计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010131] [CSTR: 32024.14.csa.010131]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合反事实语义增强与因果注意力的领域泛化①

魏成亮, 刘进锋

(宁夏大学 信息工程学院, 银川 750021)
通信作者: 刘进锋, E-mail: jfliu@nxu.edu.cn

摘　要: 针对深度学习模型在分布偏移场景中泛化能力不足的问题, 在 Mamba 状态空间模型的基础上, 提出一种融合反事实语义增强和因果注意力机制的领域泛化方法, 通过设计反事实语义增强模块, 实现前景-背景解耦与重组生成反事实特征, 显式构建“前景保持、背景干预”的因果情境, 有效削弱背景-标签的伪相关性, 强化模型对因果语义前景的挖掘能力, 引导其关注稳定可靠的语义关联; 进一步提出因果注意力机制, 将上述模块提取到的因果语义信息显式嵌入 Mamba 状态更新过程, 以提高特征的因果一致性. 整体模型结构实现了对前景与背景信息的动态区分与融合. 在标准领域泛化基准上的实验结果表明, 本文方法在 PACS、OfficeHome、VLCS 和 TerraIncognita 数据集上平均准确率分别达到 91.9%、77.0%、81.1% 和 54.9%, 均优于现有 SOTA 方法, 证实本文方法显著提高了模型对前景语义区域的关注一致性, 展现出优越的可解释性与泛化性能.

关键词: 领域泛化; 反事实语义增强; 因果注意力机制; 状态空间模型; Mamba; 可解释性

引用格式: 魏成亮,刘进锋.融合反事实语义增强与因果注意力的领域泛化.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10131.html

Domain Generalization via Counterfactual Semantic Enhancement and Causal Attention

WEI Cheng-Liang, LIU Jin-Feng

(School of Information Engineering, Ningxia University, Yinchuan 750021, China)

Abstract: To address the limited generalization capability of deep learning models under distribution shifts, this study proposes a domain generalization method based on the Mamba state-space model that integrates counterfactual semantic enhancement with a causal attention mechanism. By designing a counterfactual semantic enhancement module, foreground-background decoupling and recombination are achieved to generate counterfactual features, explicitly constructing a causal scenario of “foreground preservation and background intervention”. This effectively mitigates spurious background-label correlations, enhances the model’s ability to extract causal semantic foreground representations, and guides it to focus on stable and reliable semantic associations. Furthermore, a causal attention mechanism is introduced to explicitly embed the causal semantic information extracted by the module into the Mamba state update process, improving the causal consistency of features. The overall architecture enables dynamic discrimination and integration of foreground and background information. Experimental results on standard domain generalization benchmarks demonstrate that the proposed method achieves average accuracy rates of 91.9%, 77.0%, 81.1%, and 54.9% on the PACS, OfficeHome, VLCS, and TerraIncognita datasets, respectively, outperforming existing state-of-the-art methods. These results confirm that the proposed method significantly improves the consistency of the model’s focus on foreground semantic regions, thus demonstrating superior interpretability and generalization performance.

Key words: domain generalization (DG); counterfactual semantic enhancement; causal attention mechanism; state space model (SSM); Mamba; interpretability

① 基金项目: 宁夏自然科学基金 (2025AAC030154)
收稿时间: 2025-09-09; 修改时间: 2025-10-09, 2025-11-11; 采用时间: 2025-11-24; csa 在线出版时间: 2026-03-02

随着深度学习技术的发展, 各类视觉识别模型在源域 (source domain) 上取得了优异的性能, 但在跨领域视觉任务中面对分布外 (out-of-distribution, OOD) 数据时往往表现不佳[1]. 由于训练集与测试集在图像风格、拍摄条件、背景分布等方面存在差异, 模型常因背景风格与前景类别之间的虚假共现 (如“飞机常出现在天空”) 学习到伪相关的背景偏倚特征, 导致分布外性能显著下降. 领域泛化 (domain generalization, DG) 研究旨在解决这一问题, 使模型在未见过的目标域 (unseen domain) 中具有良好的泛化能力[2]. DG 研究者针对不同的数据分布和视觉场景提出了多种方法, 如数据增强、域对齐和元学习等策略, 并在人脸、医学等领域取得了一定成果[3]. 然而, 传统方法往往只考虑了概率层面的特征对齐, 缺乏对语义因果关系的建模[4]. 例如, 在视觉分类任务中, 船舶的背景通常是湖面, 这种背景和目标高相关性会使模型过于依赖背景信息进行判别, 而当背景改变时 (如普通路面), 模型性能急剧下降. 与此不同的是, 人类可以忽略背景干扰, 仅凭船舶形状等因果特征进行识别, 这揭示了因果语义在泛化中的重要性.

近年来, 状态空间模型 (state space model, SSM)[5] 在视觉领域的兴起, 为代替自注意力机制提供了新的可能. 作为其中的代表, Mamba 模型以线性时间复杂度实现长距离依赖建模, 并通过选择性状态更新机制 (selective state update) 实现了序列特征的动态传播, 这种结构中的状态传递过程可被视为隐式的因果传递链. 然而, 现有的 Mamba 模型多聚焦于效率和表达能力的提升, 缺乏对语义层面因果关系的显式建模与利用. 在领域泛化任务中, 不同域间的风格转移、背景偏倚等因素常破坏这种潜在的状态因果性, 导致模型学习到非稳健的伪相关特征. 如何让 Mamba 从时序建模结构转化为因果驱动结构, 是当前尚未解决的重要挑战. 基于状态空间模型的 DGMamba[6]方法引入了语义感知 (semantic prompt rebuilding, SPR) 模块, 通过特征采样增强和领域不变分类器增强等机制提升了模型对语义区域的关注, 然而该类方法在本质上仍未摆脱对特定提示结构的依赖, 缺乏对“语义因果结构”的显式建模能力, 导致其在强分布偏移场景下性能仍受限制. Wang 等[7]提出基于非线性因果框架的平衡小批量采样方案, 既为领域泛化提供了“从因果理论到工程实现”的完整链路, 又通过轻量采样策略规避了传统方法对模型结构的强约束, 为后续复杂场景下的泛化方法设计提供了新范式. 另一方面, 反事实学习与因果推理[8]作为近年来在视觉表示学习中逐渐兴起的新范式, 提供了一种剥离背景干扰、挖掘稳定语义因果关系的有效路径. 通过在特征或图像层面构造前景不变而替换背景的反事实样本, 模型得以学习稳定的领域不变特征. 已有研究表明, 该类方法在跨域分类、视觉问答和零样本识别等任务中显著提升了泛化性能, 但这类方法多停留在样本生成层面或损失约束层面, 并未深入到主干网络的结构层, 仍依赖外部模块实现因果干预. 因此, 反事实学习与深度结构设计之间的协同关系尚未被系统探索. 鉴于上述问题, 本文的核心动机在于将反事实干预思想与 Mamba 的状态建模机制有机融合, 构建结构层与语义层的双重因果约束体系. 该融合思路不仅在理论上体现了因果传播与特征建模的统一, 也在实践中显著提升了模型的泛化性能和可解释性. 具体而言: 在语义层面, 设计了反事实语义增强模块 (counterfactual feature module, CF-module), 借助生成与原始样本语义一致但背景或风格不同的对抗样本, 引导模型学习具有因果稳定性的目标语义特征; 在结构层面, 我们在 Mamba 状态空间模型中引入了因果注意力机制, 建模模块 Causal-Mamba, 将反事实语义增强模块中的前景语义权重引入到 Mamba 模块的状态更新路径中, 引导特征传播过程中更关注领域不变的前景区域.

