arXiv:2602.06787v1 [cs.LG] 6 Feb 2026

Weisfeiler 和 Lehman 的范畴化

Seongjin Choi
JINCS10@POSTECH.AC.KR浦项科技大学，庆尚北道，韩国与几何与物理中心，基础科学研究院（基础科学研究院），浦项市南区竹谷路*127*番街*79*号，庆尚北道，韩国 *37673*

Gahee Kim
GAHEEKIM@KAIST.AC.KR金载哲人工智能研究生院，韩国科学技术院（韩国科学技术院），首尔东大门区回基路*85*号，韩国 *02455*

Se-Young Yun YUNSEYOUNG@KAIST.AC.KR
韩国科学技术院金载哲人工智能研究生院，韩国首尔市东大门区回基路*85*号，邮编*02455*

摘要

尽管提升映射显著增强了图神经网络的表达能力，但将这一范式扩展到超图领域仍缺乏系统性。为此，我们提出了范畴化 Weisfeiler-Lehman 框架，将提升形式化为从任意数据范畴到统一的分次偏序集范畴的函子映射。将这一视角应用于超图时，我们可以系统地推导出超图同构网络——这是一类神经网络架构，其消息传递拓扑完全由所选函子决定。我们从超图范畴引入了两个不同的函子：关联函子和对称单纯复形函子。关联架构在结构上与标准二部图方案相似，但我们的函子推导方式在生成的偏序集上实现了更丰富的信息流，能够捕获现有方法常常忽略的复杂交几何。我们从理论上刻画了这些模型的表达能力，并证明基于关联的方案和对称单纯复形方案都涵盖了标准超图 Weisfeiler-Lehman 测试的表达能力。在真实世界基准上的大量实验验证了这些理论发现。

关键词： TopologicalDeepLearning、范畴深度学习、图神经网络、超图神经网络、Weisfeiler-Lehman 测试、表达能力

1. 引言

图神经网络（GNN）的表达能力长期以来受限于 1-Weisfeiler-Lehman（1-WL）测试（Weisfeiler and Leman，1968；Xu 等人，2019；Morris 等人，2019）。为突破这一局限，近期研究逐渐转向高阶图学习，将图提升至更丰富的定义域，例如单纯复形（Bodnar 等人，2021b）或正则胞腔复形（Bodnar 等人，2021a；Giusti等人，2024），以捕捉团和圈等标准消息传递无法识别的结构（Arvind 等人，2020；Chen 等人，2020）。这些方法成功统一了几何直觉，为高表达能力图学习确立了新标准（Ebli等人；Wu 等人，2023；Hajij 等人，2022）。

然而，将这一范式扩展到超图（Bretto，2013）仍然是一个重大的开放性挑战。不同于图中的团可以自然地提升为单纯形，超图由节点的任意子集构成，并不存在规范的几何实现。因此，当前

© S. Choi、G. Kim 和 S.-Y. Yun。

超图神经网络（Feng 等人，2019年）通常依赖基于图的归约，例如二部图展开（Feng 等人，2019年；Huang and Yang，2021年；Feng 等人，2024年；Zhang 等人，2025年）。

在本研究中，我们提出了一个统一且严谨的框架，以弥合这一差距。我们认为，当前方法之所以支离破碎，根源在于缺乏一种通用的数学语言来描述提升。我们引入范畴论（Mac Lane，1998年），并非将其视为抽象理论，而是将其作为解决超图提升问题的实用工程工具。我们提出范畴化 WL（CatWL）框架，这是一种广义同构检验机制，在通用域——分次偏序集（Stanley，2011）上运行。我们的核心洞见是，尽管输入数据类型（图、超图、单纯复形）各不相同，但其结构本质始终能够忠实地编码为分次偏序集。通过将提升映射（Bodnar 等人，2021a）定义为从数据范畴到分次偏序集范畴的函子，我们严格保证提升过程保留原对象的同构类。这种函子视角使我们能够系统地推导神经网络架构。我们证明，函子的选择决定了消息传递拓扑，由此形成一系列我们称为范畴化消息传递网络（CatMPN）的模型。具体而言，我们研究了两种不同的函子：关联偏序集函子将超边视为一维对象，从而还原出标准的二部图展开（Huang 和 Yang，2021年；Feng 等人，2024年；Zhang 等人，2025年）；另一种是对称单纯复形（Choi 等人，2025）函子，它将超边视为完全结构化的单纯形，并采用基数感知消息传递方案。我们的贡献总结如下：

- 范畴框架：我们提出 CatWL 框架，从范畴论的视角统一各种提升策略（Bodnar 等人，2021b，a）。
- 通用域：我们确定分次偏序集是 WL 测试的最优通用域，能够在单一计算结构中涵盖超图（Xu等人，2019；Feng 等人，2024）、单纯复形（Bodnar 等人，2021b）以及正则胞腔复形（Bodnar 等人，2021a）。
- 神经网络架构：我们将 CatMPN 推导为 CatWL 检验的神经网络版本，并从理论上刻画了两个具体实例的表达能力。
- 实证验证：我们进行了大量实验，证实在超图范畴中，CatMPN 在真实世界基准上的表现优于基线方法。

2. 为什么是范畴论

近年来，图神经网络的发展很大程度上由突破 1-WL 局限、提升表达能力的追求所推动（Morris等人，2019；Bodnar 等人，2021b，a）。尽管这促成了针对图的强大提升映射策略（Bodnar 等人，2021b，a），但由于结构不兼容性，这些方法向超图的扩展仍然较为零散。

提升失配本节将指出当前将高阶数据化归约为图的方法所存在的局限，并提出一个统一框架。我们认为，分次偏序集是提升操作的最佳通用域，而范畴论——尤其是函子性——提供了必要的严格条件，以确保这些提升保留同构。在图的领域，打破 1-WL 屏障的策略已相当成熟：通过为标准图神经网络引入新的几何或代数

构造来对图进行升级。单纯形 WL（SWL）和胞腔 WL（CWL）等方法将输入图提升到更高维度，例如构造团复形，或通过粘合二维圆盘来填充图中的环。这些构造拓展了交互的丰富性，使模型能够显式处理边界邻接和余边界邻接等拓扑特征。然而，这些特定的几何方法并不适用于超图。与图不同，图的团复形或环填充可提供规范的拓扑丰富化，而超图由节点的任意不规则子集构成，不存在与之等价的规范粘合过程。因此，当前尝试应用类似 WL 的过程时，往往依赖二部图展开，将超边仅视为次要的节点类型。这会扁平化高阶结构，掩盖对高阶表达能力至关重要的内在包含关系与交几何。

通用域 超越标准 WL 的一个常见策略，是将图提升到更丰富的高阶域中，并在那里进行细化。然而，目前这一领域较为分散：SWL 面向单纯复形（通过团复形），而 CWL 面向正则胞腔复形（通过为图中的每个环附加一个二维圆盘）。由于这些方法映射到不同的数学空间中，因此它们所产生的邻域、更新和表达能力概念很难在统一的理论框架内进行比较。对于超图而言，这一困难更加突出。为了解决这一问题，我们提出了一个统一域。我们将分次偏序集的空间确立为一个通用目标，它既能够编码超图的关联结构，也能够编码单纯复形和正则胞腔复形的面关系。通过将定义域固定为分次偏序集，我们确保着色更新规则及其神经松弛形式只需定义一次，便可在不同输入模态之间复用，而不必针对每一种新的提升构造重新设计。

函子性提升 在确立分次偏序集这一统一域之后，剩下的挑战就是定义提升映射本身。我们认为，仅仅将数据域视为对象的集合是不够的。为了严格捕捉同构检验所需的结构不变量，我们必须将该域视为一个范畴。因此，提升映射的概念自然应升级为函子的概念。这一视角提供了必要的有效性条件：函子性。虽然团复形之类的启发式方法隐含地满足这一条件，但函子性明确限制了超图的设计空间。它要求提升策略必须将超图的态射映射为诱导分级偏序集的态射。这一要求保证提升过程尊重输入数据的对称性，并为在偏序集域中设计规范的、保同构表示提供严格指导。

3. 范畴 WL 测试

在确立了对统一定义域和函子性提升机制的需求之后，我们现在正式定义 CatWL 框架。本节将介绍分次偏序集作为通用计算定义域，形式化提升所需的范畴论工具，并最终将 CatWL 检验定义为一种由函子引导的着色细化过程。

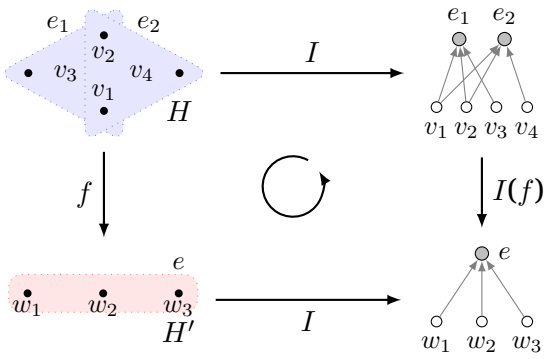


图 1： I 是一个函子，它通过将超图映射到其关联偏序集，将超图范畴映射到分次偏序集范畴。 I 将超图的一个态射 f 映射为分次偏序集的一个态射 $I(f)$，并且 I 保留恒等态射。这意味着 I 是一个提升映射。

3.1. 分次 WL 测试

我们首先定义目标定义域。选择分次偏序集（Stanley，2011）并非随意之举；它概括了图学习中使用的多种离散高阶表示。具体而言，超图、单纯复形和正则胞腔复形的关联偏序集在配备适当的维数函数后，都会自然地导出分次偏序集结构。分次偏序集提供了一个灵活的通用定义域，能够表示所有这些对象。

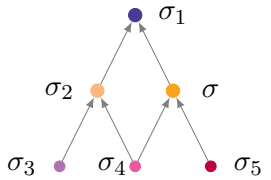
定义 1（分次偏序集）分次偏序集 P *(Stanley, 2011)* 是一个偏序集$(P, <)$，其维数函数为$\dim\colon P \to \mathbb{N}$，满足
1. $p \le p'$ 蕴含 $\dim(p) \le \dim(p')$*2.* 如果不存在满足 $p < q < p'$ 的 q，我们记为 $p \prec p'$。在这种情况下，$\dim(p') - \dim(p) = 1$。

直观而言，分次偏序集描述了一组元素，这些元素带有层次包含关系以及与之相关的维数概念。偏序规定了哪些元素包含于其他元素之中，或对其他元素进行细化；维数函数则为每个元素赋予一个整数级别。正如附录A所详述的，关联偏序集、单纯复形和正则胞腔复形会自然地导出分次偏序集结构。由于分次偏序集推广了用于 WL 测试的图的关联偏序集（Weisfeiler andLeman，1968），因此将 WL 测试推广到分次偏序集是很自然的。

定义 2（分次 Weisfeiler-Lehman（GWL））给定一个分次偏序集 P，我们定义如下着色细化算法：
1. c_0^P 在 P 上是常量着色。*2.* 给定 c_t^P 时，$c_{t+1}^P(\sigma)$ 在第 t 次迭代时，使用完美哈希函数根据 σ 的 *4*-邻接的颜色进行更新

$$c_{t+1}^P(\sigma) = \mathrm{HASH}\Big(c_t^P(\sigma), c_t^P(\mathcal{B}(\sigma)), c_t^P(\mathcal{C}(\sigma)), c_t^P(\mathcal{N}_{\downarrow}(\sigma)), c_t^P(\mathcal{N}_{\uparrow}(\sigma))\Big) \tag{1}$$

