计算机科学
COMPUTER SCIENCE

融合稀疏编码的因果解耦表征学习

黄贝贝, 刘进锋

引用本文

黄贝贝, 刘进锋. 融合稀疏编码的因果解耦表征学习[J]. 计算机科学, 2026, 53(4): 66-77.

HUANG Beibei, LIU Jinfeng. Causal Disentangled Representation Learning with Integrated Sparse Coding [J]. Computer Science, 2026, 53(4): 66-77.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.251000012

融合稀疏编码的因果解耦表征学习

黄贝贝　刘进锋
宁夏大学信息工程学院　银川 750021
(12023132024@stu.nxu.edu.cn)

摘　要　深度学习模型由于其“黑盒”特性,特征表示缺乏可解释性。现有的解耦表征学习方法虽然在一定程度上能够通过识别数据中的独立因素来增强模型的解释能力,但它们通常忽视了数据中的复杂关联性及潜在因果结构,从而限制了模型在自动驾驶、医疗诊断等关键领域的应用,特别是在需要理解和干预因果关系的场景中表现不佳。针对当前解耦表征学习中因果关系建模不足的问题,提出了一种融合稀疏编码与因果推断的解耦表征学习框架。该框架在适当监督下通过因果推断机制精准建模数据中的因果关系,不仅能够生成高质量结构化表征,更具备对潜在因果机制的建模与干预能力,进而显著提升模型在因果任务中的适应性与鲁棒性;同时通过嵌入的卷积稀疏编码层施加稀疏性约束,有效筛选与因果结构高度相关的关键表征,进一步强化模型对高阶因果关系的敏感度与表达能力。实验结果表明,该框架在 Pendulum 和 CelebA 数据集上表现出色。样本效率在 Pendulum 数据集上达 98.65%,在 CelebA 数据集上达 99.55%,此外,在因果干预有效性和分布鲁棒性方面优于现有方法,证实了该方法在复杂因果场景下的优越性。

关键词:稀疏编码;因果推断;解耦表征学习;样本效率;分布鲁棒性
中图分类号　TP181

Causal Disentangled Representation Learning with Integrated Sparse Coding

HUANG Beibei and LIU Jinfeng
School of Information Engineering, Ningxia University, Yinchuan 750021, China

Abstract　Deep learning models often lack of interpretability in their feature representations due to their “black-box” nature. Although existing disentangled representation learning methods can enhance interpretability to some extent by identifying independent factors within the data, they usually neglect complex correlations and potential causal structures, which limits their applicability in critical domains such as autonomous driving and medical diagnosis, especially in scenarios that require understanding and intervention of causal relationships. To address the insufficient causal modeling in current disentangled representation learning, a disentanglement framework integrating sparse coding with causal inference is constructed. Under appropriate supervision, this framework leverages a causal inference mechanism to precisely model causal relationships within the data, thereby not only generating high-quality and structured representations but also enabling the modeling and intervention of potential causal mechanisms, which significantly improves the model's adaptability and robustness in causal tasks. Meanwhile, the embedded convolutional sparse coding layer imposes sparsity constraints to effectively filter key representations highly relevant to causal structures, further enhancing the model's sensitivity and expressive capacity for higher-order causal relationships. Experimental results demonstrate that the proposed framework performs excellently on both the Pendulum and CelebA datasets, achieving a sample efficiency of 98.65% on the Pendulum dataset and 99.55% on the CelebA dataset. Moreover, it outperforms existing methods in terms of causal intervention effectiveness and distribution robustness, confirming its superiority in complex causal scenarios.

Keywords　Sparse coding, Causal inference, Disentangled representation learning, Sample efficiency, Distribution robustness

1　引言

近年来,深度学习在图像识别、自然语言处理、自动驾驶等领域取得了突破性进展。然而,现有深度学习模型主要基于统计相关性进行端到端学习,其高度复杂的非线性结构使得模型呈现出显著的“黑盒”特性,内部决策机制难以解释,且在分布外样本或对抗扰动下,易出现不稳定表现,难以揭示变量间的真实因果机制。在对模型决策可靠性、可解释性以及干预预测能力要求极高的关键应用场景(如医疗影像诊断中的病理特征归因、自动驾驶中的场景理解与决策归因)中[1],对模型底层生成机制与因果关系的认知不足已成为重大挑战,严重限制了模型在真实世界中的可靠部署。

表征学习作为深度学习中的核心任务之一,旨在从高维观测数据中提取低维、语义丰富的特征表示。传统的表征学

到稿日期:2025-10-09　返修日期:2026-01-19
基金项目:宁夏自然科学基金(2025AAC030154)
This work was supported by the Natural Science Foundation of Ningxia(2025AAC030154).
通信作者:刘进锋(jfliu@nxu.edu.cn)

习往往侧重于“有效编码”，而缺乏对潜在语义与生成因素的显式建模，导致所学特征缺乏可解释性与可控性，难以支持跨域迁移、因果推理和反事实推断等更高层次的智能任务。解耦表征学习正是在此背景下提出的，其核心目标是识别并分离出数据背后独立的、可解释的生成因子，从而获得语义清晰、结构紧凑且因子间可控性强的潜在表示[2]。解耦表征不仅能够提升模型的泛化性和数据效率，更重要的是，它为实现因果推理、可解释生成与稳健决策提供了结构化基础。

回顾现有研究，可以将解耦表征学习发展脉络概括为两类主流路线并分析其核心局限。1）基于独立性假设的解耦阶段：为解决潜变量纠缠问题，以 VAE[3]，GAN[4] 为代表的生成模型通过拟合观测分布获取潜空间表示，显式引入独立性或互信息约束[5]，强化潜变量间的分离性，提升可解释性。然而，该类方法假定潜在因子完全独立，忽视了现实世界中普遍存在的依赖乃至因果关系，因而在复杂场景和分布变化下表现有限。2）因果驱动的解耦尝试：研究者开始在潜空间中引入结构因果模型[6]（Structural Causal Model，SCM）或可学习的因果图，以刻画变量间的方向性依赖和因果强度[7]。此类方法不仅在理论上弥补了独立性假设的不足，也表明将因果约束融入表征学习能够显著提升模型的鲁棒性与可控性。这些方法在实践中仍面临两大瓶颈：一是因果关系建模能力有限，难以捕捉复杂非线性因果机制；二是高维观测数据中普遍存在大量与因果无关的冗余或伪相关特征，会干扰因果结构的有效学习[8]。突破这一瓶颈的关键在于重塑解耦表征学习的目标：从“拟合统计相关”转向“捕捉因果本质”。将解耦表征与因果机制深度融合，能够帮助模型揭示变量间“因 → 果”的本质依赖，使其学习到的特征具备干预稳定性。这正契合当前对模型可解释性、鲁棒性与可靠性的核心需求。因此，如何学习到具备因果结构的表征，成为本文研究的主要切入点。

为应对这一挑战，本文提出一种融合稀疏编码的因果解耦表征学习模型 SCD-VG（Sparse Causal Disentangled VAE-GAN）。该模型的核心创新在于，将稀疏编码技术与因果推断机制系统性地融入一个基于 VAE-GAN 的解耦表征学习范式中。通过实验比较了本文模型与现有经典模型在下游任务中的分类准确度、因果干预效果等多维度实验，证明了本文模型的有效性。

本文的主要贡献如下：

1）融合稀疏编码，通过稀疏约束提取数据的核心特征，实现信息的高效压缩表示，提升模型的可解释性和泛化能力；

2）引入因果推断，能够识别潜在变量间的因果关系，突破传统解耦方法忽视数据因果结构的局限，实现对复杂因果关系的建模与有效干预；

3）通过在合成数据集和真实数据集上进行的实验验证，证明了所提方法在干预准确性、样本效率和分布鲁棒性方面优于现有解耦表征学习方法，为资源受限环境下的学习提供了新的范式。

2 相关工作

解耦表征学习可以分为两种类型：基于独立假设与基于因果假设[9]。

2.1 基于独立性假设的解耦方法

基于独立假设的解耦表征学习方法通过在 VAE 中引入归纳偏好提高解耦能力，并采用各种正则化器，实现有效的解耦。例如，Higgins 等[10] 提出的 β-VAE，通过在变分自编码器目标函数中引入超参数 β 来加强 KL 散度约束，从而在重构精度与潜变量独立性之间实现平衡。该模型在 3D faces 和 3D chairs 等数据集上展示了较强的无监督解耦能力，其潜在空间遍历结果表明，模型能够自动分离出方位角、光照、物体宽度等语义因子，其表现优于原始 VAE 与同期生成模型 InfoGAN[11]。然而，β-VAE 所依赖的强独立性假设未能刻画现实数据中普遍存在的潜在因果依赖关系，在处理复杂或高相关性的场景时，其解耦效果仍然受限。Chen 等[12] 在 β-VAE 的基础上，提出了 β-TCVAE，通过分解 KL 散度项并单独惩罚总相关性项，来更有效地约束潜变量之间的统计独立。实验结果显示，β-TCVAE 在 dSprites 数据集、3D Faces 数据集上的 MIG 均优于 β-VAE。这表明显式约束潜变量的统计独立性可以带来更高的解耦质量。

InfoMax-VAE[13] 方法通过增强潜在表示与观测数据之间的互信息，来学习信息更加充分、语义可判别的潜在表示。在优化目标中，依赖传统的重构项和 KL 散度项，引入互信息最大化项。Locatello 等[14] 训练了超过 12 000 个模型，覆盖 6 种无监督解耦方法、6 种评估指标，以及 7 个数据集，其结论表明：在没有适当的归纳偏置或监督信号的情况下，β-TCVAE 等无监督方法无法稳定地学习到解耦表示，其解耦性能对随机种子和模型初始化的依赖性甚至超过了算法本身。总体而言，基于独立性假设的方法虽为解耦表征学习奠定了基础，但其强独立性假设与现实世界中普遍存在的因子依赖及因果关系存在矛盾。在处理具有内在因果结构的复杂数据时，其解耦效果和泛化能力均受到限制。

2.2 因果解耦表征学习

为克服传统方法难以捕捉数据间复杂因果关系的根本局限，近年来研究焦点逐渐转向因果解耦表征学习。因果解耦表征学习方法可以捕捉数据生成过程的潜在因果机制，并通过分解因果因素实现更可解释和更稳健的表示。

基于 Suter 等[15] 的论述，Reddy 等[16] 提出了生成式潜变量模型（如 VAE），用于实现因果解耦应满足的基本属性。Yang 等[17] 提出的 CausalVAE 从因果关系的角度考虑数据中变异因素之间的关系，用结构因果模型描述这些关系，实现了对数据中因果关系的有效解耦表示学习。同时在 Pendulum 数据集和 CelebA 数据集上的实验中，CausalVAE 成功实现了对“摆角”因子的干预，并观测到“阴影长度”随之发生符合物理规律的改变，初步验证了因果建模的可行性。