本文的主要贡献包括: (1) 提出反事实语义增强模块, 通过前景-背景特征解耦与反事实特征合成, 对比训练下拉近原始样本与其语义变换后的表示距离, 增强模型对核心语义特征的不变性, 从而提升对场景内在语义的捕获能力; (2) 引入因果注意力机制的 Mamba 模型, 结合 Mamba 状态空间模型的全局记忆特性和线性计算优势, 对隐藏状态进行注意力加权, 过滤领域特异信息, 实现对语义与领域变化之间因果关系的显式建模, 提高模型的语义可解释性和泛化能力; (3) 通过丰富的实验验证所提方法, 结果表明本文方法面对未知领域时具备良好的跨域泛化能力, 在多个领域泛化广泛使用的基准数据集上性能优于现有 SOTA 方法.

1 相关工作

1.1 领域泛化研究

领域泛化的目标是学习因果不变特征, 解决训练与测试域之间的分布偏移问题, 使模型能够在未见过

的目标域上保持良好的性能. 数据增强方法通过生成新样本扩展数据多样性, Yan 等[9]提出集成域内与域间混合训练的无监督域自适应框架, 通过 Mixup 构建跨域插值样本与虚拟标签, 解决单独在源域或目标域施加约束、忽略域间交互的问题. Zhou 等[10]通过混合两个随机实例的特征统计量来合成特征空间中的新域以实现数据增强, 有效提升模型的分布外泛化性能. 表示学习方法旨在学习到一个更好的特征提取函数来提取域间不变的特征表示, Ganin 等[11]提出领域对抗神经网络 (DANN), 一种用于无监督领域自适应的表示学习方法, 通过在神经网络中引入梯度反转层, 让特征提取器学习既对源域标签具有判别性, 又无法区分源域与目标域的领域不变特征. Arjovsky 等[12]提出不变风险最小化 (IRM) 范式, 通过从多训练环境中学习跨环境稳定的因果预测器, 寻找使最优分类器在各环境一致的数据表示, 为因果与泛化结合提供新方向. Blanchard 等[13]提出了 MTL 方法, 通过用特征向量的边际分布来增强原始特征空间. Nam 等[14]提出了 SagNet 方法, 将风格编码从类别中分离出来, 从而避免了预测的风格偏差, 并更加关注内容而非风格以缩小领域之间的差距. Krueger 等[15]提出了 VREx 方法, 通过对训练风险方差的惩罚, 适当权衡对因果关系引起的分布变化和协变量变化的稳健性, 可以恢复目标域的因果机制, 同时还能提高对于输入分布偏移的鲁棒性. 此外, Zhang 等[16]提出的 ARM 方法、Li 等[17]提出的 MLDG 框架、Cha 等[18]提出的 SWAD 方法、Wang 等[19]提出的 SAGM 方法都为领域泛化研究做出了贡献.

由于上述方法以卷积神经网络 (CNN) 作为特征提取主干结构, 而 CNN 对全局上下文信息的感知能力不足势必会影响模型的泛化能力, 为此, Sultana 等[20]提出 ERM-SDViT 方法, 将基于视觉 Transformer 的模型引入领域泛化, 利用其全局感受野保证对全局信息的捕捉. Li 等[21]提出的 GMoE 模型结合了稀疏混合专家 (sparse MoEs) 和视觉 Transformer, 旨在更好地与不变相关性对齐, 从而提高领域泛化性能. 由于 Transformer 模型存在计算复杂度高、参数量大的问题, Long 等[6]基于线性时间复杂度的 Mamba 提出了 DGMamba 框架用于领域泛化任务, DGMamba 首次将状态空间建模引入 DG 任务, 并通过 SPR 模块引导模型聚焦于前景区域. 虽然该方法在多个数据集上取得了良好效果, 但 SPR 模块依赖人工构造提示的方式在语义泛化能力和结构设计灵活性方面仍存在限制. 综上, 现有 DG 方法在提升模型泛化能力方面取得了一定进展, 但仍缺乏对语义因果关系的显式建模, 特别是在主干结构中对领域不变语义的关注.

1.2 状态空间建模与 Mamba 结构

状态空间模型 (SSM) 是一类用于建模时间序列中隐状态动态变化的经典方法, 在序列建模中具有全局感受野和线性计算优势, 随着结构化序列建模技术的发展, SSM 被引入深度学习框架[5]. 受到经典的状态空间模型启发, 近年来提出的 Mamba 架构是一种兼具线性时间复杂度和全局依赖建模能力的状态空间结构, 在语言建模、图像分类、语音识别等任务中取得了与 Transformer 相当甚至更优的性能, 其核心在于通过对连续输入的特征序列进行状态转移更新. DGMamba 是最早将 Mamba 引入领域泛化任务的工作之一, 其整体结构以 Mamba 为主干网络, 通过 SPR 模块增强语义区域建模能力, 在多个跨域数据集上超过了基于 Transformer 的 SOTA 方法. 然而, 其 Mamba 主干仍采用标准结构, 未利用语义信息引导状态传播路径, 对前景语义的强化依赖 SPR 路径, 结构表达能力仍较为有限. 因此, 如何在 Mamba 状态传播路径中引入语义偏好机制, 增强其对领域不变前景区域的建模能力, 仍是当前尚未充分研究的重要问题.