其中 σ 的 *4*-邻接的颜色为



$c(\mathcal{B}(\sigma)) = \{\!\{ \circ, \circ \}\!\}$
$c(\mathcal{C}(\sigma)) = \{\!\{ \circ \}\!\}$
$c(\mathcal{N}_{\downarrow}(\sigma)) = \{\!\{ (\circ, \circ) \}\!\}$
$c(\mathcal{N}_{\uparrow}(\sigma)) = \{\!\{ (\circ, \circ) \}\!\}$

图2：在示例分次偏序集 P上单次GWL细化步骤的可视化。哈斯图描绘了覆盖关系 $\prec$，其中箭头从被覆盖元素指向覆盖元素（例如， $\sigma_4 \prec \sigma$）。节点根据着色 c进行着色，并明确列出了元素 σ 的4-邻接多重集。

- $c_t^P(\mathcal{B}(\sigma)) = \{\!\{ c_t^P(\tau) \mid \tau \prec \sigma \}\!\}$， σ 的边界邻接颜色
- $c_t^P(\mathcal{C}(\sigma)) = \{\!\{ c_t^P(\tau) \mid \sigma \prec \tau \}\!\}$， σ的余边界邻接的颜色
- $c_t^P(\mathcal{N}_{\downarrow}(\sigma)) = \{\!\{ (c_t^P(\sigma')\ c_t^P(\tau)) \mid \tau \prec \sigma, \sigma', \sigma \neq \sigma' \}\!\}$， σ的下邻接的颜色
- $c_t^P(\mathcal{N}_{\uparrow}(\sigma)) = \{\!\{ (c_t^P(\sigma')\ c_t^P(\tau)) \mid \sigma, \sigma' \prec \tau, \sigma \neq \sigma' \}\!\}$， σ的上邻接的颜色

若对任意 $t \in \mathbb{N}$， c_t^P、 $c_t^{P'}$ 的着色直方图相等，则称两个分次偏序集 P、 P' 在*GWL*检验下同构。

图2展示了GWL着色细化的机制。我们考虑一个分次偏序集 $P = \sigma$、 σ_1、……、 σ_5，其覆盖关系由向上的箭头表示。具体的包含结构包括 σ_2、 $\sigma \prec \sigma_1$； σ_3、 $\sigma_4 \prec \sigma_2$；以及 σ_4、 $\sigma_5 \prec \sigma$。该偏序集以着色 c初始化，为每个元素分配不同的颜色： $c(\sigma) =$ 、 $c(\sigma_1) =$ 、$c($ $\sigma_2) =$ 、 $c(\sigma_3) =$ 、 $c(\sigma_4) =$ ， 以及 $c(\sigma_5) =$ 。在细化步骤中， σ ( )的颜色通过聚合图2中 σ 的4-邻接的颜色进行更新。

3.2. 范畴与函子

要将 GWL 机制应用于超图等任意数据类型，我们必须在形式上将数据与分次偏序集联系起来。我们通过范畴论实现这一点。

定义3（范畴） 范畴 $\mathcal{C}$ *(Mac Lane, 1998年)* 由对象类 $\mathcal{C}$和态射类 $Hom(\mathcal{C})$构成。对于每一对对象 $X, Y \in \mathcal{C}$，都存在一个态射集合 $\mathcal{C}(X, Y)$，其中包含所有从 X 到 Y的态射。范畴必须满足以下两条公理：

1. 复合：对于任意态射 $f \in \mathcal{C}(X, Y)$ 和 $g \in \mathcal{C}(Y, Z)$，都存在一个复合态射 $g \circ f \in \mathcal{C}(X, Z)$，且复合满足结合律。*2.* 恒等：对于每个对象 X，都存在一个恒等态射 $Id_X \in \mathcal{C}(X, X)$，使得$f \circ Id_X = f$ 且 $Id_Y \circ f = f$，对任意 $f \in \mathcal{C}(X, Y)$均成立。

如果存在满足$g \circ f = Id_X$、 $f \circ g = Id_Y$的 $f \in \mathcal{C}(X$、 $Y)$ $g \in \mathcal{C}(Y$、 $X)$，则称两个对象 X、Y同构。

常见的例子包括图范畴（**Gph**）、超图范畴（**Hyp**）、单纯形范畴和正则胞腔复形范畴，以及分次偏序集范畴（**Poset**）（形式化定义和证明见附录A）。

定义 *4*（函子）设 $\mathcal{C}$ 和 $\mathcal{D}$ 为范畴。函子 $F:\mathcal{C}\to\mathcal{D}$ （*Mac Lane，1998*）是一种映射，它为每个对象 $X\in\mathcal{C}$ 指定一个对象 $F(X)\in\mathcal{D}$，并为 $\mathcal{C}$ 中的每个态射 $f\in\mathcal{C}(X,Y)$ 指定一个态射 $F(f)\in\mathcal{D}(F(X)\ F(Y))$ $\mathcal{D}$。要成为有效的函子，F 必须保留范畴结构：

1. 保持复合：对于所有可复合的态射 f,g，$F(g\circ f)=F(g)\circ F(f)$。*2.* 保持恒等：对于每个对象 $X\in\mathcal{C}$，$F(id_X)=id_{F(X)}$。

函子推广了拓扑提升，因为如果 $X\cong Y$ 在 $\mathcal{C}$ 中同构，那么根据复合和恒等的保持性，$F(X)\cong F(Y)$ 在 $\mathcal{D}$ 中也应当同构。在我们的框架中，我们特别关注目标范畴为 $\mathcal{D}=$**偏序集** 的函子。函子 $F:\mathcal{C}\to$ **偏序集** 表示将数据类 $\mathcal{C}$ 系统地转换为分次偏序集。

3.3. 范畴 WL 测试

给定范畴 $\mathcal{C}$，判断 $\mathcal{C}$ 中的两个对象是否同构是一项困难的任务。我们通过函子 $F:\mathcal{C}\to$ **偏序集**，将 $\mathcal{C}$ 中的对象提升为分次偏序集，从而解决这一问题。借助函子 F，我们可以将 **偏序集** 上的 GWL 提升到 $\mathcal{C}$，以判断 $\mathcal{C}$ 中的两个对象是否同构。这就是我们将要介绍的 F-CatWL 检验。

定义 5（CatWL）令 $F:\mathcal{C}\to$ **偏序集**为一个函子。对于任意对象 $X\in\mathcal{C}$，F-范畴化 *WL*（*F-CatWL*）分两步进行：
1. 通过函子 F，将 $\mathcal{C}$ 中的对象 X 关联到其分次偏序集 $F(X)$。
2. 对 $F(X)$ 应用 *GWL*（定义 *2*）。
若且唯若 X、$X'\in\mathcal{C}$ 在 *GWL* 下同构，称它们在 ***F-CatWL*** 检验下同构，记作 $X\cong_F X'$。

由于任意函子 $F:\mathcal{C}\to$ **偏序集**都保持同构结构，我们得到蕴含链 $X\cong X'\implies F(X)\cong F(X')\implies F(X)\cong_{\mathrm{GWL}}F(X')$。最后这个等价关系定义了我们的 F-CatWL 可区分性，记作 $X\cong_F X'$。因此，F-CatWL 检验为 $\mathcal{C}$ 中的同构提供了一个必要的算法条件。通过将细化锚定在 F 中，我们有效地将分次偏序集定义域严谨的邻接结构拉回，从而促进源范畴中的同构检验。

4. 超图同构网络

我们已经定义了 F-CatWL，它依赖于通过函子 F 提升后的分次偏序集上的离散颜色细化过程。尽管理论上稳健，但对于连续特征学习而言，使用 HASH 在计算上难以处理。本节放宽这些限制，定义 F-范畴化消息传递网络，这是一类使用连续向量表示、对应于 F-CatWL 的神经网络架构。我们推导出不同的超图神经网络架构，统称为 F-超图同构网络（F-HIN）；每种架构都具有独特的归纳偏置和表达能力。

4.1. 范畴化消息传递网络

从 WL 测试转向神经网络，遵循几何深度学习的标准范式：用欧几里得特征向量替代离散颜色，并用通用消息函数和聚合器替代 HASH。

定义 6（F-CatMPN）设 $F\colon \mathcal{C} \to$ **偏序集** 是一个函子。对于任意对象 $X \in \mathcal{C}$，F-范畴化消息传递网络（*F-CatMPN*）的运作方式如下：
1. 通过函子 X，将对象 X 关联到 $\mathcal{C}$ 中的分次偏序集 $F(X)$ 通过函子 F。
2. 对于每个元素 $\sigma \in F(X)$，维护一个特征向量 $h_t^X(\sigma) \in \mathbb{R}^d$，在迭代 t时，计算对应于 *4*-邻接特征的四种类型的消息：

$$m^X_{\mathcal{B},t}(\sigma) := \mathrm{AGG}_{\tau \prec \sigma}\left(M_{\mathcal{B}}\left(h_t^X(\sigma), h_t^X(\tau)\right)\right) \tag{2}$$
$$m^X_{\mathcal{C},t}(\sigma) := \mathrm{AGG}_{\sigma \prec \tau}\left(M_{\mathcal{C}}\left(h_t^X(\sigma), h_t^X(\tau)\right)\right) \tag{3}$$
$$m^X_{\mathcal{N}_\downarrow,t}(\sigma) := \mathrm{AGG}_{\tau \prec \sigma,\sigma'}\left(M_{\mathcal{N}_\downarrow}\left(h_t^X(\sigma), h_t^X(\sigma'), h_t^X(\tau)\right)\right) \tag{4}$$
$$m^X_{\mathcal{N}_\uparrow,t}(\sigma) := \mathrm{AGG}_{\sigma,\sigma' \prec \tau}\left(M_{\mathcal{N}_\uparrow}\left(h_t^X(\sigma), h_t^X(\sigma'), h_t^X(\tau)\right)\right) \tag{5}$$

其中，$M_{\mathcal{B}}$、$M_{\mathcal{C}}$、$M_{\mathcal{N}_\downarrow}$、$M_{\mathcal{N}_\uparrow}$是可学习消息函数，聚合函数是一个聚合函数，而 U 是一个可学习更新函数。

这种神经近似与理论测试之间的关系由以下定理形式化，该定理将标准的 WL-MPN 等价性（Xu等人，2019年；Bodnar 等人，2021b，a）推广到了范畴设置。

定理 7（ F-CatMPN 的表达能力）令 $F\colon \mathcal{C} \to$ **偏序集**为一个函子。 *F-CatWL* 的能力至少与 *F-CatMPN* 一样强。如果更新函数和聚合函数均为单射，且网络具有足够的深度和宽度，则 *F-CatMPN* 与 *F-CatWL* 具有相同的表达能力。

特别地，对于函子 F : **超图** $\to$ **偏序集**，我们将 F-超图同构网络（F-HIN）定义为 F-CatMPN 的一种具体实例，它采用附录C中详述的消息、聚合和更新函数。

4.2. 来自两个函子的两种架构

我们的范畴框架的优势在于其模块化：消息传递架构完全由提升函子 F 的选择决定。不同的函子会导出不同的分次偏序集结构，从而生成在拓扑上各不相同的消息传递方案。本节将介绍两个主要函子——关联偏序集函子（I）和对称单纯复形函子（S）——并分析它们所导出的具体神经网络架构。

