过Shapley值衡量。为提高识别出的解释的忠实性，一些研究引入了因果分析领域的术语，例如估计每条边的个体因果效应 [56] ，或设计干预以避免发现伪相关 [57]。文献 [58] 将识别解释中最小预测性部分的思路与信息瓶颈原理 [59] 联系起来，并为图神经网络解释设计了一个端到端优化框架。

尽管上述图神经网络解释方法取得了进展，但其中大多数方法仅根据扰动输入所导致的输出变化来发现关键子结构。由于这一内在的归纳偏置，现有的标签保持方法极易受到环境中混杂因素所引发的伪相关影响。另一方面，我们的方法通过对齐图神经网络中的中间嵌入，减轻了伪相关对图神经网络解释的影响，从而实现忠实且一致的解释。

### 1.3 图对比学习

近年来，对比学习（CL）因通过无监督前置任务减少了对人工标注的需求而备受关注 [60–62]。在典型的对比学习框架中，模型以成对方式进行训练，促使正样本对相互靠近、负样本对相互远离 [63, 64]。

对比学习（CL）技术近年来已扩展至图领域，通过构建多个图视图，并最大化语义相似实例之间的互信息（MI） [62, 65–69]。现有方法的主要差异在于图增强技术和对比预训练任务。图增强通常包括节点级属性掩蔽或扰动 [70–72]、边丢弃或重连 [73, 74],以及图扩散[75, 76]。对比预训练任务可分为两类：节点（图）对嵌入之间的同尺度 对比 [75, 77, 78] ，以及全局图嵌入与局部节点表示之间的跨尺度 对比 [71, 72]。近期研究还探索了动态对比目标 [79, 80] ，通过自适应地学习增强策略来实现。

## 2 预备知识

### 2.1 问题定义

我们用 $G = \{V, E; F, A\}$ 表示一个图，其中 $V = \{v_1, \ldots, v_n\}$ 是 $n$ 个节点的集合，$E \in V \times V$ 是边的集合。节点由属性矩阵 $F \in \mathbb{R}^{n \times d}$ 描述，$F[i,:] \in \mathbb{R}^{1 \times d}$ 是节点 $v_i$ 的 $d$ 维节点属性。E 由邻接矩阵 $A \in \mathbb{R}^{n \times n}$ 描述。$A_{ij} = 1$ 表示节点 $v_i$ 和 $v_j$ 之间存在一条边；否则，$A_{ij} = 0$。对于图分类，每个图 $G_i$ 都有一个标签 $Y_i \in C$，训练图神经网络模型 $f$ 将 G 映射到其类别，即 $f$: {F, A} $\mapsto$ $\{1, 2, \ldots, C\}$。类似地，对于节点分类，每个图 $G_i$ 表示以节点 $v_i$ 为中心的 $K$ 跳子图，并训练图神经网络模型 $f$，根据从 $G_i$ 学得的节点表示 $v_i$ 来预测 $v_i$ 的标签。解释的目的是找到一个子图 $G'$，并分别在邻接矩阵和节点属性上标记二值重要性掩码 $M_A \in [0,1]^{n \times n}$ 和 $M_F \in [0,1]^{n \times d}$，例如 $G' = \{A \odot M_A; F \odot M_F\}$，其中 $\odot$ 表示逐元素乘法。这两个掩码突出显示了 G 中对 $f$ 预测其标签至关重要的组成部分。根据上述记号，事后实例级图神经网络解释任务为：

给定训练后的图神经网络模型 $f$，对于任意输入图$\mathcal{G} = \{\mathcal{V}、\mathcal{E}; \mathbf{F}、\mathbf{A}\}$，找到一个能够解释 $f$ 在$\mathcal{G}$上的预测的子图$\mathcal{G}'$。所得解释$\mathcal{G}'$ 由节点属性上的重要性掩码$\mathbf{M}_F$和邻接矩阵上的重要性掩码$\mathbf{M}_A$ 表示。

### 2.2 基于最大互信息的解释框架

针对事后实例级图神经网络解释，已有许多方法。由于边的存在性具有离散性，且图结构并非网格结构，大多数研究采用基于扰动的策略来搜索解释。一般而言，这些方法可概括为最大化预测标签 $\hat{Y}$ 与解释 $G'$ 之间的互信息（MMI），即

$$\begin{aligned} &\min_{\mathcal{G}'} -I(\hat{Y}, \mathcal{G}'), \\ &\text{s.t.} \quad \mathcal{G}' \sim \mathcal{P}(\mathcal{G}, \mathbf{M}_A, \mathbf{M}_F), \quad \mathcal{R}(\mathbf{M}_F, \mathbf{M}_A) \le c \end{aligned} \tag{1}$$

其中，$I()$ 表示互信息，P 表示对带有重要性掩码 $\{\mathbf{M}_F、\mathbf{M}_A\}$ 的原始输入进行的扰动。例如，令 $\{\hat{}\mathbf{A}、\hat{}\mathbf{F}\}$ 表示经过扰动的 {A、F}。那么，在 GNNExplainer [19], 中，$\hat{}\mathbf{A} = \mathbf{A} \odot \mathbf{M}_A$，且 $\hat{}\mathbf{F} = \mathbf{Z} + (\mathbf{F} - \mathbf{Z}) \odot \mathbf{M}_F$，其中 Z 是从节点属性 F 的边缘分布中采样得到的。R 表示对解释施加的正则化项，用于在搜索过程中引入先验知识，例如对预算或连通性分布的约束 [20]。

互信息 $I(\hat{Y}\ G')$ 用于量化原始预测 $\hat{Y} = f(G)$ 与候选解释的预测 $f(G')$ 之间的一致性，这有助于提升所发现解释 G′ 的正向性。由于互信息衡量 G′ 对 Ŷ的预测能力，该框架本质上是在尝试寻找一个能够最佳预测原始输出 Ŷ的子图ˆ。训练过程中通常采用其放宽形式 [19]，如下所示：

$$\begin{aligned} &\min_{\mathcal{G}'} H_C\big(\hat{Y}, P(\hat{Y}' \mid \mathcal{G}')\big), \\ &\text{s.t.} \quad \mathcal{G}' \sim \mathcal{P}(\mathcal{G}, \mathbf{M}_A, \mathbf{M}_F), \quad \mathcal{R}(\mathbf{M}_F, \mathbf{M}_A) \le c \end{aligned} \tag{2}$$

其中 $H_C$ 表示交叉熵。在采用相同目标的情况下，现有方法的主要区别在于优化和搜索策略。

可以从不同方面评估解释的质量 [26]。其中，最重要的两个标准是**忠实性** 和**一致性**。忠实性衡量解释的描述准确性，即解释与目标模型行为的吻合程度。一致性关注解释不变性，检验相同的输入是否应得到相同的解释。然而，如图 1 所示，现有的基于 MMI 的框架在这些标准上的表现并非最优。问题根源在于其学习目标仅以预测作为搜索解释的指导。由于图结构复杂，仅凭预测作为指导可能会导致虚假解释。下一节将对此进行详细分析。

## 3 分析伪解释

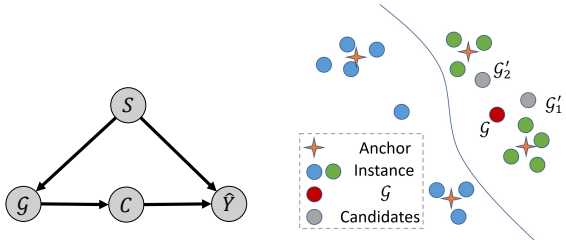
我们将“伪解释”定义为偏离图 G 上预测真实依据的解释，使得将 G′ 用作解释显得毫无根据。以图 1 为例，尽管图神经网络解释方法发展迅速，但检测到的解释在忠实性和一致性方面仍存在问题。为了深入理解这一问题的成因，我们将从因果视角考察目标图神经网络模型的行为。图 2(a) 展示了结构方程模型（SEM），其中变量 $C$ 表示判别性因果因素，变量 $S$ 表示环境混杂因素。G 与预测的 ˆ标签 $Y$ 之间存在两条路径。

• G $\rightarrow C \rightarrow \hat{Y}$ 该路径展示了目标图神经网络模型的推理过程，即从输入图中提取对预测 $Y$ 具有信息性和判别性的关键模式 $C$ˆ，目标模型据此进行预测。因果变量由输入图和目标图神经网络模型学到的知识共同决定。

• G $\leftarrow S \rightarrow \hat{Y}$ 我们将 $S$ 视为混杂因素，例如用于描述图的整体分布。它与输入图的出现以及目标图神经网络模型的预测都存在因果关系。G 的掩码版本可能会产生分布外（OOD）样本，从而对预测输出造成伪因果关系。例如，在化学领域，移除边（键）或节点（原子）可能会得到训练期间从未出现过的无效分子图。在存在分布偏移的情况下，模型预测的可靠性会降低。

图 2(a) 为我们提供了一种分析 $f$行为的方法。从因果结构中可以看出，伪解释可能源于未能恢复原始因果′ ˆ依据。根据公式1学习得到的 G′ 可能会保留预测 $Y$，这是由于分布偏移或与原始 G 相比因果变量不同所产生的混杂效应 $C$。训练不足的图神经网络 $f(\cdot)$对噪声不稳定或缺乏鲁棒性，会进一步加剧这一问题，因为其预测不可靠。

为进一步理解这一问题，我们建立图 2(a) 中结构方程模型与图神经网络 $f$ 推理过程之间的对应关系。具体而言，我们首先将 $f()$ 分解为特征提取器 $f_{ext}()$ 和