与现有方法强制隐变量保持独立的假设不同，Shen 等[18] 在 DEAR 模型中考虑了更通用的情况，潜在的兴趣变量之间可能存在因果关系，从而同时实现因果可控生成和因果解耦表征学习。该模型突破了传统方法中“隐变量独立”这一假设，允许潜在变量之间存在潜在因果关系。其关键机制是将结构因果模型作为先验，结合 GAN 对抗机制联合优化生成器与编码器，并辅以适度监督信号，以提升可识别性与一致性。DEAR 框架不仅实现了因果可控的表示学习，还能从干预分布中生成样本，适用于反事实推理、因果发现等任务场景。具体模型如图 1 所示。

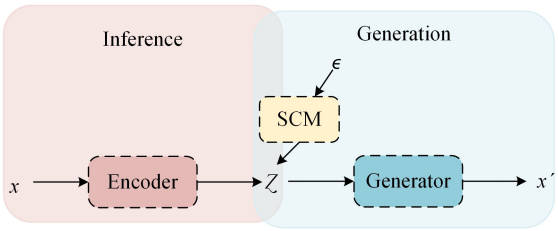


图 1 DEAR 模型的框架

Fig. 1 Framework of DEAR model

Komanduri 等[19]提出的 ICM-VAE 基于独立因果机制原则(Independent Causal Mechanisms, ICM),认为每个因果变量由相互独立的生成机制控制,机制之间彼此独立且不受外部干扰。ICM-VAE 在框架上引入了两个关键模块:其一是结构因果流,通过基于可逆流的非线性映射,对每个因果变量的生成机制进行参数化,能够有效刻画潜变量之间的复杂非线性关系;其二是因果解耦先验,利用标签监督信息建立从外部观测变量到因果潜变量的双射映射,保证每个潜变量与唯一的因果机制相对应,从而实现机制层面的可辨识性。理论上,ICM-VAE 在 iVAE 的可识别性定理基础上进行了扩展,证明了在满足充分可变性与光滑条件的情况下,模型能够在置换与逐元素重参数化意义下实现因果机制可识别。在实验方面,ICM-VAE 在 Pendulum, Flow 以及 CausalCircuit 等数据集上进行了系统评估:在 Pendulum 数据集上,模型能够准确识别"摆角→阴影位置与长度"的因果链,并生成符合物理规律的反事实图像;在更复杂的 CausalCircuit 数据集中,ICM-VAE 依然保持较高的因果可识别性与可控生成性能,验证了其在多变量非线性场景下的可推广性。各方法间的具体差异如表 1 所列。

表 1 不同方法对比

Table 1 Comparison of different methods

	β-VAE	CausalVAE	DEAR	ICM-VAE	SCD-VG
核心假设	潜在因子独立	潜在因子存在线性关系	潜在因子间存在一般非线性因果关系	潜在因子满足独立因果机制原则	潜在因子间存在复杂非线性因果关系
因果建模方式	无显式因果建模	线性结构方程模型	结构因果模型	基于流的非线性结构因果模型	基于流的非线性可学习的结构因果模型
因果函数灵活性	不适用(假设高斯先验)	线性函数,表达能力有限	神经网络,表达能力中等	非线性函数	仿射自回归流函数,能捕捉复杂、多模态因果机制
特征选择机制	无	无	无	无	有,卷积稀疏编码层,主动筛选关键因果特征
是否显式学习因果图	否	是	是	否	是
关键创新	引入可调节超参数 β,平衡 KL 散度与重建精度	首次将 SCM 作为先验	将 SCM 与 GAN、弱监督结合	以 ICM 原则为核心,设计因果解耦先验	稀疏编码 + 因果流,实现特征净化与机制深化的协同

You 等[20]在 DEAR 的基础上提出了 DRL_{CET} 模型,从"因果效应传播"的视角出发,将结构因果模型与图注意力网络及层级特征损失结合,实现因果结构学习与效应传递的协同优化。实验显示,DRL_{CET} 在 CelebA 上的因果可控性优于 DEAR 与 CausalVAE,证明了引入效应传播机制的有效性。

尽管因果解耦方法在理论上突破了独立性假设的局限,但仍面临两大核心问题:1)特征层面缺乏有效的冗余抑制机制,高维数据中大量与因果无关的背景特征、伪相关因素会干扰因果结构的学习,导致潜变量解耦不彻底、语义模糊;2)部分方法的因果建模能力受限,或依赖线性假设,或因果函数的非线性表达不足,难以捕捉复杂场景下的非线性因果机制,最终影响模型的可解释性与分布鲁棒性。这些瓶颈也成为本文提出融合稀疏编码与因果推断框架的核心动机。

为进一步提升因果解耦能力,本文构建了一种兼具特征稀疏性约束与非线性因果建模能力的新型解耦框架。该框架通过在深度生成模型中引入稀疏特征提取与结构化因果推断机制,从源头净化潜在表示,强化因果依赖的建模与识别能力,实现对复杂生成因素的可解释解耦。

3 融合稀疏编码的因果解耦表征学习

本章详细介绍所提出的融合稀疏编码的因果解耦表征学习框架,该框架通过融合卷积稀疏编码层与结构化因果模型,来实现对数据的有效解耦和因果干预。模型的整体架构如图 2 所示。

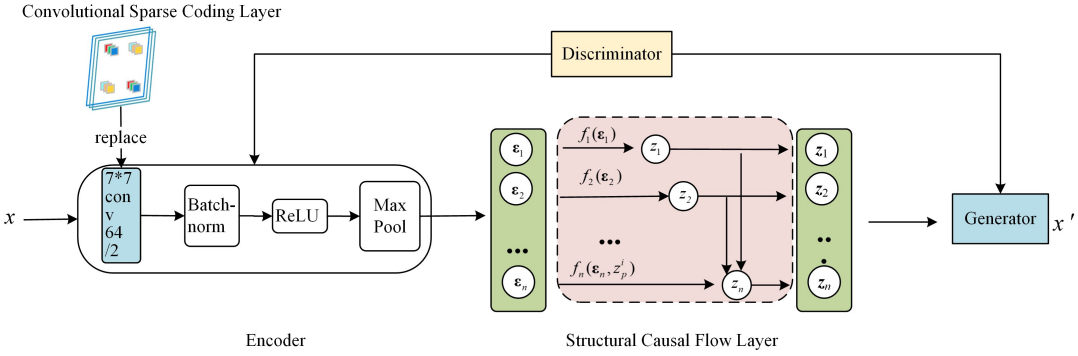


图 2 融合稀疏编码的因果解耦表征学习模型框架

Fig. 2 Causal disentangled representation learning framework integrating sparse coding

3.1 模型概述

SCD-VG 因果解耦框架通过在 VAE-GAN 解耦架构的编码器中引入卷积稀疏编码层，可显著增强模型对数据中关键特征的捕捉能力。同时，SCD-VG 模型利用条件流来参数化因果关系模型，学习将独立噪声$(\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\cdots,\boldsymbol{\varepsilon}_n)$分布映射到因果变量$(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n)$的分布。这种基于流的映射比简单的线性映射更能真实地捕捉潜在因果变量的复杂分布，有效实现反事实可识别性[21]。

SCD-VG 解耦模型主要包含两个阶段：推理阶段和生成阶段。在推理阶段，编码器 E 将输入数据 $\boldsymbol{x}$ 映射到潜在空间的表示 $\boldsymbol{z}$，潜在表示 $\boldsymbol{z}$ 经过结构因果流层的非线性变换，生成富含语义信息的潜在表示$(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n)$。具体而言，外生噪声变量$(\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\cdots,\boldsymbol{\varepsilon}_n)$经过非线性变换 f 和结构矩阵 $\boldsymbol{A}$ 的作用，生成具有因果关系的潜在表示 $\boldsymbol{z}=f((\boldsymbol{I}-\boldsymbol{A}^{\mathrm{T}})^{-1}h(\boldsymbol{\varepsilon}))$。编码器 E 输出 $\boldsymbol{z}$ 是通过最小化监督正则化项$\mathcal{L}_{\sup}$来进行优化的，如式(1)所示：

$$\mathcal{L}_{\sup}=\mathbb{E}_{(x,y)}\left[\sum_{i=1}^{m}\mathrm{CE}([\bar{E}(\boldsymbol{x})]_i,[\boldsymbol{y}]_i)\right] \tag{1}$$

其中，$\mathbb{E}_{(x,y)}$表示对输入数据 $\boldsymbol{x}$ 和标签 $\boldsymbol{y}$ 的联合分布求期望，编码器 E 的确定性部分 $\bar{E}(\boldsymbol{x})=\mathbb{E}(E(\boldsymbol{x})\mid\boldsymbol{x})$用于表征学习，$[\boldsymbol{y}]_i\in\{0,1\}$为第 i 个因果因子的真实标签，则 $CE(\boldsymbol{l},\boldsymbol{y})=-\boldsymbol{y}\log\sigma(\boldsymbol{l})-(1-\boldsymbol{y})\log(1-\sigma(\boldsymbol{l}))$ 为交叉熵损失，$\sigma(\cdot)$为 sigmoid 函数。通过监督正则化项，能够确保潜在表示 $\boldsymbol{z}$ 与真实因果因子之间的对应关系。

在生成阶段，通过生成器 G 基于潜在表示 $\boldsymbol{z}$ 生成新的假数据样本。通过提高判别器对假样本对的打分，从而使生成的联合分布 $p_{G,F}(\boldsymbol{x},\boldsymbol{z})=p_F(\boldsymbol{z})p_G(\boldsymbol{x}\mid\boldsymbol{z})$逼近真实的联合分布。

$$\mathcal{L}_{\mathrm{gen}}(E,G,F)=\mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})\sim p_{G,F}(\boldsymbol{x},\boldsymbol{z})}[softplus(-D(\boldsymbol{x}_{\mathrm{fake}},\boldsymbol{z}))] \tag{2}$$

3.2 卷积稀疏编码层

稀疏编码作为一种高效的特征表示方法，在图像分类与特征提取等任务中被广泛应用，并展现出优越的建模能力。Li 等[22]提出的卷积稀疏编码方法为将稀疏表示引入深度神经网络提供了有效路径。其基本思想是将传统卷积层替换为具备稀疏性约束的卷积稀疏编码层，从而使网络能够从高维输入中自动学习稀疏、判别性强的低维结构，提升模型的表示能力与泛化性能。