4.2.1. The Incidence Poset Functor (I)

提升超图最简单的方法，是将超边视为一维对象，将节点视为零维对象，并规定：当且仅当节点 v 包含在超边 e 中时，才存在覆盖关系 $v \prec e$。由此得到关联偏序集函子。

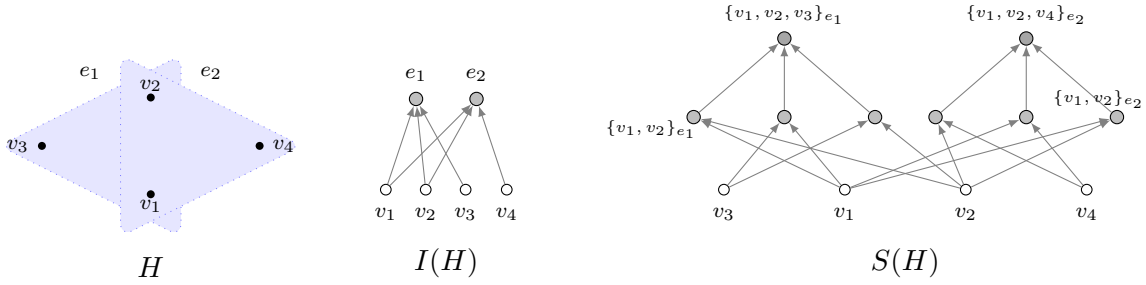


图 3：样例超图 H （左）上提升函子的比较。关联偏序集 $I(H)$（中）将超边视为一维对象，而对称单纯复形 $S(H)$（右）则显式表示所有子关系。

定义 8（关联偏序集函子）关联偏序集函子 I： **Hyp** $\to$ **偏序集**定义如下：对于对象 $H = (V, E, f)$， $I(H)$ 是关联偏序集$(V \amalg E, \prec)$，其中关系由节点-超边关联定义。对于态射 $\phi = (a, b) : H \to H'$，映射 $I(\phi)$ 将 $v \in V$ 映射到 $a(v)$，并将 $e \in E$ 映射到 $b(e)$，同时保持偏序结构。

在函子 I作用下，超图被映射为一个两层分次偏序集，其中节点位于维度 0，超边位于维度 1。所得分次偏序集中的颜色细化对应于一种自然的二部交互：节点更新会聚合来自余边界 $\mathcal{C}(v)$ 和上邻接 $\mathcal{N}_\uparrow(v)$ 的颜色（因为维度 0 的节点没有边界集和下邻接集）。反过来，超边更新会聚合来自边界 $\mathcal{B}(e)$ 和下邻接 $\mathcal{N}_\downarrow(e)$ 的颜色（因为维度 1 的超边没有余边界集和上邻接集）。

从结构上看， I-CatWL 的架构反映了标准超图 Weisfeiler-Lehman（HWL）测试（Feng 等人，2024年；Zhang 等人，2025年）背后的二部交互方案。然而，通过关联函子 I 推导该架构会带来更丰富的信息流。HWL 通常只利用可能交互的一部分，而我们的 I-CatWL 框架会自然地聚合 4-邻接（边界、余边界、下邻接和上邻接）的颜色。由于标准 HWL 更新实际上是对这一完整邻接聚合的限制，我们的方法能够捕获更大范围的几何信息。因此，正如我们将在下一个定理中正式证明的（证明推迟至附录D）， I-CatWL 的能力不弱于 HWL。

定理 9 *I-CatWL 的能力不弱于 HWL。*

4.2.2. The Symmetric Simplicial Complex Functor (S)

受近期关于超图高阶学习的研究启发（Choi 等人，2025年），我们考虑一种替代性的提升策略，它能明确保留每条超边的内部子集几何。对称单纯复形函子 S 将超图映射为一种结构，其中超边的每个子集都成为一个独立的几何对象，并按包含关系排序。这种方式会自然地为较大的超边赋予更高的维数，从而刻画子关系的完整格结构。

定义 *10*（对称单纯复形函子 S）设 $H = (V, E, f)$ 为一个超图。 H的对称单纯复形记作 $S(H)$，由 *0*-单纯形 $S(H)_0 := \{(\{v\}, v) \mid v \in V(H)\}$ 以及 n-单纯形 $S(H)_n := \{(\{v_{i_0}, \ldots, v_{i_n}\}, e) \mid \{v_{i_0}, \ldots, v_{i_n}\} \subset f(e)\}$（其中 $n \geq 1$）组成。为简化记号，我们将 $\{v_{i_0}, \ldots, v_{i_n}\}_e$ 记作 $(\{v_{i_0}, \ldots, v_{i_n}\}, e)$，将 $\{v\}_v$ 记作 $(\{v\}, v)$。该集合具有由以下两个条件定义的分次偏序集结构：*(1)* 若 $v \in e$，则偏序满足 $\{v\}_v < \sigma_e$；若 $e = e'$ 且 σ 的底层集合是 τ 的真子集，则 $\sigma_e < \tau_{e'}$；*(2)* 维数函数定义为 $\dim(\{v\}_v) = 0$，且对于 $\sigma_e \in S(H)_n$，$\dim(\sigma_e) = n$。

图 3 展示了对称单纯复形：关联偏序集 $I(H)$ 将结构扁平化为简单的二分图，而对称单纯复形 $S(H)$ 则揭示了嵌套于同一超边内的单纯形的复杂层级，例如三角形$\{v_1, v_2, v_3\}_{e_1}$覆盖边 $\{v_1, v_2\}_{e_1}$。

其背后的直觉在于， S是将超边 $e = \{v_0, \ldots, v_k\}$ 视为形式化的 k-单纯形，从而生成由子关系构成的完整面格。与 $I(H)$ 不同——其中所有超边都位于维度 1—— $S(H)$ 按基数对结构进行分层，将大小为 $k+1$的超边置于维度 k。由此产生的颜色细化在这一多层层级结构上运作：维度 k 处的更新会聚合来自（$k+1$）维余边界（超关系）以及（$k-1$）维边界（子关系）的颜色。这使得细粒度交互建模成为可能，例如节点 v 能够区分超边内部特定的成对关系 $\{v, w\}$，或者超边通过特定共享面进行通信以刻画交几何。与关联情形类似，下述定理确认这一对称单纯复形方案的能力不低于 HWL（证明详见附录 D）。

定理 11 *S-CatWL* 的能力不弱于 *HWL*。

5. 相关工作

图神经网络与拓扑提升 众所周知，标准消息传递神经网络（MPNN）受限于 1维WL（1-WL）测试所界定的表达能力 Xu等人(2019年)； Morris等人 (2019年)。尽管高阶 k-WL 变体具有更强的判别能力 Morris等人 (2023年)，但其计算成本高得难以承受。因此，近期研究转向拓扑深度学习，将图提升至高阶定义域，以捕捉圈等几何不变量。 Bodnar等人(2021b)率先通过将图提升为单纯复形来实现这一点，从而能够在更高维的单纯形上进行消息传递。随后， Bodnar 等人 (2021a)将这一方向拓展至正则 CW 复形，从而能够处理灵活的拓扑结构。同期研究也探索了类似的单纯形架构，用于轨迹预测和信号处理 Roddenberry 等人 (2021年)； Ebli等人； Bunch等人。近期，人们提出了拓扑消息传递的通用框架，以统一这些定义域 Hajij等人 (2022年)； Giusti等人 (2024年)。

超图学习与高阶 WL 超图神经网络（HNNs）将图学习扩展到捕捉多路交互。早期方法严重依赖谱方法或团扩展（Yadati 等人，2019年）。Feng 等人 （2019年）提出了

基于超边卷积的超图神经网络，以高效编码高阶相关性，而 Bai 等人 （2021年）进一步提出了超图卷积和超图注意力算子，以自适应地加权高阶交互。Dong 等人（2020）设计了一种超图卷积网络，其中包含独立的超节点和超边神经元，以及一种灵活的归一化方案，用于控制高基数超边和高阶顶点的影响。为应对传播规则的差异，Huang 和 Yang （2021年）以及 Chien 等人 （2022年）提出了可推广任意聚合函数的统一框架。最近，Wang 等人 （2022年）提出了一种等变超图扩散架构，该架构可被证明能够逼近任意连续等变超图扩散算子，从而为节点与超边的置换如何约束消息传递提供了原则性视角。然而，这些架构的理论表达能力仍是一个微妙的问题。B¨oker（2019年）建立了超图同构与颜色细化之间的基本联系，而Feng 等人 （2024年）提供了同构计算算法。Luo 等人 （2022年）进一步刻画了超图神经网络相对于标准图基线模型的区分能力。最近，研究关注点转向专门为超图构建高阶 WL 层级。Zhang 等人 （2025年）提出了高维广义 WL 算法，以将表达能力提升到标准 1-WL 约化之外。

6. 数值实验

我们在六个真实世界的超图分类基准数据集上，对采用两种提升函子实例化的 CatMPN 进行了实证评估。

6.1. 实验设置

数据集 我们在六个来自不同定义域的真实世界超图基准数据集上验证了我们的框架，这些数据集包括四种 IMDb 协作网络、Steam 玩家和 Twitter 好友 Feng 等人 (2019年) 的变体。这些数据集旨在体现广泛的拓扑复杂度，尤其是在超边基数方面。 IMDb 编剧数据集具有细粒度结构，合作单元较小，平均超边大小低于 5，因此模型需要检测细微的相关性。相比之下，Twitter 好友 数据集呈现出密集的高阶交互，每张图中包含许多超边；平均每张图约有 84.8 条超边，最多可达 603 条，而单条超边仍能捕捉基数高达 64 的高阶相互作用。附录中提供了全面的统计数据，包括分布详情和类别平衡情况F.1。

任务与评估 我们采用 10 折交叉验证对超图分类进行评估，遵循小规模超图基准数据集的标准流程 Xu等人(2019年）。我们报告各折验证准确率的最佳均值和标准差。

基线模型 我们将模型与一系列全面的基线模型进行比较：(1) 标准神经网络：多层感知机；(2) 超图神经网络：HCHA（Bai 等人，2021年）、HNHN（Dong 等人，2020）、HyperGCN（Yadati 等人，2019年）、UniGCNII（Huang 和 Yang，2021）以及 ED-HNN（Wang 等人，2022年）；(3) 基于 Transformer 的模型：AllSetTransformer（Chien 等人，2022）；(4) 基于 WL 的理论模型：HIC（Feng 等人，2024年）和