（a）结构因果模型 （b）对齐

图 2.（a）{v1}的预测规则，以结构因果模型的形式表示。（b）基于锚点的嵌入对齐示例。

分类器 $f_{cls}()$。随后，其推理过程可概括为两个步骤：(1) 使用 $f_{ext}()$ 进行编码步骤，以 G 为输入，并在表示空间 $E_C$ 中生成其嵌入；(2) 使用 $f_{cls}()$ 进行分类步骤，根据输入的嵌入预测输出标签。将这些推理步骤与图 2(a) 中的结构方程模型联系起来，我们可以发现：

• 因果路径 G $\rightarrow C \rightarrow Y$ 支撑着推理过程，并通过表示空间 $E_C$ 对关键变量进行编码 $C$；

• 分布偏移的混杂效应 $S$ 会通过影响 $E_C$中图嵌入的分布作用于推理过程。当掩蔽后的输入 G′ 属于分布外数据时，其嵌入将无法反映其判别特征，并偏离真实分布，进而使其分类步骤发生偏差。

总而言之，伪解释通常由以下两个原因导致：

(1) 得到的 G′ 是分布外图。在目标图神经网络模型的推理过程中，G′ 的编码表示与训练集中见过的表示相距甚远，导致预测不可靠；(2) 编码后的判别表示与原始图的判别表示不一致。G′ 与 G 提取出的因果因素（$C$）不同，从而导致错误的解释。

## 4 方法论

基于上述讨论，本节重点提升图神经网络解释的忠实性和一致性，并纠正仅依赖预测输出所导致的归纳偏置。我们首先直观介绍所提出的对策，该对策将内部推理过程纳入考量。随后，我们设计四种具体算法，在潜在空间中对齐 G 和 G′，以促使二者以相同的方式被观察和处理。最后，我们提供理论分析，为所提出的策略提供依据。

### 4.1 缓解虚假解释

实例级事后解释旨在找出目标模型 $f$所依赖的判别性子结构。公式 2 中的传统目标可以识别输入中最小的预测部分，但直接将其视为解释是危险的。由于图拓扑具有多样性，且子图具有组合性，可能会识别出多个不同的子结构，并得到相同的预测，正如第 3 节所述。

若要使解释子结构 G′ to忠实，它应当遵循待解释模型 $f$内部推理中原始图 G 所采用的相同依据。为实现这一目标，解释 G′ 应在决策机制方面与 G 对齐，如图 2(a) 所示。然而，由于待解释的目标图神经网络模型具有黑箱特性，提取并比较关键因果变量 $C$ 和混杂变量 $S$ 并非易事。

根据第 3 节的因果分析，我们提出一种替代方法：考察 $f$学习到的内部嵌入。因果变量 $C$编码在由 $f$提取的表示空间中，而分析嵌入分布也能反映分布外效应。我们可以合理地作出如下假设：如果两个图经图神经网络层映射到彼此接近的嵌入，那么该层会将这两个图视为相似，并由后续层以相似的方式处理。基于这一假设，可以通过对齐 G′ 与 G 的内部图嵌入来设计一个代理任务。这项新任务可以作为辅助优化目标纳入框架 1。

设$\mathbf{h}_v^l$ 表示节点 $v$ 在第 $l$层图神经网络中的表示，其中$\mathbf{h}_v^0 = \mathbf{F}[v,:]$。一般而言，图神经网络层内的推理过程可概括为一个消息传递框架：

$$\begin{aligned} \mathbf{m}_v^{l+1} &= \sum_{u \in \mathcal{N}(v)} \text{Message}_l(\mathbf{h}_v^l, \mathbf{h}_u^l, A_{v,u}), \\ \mathbf{h}_v^{l+1} &= \text{Update}_l(\mathbf{h}_v^l, \mathbf{m}_v^{l+1}), \end{aligned} \tag{3}$$

其中，$\text{Message}_l$和 $\text{Update}_l$分别是第 $l$层的消息函数和更新函数。$\mathrm{N}(v)$ 是节点 $v$的邻居集合。一般性不失，图池化层也可表示为：

$$\mathbf{h}_{v'}^{l+1} = \sum_{v \in \mathcal{V}} P_{v,v'} \cdot \mathbf{h}_v^l. \tag{4}$$

其中，$P_{v,v'}$ 表示用于图分类的众多图神经网络中，从第 $l$层的节点 $v$映射到第 $l+1$ 层的节点 $v'$ 的映射权重。我们提出对每一层的嵌入$\mathbf{h}_v^{l+1}$进行对齐，该嵌入同时包含节点和局部邻域信息。

### 4.2 分布感知对齐

实现嵌入空间中的对齐并非易事，其中存在若干不同的难点。(1) 在该嵌入空间中，难以评估 G 与 G′ 之间的距离。不同维度可能编码不同特征，且重要性各异。此外，G′ 是原始 G 的子结构，因此不重要维度上的偏移自然会存在。(2) 由于图/节点分布的复杂性，设计一种既便于计算、又能很好地反映分布流形上距离的对齐度量并非易事。

为应对这些挑战，我们设计了一种以分布感知的方式识别解释性子结构，并保持其与原始图对齐的策略。其基本思想是利用其他图获取嵌入分布密度的全局视图，从而更好地衡量对齐程度。具体而言，我们获得代表性节点/图嵌入作为锚点，并使用其与这些锚点之间的距离作为图的分布级表示。在获得的图对表示上进行对齐。下面将详细介绍这一策略。

- 首先，利用来自同一数据集的图 $\{\mathrm{G}_i\}_{i=1}^m$，对于每个层 $l$，可获得一组节点嵌入 $\{\{\mathbf{h}_{v,i}^l\}_{v\in \mathcal{V}_{ri}}\}_{i=1}^m$，其中 $\mathbf{h}_{v,i}$表示图 $\mathrm{G}_i$中节点 $v$的嵌入。对于节点级任务，我们设置 $\mathrm{V}_{ri}$ 仅包含图 $\mathrm{G}_i$ 的中心节点。对于图级任务，$\mathrm{V}_{ri}$ 包含图池化层之后的节点集合，并按照 $\{\sum_{v\in\mathcal{V}_{ri}} \mathbf{h}_{v\,i}^{l+1}/|\mathrm{V}_{r_i}|\}_{i=1}^m$对其进行处理，以获得全局图表示。
- 然后，对得到的嵌入集合应用聚类算法，以获得 $K$ 个组。这些组的聚类中心被设为锚点，记作 $\{\mathbf{h}^{l+1,k}\}_{k=1}^K$。在实验中，我们选择 DBSCAN [81] 作为聚类算法，并调整其超参数，使组数约为 20。

在第 $l$层，$\mathbf{h}_v^{l+1}$表示为其与这些 $K$ 个锚点之间的相对距离，即$\mathbf{s}_v^l \in \mathbb{R}^{1\times K}$，其中第 $k$个元素计算为$\mathbf{s}_v^{l+1\,k} = \|\mathbf{h}_v^{l+1} - \mathbf{h}_v^{l+1\,k}\|_2$。
可以通过比较 G′ 和 G 在每一层的表征来实现对齐。对齐损失的计算方式如下：

$$\mathcal{L}_{align}(f(\mathcal{G}), f(\mathcal{G}')) = \sum_l \sum_{v\in\mathcal{V}'} \|\mathbf{s}_v^l - \mathbf{s}_v'^l\|_2^2. \tag{5}$$

该指标通过比较相对于代表性聚类中心的位置，为评估嵌入分布流形中的对齐提供了一种轻量级策略。这种策略能够自然地编码各维度重要性的差异。图 2(b) 给出了一个示例，其中 G 是待解释的图，红色星形表示锚点。就绝对距离而言，$G_1'$ 和 $G_2'$ 都与 G 相似；但很容易看出，就锚点而言，$G_1'$ 与 G 更相似。换言之，锚点能够更好地衡量对齐，从而滤除虚假解释。

该对齐损失作为辅助任务纳入公式 2 中基于 MMI 的框架，以获得忠实的解释：

$$\begin{aligned} &\min_{\mathcal{G}'} H_C\big(\hat{Y}, P(\hat{Y}' \mid \mathcal{G}')\big) + \lambda \cdot \mathcal{L}_{Align}, \\ &\text{s.t.} \quad \mathcal{G}' \sim \mathcal{P}(\mathcal{G}, \mathbf{M}_A, \mathbf{M}_F), \quad \mathcal{R}(\mathbf{M}_F, \mathbf{M}_A) \le c \end{aligned} \tag{6}$$

其中，$\lambda$ 控制预测保持与嵌入对齐之间的平衡。$\mathrm{L}^{Align}$ 可以灵活地纳入各种现有的解释方法中。

### 4.3 直接对齐

作为一种更简单、直接的实现方式，我们还设计了一种基于绝对距离的变体。对于不含图池化的层，其目标函数可写为 $\sum_l \sum_{v\in\mathcal{V}} \|\mathbf{h}_v^l - \mathbf{h}_v'^l\|_2^2$。对于含图池化的层，由于结构可能不同，我们在全局表示$\sum_{v\in\mathcal{V}'} \mathbf{h}_v^{l+1}/|\mathcal{V}'|$上进行对齐，其中 $\mathcal{V}'$ 表示池化后的节点集。

## 5 扩展方法

本节进一步探讨对齐策略的更多设计选择，以获得忠实且一致的解释。我们不采用启发式方法，而是探索两个新方向：（1）基于高斯混合模型、在统计上合理的距离度量；（2）充分利用深度神经网络的能力，捕捉潜在嵌入空间中的分布。下面将介绍这两种对齐策略的具体细节。