在 SCD-VG 模型中，卷积稀疏编码被嵌入编码器部分。该稀疏层通过限制潜在变量的激活数量，从压缩表示的角度增强了潜变量空间的可分性和语义一致性。卷积稀疏编码层是 SCD-VG 框架的核心组件之一，它以隐式网络层的形式引入，通过稀疏表示捕捉数据的关键特征。具体来说，给定一个多维输入信号 $\boldsymbol{x}\in\mathbb{R}^{M\times H\times W}$，其中 H,W 是空间维度，M 是 $\boldsymbol{x}$ 的通道数。假设信号 $\boldsymbol{x}$ 是由一个多通道稀疏码 $\boldsymbol{z}\in\mathbb{R}^{C\times H\times W}$ 与一个多维核卷积 $\boldsymbol{A}\in\mathbb{R}^{M\times C\times k\times k}$ 生成的，这个核被称为卷积字典。这里 C 是 $\boldsymbol{z}$ 和卷积核 $\boldsymbol{A}$ 的通道数。更准确地说，将 $\boldsymbol{z}$ 记为 $\boldsymbol{z}\doteq(\boldsymbol{\zeta}_1,\cdots,\boldsymbol{\zeta}_c)$，其中假定每个 $\boldsymbol{\zeta}_c\in\mathbb{R}^{H\times W}$是稀疏的，则核 $\boldsymbol{A}$ 记为：

$$\boldsymbol{A}\doteq\begin{pmatrix}\boldsymbol{\alpha}_{11}&\boldsymbol{\alpha}_{12}&\boldsymbol{\alpha}_{13}&\cdots&\boldsymbol{\alpha}_{1C}\\ \boldsymbol{\alpha}_{21}&\boldsymbol{\alpha}_{22}&\boldsymbol{\alpha}_{23}&\cdots&\boldsymbol{\alpha}_{2C}\\ \vdots&\vdots&\vdots&\ddots&\vdots\\ \boldsymbol{\alpha}_{M1}&\boldsymbol{\alpha}_{M2}&\boldsymbol{\alpha}_{M3}&\cdots&\boldsymbol{\alpha}_{MC}\end{pmatrix}\in\mathbb{R}^{M\times C\times k\times k} \tag{3}$$

其中，每个 $\boldsymbol{\alpha}_{ij}\in\mathbb{R}^{k\times k}$ 是一个大小为 $k\times k$ 的核。然后算子 $\mathcal{A}(\cdot)$生成信号 $\boldsymbol{x}$，如式(4)所示。生成 $\mathcal{A}(\cdot)$的示意图如图 3 所示。

$$\boldsymbol{x}=\mathcal{A}(z)\doteq\sum_{c=1}^{C}(\boldsymbol{\alpha}_{1c}*\boldsymbol{\zeta}_c,\cdots,\boldsymbol{\alpha}_{Mc}*\boldsymbol{\zeta}_c)\in\mathbb{R}^{M\times H\times W} \tag{4}$$

其中，每个 $\boldsymbol{\alpha}_{ij}$ 是一个标准卷积核；生成算子 $\mathcal{A}(\cdot)$的作用等价于标准卷积运算；$*$ 表示 2D 卷积，且步幅、填充方式与标准卷积层一致。因此，CSC 层在数学上等价于带稀疏性约束的卷积层，区别在于卷积核 $\boldsymbol{A}$ 与稀疏码 $\boldsymbol{z}$ 同时被优化，而不是仅优化卷积核。

综上所述，给定一个多维输入信号 $\boldsymbol{x}\in\mathbb{R}^{M\times H\times W}$，将稀疏的输出 $\boldsymbol{z}_*\in\mathbb{R}^{C\times H\times W}$ 执行一个(逆)映射，其中 C 是输出通道的数量。在上述稀疏生成模型下，可以通过解决 Lasso 类型优化问题来寻找最优的稀疏解 $\boldsymbol{z}$。

$$\boldsymbol{z}_*=\underset{\boldsymbol{z}}{\operatorname{argmin}}\,\boldsymbol{\lambda}\|\boldsymbol{z}\|_1+\frac{1}{2}\|\boldsymbol{x}-\mathcal{A}(\boldsymbol{z})\|_2^2\in\mathbb{R}^{C\times H\times W} \tag{5}$$

卷积稀疏编码层旨在通过稀疏编码 $\boldsymbol{z}$ 与卷积字典 $\boldsymbol{A}$ 的线性组合，来近似重构输入信号 $\boldsymbol{x}$。在此过程中，特征图 $\boldsymbol{z}$ 决定了 $\boldsymbol{A}$ 中哪些卷积滤波器及其响应的位置和强度被用来重建信号，这一过程模型捕捉到输入数据的关键特征。模型的重构精度不必完全精确，以便能够容忍实际数据与模型假设之间的差异。这种差异通过 $\boldsymbol{x}$ 和$\mathcal{A}(\boldsymbol{z})$之间的 ℓ_2-范数来度量，即两者差的平方和。稀疏建模是通过目标函数中 $\boldsymbol{z}$ 的逐项 ℓ_1-范数引入的，这强制 $\boldsymbol{z}$ 具有稀疏性。参数 $\lambda>0$ 用于控制 $\boldsymbol{z}$ 的稀疏度与残差 $\boldsymbol{x}-\mathcal{A}(\boldsymbol{z})$之间的权衡。

具体而言，稀疏层采用卷积字典学习结构，卷积核大小设置为 7×7，输出通道数为 64，步幅为 2，填充为 3；稀疏正则化系数 λ 取 0.1，迭代更新步数为 10，字典扩张系数为 2。为防止训练不稳定和梯度爆炸，稀疏层权重采用归一化策略。该稀疏卷积层嵌入至 ResNet 编码器的首个卷积阶段，用于对输入图像进行高稀疏度特征筛选，从而抑制与因果关系无关的背景纹理与噪声信息，保留关键的因果特征表征。

卷积稀疏编码层不仅能有效捕捉数据的关键特征，还能通过压缩冗余信息，使模型在解耦因果机制时具备更强的灵活性与可解释性。特别是，稀疏表示的解耦特性为后续“结构因果流层”提供了高质量的因果变量初始表示，进而使得因果机制的建模更加精确。

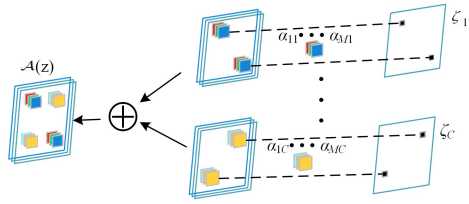


图 3 CSC 层卷积稀疏编码模型中算子$\mathcal{A}$的示意图

Fig. 3 Illustration of operator $\mathcal{A}$ in the convolutional sparse coding(CSC) model

3.3 结构因果流层

ICM-VAE 模型采用独立因果机制正则化方法，利用可逆流模型对复杂非线性因果函数进行高精度建模，为本文工作提供了有益启发。

与之相比，本文所引入的因果推断在思想上与 ICM-VAE 的因果流模块一致，均采用逐维的可逆仿射变换，并累

积对数雅可比行列式用于可识别的密度建模；父集信息均通过“掩码”注入到网络中，实现对因果父变量的条件化。然而，两者在结构学习中存在差异：ICM-VAE 假设因果结构 C 已知；本文模型中的结构因果流层（Structural Causal Flow Layer，SCF）支持在训练过程中自适应学习 C，同时以卷积稀疏编码层提取的高信息密度特征为输入，从源头减少了无关噪声对因果函数学习的干扰，进一步在优化目标中引入针对因果解耦的联合约束，使得所学映射不仅拟合因果机制，还能增强潜变量的可解释性与可控性。

3.3.1 具体实现

结构因果流层是通过引入流式[23]模型学习到更为复杂的因果分布，其具体结构如图 2 所示。

SCD-VG 模型中，假设通过非线性函数描述因果关系。结构因果模型通过三元组 $M=\langle Z,E,F\rangle$ 进行定义。其中，Z 表示由 n 个内生因果变量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n$ 构成的集合；E 表示由 n 个外生噪声变量 $\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\cdots,\boldsymbol{\varepsilon}_n$ 构成的集合，这些噪声变量通常作为中间潜在变量进行学习；$F=\{f_1,\cdots,f_n\}$ 是由 n 个独立的因果模型构成的集合，每个因果关系模型的形式为 $\boldsymbol{z}_i=f_i(\boldsymbol{\varepsilon}_i,\boldsymbol{z}_p^i)$，其中 f_i 是一个函数，将噪声变量 $\boldsymbol{\varepsilon}_i$ 和父变量 z_p^i 映射到因果变量 $\boldsymbol{z}_i$。

为了直观地表示因果关系，采用有向无环图（DAG）来展示变量间的因果结构，其中每条有向边指示因果关系的方向，其邻接矩阵为 $\boldsymbol{C}\in\{0,1\}^{n\times n}$。与固定因果图不同，这里的 $\boldsymbol{C}$ 是可学习的参数矩阵，自动去除自环约束，在训练过程中自适应地优化因果关系。根据 SCM 理论，每个内生变量通过其父变量和噪声变量应用相应的因果机制生成。这一递归过程遵循因果图的拓扑顺序，即每个变量仅依赖于其前驱节点的因果变量。具体过程如下：模型中邻接矩阵 $\boldsymbol{C}$ 采用可微参数化的自适应学习机制。训练初期通过 Sigmoid 平滑化保持梯度流动，后期引入 Gumbel-Softmax 实现硬结构学习，最终形成近似二值的有向无环图。同时通过稀疏约束与去自环操作消除无效边连接，在整体优化目标中引入结构稀疏正则项，以促进图的可解释性与稀疏性，通过可逆流网络实现端到端优化。

为了高效地实现这一递归因果推理过程，SCF 层采用仿射自回归流[24]。仿射自回归流是一种流模型，通过将变量依次生成，使每个因果变量只依赖于先前生成的因果变量的子集（即其父变量）。具体计算式如式(6)所示：

$$\boldsymbol{z}_i=\exp(s_i)\cdot\boldsymbol{\varepsilon}_i+t_i \tag{6}$$

其中，s_i 和 t_i 是可学习的函数，表示缩放和平移变换，分别是由神经网络模型学习得到的斜率和偏移量。具体结构如表 2 所列。由于缩放因子始终取指数形式且经过 clamp(·)限幅，因此其数值不会出现零或无穷大，保证了变换的非奇异性和可逆性。其逆映射可表示为 $\boldsymbol{z}_i=\dfrac{\boldsymbol{z}_i'-t_i}{\exp(clamp(s_i,-5,5))}$。这是一种可逆变换，并且变换的对数行列式可以高效计算，如式(7)所示：

$$\log\prod_i\left|\frac{\partial\boldsymbol{\varepsilon}}{\partial\boldsymbol{z}}\right|=\sum_i\log\left|\frac{\partial f_i(\boldsymbol{\varepsilon}_i;\boldsymbol{z}_p^i)}{\partial\boldsymbol{\varepsilon}_i}\right|^{-1} \tag{7}$$