表 1 ：超图分类的评估结果。最佳结果以**粗体**显示。

	IMDB导演形式	IMDB导演类型	IMDB编剧形式	IMDB编剧类型	Steam 玩家	Twitter 好友
MLP	64.42 ± 4.50	72.10 ± 2.76	54.87 ± 3.81	35.24 ± 3.95	59.43 ± 2.86	62.89 ± 2.76
HCHA	61.65 ± 3.52	67.80 ± 2.83	52.18 ± 3.18	37.55 ± 3.81	65.94 ± 4.07	62.43 ± 3.52
HNHN	67.04 ± 3.83	78.07 ± 1.29	54.30 ± 3.11	54.17 ± 3.63	63.50 ± 3.70	66.95 ± 4.01
HyperGCN	62.81 ± 3.30	72.91 ± 1.38	54.33 ± 3.27	39.85 ± 3.50	65.98 ± 4.28	64.19 ± 6.53
UniGCNII	60.55 ± 3.50	72.39 ± 2.82	47.84 ± 2.41	43.78 ± 4.11	61.09 ± 4.06	61.31 ± 1.15
ED HNN	60.44 ± 3.24	72.73 ± 3.18	47.56 ± 2.61	30.46 ± 1.40	61.07 ± 6.87	59.76 ± 1.74
AllSetTransformer	60.83 ± 3.49	76.59 ± 1.46	46.51 ± 2.65	40.60 ± 3.00	60.51 ± 3.26	61.30 ± 1.09
HIC	67.30 ± 3.75	79.49 ± 1.74	56.70 ± 4.15	51.45 ± 5.38	64.89 ± 4.40	68.17 ± 6.46
2-FHNN*	68.11 ± 2.46	78.52 ± 1.07	55.36 ± 3.30	45.44 ± 1.24	**67.53 ± 1.23**	62.37 ± 1.91
I-HIN (u, b)	68.12 ± 2.66	**81.41 ± 2.32**	58.83 ± 4.37	59.73 ± 4.55	63.93 ± 3.79	72.52 ± 4.74
I-HIN (u, b, l)	**69.02 ± 3.32**	80.97 ± 2.43	**60.17 ± 2.33**	59.05 ± 4.05	63.71 ± 3.25	**74.66 ± 3.64**
S-HIN（u, b）	68.55 ± 3.53	80.73 ± 2.30	58.86 ± 6.37	60.83 ± 4.67	62.12 ± 2.79	65.57 ± 3.66
S-HIN (u, b, l)	68.81 ± 3.11	80.92 ± 2.44	57.24 ± 3.52	**61.02 ± 4.09**	62.65 ± 2.66	65.56 ± 3.82

* 直接摘自原论文的结果 Zhang 等人(2025年)。

2-FHNN（Zhang 等人，2025年）。所有基线模型均在相同设置下进行评估，以确保比较公平。

模型变体 我们使用关联偏序集函子 I（定义 8）和对称单纯复形函子 S（定义10），实例化 CatMPN，由此得到两个主要模型系列：I-HIN 和 S-HIN。在每个系列中，我们考察定义6中的不同消息类型组合。我们的主要变体使用上邻接（$\mathcal{N}_\uparrow$）和边界（$\mathcal{B}$）消息，记为 (u, b) 变体。引理 24 （附录 B）从理论上证明了这一选择的合理性：完整的 4-邻接更新规则与仅使用边界邻接与上邻接的简化规则具有相同的表达能力，因此能够在不损失表达能力的情况下采用更高效的架构。此外，我们还考察包含下邻接（$\mathcal{N}_\downarrow$）消息的 (u, b, l) 变体；尽管理论上二者等价，这些变体仍可能带来实际优势。对于 S-HIN，我们不将基数超过 20 的超边纳入提升范围；附录 F.1.2中的拓扑统计表明，该阈值涵盖了各数据集中绝大多数超边。包括超参数配置在内的实现细节见附录 F.2。

6.2. 结果与分析

现在，我们展示 CatMPN 在采用我们的两个范畴函子后的实证结果，并结合基准数据集的拓扑性质对其进行分析。评估围绕三个问题展开：CatMPN 是否优于标准超图神经网络和基于 WL 的理论模型；函子的选择如何与底层超图拓扑相互作用；以及在消息传递拓扑中加入额外的范畴邻接会如何影响性能。

相较于基线模型的整体性能 表 1显示，CatMPN的变体在六个基准数据集中的五个上取得了最佳性能。在IMDB导演类型数据集上， I-HIN比最强的基于WL的基线模型HIC高出1.92%；而在Twitter好友数据集上，这一优势扩大至6.49%。在颇具挑战性的IMDB wri genre任务上，提升尤为显著， S-HIN达到了61.02%，比该任务上最强的基线模型HNHN高出6.85%。这些改进从实证上验证了定理9 和11： F-CatWL在表达能力上相较于HWL的优势，能够转化为可测量的性能提升。

唯一的例外是Steam 玩家数据集，此时2-FHNN仍保持领先。该数据集的超边基数高度均一，第99百分位数仅为15；在这种情况下，相较于2-FHNN的专门化设计，我们的函子所提供的额外结构编码收益有限。

函子与超图拓扑的影响 两种函子的相对性能取决于超图拓扑，主要由图级密度而非超边基数驱动。在每个图包含大量超边的稠密基准数据集上， I-HIN 占优：在平均每个图有 84.8 条超边的 Twitter 好友上，它比 S-HIN 高出 9.09%；在平均为 46.4 的 Steam 玩家上高出 1.28%。我们将这一差距归因于判别信号的性质。当图包含大量超边时，分类依赖于超图上的全局连通模式，而 I-$\bar{\text{H}}$IN 的二部消息传递能够高效地聚合这些模式。对称单纯复形 $S(H)$ 在此情形下引入了组合复杂性，却没有带来相应比例的表示增益。

相反，在 MDB wri 流派上， S-HIN 取得了最佳总体结果，为 61.02%，而 I-HIN 为 59.73%。该数据集的特点是每个图的超边很少，平均为 4.5，但超边大小较大，第 99 百分位基数为 36。当超图稀疏、但单个超边结构丰富时，判别信号转向每条超边内部的子关系。 S-HIN 的单纯分解显式捕获了成对及高阶面包含关系，而 I-HIN 的扁平关联结构无法表示这些关系。这一发现证实了一项关键的范畴论洞见：最优函子并非普遍适用，而是取决于任务相关信息位于全局超边连通性还是局部交几何之中。

范畴邻接设计的影响 引理 24证明，(u,b) 变体的表达能力与完整的四邻接规则相同，因此 (u,b) 与 (u,b,l) 之间的任何性能差距都反映了优化动态，而非表征能力。实证结果表明，I-HIN 可从额外的下邻接中受益 $\mathcal{N}_{\downarrow}$，在 Twitter 好友和 IMDB wri form 上分别提升了 2.14% 和 1.34%。我们推测，在扁平的两层关联偏序集$I(H)$ 中，下邻接消息为聚合来自共享公共节点的同级超$\bar{\text{边}}$信$\bar{\text{息}}$提供了明确的捷径，即使原则上仅凭边界邻接与上邻接也能恢复相同信息，这一捷径仍有助于优化。

相比之下， S-HIN 对这种消融几乎不敏感，在大多数基准上的性能差异均低于 0.6%。$S(H)$ 的层次化面格结构已为相同信息提供了直接通路，因此在实际应用中，额外的消息是冗余的。这证实，理论表达能力由函子的选择决定，而实际学习效率则取决于提升偏序集以多明确的方式揭示与任务相关的结构。

7. 结论

在这项工作中，我们通过引入 F-CatWL 框架，统一了高阶表示学习，并确立分次偏序集是实现严谨函子性提升的通用域。将其应用于超图后，我们推导出 F-HIN，其中消息传递拓扑完全由所选函子决定。我们从理论上证明，关联（I）和对称单纯复形（S）两种具体实现的表达能力都严格超越标准

HWL 检验。在实证研究中，我们的模型始终优于基线模型，其中 I-HIN 在稠密社交网络上表现出色，而 S-HIN 则擅长处理细粒度结构。这些结果表明，范畴论不仅是一种抽象理论，更是设计高表达能力、拓扑感知神经网络架构的实用工程蓝图。

致谢

我们感谢许多人士和资助机构。

参考文献

Vikraman Arvind、Frank Fuhlbr¨uck、Johannes K¨obler 和 Oleg Verbitsky。关于 Weisfeiler–Leman 不变性：子图计数及相关图性质。《计算机与系统科学杂志》，113：42–59，2020。

Song Bai、Feihu Zhang 和 Philip HS Torr。超图卷积与超图注意力。《模式识别》，110：107637，2021年。

Cristian Bodnar、Fabrizio Frasca、Nina Otter、Yuguang Wang、Pietro Li´o、Guido F Montufar 和 Michael Bronstein。Weisfeiler 和 Lehman 走向细胞化：CW 网络。载于神经信息处理系统进展，第34卷，第2625–2640页，2021a。

Cristian Bodnar、Fabrizio Frasca、Yuguang Wang、Nina Otter、Guido F Montufar、Pietro Li´o 和 Michael Bronstein。Weisfeiler 与 Lehman 走向拓扑：消息传递单纯形网络。载于国际机器学习会议，第 1026–1037 页。PMLR，2021b。

Jan B¨oker。颜色细化、同态与超图。载于计算机科学图论概念国际研讨会，第 338–350 页。施普林格，2019年。

Alain Bretto。超图理论。 导论。数学工程。*Cham*：施普林格，1:209–216，2013。

Eric Bunch、Qian You、Glenn Fung 和 Vikas Singh。单纯形二复形卷积神经网络。载于拓扑数据分析 {\与} 超越。

Zhengdao Chen、Lei Chen、Soledad Villar 和 Joan Bruna。图神经网络能计数子结构吗？神经信息处理系统进展，33:10383–10395，2020。

Eli Chien、Chao Pan、Jianhao Peng 和 Olgica Milenkovic。You are AllSet：一种用于超图神经网络的多重集函数框架。载于国际学习表征会议，2022年。

Seongjin Choi、Gahee Kim 和 Yong-Geun Oh。超图神经层扩散：一种用于高阶学习的对称单纯集合框架。*IEEE Access*，2025年。

Yihe Dong、Will Sawin 和 Yoshua Bengio。HNHN：采用超边神经元的超图网络。*arXiv* 预印本 *arXiv:2006.12278*，2020。

Stefania Ebli、Micha¨el Defferrard 和 Gard Spreemann。单纯形神经网络。载于 $TDA \ \{\backslash\&\}Beyond$。Yifan Feng、Haoxuan You、Zizhao Zhang、Rongrong Ji 和 Yue Gao。超图神经网络。载于 *AAAI*人工智能大会论文集，第 33 卷，第 3558–3565 页，2019年。Yifan Feng、Jiashu Han、Shihui Ying 和 Yue Gao。超图同构计算。*IEEE*模式分析与机器智能汇刊，46(5)：3880–3896，2024年。Lorenzo Giusti、Teodora Reu、Francesco Ceccarelli、Cristian Bodnar 和 Pietro Li`o。用于高阶和远程相互作用的拓扑消息传递。载于*2024*年国际神经网络联合会议 *(IJCNN)*，第 1–8 页。IEEE，2024年。Mustafa Hajij、Ghada Zamzmi、Theodore Papamarkou、Vasileios Maroulas 和 Xuanting Cai。单纯复形表示学习。载于图上的机器学习（*MLoG*）研讨会，第 *15* 届 *ACM* 国际 *WSDM*（*2022*）会议，2022年。

Jing Huang 和 Jie Yang。Unignn：图与超图神经网络的统一框架。载于第三十届国际人工智能联合会议论文集，第 2563–2569 页。国际人工智能联合会议组织，2021年。

Zhezheng Luo、Jiayuan Mao、Joshua B Tenenbaum 和 Leslie Pack Kaelbling。论超图神经网络的表达能力与泛化。载于首届图学习会议，2022年。