### 5.1 基于高斯混合的对齐

在这种策略中，我们使用高斯混合分布对节点（或图）的潜在嵌入进行建模，并将代表性节点/图嵌入作为锚点（高斯中心）。每个输入产生的嵌入都可以与这些原型锚点进行比较，而目标模型所接收输入的语义信息则由其与这些锚点之间的相对距离编码。

具体而言，我们首先按照第 4.2 节介绍的相同策略，对从图 $\{G_i\}_{i=1}^m$ 中收集的嵌入 $\{\{\mathbf{h}_{v,i}^l\}^{v\in\mathcal{V}_i}\}_{i=1}^m$ 运行聚类算法，得到原型表示 $\{\mathbf{h}^{l,k}\}_{k=1}^K$。我们采用聚类算法 DBSCAN [81]，并调整其超参数，使分组数量约为 20。

接下来，可以计算编码表征 $\mathbf{h}^l$ 属于各个原型高斯中心 $\{\mathbf{h}^{l,k}\}_{k=1}^K$ 的概率：

$$p_v^{l,k} = \frac{\exp(-\|\mathbf{h}_v^l - \mathbf{h}^{l,k}\|_2^2/2\sigma^2)}{\sum_{j=1}^K \exp(-\|\mathbf{h}_v^l - \mathbf{h}^{l,k}\|_2^2/2\sigma^2)} \tag{7}$$

该分布概率可以作为一种自然工具，用于描绘图神经网络模型所学习到的输入图的语义。因此，$\mathbf{h}_v^{'l}$ 与 $\mathbf{h}_v^l$ 之间的距离可以根据该分布概率，直接通过 KL 散度进行测量：

$$d(\mathbf{p}_v^{'l}, \mathbf{p}_v^l) = \sum_{k\in[1,\ldots,K]} p_v^{'l,k} \cdot \log(\frac{p_v^{'l,k}}{p_v^{l,k}}), \tag{8}$$

其中，$\mathbf{p}_v^{'l} \in \mathbb{R}^K$ 表示候选解释嵌入 $\mathbf{h}_v^{'l}$ 的分布概率。采用该策略，原始图与候选解释之间的对齐损失计算如下：

$$\mathcal{L}_{align}\big(f(\mathcal{G}), f(\mathcal{G}')\big) = \sum_l \sum_{v\in\mathcal{V}'} d(\mathbf{p}_v^{'l}, \mathbf{p}_v^l), \tag{9}$$

该结果可以纳入公式 6 中提出的修订版解释框架。

**比较** 与公式 5 中基于相对于锚点的相对距离的对齐损失相比，这一新目标在考虑分布方面提供了更好的策略。具体而言，我们可以说明以下两个优势：

• 在获取每个实例的分布感知表示时，该变体使用高斯距离核（公式 7），而另一种变体在第 4.2 节中使用欧氏距离，这可能会放大远距离锚点的影响。我们可以通过考察表示相对于图神经网络嵌入 $\mathbf{h}_v$ 变化的梯度来证明这一点。在第 $l$ 层的维度 $k$ 中，前一种变体的梯度可以计算为：

$$\frac{\partial s_v^{l,k}}{\partial \mathbf{h}_v^l} = 2 \cdot (\mathbf{h}_v^l - \mathbf{h}_v^{l,k}) \tag{10}$$

另一方面，该变体的梯度为：

$$\frac{\partial p_v^{l,k}}{\partial \mathbf{h}_v^l} \approx -\frac{\exp(-\|\mathbf{h}_v^l - \mathbf{h}^{l,k}\|_2^2/2\sigma^2) \cdot (\mathbf{h}_v^l - \mathbf{h}_v^{l,k})}{\sigma^2 \cdot \sum_{j=1}^K \exp(-\|\mathbf{h}_v^l - \mathbf{h}^{l,k}\|_2^2/2\sigma^2)} \tag{11}$$

显然，对于前一种变体，梯度的大小会随到相应锚点的距离线性增长。另一方面，对于此变体，项 $\frac{\exp(-\|\mathbf{h}_v^l - \mathbf{h}^{l,k}\|_2^2/2\sigma^2)}{\sigma^2 \cdot \sum_{j=1}^K \exp(-\|\mathbf{h}_v^l - \mathbf{h}^{l,k}\|_2^2/2\sigma^2)}$ 会降低那些距离较远的锚点的重要性，同时提高相似锚点的重要性，这更有利于获得分布感知表示。

• 在计算两个输入的表征之间的距离时，该变体采用了式 8 中的 KL 散度。与另一种直接采用式 5 中欧几里得距离的方法相比，它不受尺度影响。同样，我们可以展示对齐损失相对于所获得的、编码分布信息的嵌入的梯度。可以观察到，对于前一种变体：

$$\frac{\partial d(\mathbf{s}_v^l, \mathbf{s}_v^{'l})}{\partial s_v^{'l,k}} = 2 \cdot (s_v^{l,k} - s_v^{'l,k}) \tag{12}$$

对于这种基于高斯混合模型的变体，梯度可计算如下：

$$\frac{\partial d(\mathbf{p}_v^l, \mathbf{p}_v^{'l})}{\partial p_v^{'l,k}} = 1 + \log(\frac{p_v^{'l,k}}{p_v^{l,k}}) \tag{13}$$

可以看出，先前的策略侧重于绝对差异较大的表示维度，但会对每个维度的尺度较为敏感。另一方面，该策略将相对差异的对数与一个常数相加，因此对各维度的尺度不敏感。

### 5.2 基于互信息的对齐

在这一策略中，我们进一步考虑利用深度模型来捕捉分布并估计两个输入的语义相似度，并将其纳入对齐损失，以发现忠实且一致的解释。具体而言，我们训练一个深度模型来估计两个输入图之间的互信息（MI），并将其预测结果作为衡量原始图与其候选解释之间对齐程度的指标。该策略避免了对启发式策略的依赖，完全由数据驱动，并且可以通过端到端的方式进行学习。

为了学习互信息估计器，我们采用神经网络，并训练其作为詹森-香农互信息估计器 [82]。具体而言，我们基于中间嵌入训练这一基于 JSD 的估计器，学习目标如下，这有助于提高优化过程的稳定性：

$$\begin{aligned}\min_{g_{mi}} \mathcal{L}_{mi} =& \mathbb{E}_{\mathcal{G}\in\{\mathcal{G}_i\}_{i=1}^m} \mathbb{E}_{v\in\mathcal{G}} \mathbb{E}_l [\mathbb{E}_{\mathbf{h}_v^{l,+}} sp(-T^l(\mathbf{h}_v^l, \mathbf{h}_v^{l,+})) \\ &+ \mathbb{E}_{\mathbf{h}_v^{l,-}} sp(T^l(\mathbf{h}_v^l, \mathbf{h}_v^{l,-}))],\end{aligned} \tag{14}$$

其中，$\mathbb{E}$ 表示期望。在该公式中，$T^l(\cdot)$是第$l$层中的兼容性估计函数，我们将$\{T^l(\cdot)\}_l$记为互信息估计器 $g_{mi}$。激活函数 $sp(\cdot)$为$softplus$函数，$\mathbf{h}_v^+$表示增强节点 $v$ 的嵌入，该节点是原始图中 $v$ 的**正样本对**。相反，$\mathbf{h}_v^-$表示增强节点 $v$ 的嵌入，该节点是原始输入的**负样本对**。正样本对通过随机丢弃中间神经元获得，相当于屏蔽原始输入的一定比例；负样本对则由不同节点的嵌入构成。该目标可以引导 $g_{mi}$ 捕捉由目标模型编码的两个输入图之间的相关性或相似性。需要注意的是，为进一步提升互信息估计器的学习效果，可以遵循图对比学习领域的最新进展，采用更先进的图增强技术来构造正样本对/负样本对 [65, 66]。本文沿用 [82] 采用的策略，因为该策略应用广泛，并将其留待未来探索。学习得到该互信息估计器后，即可轻松计算G与G′ 之间的对齐损失：

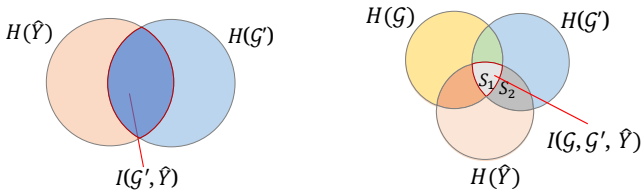
$$\mathcal{L}_{align}(f(\mathcal{G}), f(\mathcal{G}')) = \sum_l \sum_{v\in\mathcal{V}'} sp(-T^l(\mathbf{h}_v^{'l}, \mathbf{h}_v^l)), \tag{15}$$

可纳入式中提出的修订后解释框架。 6.