卷积稀疏编码层通过提取稀疏的潜在表示 $\boldsymbol{z}$，为 SCF 层提供了高质量的初始输入。这些稀疏表示的结构化性使 SCF 层能够更好地捕捉因果变量之间的独立性，在生成因果变量时具备更强的解释性和鲁棒性。

表 2 s_i 和 t_i 的具体结构

Table 2 Specific architecture of s_i and t_i

Linear→100
LayerNorm(100)
LeakyReLU(0.2)
Linear→100
LayerNorm(100)
LeakyReLU(0.2)
Linear→k

3.3.2 可识别性分析

SCF 层通过仿射自回归流在潜在语义子空间实现从外生噪声到内生因子的可逆映射。为保证因果方向在机制层面的可识别，本文在以下充分条件下给出结论。

1）观测噪声密度的特征函数非退化；

2）解码器 G 在其像空间上是可微同胚映射；

3）因果机制的充分统计量 T_i 可微同胚；

4）存在足够多的不同辅助标签向量 $\boldsymbol{u}_0,\boldsymbol{u}_1,\cdots,\boldsymbol{u}_{nk}$，使得由标签及父变量扰动形成的变异性矩阵 $\boldsymbol{L}=(\lambda(\boldsymbol{z}_p^1,\boldsymbol{u}^{(1)})-\lambda(\boldsymbol{z}^{(0)},\boldsymbol{u}^{(0)}),\cdots,\lambda(\boldsymbol{z}_p^{nk},\boldsymbol{u}^{(nk)})-\lambda(\boldsymbol{z}^{(0)},\boldsymbol{u}^{(0)}))$ 为满秩（可逆），且对任意 i，都有 $\lambda_i(\boldsymbol{z}_p^i,\boldsymbol{u}_i)\neq0$。

在潜在变量上施加结构因果流，在上述条件下，若两组参数诱导的观测-标签联合分布相同，则其在语义子空间内的因果机制在“置换与逐元素重参数化”等价意义下相同。换言之，因果方向在语义子空间可识别。直观上，雅可比项把可逆流的体积/基底约束直接写入目标，避免退化变换；外源标签提供足够条件扰动以打破混叠；稀疏且无自环的结构学习进一步收紧自由度，共同提升机制层面的可分辨性。

3.4 模型优化

3.4.1 损失函数设计

SCD-VG 模型的损失函数由 3 个部分组成。

1）判别器损失

判别器损失用于训练判别器 D，使其能够区分真实数据对 $(\boldsymbol{x},\boldsymbol{z})$ 和生成数据对 $(\boldsymbol{x}_{\text{fake}},z)$，如式(8)所示：

$$\mathcal{L}_D=\mathbb{E}[softplus(D(\boldsymbol{x}_{\text{fake}},\boldsymbol{z}))]+\mathbb{E}[softplus(-D(x,E(\boldsymbol{x}))] \tag{8}$$

其中，$softplus$ 是函数 $softplus(\boldsymbol{x})=\log(1+e^x)$；$E(\boldsymbol{x})$ 是输入 $\boldsymbol{x}$ 的编码器输出。

2）编码器损失

编码器是 $\mathcal{L}(E,G)$ 由生成模型损失和监督正则化损失 $\mathcal{L}_{\text{sup}}$ 组成的加权和组成，具体如式(9)所示：

$$\mathcal{L}_E=\mathbb{E}_{\boldsymbol{x}\sim q_x}[D(\boldsymbol{x},E(\boldsymbol{x}))]+\lambda\mathcal{L}_{\text{sup}}+\alpha\mathcal{L}_{\text{jac}} \tag{9}$$

其中，$\mathcal{L}_{\text{jac}}=-\mathbb{E}_x[\log|\det\boldsymbol{J}_{T_\psi}(\bar{\boldsymbol{z}})|]$ 相当于把可逆流的变换体积项显式放入目标，促使流在数据诱导的潜在子空间上学习到信息更充足、退化更少的结构；$\mathbb{E}_{x\sim q_x}[D(x,E(x))]$ 是通过双向判别器 D 计算在真样本对上的得分；λ 是一个正则化参数，用于平衡生成模型损失和监督正则化损失之间的权重；监督正则化损失 $\mathcal{L}_{\text{sup}}$ 是通过少量标注数据来优化的，其目标是确保潜在表示 $\boldsymbol{z}$ 与真实因果因子之间的对应关系，监督类型是交叉熵损失。

3）生成器损失

生成器损失用于训练生成器生成能够欺骗判别器的高质量输出。

$$\mathcal{L}_G=-\mathbb{E}[\boldsymbol{s}\cdot\boldsymbol{D}(\boldsymbol{x}_{\text{fake}},\boldsymbol{z})] \tag{10}$$

其中，$\boldsymbol{D}(\boldsymbol{x}_{\text{fake}},\boldsymbol{z})$计算假样本对的判别器得分；$s$ 是一个约束在 $0.5\leqslant s\leqslant 2$ 的缩放因子，用于稳定训练。

SCD-VG 模型通过分别优化各组件损失函数而非单一综合损失函数的方式，实现了解耦表示学习，借助对抗训练实现高质量生成、监督信息整合（将标签数据融入潜在表示）以及因果结构保持（学习符合预定义因果关系的表示），从而在多重学习任务间达成平衡，防止单一目标主导整个训练过程。

3.4.2 收敛性与误差上界

在上述损失函数基础上，本节给出在所采用训练策略下的收敛性与误差上界分析。$\boldsymbol{\theta}=(\boldsymbol{\theta}_{\text{E}},\boldsymbol{\theta}_{\text{G}},\boldsymbol{\theta}_{\text{F}},\boldsymbol{\theta}_{\text{D}})$ 分别对应编码器、解码器（生成器）、因果先验与判别器。为体现“最小-最大”对抗结构，定义梯度映射为：

$$g(\boldsymbol{\theta})=(\nabla_{\boldsymbol{\theta}_{\text{E}}}\mathcal{L}_{\text{E}},\nabla_{\boldsymbol{\theta}_{\text{G}}}\mathcal{L}_{G},\nabla_{\boldsymbol{\theta}_{\text{F}}}\mathcal{L}_{G},-\nabla_{\boldsymbol{\theta}_{D}}\mathcal{L}_{D}) \tag{11}$$

1）一阶期望收敛（TTUR）

若各子目标 L-smooth、小批量梯度二阶矩有界，且判别器子问题在局部满足强凸近似，取满足 TTUR 的步长（判别器步长较大，生成/编码/先验较小），则存在常数 $C>0$，任意迭代轮数 T 满足：

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\,\|\,g(\boldsymbol{\theta}^{(t)})\,\|^{2}]\leqslant\frac{C}{\sqrt{T}} \tag{12}$$

这表明参数序列在期望意义上以 $O(T^{-1/2})$ 的速度收敛。

2）泛化误差上界

若通过谱归一化、梯度裁剪、残差连接等控制使复合映射的 Lipschitz 常数有界，为 C_{Lip}，且小批量噪声为次高斯，则对任意 $\delta\in(0,1)$，经验风险 $L(\boldsymbol{\theta})$ 与 $\hat{L}(\boldsymbol{\theta})$期望风险之间以至少 $1-\delta$ 的概率满足：

$$|L(\boldsymbol{\theta})-\hat{L}(\boldsymbol{\theta})|\leqslant\frac{C_{Lip}}{\sqrt{n}}+\tilde{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) \tag{13}$$

其中，n 为样本数。这表明随着训练样本数量的增加，模型的经验性能将一致收敛于期望性能。

3）误差传播界

设潜在可逆变换与生成器分别在模型诱导的潜在支持集合上为 $\mathcal{L}_{\text{T}}-$Lipschitz 和 $\mathcal{L}_{\text{D}}-$Lipschitz，监督对齐使均值通道误差满足 $\|\,\Delta\boldsymbol{z}\,\|\leqslant\varepsilon$，则对任意样本有：

$$\|\,G(T(\boldsymbol{z}+\Delta\boldsymbol{z}))-G(T(\boldsymbol{z}))\,\|\leqslant L_{G}L_{\text{T}}\varepsilon \tag{14}$$

解码采用重参数化采样 $\boldsymbol{z}_{\text{fake}}$，取相同噪声耦合亦有几乎处处成立的路径界，从而在期望意义下同样满足式（14），即表示域的小偏差不会在像素域被无限放大，从而保证生成结果的稳定性与鲁棒性。

综上，本文模型在优化层面具有 $O(T^{-1/2})$ 的收敛速率，在统计层面具有 $O(\sqrt{\ln(1/\delta)/n})$ 的泛化误差上界，并在结构 Lipschitz 约束下满足式（14）的误差传播上界。这为后续实验部分的性能与稳定性提供了理论支撑。

4 实验

4.1 实验设置和数据集

实验在 Windows 11 操作系统、Nvidia GeForce RTX 4090 GPU 的计算环境中进行，所有实验均基于 PyTorch 深度学习框架实现。本文实验所用数据主要包括 Pendulum 数据集[17]、CelebA 数据集[25]和 MPI3D 数据集[26]。

1）Pendulum 数据集是基于物理模型的合成数据集。每个图像由 4 个连续的相位生成，即摆角、光角、阴影长度和阴影位置，具体因果关系如图 4(a)所示。参照 Shen 等的研究，在生成标签的过程中引入了随机测量噪声，以使 Pendulum 数据集更加贴近现实情况。为了模拟环境干扰，随机生成了 20%的数据样本的阴影部分，训练集和测试集的规模分别为 7500 和 2500 个样本。

2）CelebA 数据集由 200000 张名人面孔图像组成，具有 40 个离散的属性，每个属性的值都为 -1 或 1。实验选择两个因果相关的属性子集，分别是 CelebA-Smile 和 CelebA-Attractive，具体因果关系分别如图 4(b)和图 4(c)所示。

3）MPI3D 数据集是一个基于真实与模拟场景相结合的多因素视觉数据集，旨在评估表示学习方法在多维度因素解耦中的性能。该数据集包含 1036800 张分辨率为 $64\times64\times3$ 的彩色图像，涵盖 7 个相互独立的变化因素，分别为物体颜色（Object Color）、物体形状（Object Shape）、物体尺寸（Object Size）、相机高度（Camera Height）、背景颜色（Background Color）、物体自由度 1（DOF1）和物体自由度 2（DOF2）。每个因素的取值分别为 [4/6,4/6,2,3,3,40,40]，具体取决于数据模式（toy，realistic 和 real）。本文采用 mpi3d_ real 和 mpi3d_ realistic 模式。