Saunders Mac Lane。《工作数学家的范畴论》，第5卷。施普林格科学与商业媒体，1998年。Christopher Morris、Martin Ritzert、Matthias Fey、William L Hamilton、Jan Eric Lenssen、Gaurav Rattan 和 Martin Grohe。Weisfeiler and Leman 走向神经网络：高阶图神经网络。载于*AAAI*人工智能大会论文集，第33卷，第4602–4609页，2019年。Christopher Morris、Yaron Lipman、Haggai Maron、Bastian Rieck、Nils M Kriege、Martin Grohe、Matthias Fey 和 Karsten Borgwardt。Weisfeiler and Leman 走向机器学习：迄今为止的历程。《机器学习研究杂志》，24(333):1–59，2023年。T Mitchell Roddenberry、Nicholas Glaze 和 Santiago Segarra。用于轨迹预测的基于原理的单纯形神经网络。载于国际机器学习会议，第9020–9029页。PMLR，2021年。Richard P. Stanley。《枚举组合学》。剑桥高等数学研究丛书。剑桥大学出版社，第2版，2011年。Peihao Wang、Shenghao Yang、Yunyu Liu、Zhangyang Wang 和 Pan Li。等变超图扩散神经算子。《*arXiv*预印本 *arXiv:2207.06680*》，2022年。

Boris Weisfeiler 和 Andrei Leman。将图化为标准形式及其中出现的代数。《*nti，系列*》，2(9):12–16，1968年。Hanrui Wu、Andy Yip、Jinyi Long、Jia Zhang 和 Michael K Ng。单纯复形神经网络。《*IEEE*模式分析与机器智能汇刊》，46(1):561–575，2023年。Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。图神经网络的能力有多强？载于国际学习表征会议，2019年。Naganand Yadati、Madhav Nimishakavi、Prateek Yadav、Vikram Nitin、Anand Louis 和 Partha Talukdar。HyperGCN：一种在超图上训练图卷积网络的新方法。《神经信息处理系统进展》，第32卷，2019年。Detian Zhang、Chengqiang Zhang、Yanghui Rao、Li Qing 和 Chunjiang Zhu。通过高维广义 Weisfeiler-Leman 算法提升超图神经网络的表达能力。载于第四十二届国际机器学习会议，2025年。

附录 A. 提升函子的形式化定义

在本节中，我们将形式化这些范畴结构，并验证提升构造的函子性。我们将超图范畴记为 **Hyp** ，将分次偏序集范畴记为**Poset**。

定义 *12*（范畴 *Hyp* 和 *Poset*）*Hyp* 中的一个对象是三元组 $H = (V, E, f)$，其中 $f: E \to 2^V \setminus \emptyset$。态射 $\phi = (a, b): H \to H'$ 由映射 $a: V \to V', b: E \to E'$ 组成，并满足 $a_* \circ f = f' \circ b$，其中 a_* 是诱导的幂集映射。*Poset* 中的一个对象是分次偏序集 $(P, \leq, dim)$。态射 $g: P \to P'$ 是保序映射，且满足 $dim'(g(p)) \leq dim(p)$。

定理 13 对称单纯复形构造 S : **Hyp** $\to$ **Poset** 是一个函子。

证明 给定一个超图态射 $\phi = (a, b): H \to H'$，我们在单纯形上定义映射$S(\phi): S(H) \to S(H')$，其定义为 $S(\phi)(\{v_{i_0}, \ldots, v_{i_n}\}_e) := \{a(v_{i_0}) \ldots, a(v_{i_n})\}_{b(e)}$。该映射是 **偏序集** 中的合法态射，因为:

1. 若 $\sigma_e \subset \tau_e$（即 $\{v_{i_k}\} \subset \{v_{j_k}\}$），则根据集合映射 a 的定义，$\{a(v_{i_k})\} \subset \{a(v_{j_k})\}$，这意味着在 $b(e)$ 中有 $S(\phi)(\sigma_e) \subset S(\phi)(\tau_e)$。
2. 由于 $|a(\sigma)| \leq |\sigma|$，我们有 $\dim(S(\phi)(\sigma_e)) = |a(\sigma)| - 1 \leq |\sigma| - 1 = \dim(\sigma_e)$。函子性由 a和 b的逐分量复合得出。 ■

注记 14（标准几何偏序集）我们的框架自然涵盖标准几何结构。关联偏序集 $I(H)$ 是一个有效的分次偏序集，其中节点被赋予维数 *0*，超边被赋予维数 *1*。类似地，任意单纯复形 X 在包含关系下构成一个分次偏序集，且 $\dim(\sigma) = |\sigma| - 1$，而正则 *CW* 复形的面偏序集按胞腔维数分次。

附录 B. 邻接结构的等价性

在本节中，我们证明完整的 4-邻接更新规则与仅使用这些邻接关系的一个子集（通常为边界与向上，或类似关系）的简化更新规则具有相同的表达能力。该结果将拓扑和胞腔 WL 检验中的类似发现（Bodnar 等人，2021年b，a）推广到我们的范畴框架，为我们实验中所使用的高效架构提供了理论依据。

定义 15（同构测试与表达能力）设 $\mathcal{C}$ 为一个范畴。同构测试 $A = \{A_t\}_{t \in \mathbb{N}}$ 是 $\mathcal{C}$ 上的等价关系序列，满足：*(1)* 对所有 t有$X \cong X' \implies X \cong_{A_t} X'$，以及 *(2)* $X \cong_{A_{t+1}} X' \implies X \cong_{A_t} X'$。若对某个 t有 $X \ncong_{A_t} X'$，则两个对象可由 A 区分。

给定两个测试 A、A'，如果 $X \cong_A X' \implies X \cong_{A'} X'$，我们称 A 至少与 A' $(A \sqsubseteq A')$ 一样强大。若存在对象 X、X'，它们可由 A 区分，但不能由 A' 区分（即 $X \ncong_A X'$，但 $X \cong_{A'} X'$），则称 A 不弱于 A'。如果 $A \sqsubseteq A'$ 且 $A' \sqsubseteq A$，则称它们具有相同的表达能力。如果 $A \sqsubseteq A'$，但存在可由 A 区分而不能由 A'区分的对象，则称 A 严格更强大。

定义16（着色更新规则）着色 c 在 $\mathcal{C}$ 上为每个对象 X赋予一个映射 $c^X: X \to S_c$。如果 c^X、$c^{X'}$的颜色多重集（直方图）相同，我们记为 $X \cong_c X'$。着色更新规则是一个序列 $\mathbb{c} := \{\mathbb{c}_t\}_t$，其中（*1*）$\mathbb{c}_0$ 为常数，（*2*）$X \cong_{\mathbb{c}_{t+1}} X' \implies X \cong_{\mathbb{c}_t} X'$，并且（*3*）对所有 t均有 $X \cong X' \implies X \cong_{\mathbb{c}_t} X'$。

根据同构测试的定义，我们可以轻松证明以下引理。

引理 17 范畴 $\mathcal{C}$ 上的着色更新规则$\mathbb{c}$ 诱导出同构测试：当且仅当$\mathbb{c}_t^X$和$\mathbb{c}_t^{X'}$ 的两个着色直方图相等时，$X \cong_{\mathbb{c}} X'$。

因此，根据引理17，我们滥用着色更新规则及其对应的同构测试的记号。

定义 18（着色细化）如果对于所有输入，着色 c 细化 d（记作 $c \sqsubseteq d$）：$c^X(\sigma) = c^{X'}(\sigma') \implies d^X(\sigma) = d^{X'}(\sigma')$。如果两个着色等价（$c \equiv d$），则 $c \sqsubseteq d$ 且 $d \sqsubseteq c$。

引理 19（重参数化不变性）设$\mathbb{c} = \{\mathbb{c}_t\}_t$ 为着色更新规则，且 $\varphi : \mathbb{N} \to \mathbb{N}$ 为严格递增函数，并满足 $\varphi(t) \geq t + 1$。重参数化序列 $\varphi(\mathbb{c}) := \{\mathbb{c}_{\varphi(t)}\}_t$ 构成具有相同表达能力的有效同构测试：$X \cong_{\mathbb{c}} X' \iff X \cong_{\varphi(\mathbb{c})} X'$。

证明 有效性由 c 的性质得出。根据定义，条件 (1) $X \cong X' \implies X \cong_{\mathbb{c}_{\varphi(t)}} X'$ 成立。条件 ($X \cong_{\mathbb{c}_{\varphi(t+1)}} X' \implies X \cong_{\mathbb{c}_{\varphi(t)}}$) 由 X' $\varphi\ t+1\ > \varphi\ t$ 单调性得出：因为 ( ) ( )，且着色会随时间细化 (c c)。等价性$X \cong_{\mathbb{c}} X' \iff X \cong_{\varphi(\mathbb{c})} X'$ 显而易见：正向（$\implies$）显然成立，因为 $\{\mathsf{c}_{\varphi(t)}\}$是 $\{\mathsf{c}_t\}$的子序列。反向（$\impliedby$）成立，因为对于任意 k，都存在 t 使得 $\varphi(t) \geq k$，这意味着 $X \cong_{\mathbb{c}_{\varphi(t)}} X' \implies X \cong_{\mathbb{c}_k} X'$。 ■

引理20 设 X、X'为 $\mathcal{C}$中的对象，且满足 $A \subset X$、$A' \subset X'$。对于$\mathcal{C}$上的两个着色 c、d，若 $c \sqsubseteq d$，且 $\{\!\{d^X(\sigma) \mid \sigma \in A\}\!\} \neq \{\!\{d^{X'}(\sigma') \mid \sigma' \in A'\}\!\}$，则 $\{\!\{c^X(\sigma) \mid \sigma \in A\}\!\} \neq \{\!\{c^{X'}(\sigma') \mid \sigma' \in A'\}\!\}$。

证明 假设 $\{\!\{d^X(\sigma) \mid \sigma \in A\}\!\} \neq \{\!\{d^{X'}(\sigma') \mid \sigma' \in A'\}\!\}$。不失一般性，存在一种颜色 C，在第一个多重集中出现得更频繁。令 $A^* := (d^X)^{-1}(\mathsf{C}) \cap A$且 $(A')^* := (d^{X'})^{-1}(\mathsf{C}) \cap A'$。由于 $c \sqsubseteq d$，映射到 C 的 c-值集合与映射到任何其他 d-颜色的集合互不相交。因此：

$$\begin{aligned} &\{\!\{c^X(\sigma) \mid \sigma \in A^*\}\!\} \cap \{\!\{c^X(\sigma) \mid \sigma \in A \setminus A^*\}\!\} = \emptyset, \\ &\{\!\{c^{X'}(\sigma') \mid \sigma' \in (A')^*\}\!\} \cap \{\!\{c^{X'}(\sigma') \mid \sigma' \in A' \setminus (A')^*\}\!\} = \emptyset. \end{aligned} \tag{6}$$

现在，反设 $\{\!\{c^X(\sigma) \mid \sigma \in A\}\!\} = \{\!\{c^{X'}(\sigma') \mid \sigma' \in A'\}\!\}$。等式 (6) 中的不相交性质意味着，对应于$\mathbb{C}$ 的子多重集必须相同：

$$\{\!\{c^X(\sigma) \mid \sigma \in A^*\}\!\} = \{\!\{c^{X'}(\sigma') \mid \sigma' \in (A')^*\}\!\}. \tag{7}$$

由于 $c \sqsubseteq d$， A^* 和 $(A')^*$ 中的所有元素都映射到同一个 d-颜色 C。因此， c-多重集相等意味着 d-多重集相等（它们只是由 C 构成、大小分别为 $|A^*|$ 和 $|(A')^*|$的常数多重集）：

$$\{\!\{d^X(\sigma) \mid \sigma \in A^*\}\!\} = \{\!\{d^{X'}(\sigma') \mid \sigma' \in (A')^*\}\!\}. \tag{8}$$