在这一策略中，我们通过捕捉两个输入之间的互信息，设计了一种数据驱动的方法，从而避免使用人工设计的启发式方法可能带来的偏差。

## 6 理论分析

介绍了这些对齐策略和我们新提出的解释框架后，接下来我们希望进一步探讨，并为公式 6 中提出的新损失函数提供理论依据。本节首先基于第 3 节提出一种新的解释目标，以防止产生虚假解释。随后，我们从理论上证明，在适度放宽条件后，该目标与我们的损失函数相符。




(a) 先前目标 （b）我们提出的新目标

图 3. 我们提出的新目标示意图。**6.1 新解释目标**

从前面的讨论可以看出，通过公式1得到的 G′ 不能被安全地用作解释。现有图神经网络解释方法的一个主要缺陷在于其归纳偏置：相同的结果并不保证有相同的原因，因此现有方法容易受到虚假解释的影响。图 3 给出了一个示例。公式 ˆ 1 中提出的目标旨在优化解释候选 G′ 与 $Y$之间的互信息，对应于图 3(a) 中最大化 $H$(G′) 与 $H$()的重叠部分，或图 3(b) 中的区域 $S_1 \cup S_2$。这里， $H$ 表示信息熵。然而，这一学习目标无法避免生成虚假解释的风险。给定的 G′ 可能落入区域 $S_2$，而该区域无法忠实地表示图 G。相反，更合理的目标应当是最大化图 3(b) 中的区域 $S_1$。其背后的直觉是：在导致相同结果的输入搜索空间中，所识别出的 G′ 应同时涵盖原始 G 中具有代表性和区分性的部分，从而防止因不同原因而产生相同结果的虚假解释。具体而言，寻找能够最大化 $S_1$ 的 G′，可以形式化为：

$$\begin{aligned} &\min_{\mathcal{G}'} -I(\mathcal{G}, \mathcal{G}', \hat{Y}), \\ &\text{s.t.} \quad \mathcal{G}' \sim \mathcal{P}(\mathcal{G}, \mathbf{M}_A, \mathbf{M}_F) \quad \mathcal{R}(\mathbf{M}_F, \mathbf{M}_A) \leq c \end{aligned} \tag{16}$$

**6.2 与我们的方法的联系**

由于 G 的潜在生成机制未知，$I$(G，G′，ˆ)难以求解。本节将展开这一目标，将其与公式 6 联系起来，并构造其可优化的代理形式，如下所示：

$$\begin{aligned} I(\mathcal{G}, \mathcal{G}', \hat{Y}) &= \sum_{y\sim\hat{Y}} \sum_{\mathcal{G}} \sum_{\mathcal{G}'} P(\mathcal{G}, \mathcal{G}', y) \cdot \log \frac{P(\mathcal{G}', y)P(\mathcal{G}, \mathcal{G}')P(\mathcal{G}, y)}{P(\mathcal{G}, \mathcal{G}', y)P(\mathcal{G})P(\mathcal{G}')P(y)} \\ &= \sum_{y\sim\hat{Y}} \sum_{\mathcal{G}} \sum_{\mathcal{G}'} P(\mathcal{G}, \mathcal{G}', y) \cdot \log \Big[ \frac{P(\mathcal{G}', y)}{P(\mathcal{G}')P(y)} \cdot \frac{P(\mathcal{G}, \mathcal{G}')}{P(\mathcal{G})P(\mathcal{G}')} \cdot \frac{P(\mathcal{G}, y)}{P(\mathcal{G}, y|\mathcal{G}')} \Big] \\ &= \sum_{y\sim\hat{Y}} \sum_{\mathcal{G}'} P(\mathcal{G}', y) \cdot \log \frac{P(\mathcal{G}', y)}{P(\mathcal{G}')P(y)} + \sum_{\mathcal{G}} \sum_{\mathcal{G}'} P(\mathcal{G}, \mathcal{G}') \cdot \log \frac{P(\mathcal{G}, \mathcal{G}')}{P(\mathcal{G})P(\mathcal{G}')} \\ &\quad - \sum_{\mathcal{G}'} \sum_{y\sim\hat{Y}} \sum_{\mathcal{G}} P(\mathcal{G}, y, \mathcal{G}') \cdot \log \frac{P(\mathcal{G}, y, \mathcal{G}')}{P(\mathcal{G}, y)P(\mathcal{G}')} \Big] \\ &= I(\mathcal{G}', \hat{Y}) + I(\mathcal{G}, \mathcal{G}') - \sum_{y\sim\hat{Y}} \sum_{\mathcal{G}} P(\mathcal{G}, y) \sum_{\mathcal{G}'} P(\mathcal{G}'|\mathcal{G}, y) \cdot \log P(\mathcal{G}'|\mathcal{G}, y) \\ &\quad + \sum_{\mathcal{G}'} \sum_{y\sim\hat{Y}} \sum_{\mathcal{G}} P(\mathcal{G}, y, \mathcal{G}') \cdot \log P(\mathcal{G}') \\ &= I(\mathcal{G}', \hat{Y}) + I(\mathcal{G}, \mathcal{G}') + H(\mathcal{G}'|\mathcal{G}, \hat{Y}) - H(\mathcal{G}'). \end{aligned}$$

由于 $H(G'|G, \hat{Y})$ 和 $H(G')$ 描述的是解释 $G'$ 的熵，且与扰动预算密切相关，因此我们可以忽略这两项，并得到用于 $\max_{\mathcal{G}'} I(G, G', Y)$ 的替代优化目标 $\hat{\ }\hat{\ }$，即 $\max_{\mathcal{G}'} I(Y, G') + I(G', G)$。

在 $\max_{\mathcal{G}'} I(\hat{Y}, G') + I(G', G)$ 中，第一项 $\max_{\mathcal{G}'} I(\hat{Y}, G')$ 与公式（1）相同。根据 [19]，我们将其放宽为 $\min_{\mathcal{G}'} H_C(\hat{Y}, \hat{Y}'|G')$，通过优化 $G'$ 来保留原始预测输出。第二项 $\max_{\mathcal{G}'} I(G', G)$ 对应于最大化 $G'$ 与 $G$ 之间的一致性。虽然图生成过程是潜在的，但基于一个安全的假设，即 $f$ 提取的嵌入 $\mathbf{E}_{\mathcal{G}}$ 能够代表 $G$，我们可以构建一个代理目标 $\max_{\mathcal{G}'} I(\mathbf{E}_{\mathcal{G}'}, \mathbf{E}_{\mathcal{G}})$，以提升嵌入空间中的一致性。在这项工作中，我们通过对齐它们的表征来优化这一目标，具体方式包括优化简化的距离度量或进行分布感知对齐。

## 7 实验

在本节中，我们开展了一系列实验，以评估所提出的辅助任务在提供实例级事后解释方面的益处。实验在 5 个数据集上进行，并从忠实性和一致性两个方面评估所得解释。具体而言，我们旨在回答以下问题：

- **RQ1** 所提出的框架能否在识别用于解释图神经网络的解释性子结构方面表现出色？
- **研究问题2** 现有图神经网络解释方法的一致性问题是否严重？所提出的嵌入对齐能否在这一指标上改进图神经网络解释器？
- **研究问题3** 我们提出的策略能否避免产生伪解释，并更忠实于目标图神经网络模型？

### 7.1 实验设置

*7.1.1* 数据集。 我们在五个公开可用的图神经网络可解释性基准数据集上进行了实验。这些数据集的关键统计信息汇总于表 1。

- BA-Shapes[19]：一个节点分类数据集，其基础结构为包含 300 个节点的巴拉巴西-阿尔伯特（BA）图。80 个“房屋”模体随机连接到基础图上。基础图中的节点标签为 0，模体中的节点则根据其位置赋予标签。解释针对这些连接的节点进行，真实标注为相应模体内部的边。
- Tree-Grid [19]：一个节点分类数据集，由将 80 个网格模体连接到一棵 8 层平衡二叉树上构建而成。基础图中的节点标记为 0，模体中的节点标记为 1。同一模体内的边用作类别 1 节点的真实解释。
- 感染 [83]：一个以 ER随机图初始化的单一网络。5% 的节点标记为感染，其余节点根据它们到这些感染节点的最短距离进行标记。大于 4 的标签会被截断。按照 [83], 的设定，感染节点及具有多条最短路径的节点不予考虑。对于每个节点，其最短路径用作真实解释。
- Mutag [19]：一个图分类数据集。每个图对应一个分子，其中节点表示原子，边表示化学键。分子的标签依据其化学性质确定，并利用先验领域知识识别判别性化学基团。按照 PGExplainer [20], 的做法，化学基团 $NH_2$ 和 $NO_2$ 用作真实解释。
- Graph-SST5 [25]：一个由文本构建的图分类数据集，其标签来自情感分析。每个节点代表一个词，边表示词语依存关系。该数据集不提供真实解释，通常采用启发式指标进行评估。

表 1. 数据集统计

| | BA-Shapes | Tree-Grid | 感染 | Mutag | SST-5 |
|---|---|---|---|---|---|
| 层级 | Node | Node | Node | 图 | 图 |
| 图 | 1 | 1 | 1 | 4，337 | 11，855 |
| 平均节点数 | 700 | 1，231 | 1，000 | 30.3 | 19.8 |
| 平均边数 | 4,110 | 3，410 | 4，001 | 61.5 | 18.8 |
| 类别数 | 4 | 2 | 5 | 2 | 5 |

*7.1.2* 基线方法。为评估所提出框架的有效性，我们选择了一组具有代表性的最先进实例级事后 GNN 解释方法作为基线方法。详情如下：

• GRAD [20]：一种基于梯度的方法，通过计算 GNN 预测相对于邻接矩阵的梯度，为边分配重要性权重。
• ATT [20]：利用自注意力层中的平均注意力权重来区分重要的边。

• GNNExplainer [19]：一种基于扰动的方法，会为每个实例单独学习一个重要性矩阵。

• PGExplaienr [20]：一种参数化解释器，通过学习图神经网络来预测每个图中的重要边，并通过测试不同的扰动进行训练；
• Gem [56]：与 PGExplainer 类似，但采用因果视角，基于估计的个体因果效应。