1.3 反事实学习与因果推理

因果推理理论在弱监督学习、表示学习、领域泛化等任务中展现出强大潜力, 它引导模型关注稳定因果特征, 降低对虚假相关性特征的依赖. Tang 等[22]提出基于因果学习的全流程领域泛化框架, 将因果干预融入训练阶段. Wang 等[23]提出基于因果的对比增量学习框架, 从因果和增量双视角出发提取域内与域间不变知识. 反事实学习强调“前景语义保持及背景扰动”对于泛化的重要性, 通过构造与真实世界事件相对立的“如果…, 则…”情景, 能够打破训练集的共现偏倚并更直接地估计因果效应. Shao 等[24]在弱监督目标定位任务中提出的反事实共现学习 (CCL) 通过前景、背景分离并合成组合式反事实样本来抑制共现背景的影响, 为视觉任务的因果增强提供了有效范式. 这启发本文在分类与 DG 任务中采用类似思路, 但不同于其主要面向定位任务的结构, 我们将反事实合成与 Mamba 状态传播相结合, 从数据层与模型层双重路径同时引导模型学习领域不变的因果语义特征.

2 方法

本文提出的领域泛化框架包含两个关键模块: 反事实语义增强模块和融合因果注意力机制的 Mamba 建模模块, 整体结构如图 1 所示. 其中, 反事实语义增强模块负责在样本层面进行因果干预, 通过前景保持与背景替换生成反事实样本, 以模拟潜在的未见域分布; 融合因果注意力机制的 Mamba 模块则在主干网络的状态传播路径中引入因果注意权重, 实现领域不变特征的结构级建模. 整体训练过程包括 3 个阶段: 输入图像首先通过特征提取网络得到语义表示; 然后, 反事实语义增强模块通过前景-背景解耦与特征级反事实样本生成, 引导模型关注因果稳定因素, 减少对共现偏倚信息的依赖; 最后, 融合因果注意力机制的 Mamba 模块则根据反事实语义模块产生的前景语义特征, 提出一种基于因果语义的隐藏状态抑制模块, 在 Mamba 原始状态空间建模路径中引入因果语义注意机制, 引导特征演化过程中优先保留领域不变的前景语义内容. 下文分别介绍两个模块的设计.

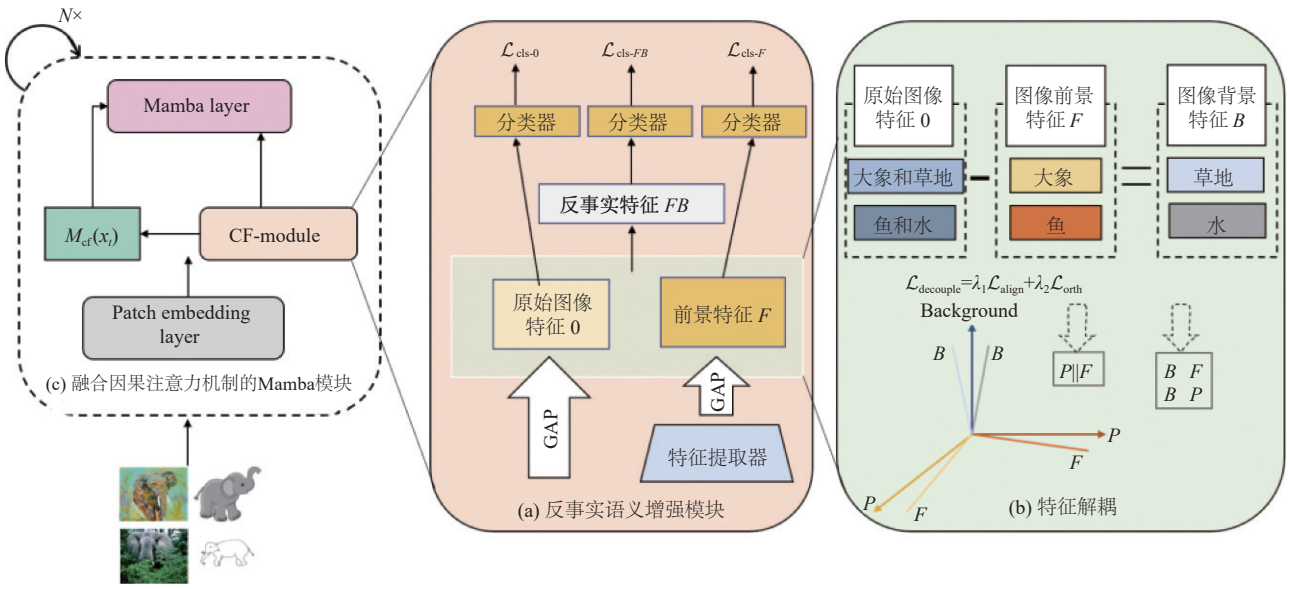


图 1 本文方法的整体架构图

2.1 反事实语义增强模块

DGMamba 原始框架中通过 SPR 模块实现语义感知增强, 图 2 进一步展示了 SPR 模块结构, 通过将当前样本的上下文特征替换为其他域的上下文特征来生成反事实样本, 其核心在于人工设计的提示结构, 根据 Grad-CAM[25]得分区分语义因果特征和语义无关上下文特征. 随后, 将目标样本的上下文特征与来自其他训练域的上下文特征进行替换, 构造反事实样本, 实现跨域语义增强的效果, 从而在特征层进行语义注入. 这种机械式的、无法让模型自己学习的提示结构对领域泛化能力的提升存在局限.

为此, 本文提出一种反事实语义增强模块, 该模块借助特征级反事实构造策略, 显式建模“在背景变化下语义保持不变”的因果假设, 从而引导模型学习领域不变的核心语义特征. 反事实语义增强模块主要包含 3 个阶段: ① 前景-背景特征解耦: 从输入图像特征中分离语义相关 (前景) 与语义无关 (背景) 部分; ② 反事实特征合成: 随机重组前景与背景特征, 构造具有因果干预意义的反事实样本; ③ 语义保持约束: 对合成样本施加标签一致性损失, 提升判别器鲁棒性, 逼迫网络对改变背景的不变前景做出正确分类. 该模块可以无缝替换 DGMamba 原有的 SPR 模块, 并在训练中与主干联合反向传播以学习稳健语义表示. 模块结构如图 1(a) 所示, 每一步具体介绍如下.