这意味着 $|A^*| = |(A')^*|$，与 C 在第一个多重集中出现次数更多这一假设矛盾。因此， c-多重集必不相同。 ■

推论21 对于 $\mathcal{C}$ 上满足 $c \sqsubseteq d$ 的着色 c, d，若 $X \ncong_d X'$，则 $X \ncong_c X'$。

证明 根据定义 16， $X \ncong_d X'$ 意味着多重集 $\{\!\{d^X(\sigma) \mid \sigma \in X\}\!\}$ 和 $\{\!\{d^{X'}(\sigma') \mid \sigma' \in X'\}\!\}$ 彼此不同。应用引理 20，取 $A = X$ 和 $A' = X'$，可知相应的 c-多重集也彼此不同，这意味着 $X \ncong_c X'$。

■

哈希函数的单射性意味着一个基本不变量：不同的边界大小会产生不同的颜色。

引理 22 设 $\mathbb{c}_t^{F\,X}$,表示迭代 t 时的 F-*CatWL* 着色。对于任意 $\sigma \in F(X)$ 和$\sigma' \in F(X')$，若 $|\mathcal{B}(\sigma)| \neq |\mathcal{B}(\sigma')|$，则对于所有 $t > 0$，$\mathbb{c}_t^{F\,X}(\sigma) \neq \mathbb{c}_t^{F\,X'}(\sigma')$。

现在我们证明，在关于 F 的温和条件下，完整的 4-邻接更新规则等价于一种简化的更新规则，该规则仅使用边界、上邻接和下邻接（省略余边界）。

引理23 设 $F\colon\ \mathcal{C} \to \mathcal{P}$ 是一个函子，使得对于任意 $\sigma \in F(X)$，都有 $|\mathcal{B}(\sigma)| > 1$，其中$\dim(\sigma) > 0$。使用完整 *4*-邻接更新规则的 F-*CatWL*，与仅使用边界、下邻接和上邻接（省略余边界）的受限测试具有相同的表达能力。

证明 令 $\mathbb{c}_t^F$和 $\mathbb{b}_t^F$分别表示使用 4-邻接和 3-邻接的着色更新规则。显然，$\mathbb{c}_t^F \sqsubseteq \mathbb{b}_t^F$，因为 $\mathbb{c}$ 利用了严格更多的信息（余边界）。只需证明 $\mathbb{b}_{t+1}^F \sqsubseteq \mathbb{c}_t^F$，对此按 t 归纳法（应用引理 19）。

基本情况（$t = 0$）： 显然，因为 a_0 是常数。

归纳步骤： 假设 $\mathsf{b}_{k+1}^F \sqsubseteq \mathsf{a}_k^F$对于 $k = 0, \ldots, t$成立。设 $\mathsf{b}_{t+2}^{F\,X}(\sigma) = \mathsf{b}_{t+2}^{F\,X'}(\sigma')$。根据更新函数的单射性，在时间$t+1$， σ、 $\mathcal{B}(\sigma)$、 $\mathcal{N}_\downarrow(\sigma)$ 和 $\mathcal{N}_\uparrow(\sigma)$ 的颜色必须与 σ' 的颜色相同。我们证明这意味着余边界颜色也相同：$\mathsf{b}_{t+1}^{F\,X}(\mathcal{C}(\sigma)) = \mathsf{b}_{t+1}^{F\,X'}(\mathcal{C}(\sigma'))$。

上邻接多重集 $\mathsf{b}_{t+1}^{F\,X}(\mathcal{N}_\uparrow(\sigma)) = \mathsf{b}_{t+1}^{F\,X'}(\mathcal{N}_\uparrow(\sigma'))$ 相等，这意味着元组集合 $\{(b^{F,X}_{t+1}(\delta)\ b^{F,X}_{t+1}(\sigma)) \mid \delta \in \mathcal{C}(\sigma)\}$ 相等。利用引理 22，我们可以根据边界 $|\mathcal{B}(\delta)| = n$ 的大小对这些多重集进行划分。对于固定的 n，颜色 $b^{F,X}_{t+1}(\delta)$ 在 σ 的上邻接多重集中恰好出现 $n-1$ 次（因为 δ 有 n 个边界元素，其中一个是 σ）。由于 $|\mathcal{B}(\delta)| > 1$，我们有 $n \geq 2$，因此 $n - 1 \geq 1$。这样，将重数除以 $n-1$，即可唯一地恢复共边界邻居的多重集。这意味着：

$$\{\!\{b^{F,X}_{t+1}(\delta) \mid \delta \in \mathcal{C}(\sigma)\}\!\} = \{\!\{b^{F,X'}_{t+1}(\delta') \mid \delta' \in \mathcal{C}(\sigma')\}\!\}.$$

因此，四个邻接输入均相同。根据归纳假设，a_t^F 被 b_{t+1}^F 细化，因此在 t 处进行 a-更新时，σ 和 σ' 的输入相同。因此，$\mathsf{a}_{t+1}^{F\,X}{}_{,}(\sigma) = \mathsf{a}_{t+1}^{F\,X'}{}_{,}(\sigma')$。由此可得 $\mathsf{b}_{t+2}^F \sqsubseteq \mathsf{a}_{t+1}^F$。

■

利用引理23，我们可以证明 F-CatWL 使用边界邻接与上邻接时，其表达能力与 F-CatWL 使用 4-邻接时相同。

引理 24 设 $F\colon \mathcal{C} \to \mathcal{P}$ 是一个函子，使得对任意 $\sigma \in F(X)$，都有 $|\mathcal{B}(\sigma)| > 1$，且 $\dim(\sigma) > 0$。使用全部四种邻接（边界、余边界、下邻接、上邻接）的 *F-CatWL* 检验，与仅使用边界邻接与上邻接（省略余边界和下邻接）的受限测试具有相同的表达能力。

证明 令 b_t^F 表示受限的着色更新规则（使用边界邻接与上邻接），令 a_t^F 表示使用全部四种邻接的规则。显然，$\mathsf{a}_t^F \sqsubseteq \mathsf{b}_t^F$，因为 a 使用了严格更多的信息（余边界和下邻接）。只需通过对 t 进行归纳证明 $\mathsf{b}_{2t+1}^F \sqsubseteq \mathsf{a}_t^F$。

基本情况（$t = 0$）显然成立，因为 a_0^F 是常数。对于归纳步骤，假设 $\mathsf{b}_{2k+1}^F \sqsubseteq \mathsf{a}_k^F$ 对 $k = 0, \ldots, t$ 成立。为导出矛盾，假设 $\mathsf{b}_{2t+3}^{F\,X}{}_{,}(\sigma) = \mathsf{b}_{2t+3}^{F\,X'}{}_{,}(\sigma')$，但下邻接多重集不同：$\mathsf{b}_{2t+1}^{F\,X}{}_{,}(\mathcal{N}_{\downarrow}(\sigma)) \neq \mathsf{b}_{2t+1}^{F\,X'}{}_{,}(\mathcal{N}_{\downarrow}(\sigma'))$。

根据更新函数的单射性，σ、$\mathcal{B}(\sigma)$ 和 $\mathcal{N}_{\uparrow}(\sigma)$ 在时刻 $2t+1$ 的颜色必须相同。由于下邻接多重集不同，因此存在某个颜色对 $(\mathsf{C}_0, \mathsf{C}_1)$，它在 σ 和 σ' 的下邻接中出现的重数不同。对于任意 δ，我们将辅助多重集 $A_X(\delta)$ 定义如下：

$$A_X(\delta) := \{\!\{(\mathbb{b}_{2t+1}^{I,X}(\psi), \mathbb{b}_{2t+1}^{I,X}(\delta)) \mid \psi \in \mathcal{C}(\delta)\}\!\}$$

并在 γ的边界上对这些值进行聚合：

$$C_X(\gamma) := \{\!\{|A_X(\delta)| \mid \delta \in \mathcal{B}(\gamma)\}\!\}.$$

$C_X(\sigma)$ 中元素的总和对应于 $\mathsf{b}_{2t+1}(\mathcal{N}_{\downarrow}(\sigma))$ 中元组的总数。因此，下邻接多重集中的差异意味着 $C_X(\sigma) \neq C_{X'}(\sigma')$，这与假设的第 $2t+3$ 步 σ 和 σ' 颜色相同相矛盾，因为这些信息会通过边界聚合传播。

命题 25 若 $C_X(\sigma) \neq C_{X'}(\sigma')$，则 $\mathbb{b}_{2t+3}^{F\,X}{}_{,}(\sigma) \neq \mathbb{b}_{2t+3}^{F\,X'}{}_{,}(\sigma')$。

证明 定义一种新的着色 $c_X(\delta) := |A_X(\delta)|$。我们先证明 $\mathsf{b}_{2t+2}^F \sqsubseteq c$。设$\delta \in F(X)$ 和 $\delta' \in F(X')$ 满足 $|A_X(\delta)| > |A_{X'}(\delta')|$。若 $\mathsf{b}_{2t+1}^{F\,X'}{}_{,}(\delta') \neq \mathsf{C}_1$，则由于$|A_X(\delta)| > 0$，我们有 $\mathsf{b}_{2t+1}^{F\,X}{}_{,}(\delta) = \mathsf{C}_1$，因此它们在第 $2t+1$ 步的颜色不同（从而在 $2t+2$ 时也不同）。若 $\mathsf{b}_{2t+1}^{F\,X'}{}_{,}(\delta') = \mathsf{C}_1$，则 $\mathsf{b}_{2t+1}^{F\,X}{}_{,}(\delta) = \mathsf{C}_1$ 也成立。根据引理 22 以及关于 F 的假设，$|\mathcal{B}(\delta)| = |\mathcal{B}(\delta')| = n > 1$。颜色 C_0 在 δ 的上邻接多重集中出现 $|A_X(\delta)| \times (n-1)$ 次，而在 δ' 的上邻接多重集中的 $|A_{X'}(\delta')| \times (n-1)$ 次。由于 $n > 1$ 且计数不同，上邻接多重集也不同，这意味着 $\mathsf{b}_{2t+2}^{F\,X}{}_{,}(\delta) \neq \mathsf{b}_{2t+2}^{F\,X'}{}_{,}(\delta')$。

建立了 $\mathsf{b}_{2t+2}^F \sqsubseteq c$ 后，我们将细化性质应用于不等式$C_X(\sigma) \neq C_{X'}(\sigma')$。边界上的 c-多重集不同，这意味着

边界上的 b^F_{2t+2}-多重集不同。因此，更新规则在 σ 和 σ'处的输入不同，从而得到 $\mathrm{b}^{F,X}_{2t+3,}(\sigma)$ $\neq \mathrm{b}^{F,X'}_{2t+3,}(\sigma')$。 ■

命题 25 表明，$C_X(\sigma) \neq C_{X'}(\sigma')$ 意味着 $\mathrm{b}^{F,X}_{2t+3,}(\sigma) \neq \mathrm{b}^{F,X'}_{2t+3,}(\sigma')$，这与我们的初始假设矛盾。因此，下邻接多重集必须相等：$\mathrm{b}^{F,X}_{2t+1,}(\mathcal{N}_\downarrow(\sigma)) = \mathrm{b}^{F,X'}_{2t+1,}(\mathcal{N}_\downarrow(\sigma'))$。既然我们现已证明，对于四种邻接，$\sigma$ 和 σ' 具有相同的 b^F_{2t+1}-颜色多重集，并且 $\mathrm{b}^F_{2t+1} \sqsubseteq \mathrm{a}^F_t$（根据归纳假设），便可推出它们的 a^F_t-颜色多重集也相同。因此，$\mathrm{a}^{F,X}_{t+1,}(\sigma) = \mathrm{a}^{F,X'}_{t+1,}(\sigma')$，由此完成了 $\mathrm{b}^F_{2t+3} \sqsubseteq \mathrm{a}^F_{t+1}$ 的归纳步骤。 ■