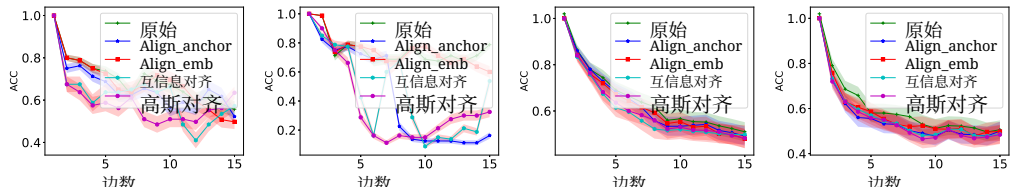
• RG-Explainer [54]：一种用于图神经网络的强化学习（RL）增强型解释器，它通过强化学习智能体依次添加节点来构建 $G_s$。

我们在第 4.2 节提出的算法已实现，并集成到两个具有代表性的图神经网络解释框架中，即 GNNExplainer [19]和 PGExplainer [20]。

*7.1.3* 配置。根据已有工作 [20], ，在每个数据集上训练一个三层图卷积网络 [8] 作为目标模型，并按 8:1:1 的比例划分训练/验证/测试数据。所有数据集的潜在维度均设为 64，激活函数采用 ReLU。对于图分类任务，将全局最大池化和全局平均池化的输出拼接，作为图表示。所有解释器均使用 Adam优化器训练，权重衰减设为 $5e$-4。对于 GNNExplainer，学习率初始化为 0。01，训练轮次为 100。对于 PGExplainer，学习率初始化为 0。003，训练轮次设为 30。控制 $L^{align}$ 权重的超参数 $\lambda$，针对不同数据集分别在范围 $[0.1, 10]$ 内通过网格搜索进行调优。在所有实例上测试解释结果。

*7.1.4* 评估指标。为了评估不同方法的忠实性，我们采用以下 [25], 两项指标：（1）边重要性的 AUROC 分数；（2）解释的忠实度。在有预言机解释可用的基准测试中，我们可以计算所识别边的 AUROC 分数，因为经过充分训练的目标图神经网络应遵循这些预定义的解释。对于没有真实解释的数据集，我们按照 [25]通过忠实度度量来评估解释质量。具体而言，我们按照分配的重要性权重依次移除边，并观察预测的变化；性能下降越快，表示忠实度越高。

为评估解释的一致性，我们对每种方法随机运行 5 次，并报告所获得解释之间的平均结构汉明距离（SHD）[84] 。SHD 分数越小，表示一致性越强。




(a) Tree-Grid，GNNExplainer (b) Tree-Grid，PGExplainer (c) Graph-SS5，GNNExplainer (d) Graph-SS5，PGExplainer

图4. 解释忠实度（彩色效果最佳）。

## 7.2 解释忠实度

为回答**RQ1**，我们从 AUROC分数和解释忠实度两方面比较解释方法。

*7.2.1* 边上的 *AUROC*。本小节通过比较分配的边重要性权重与真实解释，报告不同方法的 AUROC 分数。对于基线方法 GRAD、ATT、Gem 和 RG-Explainer，本文展示其原始论文中报告的性能。本文重新实现了 GNNExplainer 和 PGExplainer，并在此基础上实例化并测试了四种对齐策略。每项实验均进行 5 次，并在表2中汇总平均性能。AUROC 分数越高，表示解释越准确。根据结果，我们可以得出以下观察：

- 在全部四个数据集上，无论以 GNNExplainer 还是 PGExplainer 作为基础方法，加入嵌入对齐都能提升所得解释的质量；
- 在所提出的对齐策略中，考虑分布的方案表现尤为出色，尤其是基于高斯混合模型的变体。在大多数情况下，采用潜在高斯分布的变体带来了更显著的提升，并在 4 个数据集中的 3 个上取得最佳结果；
- 对于 Mutag 这类更复杂的数据集，引入嵌入对齐带来的益处更加显著。例如，使用 Align_Gaus 后，PGExplainer 的性能从 83.7% 提升至 95.9%。这一结果还表明，随着数据集复杂度增加，伪解释问题会更加严重。

*7.2.2* 解释忠实度。除了与真实解释进行比较外，我们还从忠实度角度评估所获得的解释。具体而言，我们按照解释模型学习到的重要性权重，依次从图中移除边，并测试分类性能。一般来说，移除真正重要的边会使分类性能显著下降。因此，性能下降越快，表明忠实度越强。我们在 Tree-Grid 和 Graph-SST5 上开展实验。每项实验重复进行 3 次，并报告每个数据集上所有实例的平均结果。我们使用 PGExplainer 和 GNNExplainer 作为骨干方法。在图 4 中，我们绘制了预测准确率随移除边数变化的曲线。从图中可以观察到，当引入所提出的嵌入对齐后，移除边所导致的分类准确率下降得更快，这表明所提出的嵌入对齐能够帮助识别 GNN 用于分类的更重要的边，从而提供更好的解释。此外，分布感知对齐策略（如基于高斯混合模型的变体）在大多数情况下表现出更强的忠实度。另外值得注意的是，在 Tree-Grid 上，基于互信息的对齐的忠实度取决于边数，并且当边数处于 $[8, 15]$ 范围内时可取得更好的结果。

表2. 边上的曲线下面积（AUC）解释忠实度

| | BA-Shapes | Tree-Grid | 感染 | Mutag |
|---|---|---|---|---|
| GRAD | 88.2 | 61.2 | 74.0 | 78.3 |
| ATT | 81.5 | 66.7 | – | 76.5 |
| Gem | 97.1 | – | – | 83.4 |
| RG-Explainer | 98.5 | 92.7 | – | 87.3 |
| GNNExplainer | $93.1_{\pm 1.8}$ | $86.2_{\pm 2.2}$ | $92.2_{\pm 1.1}$ | $74.9_{\pm 1.9}$ |
| + 对齐_嵌入 | $95.3_{\pm 1.4}$ | $91.2_{\pm 2.3}$ | $93.0_{\pm 1.0}$ | $76.3_{\pm 1.7}$ |
| + 对齐_锚点 | $97.1_{\pm 1.3}$ | $92.4_{\pm 1.9}$ | $93.1_{\pm 0.8}$ | $78.9_{\pm 1.6}$ |
| + 对齐_互信息 | $97.4_{\pm 1.6}$ | $92.2_{\pm 2.5}$ | $93.2_{\pm 1.0}$ | $78.2_{\pm 1.8}$ |
| + 对齐_高斯 | $96.7_{\pm 1.3}$ | $92.5_{\pm 1.9}$ | $93.1_{\pm 0.9}$ | $79.3_{\pm 1.5}$ |
| PGExplainer | $96.9_{\pm 0.7}$ | $92.7_{\pm 1.5}$ | $89.6_{\pm 0.6}$ | $83.7_{\pm 1.2}$ |
| + 对齐_嵌入 | $97.2_{\pm 0.7}$ | $95.8_{\pm 0.9}$ | $90.5_{\pm 0.7}$ | $92.8_{\pm 1.1}$ |
| + 对齐_锚点 | $98.7_{\pm 0.5}$ | $94.7_{\pm 1.2}$ | $91.6_{\pm 0.6}$ | $94.5_{\pm 0.8}$ |
| + 对齐_互信息 | $\mathbf{99.3}_{\pm 0.8}$ | $96.2_{\pm 1.3}$ | $92.0_{\pm 0.4}$ | $94.3_{\pm 1.1}$ |
| + 对齐_高斯 | $99.2_{\pm 0.3}$ | $\mathbf{96.4}_{\pm 1.1}$ | $\mathbf{92.5}_{\pm 0.8}$ | $\mathbf{95.9}_{\pm 1.2}$ |

通过这两项实验，我们可以观察到，嵌入对齐能够获得忠实性更高的解释，并且可以灵活地融入 GNNExplainer 和 PGExplainer 等各种模型中，从而回答了 RQ1。

### 7.3 解释一致性

伪解释的一个问题是，由于解释器初始化具有随机性，图神经网络解释器在不同运行中对同一实例给出的解释可能不同，这违背了解释的一致性 。为检验这一问题的严重程度并回答**研究问题2**，我们从解释一致性的角度评估所提出的框架。我们采用 GNNExplainer 和 PGExplainer 作为基线方法。具体而言，我们计算每次识别出的、重要性权重排名前-$k$ 的解释性边之间的 SHD距离。随后，对测试集中的所有实例取平均值。每项实验均进行 5 次，并分别在表3和表4中报告 Tree-Grid 和 Mutag 数据集上的平均一致性。距离越大，表示解释越不一致。从表中可以看出，现有遵循公式1的方法存在一致性问题。例如，GNNExplainer 对排名前 6 的边计算出的平均 SHD距离为 4.39。引入对齐嵌入的辅助任务，可显著提升解释器在这一指标上的表现。在这些不同的对齐策略中，基于高斯混合模型的变体在大多数情况下表现最佳。在 TreeGrid 数据集上引入 Align_Gaus 后，GNNExplainer 的排名前 6 的边的 SHD距离从 4.39 降至 2.13，PGExplainer 则从 1.38 降至 0.13。在 Mutag 数据集上引入该方法后，GNNExplainer 的排名前 6 的边的 SHD距离从 4.78 降至 3.85，PGExplainer 则从 3.42 降至 1.15。这些结果验证了我们提出的方法在获得一致性解释方面的有效性。