(1) 特征解耦

给定输入图像 I, 经主干网络提取出特征 $O\in\mathbb{R}^d$. 为了从 O 中提取语义前景, 本文引入一个轻量级前景特征提取器 $f_{\text{fg}}(\cdot)$, 计算前景特征 F:

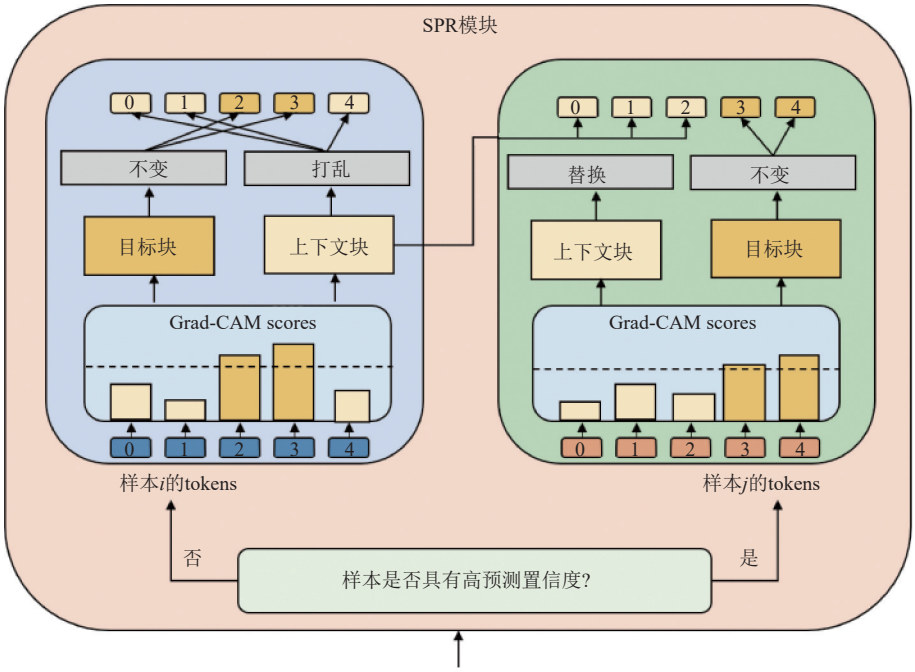


图 2 SPR 模块结构图

$$F = f_{\text{fg}}(O) \tag{1}$$

然后, 通过残差方式计算背景特征 B:

$$B = O - F \tag{2}$$

如图 1(b) 所示, 前景特征 F 被约束与对应类别的原型向量 P 对齐, 前景对齐损失:

$$\mathcal{L}_{\text{align}} = \| F - P_y \|^2 \tag{3}$$

背景特征 B 被约束与前景特征 F 及类别原型向量 P 均保持正交, 以减少其与语义中心的重合, 背景与前景和语义原型正交性损失:

$$\mathcal{L}_{\text{orth}} = \| B^{\top} \cdot F \|^2 + \| B^{\top} \cdot P_y \|^2 \tag{4}$$

其中, P_y 为当前类别 y 的语义原型向量 (可直接使用分类器权重向量替代), 最终得到解耦损失:

$$\mathcal{L}_{\text{decouple}} = \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{orth}} \tag{5}$$

其中, λ_1、λ_2 为超参数.

(2) 反事实样本构造

为了模拟“前景保持不变、背景发生干预”的因果假设, 随机重组前景与背景生成反事实样本, 如图 3 所示. 其中, $f_1, f_2 \in F$ 为前景特征, $b_1, b_2 \in B$ 为背景特征, FB 为构造出的反事实样本, $y_1, y_2 \in Y$ 为图像标签.

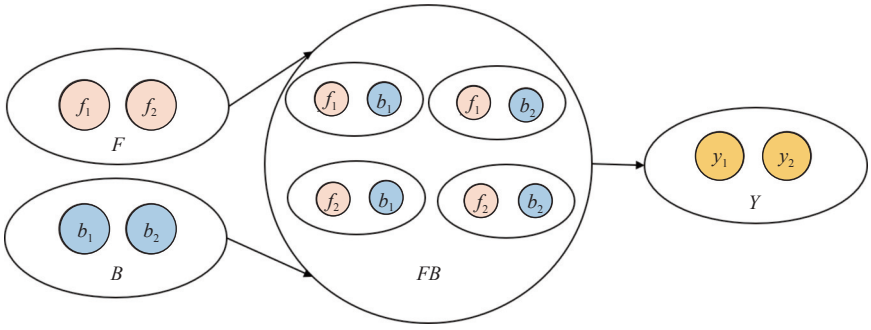


图 3 构造反事实样本示意图

本文在每个训练 batch 中构造反事实样本:

$$FB_{i,j} = F_i + B_j \tag{6}$$

其中, $i \neq j$, 即将样本 i 的语义前景与样本 j 的背景组合, 构造反事实样本 $FB_{i,j}$, 并赋予其与 F_i 相同的类别标签 y_i. 这一过程可理解为特征空间中的背景置换, 能够破除模型对乱真背景的偏倚依赖.

(3) 训练目标函数

反事实语义增强模块的训练目标综合了原始特征判别、前景判别和反事实判别这 3 种路径, 对于原始样本、纯前景样本与反事实样本, 分别计算分类损失, 同时加入解耦结构的正则项:

$$\mathcal{L}_{\mathrm{CF}} = \mathcal{L}_{\mathrm{ce}}(O,y) + \mathcal{L}_{\mathrm{ce}}(F,y) + \mathcal{L}_{\mathrm{ce}}(FB,y) + \mathcal{L}_{\mathrm{decouple}} \tag{7}$$

其中, $\mathcal{L}_{\mathrm{ce}}$为标准交叉熵损失.

反事实语义增强模块所引入的反事实合成机制可以显式模拟分布偏移, 通过背景干预实现领域变化的近似建模, 且无需额外的数据或标签, 仅依赖样本内部组合即可实现语义增强. 此外, 还具有较强的可替换性和较好的兼容性.