附录 C. F-HIN 架构

给定一个函子 $F\colon \mathbf{Hyp} \to$ **偏序集**，若 F-CatMPN 的消息传递函数具有以下特定形式，则称其为 F-HIN：

$$
\begin{aligned}
\mathbb{m}^{F,X}_t(\mathcal{B}(\sigma)) &= \mathrm{MLP}^F_{t,\mathcal{B}}\left((1+\epsilon_{\mathcal{B}})\mathbb{x}^{F,X}_t(\sigma) + \sum_{\tau\in\mathcal{B}(\sigma)} \mathbb{x}^{F,X}_t(\tau)\right)\\
\mathbb{m}^{F,X}_t(\mathcal{C}(\sigma)) &= \mathrm{MLP}^F_{t,\mathcal{C}}\left((1+\epsilon_{\mathcal{C}})\mathbb{x}^{F,X}_t(\sigma) + \sum_{\tau\in\mathcal{C}(\sigma)} \mathbb{x}^{F,X}_t(\tau)\right)\\
\mathbb{m}^{F,X}_t(\mathcal{N}_\downarrow(\sigma)) &= \mathrm{MLP}^F_{t,\downarrow}\left((1+\epsilon_{\downarrow})\mathbb{x}^{F,X}_t(\sigma) + \sum_{(\tau,\sigma')\in\mathcal{N}_\downarrow(\sigma)} \mathrm{MLP}_{t,D}(\mathbb{x}^{F,X}_t(\sigma') \,\|\, \mathbb{x}^{F,X}_t(\tau))\right)\\
\mathbb{m}^{F,X}_t(\mathcal{N}_\uparrow(\sigma)) &= \mathrm{MLP}^F_{t,\uparrow}\left((1+\epsilon_{\uparrow})\mathbb{x}^{F,X}_t(\sigma) + \sum_{(\tau,\sigma')\in\mathcal{N}_\uparrow(\sigma)} \mathrm{MLP}_{t,U}(\mathbb{x}^{F,X}_t(\sigma') \,\|\, \mathbb{x}^{F,X}_t(\tau))\right)
\end{aligned}
\tag{9}
$$

这里，$\mathrm{MLP}^F_{t,\bullet}$ 表示一个 2 层感知机，$\mathrm{MLP}_{t,\bullet}$ 表示一个 1 层感知机。可学习参数包括这些多层感知机的权重以及标量 $\epsilon_{\mathcal{B}}, \epsilon_{\mathcal{C}}, \epsilon_{\downarrow}, \epsilon_{\uparrow}$。

附录 D. HWL、 I-CatWL 和 S-CatWL 的比较

本节将证明，I-CatWL 和 S-CatWL 的能力不弱于 HWL。

定理 26 *I-CatWL 的能力不弱于 HWL。*

证明 考虑图 4 中的超图 H 和 H'，其初始着色为常数 $*$。正如 Zhang 等人（2025年）所示，H 和 H' 在 HWL 检验下同构。为证明它们在 I-CatWL 下不同构，只需证明颜色 $^{I\,H}_{1\,}(e_2)$ 和 $^{I\,H'}_{1\,}(e'_2)$ 不同，其中 e_2 和 e'_2 分别是 H 和 H' 中唯一的基数为 2 的超边。

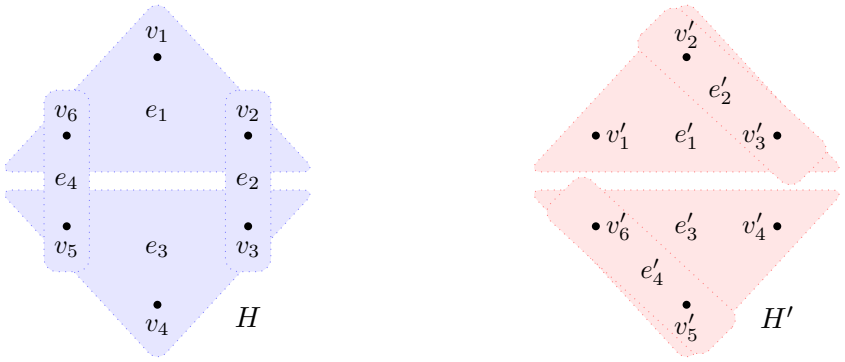


图 4：两个不同构的超图 H 和 H' 无法被 HWL 检验区分，但可被 I-CatWL 检验和 S-CatWL 检验区分。

我们首先考察在第 1 次迭代时，从 H 中 e_2 的 4-邻接聚合得到的颜色。边界颜色为 ${}_0^{I\,H}(\mathcal{B}(e_2)) = \{\!\{{}_0^{I\,H}(v_2), {}_0^{I\,H}(v_3)\}\!\} = \{\!\{*,*\}\!\}$，下邻接颜色为 ${}_0^{I\,H}(\mathcal{N}_\downarrow(e_2)) = \{\!\{{}_0^{I\,H}(e_1), {}_0^{I\,H}(e_3)\}\!\} = \{\!\{*,*\}\!\}$。余边界 $\mathcal{C}(e_2)$ 集合和上邻接 $\mathcal{N}_\uparrow(e_2)$ 集合均为空，因此颜色多重集为空。

相比之下，对于 H'中的超边 e_2'，尽管边界颜色仍为 $\mathsf{c}_0^{I\,H'}.(\mathcal{B}(e_2')) = \{\!\{*,*\}\!\}$，但下邻接有所不同。具体而言，$\mathsf{c}_0^{I\,H'}.(\mathcal{N}_\downarrow(e_2')) = \{\!\{\mathsf{c}_0^{I\,H'}.(e_1')\}\!\} = \{\!\{*\}\!\}$，仅包含一个邻居。余边界集和上邻接集仍为空。

因此，应用更新规则可得不同的哈希值：

$$\mathfrak{c}_1^{I,H}(e_2) = \mathrm{HASH}\left(*, \{\!\{*,*\}\!\}, \emptyset, \{\!\{*,*\}\!\}, \emptyset\right) \neq \mathrm{HASH}\left(*, \{\!\{*,*\}\!\}, \emptyset, \{\!\{*\}\!\}, \emptyset\right) = \mathfrak{c}_1^{I,H'}(e_2'). \quad (10)$$

根据 HASH 的单射性，得到的颜色多重集各不相同，这意味着 I-CatWL 能够成功区分 H 和 H'。这证实了 I-CatWL 的能力不弱于 HWL。

■

定理 27 *S-CatWL 的能力不弱于 HWL。*

证明 我们采用定理 26 中的相同反例，以证明超图 H 和 H' 在 S-CatWL下不同构。只需证明着色 $\mathsf{c}_1^{S\,H}(\{v_2,v_3\}_{e_2})$ 和 $\mathsf{c}_1^{S\,H'}(\{v_2',v_3'\}_{e_2'})$ 不同，因为它们分别对应于各自结构中唯一的极大1维单纯形。

首先，我们计算 $\{v_2,v_3\}_{e_2}$ 在 H 中第1次迭代时来自4-邻接的聚合颜色。边界颜色为 $\mathsf{c}_0^{S\,H}(\mathcal{B}(\{v_2,v_3\}_{e_2})) = \{\!\{\mathsf{c}_0^{S\,H}(\{v_2\}_{v_2}), \mathsf{c}_0^{S\,H}(\{v_3\}_{v_3})\}\!\} = \{\!\{*,*\}\!\}$。下邻接颜色为

$$\begin{aligned}\mathfrak{c}_0^{S,H}(\mathcal{N}_\downarrow(\{v_2,v_3\}_{e_2})) &= \{\!\{\mathfrak{c}_0^{S,H}(\{v_1,v_2\}_{e_1}), \mathfrak{c}_0^{S,H}(\{v_2,v_6\}_{e_1}), \mathfrak{c}_0^{S,H}(\{v_3,v_4\}_{e_3}), \mathfrak{c}_0^{S,H}(\{v_3,v_5\}_{e_3})\}\!\} \\ &= \{\!\{*,*,*,*\}\!\}.\end{aligned} \quad (11)$$

余边界和上邻接颜色均为空。

相反，对于 $\{v_2', v_3'\}_{e_2'}$ in H'，边界颜色相同：$\mathrm{c}_0^{S,H'}(\mathcal{B}(\{v_2', v_3'\}_{e_2'})) = \{\!\{*, *\}\!\}$。然而，下邻接颜色仅包含三个元素：

$$\begin{aligned} \mathrm{c}_0^{S,H'}(\mathcal{N}_\downarrow(\{v_2', v_3'\}_{e_2'})) &= \{\!\{\mathrm{c}_0^{S,H'}(\{v_1', v_2'\}_{e_1'}), \mathrm{c}_0^{S,H'}(\{v_1', v_3'\}_{e_1'}), \mathrm{c}_0^{S,H'}(\{v_2', v_3'\}_{e_1'})\}\!\} \\ &= \{\!\{*, *, *\}\!\}. \end{aligned} \tag{12}$$

余边界和上邻接的颜色仍为空。

因此，更新规则会产生不同的哈希值。具体而言，注意到下邻接输入不同（4 个元素与 3 个元素），我们有：

$$\mathrm{HASH}(\dots, \{\!\{*, *, *, *\}\!\}, \dots) \neq \mathrm{HASH}(\dots, \{\!\{*, *, *\}\!\}, \dots), \tag{13}$$

这意味着 $\mathrm{c}_1^{S,H'}(\{v_2, v_3\}_{e_2}) \neq \mathrm{c}_1^{S,H'}(\{v_2', v_3'\}_{e_2'})$。根据 HASH 的单射性，这证实了 S-CatWL 能够成功区分 H 和 H'，而 HWL 则不能。 ■

附录 E。 F-CatWL 与 F-CatMPN 的等价性

在本节中，我们比较 F-CatWL 和 F-CatMPN 的表达能力 .