### 7.4 避免虚假解释的能力

现有的图解释基准通常设计得较为简单，歧义较少，其中标签仅有一个真实原因，而识别出的解释性子结构则通过与真实解释进行比较来评估。然而，这一结果可能会产生误导，因为解释在更复杂场景中的忠实性尚未经过检验。真实世界数据集通常充斥着伪模式，而训练后的图神经网络可能包含各种偏差，因此对解释方法提出了更高的要求。因此，为了评估我们的框架能否缓解伪解释问题

表3. 一致 性的解释 在 Tree-Grid 上随机运行 5 轮的平均 SHD 在 Tree-Grid 上随机运行的彩争霸。

| | Top-K边 | | | | | |
|---|---|---|---|---|---|---|
| 方法 | 1 | 2 | 3 | 4 | 5 | 6 |
| GNNExplainer | 0.86 | 1.85 | 2.48 | 3.14 | 3.77 | 4.39 |
| +对齐_嵌入 | 0.77 | 1.23 | 1.28 | 0.96 | 1.81 | 2.72 |
| +对齐_锚点 | 0.72 | 1.06 | 0.99 | 0.53 | 1.52 | 2.21 |
| +对齐_互信息 | 0.74 | 1.11 | 1.08 | 1.32 | 1.69 | 2.27 |
| +对齐_高斯 | 0.68 | 1.16 | 1.13 | 0.72 | 1.39 | 2.13 |
| PGExplainer | 0.74 | 1.23 | 0.76 | 0.46 | 0.78 | 1.38 |
| +对齐_嵌入 | 0.11 | 0.15 | 0.13 | **0.11** | 0.24 | 0.19 |
| +对齐_锚点 | 0.07 | 0.12 | 0.13 | 0.16 | 0.21 | **0.13** |
| +对齐_互信息 | 0.28 | 0.19 | 0.27 | 0.15 | 0.20 | 0.16 |
| +对齐_高斯 | **0.05** | **0.08** | **0.10** | 0.12 | **0.19** | **0.13** |

表4. 在 Mutag 上随机运行 5 轮时，解释在平均 SHD距离方面的一致性。

| | Top-K边 | | | | | |
|---|---|---|---|---|---|---|
| 方法 | 1 | 2 | 3 | 4 | 5 | 6 |
| GNNExplainer | 1.12 | 1.74 | 2.65 | 3.40 | 4.05 | 4.78 |
| +对齐_嵌入 | 1.05 | 1.61 | 2.33 | 3.15 | 3.77 | 4.12 |
| +对齐_锚点 | 1.06 | 1.59 | 2.17 | 3.06 | 3.54 | 3.95 |
| +对齐_互信息 | 1.11 | 1.68 | 2.42 | 3.23 | 3.96 | 4.37 |
| +对齐_Gaus | 1.03 | 1.51 | 2.19 | 3.02 | 3.38 | 3.85 |
| PGExplainer | 0.91 | 1.53 | 2.10 | 2.57 | 3.05 | 3.42 |
| +对齐_嵌入 | 0.55 | 0.96 | 1.13 | 1.31 | 1.79 | 2.04 |
| +对齐_锚点 | **0.51** | **0.90** | **1.05** | 1.27 | 1.62 | 1.86 |
| +对齐_互信息 | 0.95 | 1.21 | 1.73 | 2.25 | 2.67 | 2.23 |
| +对齐_Gaus | 0.59 | 1.34 | 1.13 | **0.84** | **1.25** | **1.15** |

并回答**研究问题3**，我们创建了一个新的图分类数据集：MixMotif。该数据集使我们能够训练一个存在偏差的图神经网络模型，并测试解释方法能否成功揭示这些偏差。

具体而言，受 [57], 启发，我们设计了三种基础图，即树、梯形和轮形，以及三种模体，即环、房屋和网格。按照混合比例 $\gamma$，模体会优先连接到基础图上。例如，环以 $\frac{2}{3}\gamma + \frac{1}{3}$ 的概率连接到树，以 $\frac{1-\gamma}{3}$ 的概率连接到其他基础图。房屋与梯形、网格与轮形的情况也类似。将所得图的标签设为模体类型。当 $\gamma$ 设为 0 时，每种模体连接到三种基础图的概率相同。换言之，对于每种模体，连接到哪种基础图都不存在偏差。因此，我们将 $\gamma = 0$ 的数据集视为干净或无偏的数据集。我们预期，在包含 $\gamma = 0$ 的数据上训练的图神经网络会关注模体结构来进行模体分类。然而，当 $\gamma$ 变大时，基础图与标签之间就会出现伪相关；也就是说，图神经网络可能会利用基础图结构进行模体分类，而不是依赖模体结构。对于每种设置，生成的数据集都包含 3,000 个图，训练集、验证集和测试集按 5 : 2 : 3 划分。

在本实验中，我们分别将 $\gamma$ 设为 0 和 0.7，并在每种设置下训练图神经网络 $f_0$ 和 $f_{0.7}$。我们测试这两个模型的图分类性能。随后，将解释方法应用于 $f_0$并在其上进行微调。之后，使用找到的超参数将这些解释方法应用于 $f_{0.7}$ 进行解释。结果汇总于表5。

从表5中可以观察到：(1) $f_0$ 在测试期间实现了近乎完美的图分类性能。这一高准确率表明，它捕捉到了真正的模式，依靠模体进行

表5。MixMotif 上的性能。使用不同的 $\gamma$训练两个图神经网络。我们评估它们在图分类任务中的性能，然后将得到的解释与模体进行比较。

| | | $\gamma$ 在训练中 | |
|---|---|---|---|
| 分类 | | 0 | 0.7 |
| $\gamma$ 在测试中 | 0 | 0.982 | 0.765 |
| | 0.7 | 0.978 | 0.994 |
| 解释 | | PGExplainer +对齐 | PGExplainer +对齐 |
| AUROC 针对模体 | | 0.711 **0.795**（越高越好） | 0.748 **0.266**（越低越好） |

预测。观察解释结果可以发现，我们的方案提供了更忠实的解释，在模体上的 AUROC 更高。(2) $f_{0.7}$ 在 $\gamma = 0$下无法进行良好预测，表明其中存在偏差，且不再仅依赖模体结构进行预测。尽管此时未知真实解释，但成功的解释应当揭示这种偏差。然而，PGExplainer 会生成与干净模型相似的解释，仍然与模体结构高度一致。相比之下，对于嵌入对齐生成的解释，AUROC分数会从 0.795 降至 0.266，揭示预测依据的变化，从而能够暴露偏差。(3) 总而言之，我们的方案能够为干净和混合设置提供更忠实的解释，而 PGExplainer 会受到伪解释的影响，无法忠实地解释图神经网络的预测，尤其是在存在偏差的情况下。

### 7.5 超参数敏感性分析

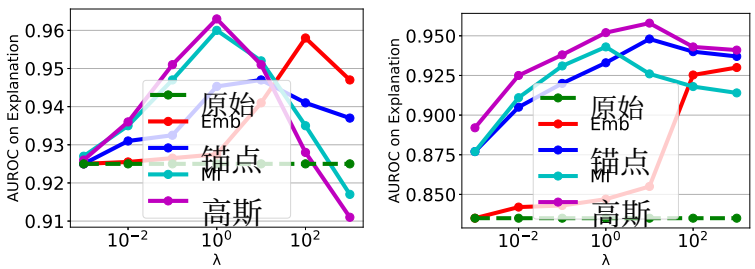
在本部分中，我们改变超参数 $\lambda$，以测试所提出框架对其取值的敏感性。$\lambda$ 控制所提出的嵌入对齐任务的权重。为简化起见，其他所有配置均保持不变，并将 $\lambda$ 的取值范围设为 $[1e-3, 1e-2, 1e-1, 1, 10, 1e2, 1e3\}$。采用 PGExplainer 作为基础方法。在 Tree-Grid 和 Mutag 数据集上分别随机进行 3 次实验。平均结果如图 5 所示。从图中可以得出以下观察结果：

• 对于所有四种变体，增大 $\lambda$ 起初会带来正面效果，但继续增大则会导致性能下降。例如，在 Tree-Grid 数据集上，基于锚点、潜在高斯混合模型和互信息分数的变体，其最佳结果均在 $\lambda$ 约为 1 时取得。当 $\lambda$ 较小时，公式 6 中的解释对齐正则化权重不足。另一方面，过大的 $\lambda$ 可能会削弱基于互信息最大化的解释框架的作用，而该框架能够保留所获解释的预测能力。

• 在这些变体中，基于潜在高斯混合模型的策略在大多数情况下表现最佳。例如，在 Tree-Grid 和 Mutag 这两个数据集上，该变体在识别出的解释性边上均取得了最高的 AUROC 分数。另一方面，直接使用欧几里得距离的变体在大多数情况下表现较差。我们将此归因于它们在分布建模和执行对齐方面能力的差异。

### 8 结论

在本文中，我们研究了一个为图神经网络获取忠实且一致的解释的新问题，而这一问题在现有的基于互信息最大化的解释框架中基本上被忽视了。通过对图神经网络推理过程进行深入分析，我们提出了一种通过对齐内部嵌入来实现的简单而有效的方法。理论分析表明，该方法在设计上更加忠实，其优化目标旨在促进原始图、图神经网络输出与识别出的解释之间的高互信息。我们设计了四种不同的策略，分别是直接采用欧几里得距离、使用锚点、采用基于高斯混合模型的 KL 散度，以及使用估计的互信息分数。所有这些算法都可以




(a) Tree-Grid (b) Mutag

图 5. PGExplainer 对嵌入对齐损失权重的敏感性。

轻松融入现有方法中。实验验证了它们在提升解释的忠实性和一致性方面的有效性。

未来，我们将寻求更鲁棒的解释。鲁棒性的提升意味着更强的泛化性，同时也能提供更好的类别级解释。此外，解释方法的评估也需要进一步研究。现有基准测试通常清晰明确，无法模拟复杂的现实场景。