2.2 因果注意力机制

Mamba 模块以线性时间复杂度对长序列特征进行建模, 其核心为状态变量的迭代更新与输入驱动函数的卷积形式. 在视觉领域泛化任务中, 输入图像可能包含大量与类别标签无关但与领域高度相关的噪声背景(如风格、色彩、纹理等), 若无语义区分机制, Mamba 仍可能在状态传播过程中无差别放大这些领域特有特征, 影响模型泛化能力. 为此, 本文提出融合因果注意力机制的 Causal-Mamba 模块, 将反事实模块中获得的前景语义注意权重引入 Mamba 的状态更新路径, 构建具备因果引导能力的因果隐藏状态抑制模块, 动态调整输入特征在状态空间中的传播强度, 提升对前景区域建模的优先级. 如图 1(c) 所示, 在标准 Mamba 状态更新过程中, 特征输入x_t被卷积驱动后传入状态更新:

$$h_{t+1} = Ah_t + Bx_t \tag{8}$$

$$y_t = Ch_t \tag{9}$$

其中, h_t为当前状态, x_t为输入特征, $A \in \mathbb{R}^{N\times N}$, B、$C \in \mathbb{R}^N$为原始 Mamba 架构的状态转移矩阵, N表示状态大小. 该形式在本质上可视作一种隐式的“因果传播过程”: 输入特征通过递归状态更新不断影响后续的预测输出. 若将x_t理解为不同语义因子的混合表达, 则状态更新过程实际上对应语义因子间的潜在因果传递. 然而, 标准 Mamba 模型未显式区分领域不变语义与领域特有噪声两类因子, 从而容易在分布偏移下学习到伪相关特征. 为实现领域不变特征的稳定传播, 本文引入因果注意力权重, 用于在状态更新过程中对不同语义因子赋予差异化的传递强度. 由反事实语义增强模块计算每个输入通道的前景语义注意权重向量 $M_{\mathrm{cf}}(x_t) \in [0,1]^d$, 表示每一维特征通道的重要性, 并将其作为因果掩码对状态更新过程加权修正, 构建新的状态更新形式. 为此, 首先给出因果传播的基本表达形式如下:

$$\overline{C} = C \odot M_{\mathrm{cf}}(x_t) \tag{10}$$

$$y_t = \overline{C}h_t \tag{11}$$

其中, 符号⊙表示哈达玛积即逐元素乘法操作, $M_{\mathrm{cf}}(x_t)$是经过前景-背景解耦与前景强化后生成的注意力掩码, 用于强调图像中的语义因果区域. 该机制实质上构建了一种因果注意门控结构, 在特征传播阶段动态抑制背景噪声信息传播, 加强对因果前景区域的建模能力. 引入因果注意权重$M_{\mathrm{cf}}(x_t)$使模型在状态传播中近似实现了对域特异因素的干预消除, 从而在结构层面达到消除伪相关信息的效果. 通过该机制, 状态传播路径中的语义因子将根据其因果贡献度被重新加权, 使网络在隐状态演化过程中自动抑制域特异性特征的干扰. 为增强训练稳定性, 我们保留标准 Mamba 的残差连接机制, 将 Causal-Mamba 模块的输出表示为:

$$z_t = x_t + CausalMamba(x_t) \tag{12}$$

其中, $CausalMamba(x_t)$表示引入注意力机制后的状态空间建模模块. 该设计能动态调节各输入维度在状态传播中的影响力, 有效抑制随传播放大的领域特异信息. 式 (10) 和式 (11) 表示依据反事实建模对原语义响应进行校正, 利用掩码$M_{\mathrm{cf}}(x_t)$隔离背景干扰信息, 对应图 1(c) 中因果注意力模块的显式选择机制; 式 (12) 实现了去偏特征的动态融合, 由此得到最终的语义响应. 综上, 式 (10)–(12) 共同构成了因果驱动的语义特征更新机制, 实现从相关性向因果性特征表达的过渡.

当前主流的因果领域泛化方法主要从损失函数或样本层面实现因果约束, 例如通过不变风险最小化或样本重组来学习稳定特征. 然而, 这些方法均未将因果机制嵌入模型的结构更新中, 因此, 在特征传播过程中仍可能受到域特异噪声的累积影响. 与之不同, 本文方法将因果注意权重直接引入状态更新方程, 使 Mamba 的隐状态演化具备显式的因果传播意义. 这一改进提升了 Mamba 在领域泛化任务中的理论解释性, 融合了因果注意力机制的 Causal-Mamba 模块通过因果引导建模路径, 以掩码显式增强前景区域建模, 避免信息平均扩散; 模块通用性强, 可作为任意 Mamba 层的增强插件, 具备较强可扩展性; 增强了模型可解释性, 可以提高模型关注区域与前景重合度. 最后, 上述两个模块

协同工作, 分别作用于样本生成路径与主干特征提取路径, 共同促进判别器在训练过程中形成对领域不变语义的更强依赖, 从而提升整体领域泛化性能.

3 实验分析

3.1 实验细节

我们在以下 4 个经典领域泛化数据集上进行实验来验证本文所提方法的有效性.

PACS[26]数据集具有显著的视觉风格迁移特征, 是 DG 研究中最常用的基准之一, 包含 7 个类别、9 991 张图片, 有 4 个风格差异明显的域: Photo (P) 是真实图片, 外观特征丰富; Art painting (A) 为包含不规则笔触和颜色的艺术风格图像; Cartoon (C) 的图像为色块鲜明轮廓简洁的卡通风格; Sketch (S) 则以黑白素描及线条为主.

OfficeHome[27]数据集来自办公室和家庭环境的日常物品的图片, 包含 65 个类别、15 500 张图片, 4 个域的领域间差异显著: Art、Clipart (Cli)、Product (Pro) 和 Real-World (Real).

VLCS[28]数据集包含 5 个类别和 10 729 张图片, 由 4 个图像分布和风格不同的域组成: PASCAL VOC (PV) 为标准的物体检测照片, LabelMe (LM) 接近日常场景, Caltech (Cal) 为标准的物体分类图片, SUN09 (SUN) 专注于场景识别.

TerraIncognita[29]数据集是一个用于监测动物种群的重要数据集, 包含了分布于 4 个固定位置 (location_100 (L100)、location_38 (L38)、location_43 (L43) 和 location_46 (L46)) 的 10 类野生动物图片. 数据集包括来自不同地理位置的图像, 其中每个位置可能对应不同的环境和动物群落.

本文所提模型采用在 ImageNet1K 上预训练的 VMamba-T[5]网络作为主干模型进行 10 000 次迭代训练, 每个源域的批量大小为 16, 使用了 β=(0.9, 0.999)、动量参数设置为 0.9 的 AdamW 优化器, 采用余弦衰减学习率调度器在 (0.000 3, 0.000 5) 间寻找初始学习率. 作为与之对比的基线方法, 特征提取主干网络选用了和 VMamaba-T 网络参数量相当的 ResNet-50[30]和 DeiT-Small[31]网络, 同样在 ImageNet1K 上进行了预训练. 本文实验遵循标准的留一域验证 (leave-one-domain-out, LODO) 设置, 每次实验选取 1 个域作为测试域, 其余域作为训练域, 依次选择每个域作为测试域, 记录每次实验的平均分类准确率 (Avg Acc, 表格中均简称为 Avg). 为降低随机性影响, 所有实验均在相同训练设置与随机种子下重复运行 3 次, 报告平均结果. 本文实验均在 PyTorch 2.0.0、CUDA 11.8、Python 3.8 环境上进行, 使用 Ubuntu 20.04 操作系统, GPU 为 NVIDIA Tesla V100 (32 GB), 内存大小为 25 GB.