定理 28 设 $F: \mathcal{C} \to$ **偏序集** 是一个函子。 *F-CatWL 至少与 F-CatMPN 一样强。如果更新函数和聚合函数均为单射，且网络具有足够的深度和宽度，则 F-CatMPN 与 F-CatWL 具有相同的表达能力。*

证明 令 $\mathrm{c}_k^{F,X}$,表示第 k 次迭代时的 F-CatWL 着色，令 $h^{F,X}_{,k}$ 表示第 k 层的F-CatMPN 节点特征。我们假设 MPN 会稳定下来，或者具有足够多的层，使得对于所有 $k > L$， $h^{F,X}_{,k} = h^{F,L,X}$ 成立。我们通过对 k 使用归纳法证明等价性，即证明 $\mathrm{c}_k^{F,X} \sqsubseteq h^{F,X}_{,k}$ 以及 $h^{F,X}_{,k} \sqsubseteq \mathrm{c}_k^{F,X}$。

- $\mathrm{c}_k^{F,X} \sqsubseteq h^{F,X}_{,k}$**：** 基本情况 $k = 0$ 显然成立。假设该假设对 $0, \dots, t$成立。设 $\mathrm{c}_{t+1}^{F,X}(\sigma) = \mathrm{c}_{t+1}^{F,X'}(\sigma')$。HASH 函数的单射性意味着，在第 t 步， σ 的颜色多重集以及所有四种邻接类型（$\mathcal{B}, \mathcal{C}, \mathcal{N}_\downarrow, \mathcal{N}_\uparrow$）的颜色多重集都相同。根据归纳假设（$h_t \sqsubseteq \mathrm{c}_t$），对于 σ, σ' 及其各自的邻域，相应的 MPN 特征 h_t 也必定相同。由于 h_{t+1} 是这些相同输入的函数，因此可得 $h^{F,X}_{,t+1}(\sigma) = h^{F,X'}_{,t+1}(\sigma')$。

- $h^{F,X}_{,k} \sqsubseteq \mathrm{c}_k^{F,X}$**：** 基本情况 $k = 0$ 显然成立。假设该假设对 0，.....， t成立。设 $h^{F,X}_{,t+1}(\sigma) = h^{F,X'}_{,t+1}(\sigma')$。我们假设神经聚合函数和更新函数（聚合函数，更新函数）均具有单射性（根据通用逼近定理，这是可能的）。这种单射性意味着，在层 t， σ， $\mathcal{B}$， $\mathcal{C}$， $\mathcal{N}_\downarrow$， $\mathcal{N}_\uparrow$的输入特征多重集必须相同。根据归纳假设（$\mathrm{c}_t \sqsubseteq h_t$），对应的 WL 颜色 c_t 也必须相同。由于 c_{t+1} 由这些颜色决定，因此有 $\mathrm{c}_{t+1}^{F,X}(\sigma) = \mathrm{c}_{t+1}^{F,X'}(\sigma')$。

■

附录 F. 实验详情

F.1. 数据集详情

F.1.1. Dataset Descriptions

我们在六个源自不同领域的真实世界超图分类基准上评估所提模型，这些领域包括社交媒体、电子商务和协作网络。这些数据集最初由 Feng 等人 (2019年) 提出，涵盖了广泛的拓扑特征，从稀疏协作图到稠密社交自我网络。

IMDB 数据集。 IMDB 数据集由电影协作网络构建而成，其中每个超图代表一部特定电影，节点对应工作人员，包括导演和编剧。结构连通性由共现关系决定，形成两种主要变体：**IMDB-Dir**，其超边连接有过合作的导演；以及**IMDB-Wri**，其构建依据是共同编剧关系。这些结构对应两种不同的分类目标：识别电影的形式，如动画片、纪录片和剧情片，记作**-Form**；或预测其类型，如冒险片、犯罪片和家庭片，记作**-Genre**。因此，我们在四个不同的基准上评估模型：IMDB导演形式、IMDB导演类型、IMDB wri form 和 IMDB wri genre。

Steam-Player. 该数据集源自 Steam 游戏平台，将每位游戏玩家建模为一个超图，其中节点对应其拥有的游戏。通过关联具有相同用户定义标签（如角色扮演游戏和动作游戏）的游戏来构建超边。相关的二元分类任务旨在判断玩家偏好单人游戏还是多人游戏。

Twitter-Friend. Twitter-Friend 数据集侧重于从 Twitter 收集的社交自我网络，其中每个超图代表一位特定用户的网络。在此情境下，节点表示该用户的朋友，超边则连接网络中彼此互动或形成圈子的朋友群体。该任务旨在根据用户的发推行为对其兴趣进行分类，区分发推谈论“全国爱犬日”的用户与发推谈论“Respect Tyler Joseph”的用户。

F.1.2. Statistical Analysis

我们从三个角度对这六个数据集进行了全面的统计分析：(1) 图规模统计，(2) 微观拓扑性质，以及 (3) 类别标签分布。

图级统计（规模与密度）表2报告了每个图中节点数（$|V|$）和超边数（$|E|$）的分布，凸显了各基准数据集之间的结构多样性。IMDb wri数据集由规模相对较小的超图组成，平均超边数少于 5。与之形成鲜明对比的是，Steam 玩家和Twitter 好友呈现出极为密集的高阶交互；例如，Steam 玩家的平均超边节点比（$|E|/|V|$）约为 3.4。与此同时，IMDb dir数据集的图规模差异显著，超边数最高可达 450，对模型跨不同规模泛化的能力构成了挑战。

表2：数据集的图级统计。 N_G 表示超图的总数。我们报告每个图的节点（$|V|$）和超边（$|E|$）数量的平均值和最大值。

数据集	N_G	**每个图的节点数** ($\|V\|$)		**每个图的超边数**($\|E\|$)	
		Mean	Max	Mean	Max
IMDB导演形式	1869	15.72	264	39.30	450
IMDB导演类型	3393	17.33	264	36.55	450
IMDB_wri_form	374	10.05	180	3.76	23
IMDB_wri_genre	1172	12.82	273	4.46	75
Steam 玩家	2048	13.83	76	46.36	140
Twitter 好友	1310	21.65	166	84.77	603

拓扑统计（度与基数）表3展示了节点度和超边大小的分布。在节点度方面，Twitter 好友的平均值最高（19.0），且具有长尾分布（最大值为 199），表明存在显著的枢纽用户。在超边基数方面，我们观察到，尽管Twitter 好友的最大超边大小达到 64，但第 99 百分位数仍为 16。这一分布为采用截断阈值为 20 的对称单纯形提升策略提供了实证依据，因为该策略在保留绝大多数局部高阶结构的同时，能够确保计算效率。

表 3：节点度和超边大小的分布。报告的百分位数为第 99 百分位数。

数据集	**节点度分布**			**超边大小分布**	
	Mean	Max	第 99 百分位	Max	第 99 百分位
IMDB导演形式	8.40	68	42	110	12
IMDB导演类型	7.30	77	39	110	14
IMDB_wri_form	1.70	19	6	179	25
IMDB_wri_genre	1.70	18	7	266	36
Steam 玩家	15.20	20	20	71	15
Twitter 好友	19.00	199	89	64	16

类别标签分布 表 4展示了每项任务的类别平衡情况。 IMDb各变体均存在显著的类别不平衡——例如， `IMDB wrigenre`中一个类别仅占数据的 2.6%，而 Steam 玩家 和 Twitter 好友 则相对平衡。

F.2. 实现细节。

所有模型均使用 PyTorch 实现，并通过 Adam 优化器进行训练。为确保比较公平，我们对所有基线模型和所提模型在以下超参数空间内进行了网格搜索：批次大小 $\{4, 32, 64\}$、隐藏维度 $\{32, 64\}$，

表 4：类别标签分布。 C 表示类别数。

数据集	C	分布（类别索引：比例）
IMDb导演形式	3	0：42.7%，1：36.8%，2：20.5%
IMDB导演类型	3	0：32.6%，1：16.1%，2：51.3%
IMDB_wri_form	4	0：45.7%，1：20.9%，2：27.3%，3：6.2%
IMDB_wri_genre	6	0：24.3%，1：21.9%，2：26.2%，3：18.3%，4：2.6%，5：6.7%
Steam 玩家	2	0：50.0%，1：50.0%
Twitter 好友	2	0：59.2%，1：40.8%

层数 $\{2, 3\}$ 和初始学习率 $\{10^{-2}, 5 \times 10^{-3}, 10^{-3}\}$。所有实验中的 dropout 率均固定为 0.5。

在训练计划中，我们采用了 ReduceLROnPlateau，这是一种在验证性能趋于饱和时动态降低学习率的机制，从而促进模型精细收敛。具体而言，如果验证损失连续 20 个训练轮次未见改善，学习率就会降低至原来的 0.5 倍，且最低阈值为1×10^{-5}。

对于对称单纯形模型（S-HIN），我们在提升过程中采用了截断策略：基数超过 20 的超边不会被提升为高阶单形，以降低计算复杂度。这一设计选择得到了表 3 中拓扑分析结果的实证支持。对于 Steam 玩家 和 Twitter 好友 等稠密社交和电子商务基准数据集，超边大小的第 99 百分位数最多为 16。因此，阈值设为 20 的截断策略能够保留这些具有挑战性的数据集中的绝大多数高阶相互作用，同时仅丢弃 IMDB 变体中的少量超大超边。

所有实验均在一台服务器上进行，该服务器配备 Intel Xeon Gold 5220R CPU @ 2.20GHz、503GB 内存和 NVIDIA GeForce RTX 3090 GPU。

F.3. 提升函子的计算复杂度

本节分析构造由两个主要函子定义的分次偏序集所需的计算复杂度：**关联偏序集函子**(I) 和 **对称单纯复形函子** (S)。我们重点关注预处理阶段，在此阶段，原始超图结构 H 会被转换为其偏序集表示。

设 $H = (\mathcal{V}, \mathcal{E})$ 是一个具有 $N = |\mathcal{V}|$ 个节点和 $M = |\mathcal{E}|$ 条超边的超图。我们将超边 e 的基数记为 $c_e = |e|$，将节点 v 的度记为 d_v =deg(v)。
关联偏序集 $I(H)$ 函子 I将超图映射到偏序集 $I(H) = (V \coprod E, \prec)$，以表示关联结构。该构造主要用于建立消息传递所需的关系结构。基本偏序 $v \prec e$ (对于 $v \in e$) 需要遍历所有节点—超边对，其规模随超图总规模线性增长， $\mathcal{O}(\sum_{e \in \mathcal{E}} c_e)$。此外，为了促进同级元素之间的信息流，我们构造了辅助邻接关系。团扩展将同一超边中的节点相连 e，需要 $\mathcal{O}(c_e^2)$ 次操作。同样，线图扩展将共享节点的

超边相连 v，其规模随 $\mathcal{O}(d_v^2)$增长。因此，总复杂度为 $\mathcal{O}\left(\sum_{e\in\mathcal{E}} c_e^2 + \sum_{v\in\mathcal{V}} d_v^2\right)$。虽然相对于超边基数和节点度而言，该复杂度呈二次增长，但对于稀疏真实世界超图而言仍然可控，因此无需进行结构截断即可构造 $I(H)$。

对称单纯复形偏序集 $S(H)$ 函子 S 将超图映射到由结构化单纯形组成的偏序集 $S(H)$。构造第 n 层集合 $S(H)_n$ 需要对各个子集进行组合枚举。具体而言，构造 $S(H)_n$需要为每条超边生成 $\binom{c_e}{n+1}$ 个单纯形。对于包含二单纯形（三角形， $n=2$）的标准配置，其复杂度为立方级，表示为 $\mathcal{T}_{gen} = \mathcal{O}\left(\sum_{e\in\mathcal{E}} c_e^3\right)$。这种多项式缩放给包含大基数超边的数据集带来了巨大挑战。例如，在 IMDB wri genre 中，超边的最大大小达到 266。显式生成这样一条超边的所有三角形，将从单次交互中产生数百万个元素，使计算过程难以承受。为确保计算可行，我们对 S 采用截断评估，设置阈值 $\tau = 20$，将定义域限制为满足 $c_e \le \tau$ 的超边。在这一约束下，复杂度上界为

$$\mathcal{T}_{S(H)} \approx \mathcal{O}\left(\sum_{e\in\mathcal{E}, c_e\le\tau} c_e^3\right) \approx \mathcal{O}\left(M\cdot\tau^3\right).$$

将 τ 视为常量后， $S(H)$ 的构造相对于超边数量 M 呈线性复杂度。这一分析证明了我们的截断策略是合理的，既能高效构造偏序集，又能为绝大多数数据保留高阶几何结构。