**致谢**

本材料基于美国国家科学基金会通过项目编号 IIS-1707548 和 IIS-1909702、美国陆军研究办公室通过项目编号 W911NF21-1-0198，以及 DHS CINA 通过项目编号 E205949D 提供的资助或部分资助。本论文中的研究结果和结论不一定反映资助机构的观点。

**参考文献**

[1] Wenqi Fan、Y. Ma、Qing Li、Yuan He、Y. Zhao、Jiliang Tang 和 D. Yin。用于社交推荐的图神经网络。万维网大会，2019年。[2] Tianxiang Zhao、Xianfeng Tang、Xiang Zhang 和 Suhang Wang。半监督图到图转换。载于第*29*届*ACM*信息与知识管理国际会议论文集，第1863–1872页，2020年。[3] Linli Xu、Wenjun Ouyang、Xiaoying Ren、Yang Wang 和 Liang Jiang。利用句法依存关系增强双语词嵌入的语义表示。载于*IJCAI*，第4517–4524页，2018年。[4] Elman Mansimov、O. Mahmood、Seokho Kang 和 Kyunghyun Cho。利用深度生成式图神经网络预测分子几何。《科学报告》，9，2019年。[5] Hryhorii Chereda、A. Bleckmann、F. Kramer、A. Leha 和 T. Beißbarth。利用图卷积神经网络中的分子网络信息预测乳腺癌转移事件。《健康技术与信息学研究》，267:181–186，2019年。[6] Daniil Sorokin 和 Iryna Gurevych。利用门控图神经网络建模语义，以实现知识库问答。*arXiv*，abs/1808.04126，2018年。[7] Zhiwei Liu、Liangwei Yang、Ziwei Fan、Hao Peng 和 Philip S Yu。基于图神经网络的联邦社交推荐。*ACM*智能系统与技术汇刊 *(TIST)*，13(4):1–24，2022年。[8] Thomas N Kipf 和 Max Welling。利用图卷积网络进行半监督分类。*arXiv*预印本 *arXiv:1609.02907*，2016年。[9] Petar Veličković、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。图注意力网络。*arXiv*预印本 *arXiv:1710.10903*，2017年。[10] Will Hamilton、Zhitao Ying 和 Jure Leskovec。大规模图上的归纳式表示学习。神经信息处理系统进展，30，2017年。[11] Weijieying Ren、Lei Zhang、Bo Jiang、Zhefeng Wang、Guangming Guo 和 Guiquan Liu。面向缺失标签多视图多标签分类的鲁棒映射学习。载于知识科学、工程与管理：第十届国际会议，*KSEM 2017*，澳大利亚维多利亚州墨尔本，*2017*年*8*月*19*日至*20*日，论文集第*10*卷，第543–551页。施普林格，2017年。

[12] Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。图神经网络的表达能力有多强？arXiv预印本 arXiv:1810.00826，2018年。[13] Muhan Zhang 和 Yixin Chen。基于图神经网络的链接预测。神经信息处理系统进展，31，2018年。[14] Zhilong Lu、Weifeng Lv、Zhipu Xie、Bowen Du、Guixi Xiong、Leilei Sun 和 Haiquan Wang。采用注意力机制进行交通速度预测的图序列神经网络。ACM 智能系统与技术汇刊 (TIST)，13(2):1–24，2022年。[15] Xiaoying Ren、Linli Xu、Tianxiang Zhao、Chen Zhu、Junliang Guo 和 Enhong Chen。追踪并预测众筹动态：一种基合成方法。载于2018年IEEE国际数据挖掘会议 (ICDM)，第1212–1217页。IEEE，2018年。[16] Jiahua Rao、Shuangjia Zheng 和 Yuedong Yang。分子性质预测中可解释图神经网络的定量评估。arXiv预印本 arXiv:2107.04119，2021年。[17] Weijieying Ren、Lei Wang、Kunpeng Liu、Ruocheng Guo、Lim Ee Peng 和 Yanjie Fu。以不平衡交互缓解推荐中的流行度偏差：梯度视角。载于2022年IEEE国际数据挖掘会议 (ICDM)，第438–447页。IEEE，2022年。[18] Weijieying REN、Jing Jiang、Ling Min Serena Khoo 和 Hai Leong Chieu。利用主题混合进行跨主题谣言检测。2021年。[19] Rex Ying、Dylan Bourgeois、Jiaxuan You、Marinka Zitnik 和 Jure Leskovec。GNNExplainer：生成图神经网络的解释。神经信息处理系统进展，32:9240，2019年。[20] Dongsheng Luo、Wei Cheng、Dongkuan Xu、Wenchao Yu、Bo Zong、Haifeng Chen 和 Xiang Zhang。用于图神经网络的参数化解释器。神经信息处理系统进展，33:19620–19631，2020年。[21] Tianxiang Zhao、Dongsheng Luo、Xiang Zhang 和 Suhang Wang。关于图神经网络解释的一致性。arXiv预印本 arXiv:2205.13733，2022年。[22] Qiang Huang、Makoto Yamada、Yuan Tian、Dinesh Singh、Dawei Yin 和 Yi Chang。GraphLIME：用于图神经网络的局部可解释模型解释。arXiv预印本 arXiv:2001.06216，2020年。[23] Minh N Vu 和 My T Thai。PGM-Explainer：图神经网络的概率图模型解释。arXiv预印本 arXiv:2010.05788，2020年。[24] Hao Yuan、Haiyang Yu、Jie Wang、Kang Li 和 Shuiwang Ji。通过子图探索实现图神经网络的可解释性。载于国际机器学习会议，第12241–12252页。PMLR，2021年。[25] Hao Yuan、Haiyang Yu、Shurui Gui 和 Shuiwang Ji。图神经网络可解释性：分类综述。arXiv预印本 arXiv:2012.15445，2020年。[26] Meike Nauta、Jan Trienes、Shreyasi Pathak、Elisa Nguyen、Michelle Peters、Yasmin Schmitt、Jörg Schlötterer、Maurice van Keulen 和 Christin Seifert。从轶事证据到定量评估方法：可解释人工智能评估的系统综述。arXiv预印本 arXiv:2201.08164，2022年。[27] Enyan Dai、Wei Jin、Hui Liu 和 Suhang Wang。面向含噪图与稀疏标签的鲁棒图神经网络。arXiv预印本 arXiv:2201.00232，2022年。[28] Tianxiang Zhao、Xiang Zhang 和 Suhang Wang。GraphSMOTE：利用图神经网络处理图上的不平衡节点分类。载于第十四届ACM网络搜索与数据挖掘国际会议论文集，2021年。[29] Yu Zhou、Haixia Zheng、Xin Huang、Shufeng Hao、Dengao Li 和 Jumin Zhao。图神经网络：分类体系、进展与趋势。ACM智能系统与技术汇刊 (TIST)，13(1):1–54，2022年。[30] Tianxiang Zhao、Wenchao Yu、Suhang Wang、Lu Wang、Xiang Zhang、Yuncong Chen、Yanchi Liu、Wei Cheng 和 Haifeng Chen。通过次优示范进行模仿学习的技能解耦。KDD，2023年。[31] Joan Bruna、Wojciech Zaremba、Arthur Szlam 和 Yann LeCun。图上的谱网络与局部连接网络。arXiv预印本 arXiv:1312.6203，2013年。[32] Ke Liang、Jim Tan、Dongrui Zeng、Yongzhe Huang、Xiaolei Huang 和 Gang Tan。AbsLearn：用于混叠和缓冲区大小信息检索的基于图神经网络的框架。《模式分析与应用》，第1–19页，2023年。[33] Ke Liang、Lingyuan Meng、Meng Liu、Yue Liu、Wenxuan Tu、Siwei Wang、Sihang Zhou、Xinwang Liu 和 Fuchun Sun。针对不同类型知识图谱的推理：静态、时序与多模态。arXiv预印本 arXiv:2212 05767，2022年。。[34] David K Duvenaud、Dougal Maclaurin、Jorge Iparraguirre、Rafael Bombarell、Timothy Hirzel、Alán Aspuru-Guzik 和 Ryan P Adams。用于学习分子指纹的图卷积网络。载于神经信息处理系统进展，第2224–2232页，2015年。[35] James Atwood 和 Don Towsley。扩散卷积神经网络。载于神经信息处理系统进展，第1993–2001页，2016年。[36] Teng Xiao、Zhengyu Chen、Donglin Wang 和 Suhang Wang。学习如何在图神经网络中传播消息。载于第27届第1903届ACM SIGKDD知识发现与数据挖掘会议论文集，第–页，2021年1894。