3.2 实验结果与分析

从表 1 可以看出, 本文方法在 PACS 数据集上的平均准确率达到 91.9%, 相比于当前最优方法 DGMamba (91.0%) 提升了 0.9 个百分点, 相较于 ResNet-50 基线方法和 DeiT-Small 基线方法也分别至少提升了 3.8 个百分点和 5.2 个百分点. 具体来看, 在 Sketch (S) 这一具有显著分布偏移的域上, 本文方法从 DGMamba 的 87.0% 提升到 88.9%, 说明反事实语义增强有效减少了模型对背景的依赖; 在 Photo (P) 和 Cartoon (C) 域上, 准确率分别达到 99.3% 和 87.5%, 均优于其他方法, 显示出模型在多风格域下的稳健性. 整体结果表明, 反事实语义增强模块与因果注意力机制协同作用, 有效引导模型关注领域不变的语义特征, 从而在风格变化显著的 PACS 数据集上表现出更强的跨域泛化能力.

表 1 在 PACS 数据集上的实验结果

方法	backbone	参数量 (M)	A (%)	C (%)	P (%)	S (%)	Avg (%)
Mixup	ResNet-50	23.5	86.9	78.8	95.0	76.7	84.3
MixStyle	ResNet-50	23.5	86.8	79.0	96.6	78.5	85.2
DANN	ResNet-50	23.5	85.9	78.2	95.9	76.0	84.0
IRM	ResNet-50	23.5	86.3	77.1	95.8	76.4	83.9
MTL	ResNet-50	23.5	87.1	77.5	96.4	77.3	84.6
SagNet	ResNet-50	23.5	87.6	80.7	97.1	80.0	86.3
VREx	ResNet-50	23.5	86.0	79.1	96.9	77.7	84.9
ARM	ResNet-50	23.5	86.8	76.8	97.4	79.3	85.1
MLDG	ResNet-50	23.5	87.3	80.6	97.8	79.7	86.4
SWAD	ResNet-50	23.5	89.3	83.4	97.3	82.5	88.1
SAGM	ResNet-50	23.5	87.4	80.2	98.0	80.8	86.8
ERM-SDViT	DeiT-Small	22	87.6	82.4	98.0	77.2	86.3
GMoE	DeiT-Small	22	89.4	83.9	99.1	74.5	86.7
DGMamba	VMamba-T	30	91.2	86.9	98.9	87.0	91.0
Ours	VMamba-T	30	**91.9**	**87.5**	**99.3**	**88.9**	**91.9**

表 2 为本文方法在 OfficeHome 数据集上与其他基线方法的对比实验结果, 可以看出, 本文方法取得了 77.0% 的平均准确率, 略高于 DGMamba 的 76.8%, 并在 Art 域上显著提升了 1.9 个百分点. 实验结果表明, OfficeHome 数据集的域间差异更复杂, 尤其是在 Art 和 Product 域, 背景风格、光照和物体布局变化较大. 本文方法通过反事实样本构造打破前景-背景的虚假

共现, 使得主干模型在面对复杂背景变化时依然能提取稳定语义特征, 虽然整体提升幅度较小于 PACS 数据集, 但结果仍证明本文方法具备一定的鲁棒性和通用性.

表 2 在 OfficeHome 数据集上的实验结果

方法	backbone	参数量 (M)	Art (%)	Cli (%)	Pro (%)	Real (%)	Avg (%)
Mixup	ResNet-50	23.5	58.9	51.6	70.9	71.4	63.2
MixStyle	ResNet-50	23.5	51.1	53.2	68.2	69.2	60.4
DANN	ResNet-50	23.5	60.2	55.3	69.5	74.6	64.9
IRM	ResNet-50	23.5	61.8	53.7	72.6	75.3	65.9
MTL	ResNet-50	23.5	61.5	52.4	74.9	76.8	66.4
SagNet	ResNet-50	23.5	63.4	54.8	75.8	78.3	68.1
VREx	ResNet-50	23.5	60.7	53.0	75.3	76.6	66.4
ARM	ResNet-50	23.5	58.9	51.0	74.1	75.2	64.8
MLDG	ResNet-50	23.5	64.2	54.5	74.6	75.9	67.3
SWAD	ResNet-50	23.5	66.1	57.7	78.4	80.2	70.6
SAGM	ResNet-50	23.5	65.4	57.0	78.0	80.0	70.1
ERM-SDViT	DeiT-Small	22	68.3	56.3	79.5	81.8	71.5
GMoE	DeiT-Small	22	69.3	58.0	79.8	82.6	72.4
DGMamba	VMamba-T	30	75.6	**61.9**	**83.8**	**86.0**	76.8
Ours	VMamba-T	30	**77.5**	**61.9**	83.4	85.1	**77.0**

表 3 给出了基线方法和本文方法在 VLCS 数据集上取得的分类准确率及均值, 本文方法的平均准确率达到 81.1%, 相比 DGMamba 的 80.7% 提升 0.4 个百分点, 依旧保持领先. 在 SUN 和 PV 域上, 准确率分别达到 79.9% 和 81.7%, 均优于所有对比方法; 这表明本文方法能够有效识别跨域场景中稳定的语义因果特征, 即使在类别和背景分布差异较大的情况下, 依旧能保持对目标语义的稳健建模. 这一结果进一步验证了因果注意力机制在状态更新路径中显式引导特征传播的有效性, 使得模型对不稳定的领域特征敏感性降低.