模型的评估从 3 个方面展开：1）在 Pendulum 数据集和 CelebA 数据集上进行干预实验，验证模型学习到的因果系统的语义正确性；2）研究 SCD-VG 模型在两个下游任务中的优势，即样本效率和分布鲁棒性；3）为全面评估模型 SCD-VG 在复杂多因子场景下的解耦能力，本文在 MPI3D 数据集上引入 Mutual Information Coefficient（MIC）与 Total Information Content（TIC）两项指标进行量化验证。

4.2 评价指标

MIC[27]衡量单个潜变量 z_i 与真实因子 v_j 之间的最大互信息值。

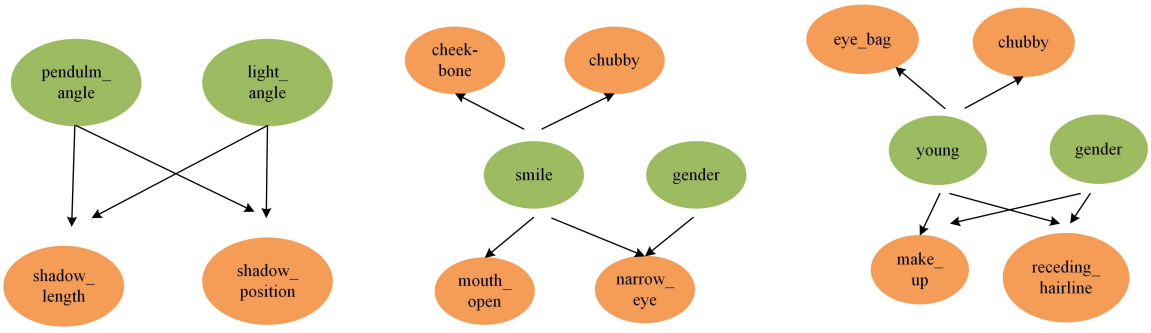
$$MIC(\boldsymbol{z}_i,\boldsymbol{v}_j)=\max_{xy\leqslant B(n)}\boldsymbol{M}_{x,y}(\boldsymbol{z}_i,\boldsymbol{v}_j) \tag{15}$$

其中，特征矩阵 $\boldsymbol{M}_{x,y}(\boldsymbol{z}_i,\boldsymbol{v}_j)=\max_{G\in G_{x,y}}\frac{I_G(z_i;v_j)}{\log\min\{x,y\}}$表示在所有 $x\times y$ 的网格划分中，潜变量与真实因子之间归一化互信息的最大值；$G_{x,y}$表示将$(\boldsymbol{z}_i,\boldsymbol{v}_j)$的取值空间进行 $x\times y$ 网格划分的集合；$\boldsymbol{I}_G(\boldsymbol{z}_i;\boldsymbol{v}_j)$ 是在网格 G 上离散后的互信息，用于衡量该分辨率下两者的信息共享程度。MIC 值越高，表明该潜变量越能集中、唯一地反映某一特定真实因子的变化模式，即单一潜变量与单一真实因子之间的对应关系越强，这直接体现了模型在因子级别解耦的纯度与清晰度。

TIC 则进一步衡量潜变量与所有真实因子之间总体信息关联的紧密程度。TIC 值越高，表明学习到的潜变量表示在整体上蕴含了更多关于真实生成因子的信息，即潜变量系统对数据生成过程的解释能力越强，反映了模型整体表征的充分性与信息保留能力。$\boldsymbol{M}_{x,y}(\boldsymbol{z}_i,\boldsymbol{v}_j)$与 MIC 的定义相同，表示在不同网格分辨率下的归一化互信息。

$$TIC(\boldsymbol{z}_i,\boldsymbol{v}_j)=\sum_{xy\leqslant B(n)}\boldsymbol{M}_{x,y}(\boldsymbol{z}_i,\boldsymbol{v}_j) \tag{16}$$

在解耦建模的语境下，高 MIC 值意味着良好的因子分离性，即每个潜变量主要控制一个语义因子；高 TIC 值则意味着良好的完备性，即所有重要的数据变异因素都能在潜空间中找到对应的表示。



(a) Pendulum (b) CelebA-Smile (c) CelebA-Attractive

图 4 Pendulum 数据集和 CelebA 数据集上的因果关系图

Fig. 4 Causal graphs of the Pendulum dataset and the CelebA dataset

4.3 基线模型

为全面评估所提模型 SCD-VG 的性能，将其与 6 种基于 VAE 框架的主流表示学习和解耦方法进行了系统性对比，包括标准 VAE（2014 年）、β-VAE（2017 年）、InfoMax-VAE（2020 年）、DEAR（2022 年）、ICM-VAE（2024 年）以及最新的 DRL_{CET}（2025 年）。需要指出的是，虽然其中部分方法提出较早，但它们在解耦表征学习研究中具有里程碑意义，并长期作为标准参照。在相关文献中，这些模型已形成被广泛接受的标准基线，能够保证比较结果的可解释性与普适性。VAE 与 β-VAE 分别代表了生成式表征学习的基础框架和独立性约束范式，InfoMax-VAE 代表了信息论驱动的解耦思路。相比之下，DEAR 首次将结构因果图和对抗训练机制结合，能够在潜在空间中学习变量间的因果依赖，是因果解耦方向的重要进展。ICM-VAE 基于独立因果机制假设，强调各潜在变量的生成过程应由一组互不干扰的独立因果机制决定，从而在结构上实现因果驱动的解耦。在 DEAR 基础上，DRL_{CET} 进一步在 VAE 框架中引入因果效应传递机制并结合图注意力网络，不仅能挖掘潜在变量的因果依赖，还能实现因果效应的高效传播与聚合，在多变量干预下保持合理的因果一致性。选择这 6 种模型作为基线，不仅覆盖了解耦研究的主要演进脉络，也能够凸显 SCD-VG 在经典框架和最新方法上的全面提升。上述方法的直观对比如表 3 所列，实验参数的对比如表 4 所列。

表 3 基线模型方法的对比

Table 3 Comparison of baseline methods

模型	核心思想	是否建模因果结构	特征选择机制	代表性
VAE	变分自编码框架，最小化重构误差与 KL 散度	否	否	生成式表征学习基础模型
β-VAE	引入 β 权重，强化因子独立性	否	否	独立性解耦范式代表
InfoMax-VAE	最大化潜在变量与输入数据互信息	否	否	信息论驱动解耦方法
ICM-VAE	引入独立因果机制约束，使潜变量对应不同的独立因果因素	是	否	强调独立因果机制的结构性假设
DEAR	引入结构因果图先验学习潜在变量因果关系	是	否	因果解耦主流模型
DRL_{CET}	基于因果效应传递与图注意力网络的结构性 VAE	是	否	最新因果效应传播模型
SCD-VG	卷积稀疏编码 + 结构因果流层，联合实现特征净化与非线性因果建模	是	是	本文提出的因果解耦新框架

表 4 对比方法参数总结

Table 4 Summary of parameters for comparative methods

模型	参数	Pendulum	CelebA
所有的模型	Discriminator Learning rate	1×10^{-4}	1×10^{-4}
	Encoder Learning rate	5×10^{-5}	5×10^{-5}
	Decoder Learning rate	5×10^{-5}	5×10^{-5}
	Latent dimension	6	100
	Batch size	128	128
	监督损失权重	5	5
VAE	β	1	1
β-VAE	β	4	4
InfoMax-VAE	MI	1	1
ICM-VAE	Latent dimension	16	128
	Learning rate	1×10^{-3}	1×10^{-3}
DRL_{CET}	Latent dimension	10	50
	GAT 层	4	4
	监督损失权重	1	1
	Attention heads	2	2

VAE 作为最早的变分自编码框架，奠定了表征学习的基础；β-VAE 在其基础上引入了可调的 KL 权重，是最具影响力的独立性约束方法之一；InfoMax-VAE 通过最大化互信息引导特征提取，是信息论驱动解耦的代表；DEAR 融合了因果图先验，是当前因果解耦领域的先进模型；ICM-VAE 从独立因果机制角度推进了解耦的因果建模；DRL_{CET} 通过显式的因果效应传递与图神经网络的结合，代表了因果解耦研究的前沿方向。它们分别代表了经典 VAE 框架、独立性约束范式、信息论范式以及因果建模范式，涵盖了近年来主要的解耦建模思路，因此具有较强的代表性。选择这 6 种模型作为对比基准，旨在从多个维度全面展示 SCD-VG 在因果解耦表示学习中的优势和独特贡献。

4.4 干预实验

为评估模型的因果解耦性能，设计了两类干预实验。第一类实验是潜在遍历，即在潜在空间中遍历一个维度，同时固定其他维度，通过改变特定维度的值来观察对生成样本的影响，从而验证单个潜在维度的可控性及其对生成数据的影响。第二类实验则是对一个潜在变量施加定向干预，观测其他潜

在变量随之产生的变化，评估模型捕捉潜在因果关系的能力。通过这两类实验，能够系统地考察模型在局部干预与全局因果反应中的表现。

本文将 SCD-VG 与最先进的因果解耦模型 DEAR 在合成数据集 Pendulum 以及真实数据集 CelebA 上进行比较，以评估模型的解耦效果。图 5(a)和图 5(b)分别是对应 DEAR 和 SCD-VG 模型对 Pendulum 数据集进行第一类干预实验所产生的结果，可以看出，DEAR 模型在干预光角过程中，其他维度如阴影长度出现了不稳定的变化。而 SCD-VG 干预光角时，其他维度基本不变。

图 6(a)和图 6(b)分别对应 DEAR 和 SCD-VG 模型对 CelebA 数据集进行第一类实验操作的结果，有 6 个干预维度。可以看出，DEAR 对于每个维度并没有很好的解耦，导致对每个维度干预时，其他维度也发生变化，而 SCD-VG 对每个维度能够产生更准确的干预措施，其他维度保持不变。例如，在图 6(b)中，对性别进行干预时，嘴巴未发生变化，说明 SCD-VG 模型正确捕捉了因果关系。在图 7(a)中，对 Pendulum 数据集进行第二类干预实验，SCD-VG 对摆角角度进行干预时，不同图像中的摆锤角度做出相应地改变，同时阴影位置和阴影长度也按照因果关系发生变化，改变阴影位置和阴影长度时，摆角和光角未产生变化，表明 SCD-VG 能够更好地实现解耦。图 7(b)为 SCD-VG 在 CelebA 数据集上进行第二类实验操作的结果。对微笑进行干预时，微笑做出相应改变，同时张嘴也按照因果关系发生变化，改变嘴的大小时，微笑的弧度未产生变化。这表明 SCD-VG 能够更好地实现解耦。