[37] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals 和 George E Dahl。《用于量子化学的神经消息传递》。载于 ICML，2017年。[38] Tianxiang Zhao、Dongsheng Luo、Xiang Zhang 和 Suhang Wang。《Topoimb：面向图学习中的拓扑层级不平衡》。载于图学习会议，第 37–1 页。PMLR，2022年。[39] Xiang Wang、Hongye Jin、An Zhang、Xiangnan He、Tong Xu 和 Tat-Seng Chua。《解耦图协同过滤》。载于第43届ACM信息检索研究与发展国际会议论文集，第 1001–1010 页，2020年。[40] Tianxiang Zhao、Xiang Zhang 和 Suhang Wang。《探索边解耦在节点分类中的应用》。载于2022 年 ACM Web 会议论文集，第 1028–1036 页，2022年。[41] Teng Xiao、Zhengyu Chen、Zhimeng Guo、Zeyang Zhuang 和 Suhang Wang。《面向非同配图的解耦自监督学习》。arXiv 电子预印本，第 arXiv–2206 页，2022年。[42] Shuai Lin、Chen Liu、Pan Zhou、Zi-Yuan Hu、Shuojia Wang、Ruihui Zhao、Yefeng Zheng、Liang Lin、Eric Xing 和 Xiaodan Liang。《原型图对比学习》。IEEE神经网络与学习系统汇刊，2022年。[43] Junjie Xu、Enyan Dai、Xiang Zhang 和 Suhang Wang。《Hp-gmn：用于异配图的图记忆网络》。arXiv 预印本 arXiv:2210.08195，2022年。[44] Muhammet Balcilar、Pierre Héroux、Benoit Gauzere、Pascal Vasseur、Sébastien Adam 和 Paul Honeine。《突破消息传递图神经网络的局限》。载于国际机器学习会议，第 599–608 页。PMLR，2021年。[45] Ryoma Sato。《图神经网络表达能力综述》。arXiv 预印本 arXiv:2003.04078，2020年。[46] Hoang Nt 和 Takanori Maehara。《重新审视图神经网络：我们拥有的只有低通滤波器》。arXiv 预印本 arXiv:1905.09550，2019年。[47] Muhammet Balcilar、Renton Guillaume、Pierre Héroux、Benoit Gaüzère、Sébastien Adam 和 Paul Honeine。《从谱视角分析图神经网络的表达能力》。载于国际学习表征会议（ICLR）论文集，2021年。[48] Xiang Wang、Yingxin Wu、An Zhang、Xiangnan He 和 Tat-seng Chua。《通过因果筛选解释图神经网络》。2020年。[49] Federico Baldassarre 和 Hossein Azizpour。《图卷积网络的可解释性技术》。arXiv 预印本 arXiv:1905.13686，2019年。[50] Zaixi Zhang、Qi Liu、Hao Wang、Chengqiang Lu 和 Cheekong Lee。《Protgnn：迈向自解释图神经网络》。arXiv 预印本 arXiv:2112.00911，2021年。[51] Enyan Dai 和 Suhang Wang。《迈向自可解释图神经网络》。载于第30届ACM信息与知识管理国际会议论文集，第 302–311 页，2021年。[52] Enyan Dai、Tianxiang Zhao、Huaisheng Zhu、Junjie Xu、Zhimeng Guo、Hui Liu、Jiliang Tang 和 Suhang Wang。《可信图神经网络综合综述：隐私、鲁棒性、公平性与可解释性》。arXiv 预印本 arXiv:2204.08570，2022年。[53] Phillip E Pope、Soheil Kolouri、Mohammad Rostami、Charles E Martin 和 Heiko Hoffmann。《图卷积神经网络的解释方法》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 10772–10781 页，2019年。[54] Caihua Shan、Yifei Shen、Yao Zhang、Xiang Li 和 Dongsheng Li。《强化学习增强的图神经网络解释器》。载于神经信息处理系统进展，第 34 卷，2021年。[55] Thomas Schnake、Oliver Eberle、Jonas Lederer、Shinichi Nakajima、Kristof T Schütt、Klaus-Robert Müller 和 Grégoire Montavon。《通过相关游走实现图神经网络的高阶解释》。arXiv 预印本 arXiv:2006 03589，2020年。。[56] Wanyu Lin、Hao Lan 和 Baochun Li。《图神经网络的生成式因果解释》。载于国际机器学习会议，第 6666–6679 页。PMLR，2021年。[57] Ying-Xin Wu、Xiang Wang、An Zhang、Xiangnan He 和 Tat-Seng Chua。《发现图神经网络的不变理由》。arXiv 预印本 arXiv:2201.12872，2022年。[58] Junchi Yu、Tingyang Xu、Yu Rong、Yatao Bian、Junzhou Huang 和 Ran He。《用于子图识别的图信息瓶颈》。arXiv 预印本 arXiv:2010.05563，2020 年。[59] Naftali Tishby 和 Noga Zaslavsky。《深度学习与信息瓶颈原理》。载于2015年IEEE信息论研讨会（ITW），第 1–5 页。IEEE，2015年。[60] Phúc H. Lê Khac、Graham Healy 和 Alan F. Smeaton。《对比表征学习：框架与综述》。IEEE Access，8：193907–193934，2020年。[61] Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。《监督式对比学习》。ArXiv，abs/2004.11362，2020年。

[62] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。《视觉表征对比学习的简单框架》。载于国际机器学习会议，第 1597–1607 页。PMLR，2020年。[63] Florian Graf、Christoph Hofer、Marc Niethammer 和 Roland Kwitt。《剖析监督式对比学习》。ArXiv，abs/2102.08817，2021年。[64] Tongzhou Wang 和 Phillip Isola。《通过超球面上的对齐与均匀性理解对比表征学习》。ArXiv，abs/2005.10242，2020年。[65] Shengyu Feng、Baoyu Jing、Yada Zhu 和 Hanghang Tong。《结合信息正则化的对抗式图对比学习》。载于2022 年 ACM Web 会议论文集，第 1362–1371 页，2022年。[66] Lianghao Xia、Chao Huang、Yong Xu、Jiashu Zhao、Dawei Yin 和 Jimmy Huang。《超图对比协同过滤》。载于第 45 届国际 ACM SIGIR 会议论文集，第 70–79 页，2022年。[67] Yanqiao Zhu、Yichen Xu、Qiang Liu 和 Shu Wu。《图对比学习的实证研究》。arXiv 预印本 arXiv:2109.01116，2021年。[68] Lei Wang、Ee-Peng Lim、Zhiwei Liu 和 Tianxiang Zhao。《解释引导的序列推荐对比学习》。载于第 31 届 ACM 国际信息与知识管理会议论文集，第 2017–2027 页，2022年。[69] Ke Liang、Yue Liu、Sihang Zhou、Wenxuan Tu、Yi Wen、Xihong Yang、Xiangjun Dong 和 Xinwang Liu。《基于关系对称结构的知识图谱对比学习》。IEEE 知识与数据工程汇刊，2023年。[70] Weihua Hu、Bowen Liu、Joseph Gomes、Marinka Zitnik、Percy Liang、Vijay S. Pande 和 Jure Leskovec。《图神经网络预训练策略》。arXiv: Learning，2019年。[71] Felix L. Opolka、Aaron Solomon、Cătălina Cangea、Petar Velickovic、Pietro Lio’ 和 R. Devon Hjelm。《时空深度图互信息最大化》。ArXiv，abs/1904.06316，2019年。[72] Petar Velickovic、William Fedus、William L. Hamilton、Pietro Lio’、Yoshua Bengio 和 R. Devon Hjelm。《深度图互信息最大化》。ArXiv，abs/1809.10341，2018年。[73] Qikui Zhu、Bo Du 和 Pingkun Yan。《图卷积网络的自监督训练》。ArXiv，abs/2006 02380，2020年。。[74] Yuning You、Tianlong Chen、Yongduo Sui、Ting Chen、Zhangyang Wang 和 Yang Shen。《结合数据增强的图对比学习》。ArXiv，abs/2010.13902，2020年。[75] Kaveh Hassani 和 Amir Hosein Khas Ahmadi。《图上的对比式多视图表征学习》。ArXiv，abs/2006.05582，2020年。[76] Ming Jin、Yizhen Zheng、Yuan-Fang Li、Chen Gong、Chuan Zhou 和 Shirui Pan。《用于自监督图表征学习的多尺度对比孪生网络》。ArXiv，abs/2105.05682，2021年。[77] Yanqiao Zhu、Yichen Xu、Feng Yu、Q. Liu、Shu Wu 和 Liang Wang。《深度图对比表征学习》。ArXiv，abs/2006.04131，2020年。[78] Liang Zeng、Lanqing Li、Zi-Chao Gao、Peilin Zhao 和 Jian Li。《Imgcl：重新审视不平衡节点分类中的图对比学习》。ArXiv，abs/2205.11332，2022年。[79] Yuning You、Tianlong Chen、Yang Shen 和 Zhangyang Wang。《图对比学习的自动化》。载于国际机器学习会议，2021年。[80] Yihang Yin、Qingzhong Wang、Siyu Huang、Haoyi Xiong 和 Xiang Zhang。《Autogcl：通过可学习视图生成器实现自动化图对比学习》。载于AAAI 人工智能会议，2021年。[81] Martin Ester、Hans-Peter Kriegel、Jörg Sander、Xiaowei Xu 等。《一种用于在含噪大型空间数据库中发现簇的基于密度的算法》。载于 KDD，第 96 卷，第 226–231 页，1996年。[82] R Devon Hjelm、Alex Fedorov、Samuel Lavoie-Marchildon、Karan Grewal、Phil Bachman、Adam Trischler 和 Yoshua Bengio。《通过互信息估计与最大化学习深层表征》。载于国际学习表征会议，2018年。[83] Lukas Faber、Amin K. Moghaddam 和 Roger Wattenhofer。《与真实标注进行比较为何有误：论图神经网络解释方法的评估》。载于第 27 届 ACM SIGKDD 知识发现与数据挖掘会议论文集，第 332–341 页，2021年。[84] Ioannis Tsamardinos、Laura E Brown 和 Constantin F Aliferis。《最大—最小爬山贝叶斯网络结构学习算法》。机器学习，65(1)：31–78，2006年。

收稿日期：2023年3月13日