表 3 在 VLCS 数据集上的实验结果

方法	backbone	参数量 (M)	Cal (%)	LM (%)	SUN (%)	PV (%)	Avg (%)
Mixup	ResNet-50	23.5	98.0	63.8	72.3	73.1	76.8
MixStyle	ResNet-50	23.5	98.6	64.5	72.6	75.7	77.9
DANN	ResNet-50	23.5	98.4	64.8	73.4	76.8	78.1
IRM	ResNet-50	23.5	98.3	64.2	73.5	76.8	78.2
MTL	ResNet-50	23.5	97.8	64.3	71.5	75.3	77.2
SagNet	ResNet-50	23.5	97.9	64.5	71.4	77.5	77.8
VREx	ResNet-50	23.5	98.4	64.4	74.1	76.2	78.3
ARM	ResNet-50	23.5	98.7	63.6	71.3	76.7	77.6
MLDG	ResNet-50	23.5	98.9	63.5	75.0	78.6	79.0
SWAD	ResNet-50	23.5	98.8	63.3	75.3	79.2	79.1
SAGM	ResNet-50	23.5	99.0	**65.2**	75.1	80.7	80.0
ERM-SDViT	DeiT-Small	22	96.8	64.2	76.2	78.5	78.9
GMoE	DeiT-Small	22	96.9	63.2	72.3	79.5	78.0
DGMamba	VMamba-T	30	97.7	64.8	79.3	81.0	80.7
Ours	VMamba-T	30	**98.6**	64.4	**79.9**	**81.7**	**81.1**

表 4 给出了 TerraIncognita 数据集上各方法的实验结果. 在 TerraIncognita 数据集上, 本文方法的平均准确率为 54.9%, 略优于 DGMamba 的 54.5%, 在 L100 和 L43 两个子域上表现最佳, 分别达到 62.8% 和 62.1%. 这一结果表明对于具有显著分布偏移、背景噪声复杂且样本差异较大的场景, 本文方法依然具备一定的泛化优势, 尽管提升幅度相对有限, 但在高噪声环境中仍能通过反事实增强生成更加稳健的语义表示, 从而保持性能领先.

表 4 在 TerraIncognita 数据集上的实验结果

方法	backbone	参数量 (M)	L100 (%)	L38 (%)	L43 (%)	L46 (%)	Avg (%)
Mixup	ResNet-50	23.5	50.2	36.5	58.3	43.4	47.1
MixStyle	ResNet-50	23.5	54.3	34.1	55.9	31.7	44.0
DANN	ResNet-50	23.5	48.5	37.2	56.4	34.7	46.7
IRM	ResNet-50	23.5	51.3	37.9	56.6	43.0	47.2
MTL	ResNet-50	23.5	49.3	39.6	55.6	37.8	45.6
SagNet	ResNet-50	23.5	53.0	43.0	57.9	40.4	48.6
VREx	ResNet-50	23.5	48.2	41.7	56.8	38.7	46.4
ARM	ResNet-50	23.5	49.3	38.3	55.8	38.7	45.5
MLDG	ResNet-50	23.5	52.6	42.3	59.1	38.4	48.1
SWAD	ResNet-50	23.5	55.4	44.9	59.7	39.9	50.0
SAGM	ResNet-50	23.5	54.8	41.4	57.7	41.3	48.8
ERM-SDViT	DeiT-Small	22	55.9	31.7	52.2	37.4	44.3
GMoE	DeiT-Small	22	59.2	34.0	50.7	38.5	45.6
DGMamba	VMamba-T	30	62.0	**47.7**	61.7	46.9	54.5
Ours	VMamba-T	30	**62.8**	47.1	**62.1**	**47.6**	**54.9**

3.3 消融实验与模块交互分析

为进一步验证本文方法中两个关键模块的性能, 本文构建了 5 个模型变体进行消融实验. 其中, A 指本文所提反事实语义增强模块, B 是使用 Causal-Mamba 结构替换原本 Mamba 结构的变体版本, 在 PACS 数据集上进行消融实验的结果如表 5 所示, 不难看出两个模块都起着重要的作用. 具体来说, 单独使用基础模型 VMamba 时性能最低; 当分别加入反事实模块或 HSS 隐藏状态抑制机制后, 模型性能都有所提升; 而同时加入反事实语义增强模块和 Causal-Mamba 结构时性能最高. 例如, 加入反事实模块后, 相较于基线模型准确率提升了 1.8 个百分点; 再加入因果注意力机制后较基线模型准确率提升了 2.5 个百分点.

表 5 在 PACS 数据集上的消融实验结果 (%)

变体	A	C	P	S	Avg
VMamba	88.2	86.2	98.4	84.9	89.4
DGMamba	91.2	86.9	98.9	87.0	91.0
Vmamba+A	91.6	86.6	98.4	88.2	91.2
VMamba+A+HSS	91.5	87.4	98.9	88.6	91.6
VMamba+A+B (Ours)	**91.9**	**87.5**	**99.3**	**88.9**	**91.9**

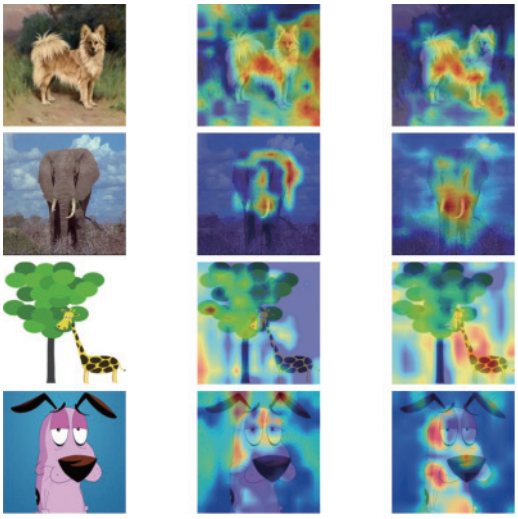
此外, 表 5 中的 VMamba+A+HSS 变体也即仅用反事实语义增强模块替换 DGMamba 中的原 SPR 模块来进行对比, 从结果可以看出, 本文的反事实语义增强模块相较 SPR 模块更具语义稳健性, 因此配合 Causal-Mamba 主干后的完整方法进一步提升了整体性能. 从模块交互的角度来看, 反事实模块与因果注意力机制在本方法中相互协作、相辅相成. 反事实模块通过生成多样化的干预样本, 引导模型关注隐含的因果结构, 从而学习到域不变特征; 因果注意力机制则进一步强化这些重要因果特征在特征表达中的权重, 有效滤除无关信息. 消融实验结果表明: 缺少任一模块都会导致性能下降, 而同时保留两者能够达到最佳效果. 这说明这两个模块的组合比单独使用任何一个都更能提升模型的泛化能力. 总的来说, 反事实模块提供了额外的监督信号, 丰富了模型学习的视角; 而因果注意力机制保证了最终提取的特征与因果关系紧密相关, 两者配合显著增强了模型对未知域的适应性.