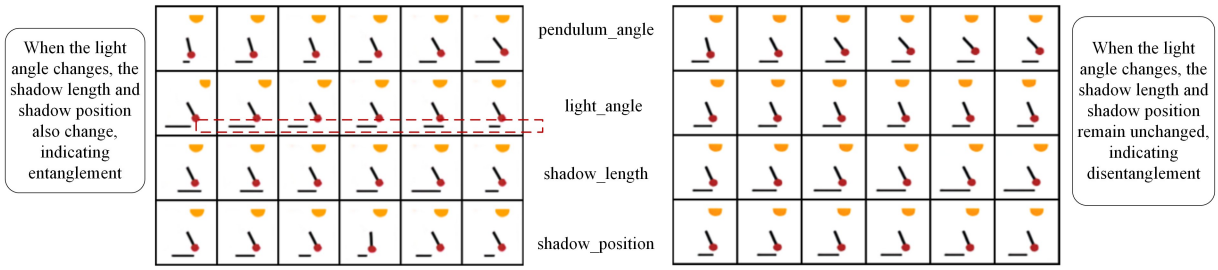


图 5 第一类干预实验在 Pendulum 数据集上 DEAR 和 SCD-VG 的对比结果

Fig. 5 Comparison results of DEAR and SCD-VG in the first type of intervention experiment on the Pendulum dataset

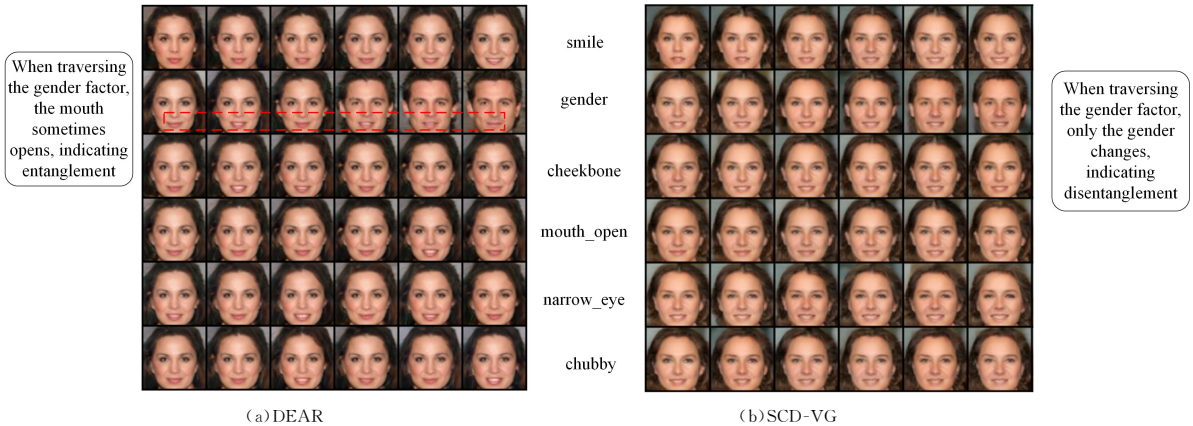


图 6 第一类干预实验在 CelebA 数据集上 DEAR 和 SCD-VG 的结果

Fig. 6 Results of DEAR and SCD-VG in the first type of intervention experiment on the CelebA dataset

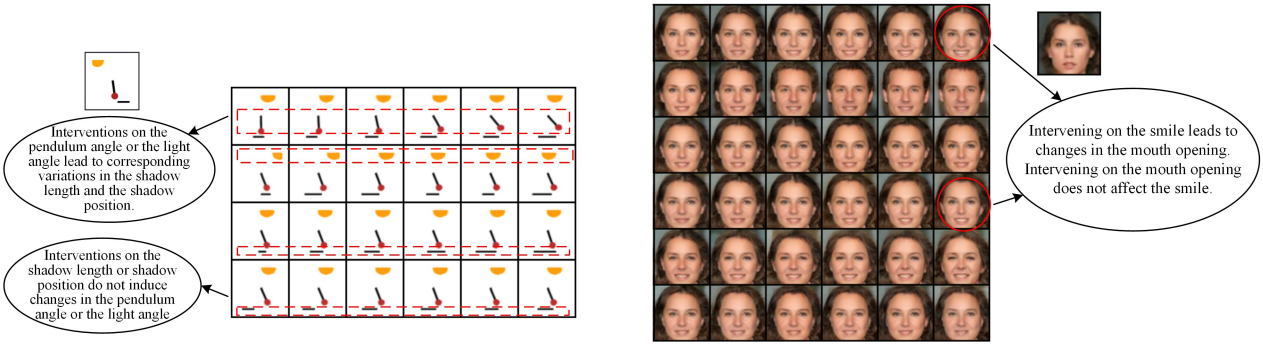


图 7 SCD-VG 在 Pendulum 数据集和 CelebA 数据集上的第二类干预实验结果

Fig. 7 Results of the second type of intervention experiment using SCD-VG on the Pendulum dataset and the CelebA dataset

4.5 下游任务

本节旨在研究并展示 SCD-VG 学习到的因果表征对下游任务的积极影响，重点关注样本效率和分布鲁棒性两个关键评估维度。主要采用二元分类任务进行评估，评估模型在因果表征上的泛化能力和实际表现。

在 Pendulum 数据集中，将数据损坏的标签视为目标标签，如果数据损坏，则 $\tau=1$，否则 $\tau=0$。在 CelebA 数据集上，参考图 4(c)中的因果结构 CelebA-Attractive，人工构建一个目标标签 τ，当满足条件：年轻(是否年轻=1)，女性(性别=0)，没有脱发(是否脱发=0)，化了妆(是否化妆=1)，苗条(是否肥胖=0)，没有眼袋(是否有眼袋=0)，设 $\tau=1$；否则 $\tau=0$。这个标签 τ 表示一种理想化的任务特征，以此来训练模型识别和理解这些特定的人物属性。τ 的构建明确依赖于一组特定的生成因子，因此假设能够有效捕捉和解耦这些因果因素的模型，在下游任务中应展现出更高的数据利用效率和对分布变化更强的鲁棒性。

为全面评估 SCD-VG 的表现，将其与几种代表性的解耦方法进行对比，包括 VAE，β-VAE，InfoMax-VAE 以及具有明确因果结构建模能力的 DEAR 等模型。由于原始 VAE 与 β-VAE 多在低分辨率数据集上采用全连接网络(MLP)编码器，InfoMax-VAE 则多采用浅层卷积网络，这些结构与本研究在高分辨率数据上的需求及网络容量存在显著差异。为消除因网络结构和参数规模不同带来的性能偏差，确保对比聚焦于方法本身的核心机制，本研究对上述基线模型进行了统一的结构适配，即将其编码器全部替换为与 SCD-VG 相同的 ResNet18 结构，生成器采用与 SCD-VG 一致的解码器架构，并保持输入分辨率为 64×64 及相同的归一化方式，同时确保网络参数规模处于同一量级，以排除容量差异的影响。这种结构适配能够避免网络容量差异带来的性能偏差，使得比较结果更能反映各方法的核心机制(损失函数与先验建模)的差异。在训练过程中，VAE，β-VAE 和 InfoMax-VAE 不仅包含 SCD-VG 所使用的损失项，同时还保留了各自原始的损失函数，以充分发挥各方法的优势。

在数据预处理方面，CelebA 数据集先将原始人脸图像中心裁剪至 128×128 像素，再缩放至 64×64 分辨率，并将像素值归一化到 $[-1,1]$ 区间；同时将属性标签由 $\{-1,1\}$ 映射为 $\{0,1\}$，以适配网络输入；Pendulum 数据集将图像缩放至 64×64 分辨率，将其标签数值归一化至 $[0,1]$ 区间。采用 Adam 优化器，其中判别器 D 的学习率设置为 1×10^{-4}，编码器 E、生成器 G 的学习率为 5×10^{-5}，$\beta_1=0$，$\beta_2=0.999$。

4.5.1 样本效率

样本效率(SE)指在有限的样本数量下模型的学习能力。为了评估这一效率，采用一个统计效率指标：以 100 个样本的测试准确率与全部样本(或 CelebA 数据集上的 10000 个样本)的测试准确率之比来衡量。该比值越高，说明模型在数据稀缺时的性能越接近于在数据充足时的表现，具有更强的数据利用能力。

为准确评估样本效率，所有方法均通过学习到的编码器将数据嵌入到潜在空间，并在这些表征之上训练一个多层感知器(MLP)分类器来预测目标标签，并且所有的模型均使用同一个 MLP 分类器进行评估。同时，在结果中给出了完整的测试准确率，以防止模型在少样本和全样本条件下都表现较差而导致效率评分被误解。表 5 列出了在 Penulum 数据集和 CelebA 数据集上，在各模型上测试准确率以及样本效率的实验结果，表中的平均值和标准差是基于 10 次重复实验的结果计算得出的。在这些指标中，数值越大，表示性能越优。

在 Pendulum 数据集上，SCD-VG 的样本效率比 DEAR 高约 5.56 个百分点，其中 100 个样本的测试准确率比 DEAR 高约 6.74 个百分点，整体表现最佳；在 CelebA 数据集上，100 个样本的测试准确率比 DEAR 高约 1.06 个百分点，尽管在 10000 个样本上不及 β-VAE 和 InfoMax-VAE，但在极低样本场景下仍保持最高准确率，因而样本效率显著优于对比方法。SCD-VG 所引入的因果建模与稀疏特征提取机制，使其在数据有限时更能识别任务相关的核心因子，从而实现高效学习；而 VAE 类模型在数据量充足时依赖其强建模能力，在测试准确率上略有优势，但其在低样本场景中无法充分激活潜在结构，因此样本效率偏低。

为进一步理解模型性能差异，具体分析 CSC 层与 SCF 层的作用。CSC 层通过稀疏性约束提升特征纯度，在小样本场景下能有效剔除冗余与噪声，使模型聚焦于核心因果因子，从而提升样本效率；在全样本场景中，数据量已足以覆盖噪声分布，即便不加 CSC 层也能学到较稳定的因果模式，因此其提升幅度有限。相比之下，SCF 层依托流模型捕捉潜在因果依赖，在大样本下优势更为明显；但在小样本条件下，由于其参数规模大且对输入敏感，即便 CSC 层改善了输入质量，这种复杂建模能力也难以完全发挥，导致增益有限甚至出现样本效率略低的情况。但这并不意味着协同效应被破坏，而是表明 CSC 与 SCF 的优势在不同数据规模下发挥程度有所差异：CSC 更适合小样本去噪，SCF 在大样本下更能建模复杂因果依赖。两者结合构成的 SCD-VG 框架，在多样因果场景中依然表现出整体的适应性与鲁棒性。

表 5 样本效率实验对比结果

Table 5 Comparison results of sample efficiency experiment

(%)

Model	Pendulum			CelebA		
	100	All	*SE*	100	10000	*SE*
VAE	$87.18_{\pm0.10}$	$90.38_{\pm0.15}$	$96.45_{\pm0.43}$	$84.15_{\pm0}$	$84.84_{\pm0.02}$	$99.18_{\pm0.03}$
β-VAE	$85.82_{\pm0.49}$	$89.30_{\pm0.10}$	$96.10_{\pm0.53}$	$81.08_{\pm0.15}$	$85.41_{\pm0.30}$	$94.93_{\pm0.81}$
InfoMax-VAE	$88.38_{\pm0.41}$	$90.34_{\pm0.08}$	$97.82_{\pm0.29}$	$83.93_{\pm0.50}$	**$86.26_{\pm0.01}$**	$97.30_{\pm0.63}$
ICM-VAE	$87.80_{\pm1.40}$	$90.64_{\pm0.06}$	$96.87_{\pm0.19}$	$76.70_{\pm0.06}$	$78.44_{\pm0.05}$	$97.78_{\pm0.28}$
DRL_{CET}	$85.71_{\pm1.06}$	$90.14_{\pm0.05}$	$95.09_{\pm1.21}$	$84.20_{\pm0.04}$	$85.04_{\pm0.07}$	$99.01_{\pm0.26}$
DEAR	$82.66_{\pm0.41}$	$88.80_{\pm0.04}$	$93.09_{\pm0.41}$	$83.62_{\pm0.01}$	$84.34_{\pm0.02}$	$99.15_{\pm0.06}$
DEAR+csc	$88.58_{\pm0.07}$	$90.30_{\pm0.03}$	$98.10_{\pm0.17}$	$84.18_{\pm0.03}$	$84.77_{\pm0}$	$99.31_{\pm0.06}$
DEAR+scf	$86.56_{\pm0.30}$	**$91.25_{\pm0.02}$**	$94.87_{\pm0.30}$	$84.13_{\pm0}$	$84.42_{\pm0.01}$	**$99.65_{\pm0.03}$**
SCD-VG	**$89.40_{\pm0.07}$**	$90.61_{\pm0.02}$	**$98.65_{\pm0.12}$**	**$84.68_{\pm0}$**	$85.07_{\pm0.02}$	$99.55_{\pm0.03}$

4.5.2 分布鲁棒性

为了评估因果表示的分布稳健性，操纵训练数据集在目标标签和图像的虚假属性之间注入虚假相关性。在 CelebA 数据集中构造目标标签并人为注入伪相关性。首先，若满足条件：年轻（是否年轻=1），女性（性别=0），没有脱发（是否脱发=0），化了妆（是否化妆=1），苗条（是否肥胖=0），没有眼袋（是否有眼袋=0），设 $\tau=1$，否则 $\tau=0$，从而生成一个依赖特定因果因子的二分类目标标签。随后，人为操控“微笑”与“张嘴”之间的关系：在训练集中，将 80% 的微笑样本强制设为张嘴、20% 设为不张嘴，制造出两者之间的强相关性；而在测试集中，将该比例调整为 50%/50%，打破训练时的伪相关。这样，在训练分布中“微笑”几乎等同于“张嘴”，而在测试分布中两者接近独立。在 Pendulum 数据集上，选择背景颜色∈{白色(0)，蓝色(1)}作为一个虚假特征。具体而言，在训练集中，当目标标签 $\tau=1$ 时，以 80% 的概率将背景设为蓝色；当 $\tau=0$ 时，则以 20% 的概率设为蓝色，从而在目标标签与背景颜色之间引入显著的相关性。相比之下，在测试集中，无论 τ 的取值为何，均以 50% 的概率将背景设为蓝色，使标签与背景颜色基本独立。以上人为注入虚假特征可以有效检验模型是否学会了与任务标签真正相关的因果特征，而非依赖训练集中存在的伪特征模式。在分布偏移的情况下，更有助于评估模型的鲁棒性。表 6 列出了各模型在因果解耦表征的分布鲁棒性（下游任务）中的表现。最坏的情况（‘TestWorst’）是，根据目标标签与虚假属性这两个二值变量将测试样本划分为 4 个子组，并选择其中准确率最低的一组作为评估指标。该子组通常具有与训练阶段相反的伪相关性，因此可有效反映模型在极端分布偏移下的表现。表中的平均值和标准差是基于 10 次重复实验的结果计算得出的，数值越高，表示性能越好。

实验结果如表 6 所列，SCD-VG 在 Pendulum 数据集上优势显著，其最坏情况下的性能接近平均性能，表明所学表征具有更强的泛化能力，对分布异常样本的敏感度更低。在 CelebA 数据集上，尽管 SCD-VG 在最坏情况下的表现最佳，但平均性能略低于 VAE，β-VAE 和 InfoMax-VAE 传统模型。

表 6 分布鲁棒性实验结果对比

Table 6 Comparison of distribution robustness experiment results

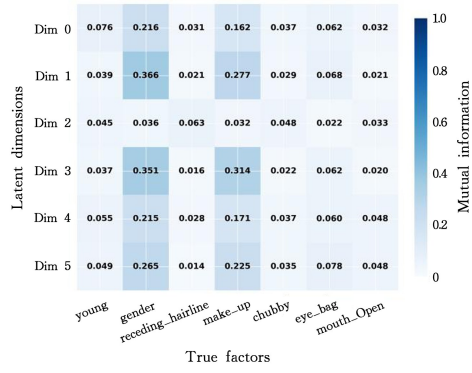
(%)

Model	Pendulum		CelebA	
	TestAvg	TestWorst	TestAvg	TestWorst
VAE	$58.68_{\pm 0.12}$	$23.47_{\pm 3.69}$	$85.05_{\pm 0}$	$80.52_{\pm 0.68}$
β-VAE	$60.67_{\pm 0.26}$	$32.33_{\pm 8.28}$	$85.30_{\pm 0.01}$	$81.55_{\pm 2.30}$
InfoMax-VAE	$60.53_{\pm 0.32}$	$31.75_{\pm 10.92}$	**$85.98_{\pm 0}$**	$81.33_{\pm 0.53}$
ICM-VAE	$52.13_{\pm 0.48}$	$28.66_{\pm 7.09}$	$78.44_{\pm 0}$	$68.25_{\pm 7.18}$
DRL$_{CET}$	$62.01_{\pm 0.31}$	$36.86_{\pm 12.79}$	$84.83_{\pm 0.01}$	$80.09_{\pm 3.74}$
DEAR	$61.50_{\pm 1.11}$	$39.16_{\pm 4.69}$	$84.50_{\pm 0.03}$	$81.97_{\pm 0.47}$
DEAR+csc	$63.72_{\pm 0.82}$	$48.44_{\pm 15.16}$	$84.68_{\pm 0.01}$	$81.78_{\pm 0.16}$
DEAR+scf	**$67.87_{\pm 0.30}$**	$61.12_{\pm 3.52}$	$84.73_{\pm 0}$	$82.04_{\pm 0.45}$
SCD-VG	$67.34_{\pm 0.19}$	**$62.14_{\pm 2.82}$**	$85.04_{\pm 0.02}$	**$82.86_{\pm 0.36}$**

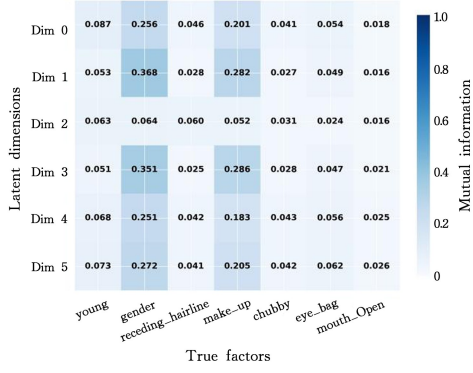
尽管 SCD-VG 的平均性能略低于 VAE，β-VAE 和 InfoMax-VAE 这些传统模型，但在最坏情况下表现最佳。这种“平均性能与鲁棒性之间的权衡”正是本文方法设计思想的体现。SCD-VG 并非单纯追求训练分布下的最优精度，而是通过因果结构建模与稀疏特征约束，实现对“因果一致性”的优先保持。一方面，卷积稀疏编码层通过稀疏性约束有效抑制高频伪相关特征，使模型更关注与目标标签存在稳定因果联系的语义因素，从而减少模型对伪特征（如 CelebA 中的“张嘴”等非因果线索）的依赖。另一方面，结构因果流层在潜空间中显式建模因果依赖关系，使得模型在分布偏移下仍能维持稳定的因果机制推理能力。

为直观佐证模型对伪相关特征的抑制效果，引入互信息矩阵图（见图 8）。该图量化了模型 6 个潜在维度与数据中 6 个真实生成因子及 1 个伪相关特征（张嘴）的信息关联强度：元素数值越接近 1，颜色越深，代表两者信息关联越强；元素数值越接近 0，颜色越浅，则关联越弱。对比可见，SCD-VG 与伪相关特征“张嘴”的互信息值显著小于 DEAR，印证了 SCD-VG 能有效抑制非因果信息干扰。

综上所述，SCD-VG 在分布鲁棒性实验中的表现证实了其设计的有效性：它通过内在的因果学习机制，主动摒弃对虚假特征的依赖，虽然可能轻微影响其在独立同分布测试中的平均性能，但显著提升了在分布偏移下的泛化能力与可靠性，这对于模型在真实世界中的安全部署至关重要。



(a)DEAR



(b)SCD-VG

图 8 CelebA 数据集互信息矩阵对比

Fig. 8 Comparison of mutual information matrices in CelebA dataset

4.6 MPI3D 数据集上的解耦性能

为全面检验所提模型 SCD-VG 在复杂多因子场景下的解耦表现，本文在 MPI3D 数据集与 CelebA 数据集上分别设计实验，引入互信息系数（MIC）与总信息含量（TIC）两项指标进行量化评估：前者重点验证模型在多生成因子场景下的解耦与跨域（真实 / 模拟数据）迁移能力，后者则聚焦高分辨率差异场景下的表征可迁移性。

1)MPI3D 实验设置与结果分析

实验设置 0.01% 的样本带有真实因子标签(约 100 个),其余样本标签置为 −1,训练使用 80%的数据,测试阶段选取剩下 20%作为测试集。在评估阶段,首先在整个测试集上提取潜在表示并进行标准化,然后计算 MIC 与 TIC 指标,以确保指标计算的有效性和可比性。所有因子值均归一化至区间[0,1]。