3.4 可视化分析

为了对本文方法取得的结果进行可视化, 进一步验证其在语义聚焦方面的有效性, 本节使用 Grad-CAM 技术对模型在测试样本上的激活区域进行可视化分析. 实验在 PACS 数据集上进行, 分别选取以 Art painting、Photo 和 Cartoon 域为目标域时的 4 张样本, 展示原图、基线模型及本文方法的注意力分布. 实验结果如图 4 所示, (a)、(b)、(c) 分别为原图、DGMamba 基线模型和本文方法的注意力热力图. 从图 4 可以看出, DGMamba 的激活热区分布较为分散, 部分集中于背景区域和无关纹理; 而本文方法在引入反事实语义增强与因果注意力机制后, 激活区域明显聚焦于前景语义目标(如动物主体、目标轮廓等), 背景响应显著减弱. 这说明本文模型能够有效过滤与类别无关的域特异特征, 实现因果层面的语义聚焦. 此外, 从整体分布趋势可以观察到, 本文方法所生成的注意力图在不同域样本上保持了一致的聚焦模式, 验证了其在领域不变特征学习上的稳定性与泛化能力. 该结果与前文定量实验结果相一致, 进一步印证了所提方法对模型泛化性的促进作用.

4 结论与展望

本文提出了一种融合反事实语义增强与因果注意力机制的领域泛化方法, 设计的反事实语义增强模块 CF-module 合成不同背景的前景样本, 遵循“背景变化下前景语义不变”的因果假设, 切断背景-标签的共现路径, 在样本层面构建了反事实干预. 利用其产生的前景语义权重, 在模型层面注入因果注意门控机制来建模 Causal-Mamba 模块, 在特征传播路径对隐藏状态加权. 本文方法以提升模型的语义可解释性和泛化能力为目标, 实验证明我们提出的方法从数据与结构两个维度共同将判别信号偏向语义因果成分, 减少由于背景偏移导致的模型误判, 提升了模型对领域不变语义的建模能力. 本文方法在领域泛化研究中广泛使用的 4 个数据集上都取得了优异性能, 且验证了反事实语义增强模块和因果注意力机制各自的重要性. 未来工作中可探索更高质量的反事实生成策略及更高效的因果注意结构, 以进一步提高模型的可解释性与计算效率, 提升模型在开放世界中的泛化鲁棒性



(a) Original (b) Baseline (DGMamba) (c) Ours

图 4 PACS 数据集上的可视化对比结果

参考文献

1 Wang JD, Lan CL, Liu C, *et al*. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 8052–8072. [doi: 10.1109/TKDE.2022.3178128]

2 Liu JS, Shen ZY, He Y, *et al*. Towards out-of-distribution generalization: A survey. arXiv:2108.13624, 2021.

3 Robey A, Pappas GJ, Hassani H. Model-based domain generalization. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1546.

4 Li SS, Zhao QJ, Zhang CC, *et al*. Deep discriminative causal domain generalization. Information Sciences, 2023, 645: 119335. [doi: 10.1016/J.INS.2023.119335]

5 Liu Y, Tian YJ, Zhao YZ, *et al*. VMamba: Visual state space

model. Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 3273.
6 Long SC, Zhou QY, Li XT, *et al*. DGMamba: Domain generalization via generalized state space model. Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne: ACM, 2024. 3607–3616. [doi: 10.1145/3664647.3681247]
7 Wang XY, Saxon M, Li JC, *et al*. Causal balancing for domain generalization. Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). Kigali: OpenReview.net, 2023.
8 Rao YM, Chen GY, Lu JW, *et al*. Counterfactual attention learning for fine-grained visual categorization and re-identification. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1005–1014.
9 Yan S, Song H, Li NX, *et al*. Improve unsupervised domain adaptation with mixup training. arXiv:2001.00677, 2020.
10 Zhou KY, Yang YX, Qiao Y, *et al*. Domain generalization with MixStyle. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
11 Ganin Y, Ustinova E, Ajakan H, *et al*. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 2016, 17(59): 1–35.
12 Arjovsky M, Bottou L, Gulrajani I, *et al*. Invariant risk minimization. arXiv:1907.02893, 2019.
13 Blanchard G, Deshmukh AA, Dogan Ü, *et al*. Domain generalization by marginal transfer learning. The Journal of Machine Learning Research, 2021, 22(2): 2.
14 Nam H, Lee H, Park J, *et al*. Reducing domain gap by reducing style bias. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8686–8695.
15 Krueger D, Caballero E, Jacobsen JH, *et al*. Out-of-distribution generalization via risk extrapolation (REx). Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 5815–5826.
16 Zhang M, Marklund H, Dhawan N, *et al*. Adaptive risk minimization: Learning to adapt to domain shift. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1812.
17 Li D, Yang YX, Song YZ, *et al*. Learning to generalize: Meta-learning for domain generalization. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 427. [doi: 10.1609/aaai.v32i1.11596]
18 Cha J, Chun S, Lee K, *et al*. SWAD: Domain generalization by seeking flat minima. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1716.
19 Wang PF, Zhang ZX, Lei Z, *et al*. Sharpness-aware gradient matching for domain generalization. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3769–3778.
20 Sultana M, Naseer M, Khan MH, *et al*. Self-distilled vision Transformer for domain generalization. Proceedings of the 16th Asian Conference on Computer Vision. Macao: Springer, 2022. 273–290.
21 Li B, Shen YF, Yang JK, *et al*. Sparse mixture-of-experts are domain generalizable learners. Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2022.
22 Tang LY, Yuan YX, Chen CQ, *et al*. Mixstyle-Entropy: Whole process domain generalization with causal intervention and perturbation. Proceedings of the 35th British Machine Vision Conference. Glasgow: BMVC, 2024. 25–28.
23 Wang X, Zhao QJ, Wang L, *et al*. Causality-based contrastive incremental learning framework for domain generalization. Tsinghua Science and Technology, 2025, 30(4): 1636–1647. [doi: 10.26599/TST.2024.9010072]
24 Shao FF, Luo YW, Chen L, *et al*. Counterfactual co-occurring learning for bias mitigation in weakly-supervised object localization. arXiv:2305.15354, 2023.
25 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 618–626.
26 Li D, Yang YX, Song YZ, *et al*. Deeper, broader and artier domain generalization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5543–5551.
27 Venkateswara H, Eusebio J, Chakraborty S, *et al*. Deep hashing network for unsupervised domain adaptation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5385–5394.
28 Fang C, Xu Y, Rockmore DN. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney: IEEE, 2013. 1657–1664.
29 Beery S, Van Horn G, Perona P. Recognition in terra incognita. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 472–489.
30 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
31 Touvron H, Cord M, Douze M, *et al*. Training data-efficient image transformers & distillation through attention. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 10347–10357.

(校对责编: 李慧鑫)