在真实数据 MPI3D-real 和模拟数据 MPI3D-realistic 上进行训练。模型先验设定为各因子独立的标准高斯分布,因果结构矩阵初始化为零矩阵,以符合"无先验因果假设"。编码器采用 ResNet-18;优化器统一为 Adam($\beta_1=0$,$\beta_2=0.999$),学习率与动量参数保持一致。训练与评估流程、超参数及随机种子均与基线方法 DEAR 严格对齐,确保公平比较。

实验部分选择当前较为先进的解耦表示方法 DEAR 作为对比基线,并在相同的训练与评估条件下计算两种指标的结果。表 7 列出了 SCD-VG 与 DEAR 在 MIC 和 TIC 上的性能对比。可以看到,在 MPI3D-real → MPI3D-realistic 迁移任务中,SCD-VG 的 TIC 略低于 DEAR。分析认为,这一差异主要源于以下两点:1)两个域在纹理、光照等视觉特征上存在明显差异,导致 SCD-VG 的稀疏编码层在迁移时可能对部分共享信息产生过度抑制,从而降低总体信息保留;2)真实域与模拟域的因果机制存在差异,SCD-VG 的非线性因果结构在迁移时需要重新适配,短期内会带来轻微的性能损失。但整体结果表明,SCD-VG 在 MIC 和 TIC 指标上整体优于 DEAR,说明 SCD-VG 的解耦效果更优,泛化迁移能力更强。

表 7 MPI3D-realistic 和 MPI3D-real 数据集在 MIC 和 TIC 上的实验对比结果

Table 7 Experimental comparison results of the MPI3D-realistic and MPI3D-real datasets in MIC and TIC

Train	Test	Model	MIC/%	TIC/%
MPI3D-realistic	MPI3D-realistic	DEAR	90.50	58.59
		SCD-VG	**92.82**	**72.98**
MPI3D-realistic	MPI3D-real	DEAR	87.89	51.81
		SCD-VG	**92.46**	**71.26**
MPI3D-real	MPI3D-real	DEAR	91.93	**67.09**
		SCD-VG	**93.88**	60.28
MPI3D-real	MPI3D-realistic	DEAR	86.48	**64.12**
		SCD-VG	**87.64**	55.71
CelebA (64×64)	CelebA (1024×1024)	DEAR	90.23	81.45
		SCD-VG	**92.51**	**85.33**

2)CelebA 实验与跨分辨率验证

为进一步验证模型在真实高维图像数据上的迁移与鲁棒性表现,本文在 CelebA 数据集上开展了从低分辨率(64×64)到高分辨率(1024×1024)的迁移实验。该任务能够更好地模拟真实视觉场景中由采样密度变化或感知尺度变化引起的分布差异,考察模型在高分辨率生成与语义一致性保持方面的能力。

具体而言,CelebA(64×64)作为训练集,共包含 162 770 张图像;CelebA(1024×1024)则包含 30 000 张高分辨率人脸图像,用于测试阶段的跨分辨率迁移与生成评估。训练与测试均采用相同的优化设置(Adam 优化器,$\beta_1=0$,$\beta_2=0.999$),并与基线方法 DEAR 保持一致,以确保公平性。

实验结果表明,SCD-VG 在 CelebA 任务中的整体表现优于 DEAR,进一步验证了 SCD-VG 在复杂真实数据场景下的稳健性与普适性。

4.7 计算效率分析

为全面评估 SCD-VG 的实用价值,本节从训练时间、显存占用以及参数量 3 个核心维度,在 CelebA 人脸数据集上与 DEAR 模型进行针对性量化对比。实验环境统一为 Nvidia GeForce RTX 3090 GPU 及 PyTorch 框架,所有模型均采用相同的批量大小(128)、训练轮数(200 轮)及优化器配置(Adam,$\beta_1=0$,$\beta_2=0.999$),确保对比的公平性。

选择 CelebA 人脸数据集作为计算效率评估的唯一对象,主要基于以下 3 点核心考量,确保评估结果的场景针对性与实用参考价值:人脸数据的复杂特性与实际应用匹配度高——CelebA 包含 20 万张高分辨率(64×64)人脸图像,涵盖 40 个离散属性,其数据维度、特征冗余度(如背景纹理、光照变化)与真实世界人脸分析场景(如身份识别、表情迁移)高度一致;相比 Pendulum(低维合成物理数据)、MPI3D(结构化物体数据),人脸数据的特征分布更复杂、语义因子更细腻,更能反映模型在高维真实数据场景下的计算效率;而人脸分析恰是因果解耦表征学习的核心应用领域,此场景下的效率评估对工程落地更具指导意义。

对比结果如表 8 所列,结果显示,SCD-VG 在训练时长、参数量以及显存占用方面均优于 DEAR。尽管计算成本略有上升,但 SCD-VG 在表征可解释性、因果干预准确性及分布鲁棒性等关键指标上显著高于 DEAR。这一结果表明,适度的计算代价换取更高的结构可解释性与稳定性是合理且必要的,尤其在面向高维复杂数据(如人脸场景)的实际应用中。

表 8 训练参数对比

Table 8 Comparison of training parameters

模型	训练时长/h	轮数	参数量	占显存量/GB
DEAR	60.98	200	Model:3.724×10^7 Discriminator:1.359×10^7	6.38
SCD-VG	68.35	200	Model:3.748×10^7 Discriminator:1.359×10^7	6.74

结束语 本文提出了一种融合稀疏编码与因果推断的解耦表征学习框架 SCD-VG,可有效缓解现有解耦表征学习方法在捕捉复杂因果关系方面的局限,但仍存在一些不足。首先,在验证范围方面,当前实验主要基于因果结构明确的合成数据集和属性标注齐全的真实数据集。未来工作将验证场景扩展至更具挑战性的现实环境,如动态视频序列、多模态医疗数据等,以评估模型在复杂、开放环境中的泛化能力。其次,在先验依赖方面,模型对因果图结构的先验知识仍有一定依赖。后续研究可探索因果发现与表征学习的深度融合,发展能够从数据中联合推断因果结构并学习解耦表征的新方法,降低对先验知识的依赖。最后,在应用广度方面,当前框架主要面向静态图像数据。未来的重要方向是将 SCD-VG 的核心思想拓展至时间序列分析、自然语言处理等领域,研究适用于序列数据的因果解耦表征学习新范式。

参 考 文 献

[1] CHENG K Y, MENG C Y, WANG W S, et al. Research Progress in Disentangled Representation Learning[J]. Journal of Computer Applications, 2021, 41(12): 3409-3418.

[2] HIGGINS I, AMOS D, PFAU D, et al. Towards a definition of disentangled representations[J]. arXiv: 1812.02230, 2018.

[3] KINGMA D P, WELLING M. Auto-encoding variational Bayes [J]. arXiv: 1312.6114, 2013.

[4] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680.

[5] SIKKA H D. A Deeper Look at the Unsupervised Learning of Disentangled Representations in β-VAE from the perspective of Core Object Recognition[M]. Harvard University, 2020.

[6] PEARL J. Models, reasoning and inference[J]. Cambridge, UK: Cambridge University Press, 2000, 19(2): 3.

[7] YANG T S. Disentangled Representation Learning Based on Causal Inference [D]. Qingdao: Qingdao University of Science and Technology, 2023.

[8] SCHÖLKOPF B, LOCATELLO F, BAUER S, et al. Toward causal representation learning[J]. Proceedings of the IEEE, 2021, 109(5): 612-634.

[9] WEN Z D, WANG J R, WANG X X, et al. A Survey on Disentangled Representation Learning [J]. Acta Automatica Sinica, 2022, 48(2): 351-374.

[10] HIGGINS I, MATTHEY L, PAL A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework [C]// International Conference on Learning Representations. 2017.

[11] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2016.

[12] CHEN R T Q, LI X, GROSSE R, et al. Isolating sources of disentanglement in VAEs[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 2615-2625.

[13] REZAABAD A L, VISHWANATH S. Learning representations by maximizing mutual information in variational autoencoders [C]// 2020 IEEE International Symposium on Information Theory(ISIT). IEEE, 2020: 2729-2734.

[14] LOCATELLO F, BAUER S, LUCIC M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations [C] // International Conference on Machine Learning. PMLR, 2019: 4114-4124.

[15] SUTER R, MILADINOVIC D, SCHÖLKOPF B, et al. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness[C]// International Conference on Machine Learning. PMLR, 2019: 6056-6065.

[16] REDDY A G, BALASUBRAMANIAN V N. On causally disentangled representations[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 8089-8097.

[17] YANG M, LIU F, CHEN Z, et al. Causalvae: Disentangled representation learning via neural structural causal models[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9593-9602.

[18] SHEN X, LIU F, DONG H, et al. Weakly supervised disentangled generative causal representation learning[J]. Journal of Machine Learning Research, 2022, 23(241): 1-55.

[19] KOMANDURI A, WU Y, CHEN F, et al. Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms[J]. arXiv: 2306.01213, 2023.

[20] YOU D, LI Z, SHEN J, et al. Disentangled representation learning with causal effect transmission in variational autoencoder [J]. Pattern Recognition, 2026, 170: 112018.

[21] NASR-ESFAHANY A, ALIZADEH M, SHAH D. Counterfactual identifiability of bijective causal models[C]// International Conference on Machine Learning. PMLR, 2023: 25733-25754.

[22] LI M, ZHAI P, TONG S, et al. Revisiting sparse convolutional model for visual recognition[J]. Advances in Neural Information Processing Systems, 2022, 35: 10492-10504.

[23] PAPAMAKARIOS G, NALISNICK E, REZENDE D J, et al. Normalizing flows for probabilistic modeling and inference[J]. Journal of Machine Learning Research, 2021, 22(57): 1-64.

[24] KHEMAKHEM I, MONTI R, LEECH R, et al. Causal autoregressive flows[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2021: 3520-3528.

[25] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 3730-3738.

[26] GONDAL M W, WUTHRICH M, MILADINOVIC D, et al. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset[C]// Advances in Neural Information Processing Systems. 2019.

[27] KINNEY J B, ATWAL G S. Equitability, mutual information, and the maximal information coefficient[C]// Proceedings of the National Academy of Sciences. 2014: 3354-3359.

HUANG Beibei, born in 2001, master candidate, is a student member of CCF (No. A06163G). Her main research interests include computer vision and representation learning.

LIU Jinfeng, born in 1971, Ph.D, professor, is a professional member of CCF (No. 75718M). His main research interests include deep learning and heterogeneous computing.

（责任编辑：何杨）