doi:10.11887/j.issn.1001-2486.26010026 http://journal.nudt.edu.cn

面向可信高效人工智能的原型学习方法研究进展

胡星辰[1,2],朱修彬[3*],刘吉元[2],申映华[4],曾泽凡[2],万欣航[2],成 清[2],黄金才[2]
(1. 国防科技大学 计算机学院,湖南 长沙 410073;2. 国防科技大学 系统工程学院,湖南 长沙 410073;
3. 西安电子科技大学 机电工程学院,陕西 西安 710071;4. 重庆大学 经济与工商管理学院,重庆 400044)

摘 要:针对深度学习等人工智能技术在可解释性、数据依赖及鲁棒性方面的严峻挑战,系统评述了原型学习(prototypical learning,PL)的理论方法与前沿进展。通过界定原型学习的基本概念与数学表示,构建涵盖统计机器学习、深度特征驱动及语义表示维度的原型生成体系。解析基于原型的单/多模态数据增强与融合机制,阐明其突破数据质量瓶颈的核心逻辑。重点论述原型学习在可解释深度网络、模糊规则推理、因果溯因及时序分析中的应用效能。进一步探索原型学习在生成式学习、大模型能力增强及图学习等交叉领域的演进动态。通过凝练原型学习在表征效率与逻辑透明度方面的独特价值,揭示其在构建可信、高效人工智能系统方面的关键技术价值。最后,展望原型学习在生成式人工智能、大模型协同及可持续学习等方向的发展趋势。

关键词:原型学习;数据挖掘;知识表示;机器学习
中图分类号:TP3-0 **文献标志码**:A **文章编号**:1001-2486(2026)03-228-24

Prototypical learning for trustworthy and efficient AI: a survey

HU Xingchen[1,2], *ZHU Xiubin*[3*], *LIU Jiyuan*[2], *SHEN Yinghua*[4], *ZENG Zefan*[2], *WAN Xinhang*[2], *CHENG Qing*[2], *HUANG Jincai*[2]
(1. College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China;
2. College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;
3. School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China;
4. School of Economics and Business Administration, Chongqing University, Chongqing 400044, China)

Abstract: In response to the significant challenges faced by artificial intelligence techniques, particularly deep learning, in terms of interpretability, heavy data dependence, and limited robustness, the theoretical methodologies and recent advances of PL (prototypical learning) were systematically reviewed. By clarifying the fundamental concepts and mathematical formulations of prototypical learning, a prototype generation framework was established that encompasses statistical machine learning, deep feature-driven representations, and semantic representation perspectives. Prototype-based mechanisms for single-modal and multimodal data augmentation and fusion were analyzed, and the underlying rationale for overcoming data quality bottlenecks was elucidated. Particular emphasis was placed on the effectiveness of prototypical learning in interpretable deep neural networks, fuzzy rule-based reasoning, causal abduction, and time-series analysis. Furthermore, the evolutionary dynamics of prototypical learning across interdisciplinary domains, including generative learning, capability enhancement of large-scale models, and graph learning, were explored. By synthesizing the distinctive advantages of prototypical learning in representation efficiency and logical transparency, its critical role as a key enabling technology for constructing trustworthy and efficient artificial intelligence systems was highlighted. Finally, future development trends of prototypical learning were discussed, including directions related to generative artificial intelligence, collaboration with large models, and sustainable learning.

Keywords: prototypical learning; data mining; knowledge representation; machine learning

收稿日期:2026-01-16
基金项目:国家自然科学基金资助项目(62376279,62306324,U24A20333);重庆自然科学基金资助项目(CSTB2023NSCQ-MSX1075);湖南省科技创新基金资助项目(2024RC3128)
第一作者:胡星辰(1989—),男,安徽合肥人,副教授,博士,硕士生导师,E-mail:xingchenhu@nudt.edu.cn
*通信作者:朱修彬(1982—),男,山东泰安人,副教授,博士,硕士生导师,E-mail:xbzhu@mail.xidian.edu.cn
引用格式:胡星辰,朱修彬,刘吉元,等. 面向可信高效人工智能的原型学习方法研究进展[J]. 国防科技大学学报,2026,48(3):228-251.
Citation:HU X C, ZHU X B, LIU J Y, et al. Prototypical learning for trustworthy and efficient AI: a survey [J]. Journal of National University of Defense Technology, 2026, 48(3): 228-251.

在以深度学习为代表的人工智能技术的快速发展过程中，虽然模型的性能不断提升，但随之而来的模型“黑箱”、模型训练依赖大规模数据，以及在数据小样本和低质量条件下表现不稳定等问题，引发了学术界和产业界的广泛关注。如何在保障模型预测性能的前提下，提升模型的可解释性、鲁棒性与数据利用效率，实现知识的高效表示与处理，已成为人工智能系统研究中的核心问题。作为一种兼具知识表示能力与直观解释性的学习范式，原型学习（prototypical learning，PL）正是在这一背景下应运而生，并逐步发展成为近年来人工智能领域的重要研究方向。

原型学习的核心思想是通过一组具有代表性的“原型”来概括数据分布，从而以典型实例或结构作为知识载体，实现模型推理、决策和知识组织的直观化。不同于传统的特征抽象方法，原型不仅承载了类别或语义的中心表示，还能以局部或全局的形式解释样本与类别之间的关系，为知识表示和处理提供天然的结构化支持，使得原型学习天然具备较强的可解释性，并在建模过程中提供可信的决策依据。与此同时，原型作为低维、紧凑的知识表示单元，能够有效缓解深度模型对大规模数据的依赖，并在小样本学习、跨域迁移、噪声与缺失数据处理等任务中展现出独特优势，从而实现对数据的高效组织与利用。

从发展脉络来看，原型学习经历了多个阶段的演进：早期研究主要借鉴聚类分析与度量学习的思想，以原型作为类别中心或典型样本进行分类与识别；随着深度神经网络的发展，原型学习逐渐与卷积网络、图神经网络（graph neural network，GNN）和注意力机制相结合，形成了端到端可训练的原型网络（prototypical networks，ProtoNet）与多视图原型表示（prototype representation，PR）方法；近年来，研究进一步扩展至跨模态表示、知识图谱推理和复杂场景的可解释预测，逐渐呈现从静态到动态、从单一模态到多模态、从分类任务到生成与推理任务的全面发展趋势，为人工智能模型构建提供了系统化方法。

原型学习的优势主要体现在三个方面：其一，可解释性强，能够通过原型实例或结构提供直观的推理过程，从而增强模型的透明性与可追溯性；其二，具备较强的鲁棒性，能够在数据存在噪声、缺失等情况下保持相对稳定性能；其三，适应小样本与跨域场景，能够利用有限数据实现高效的知识迁移与泛化。因此，原型学习在军事国防、医学诊断、金融风控等对模型可解释性与决策可信性要求较高且数据获取成本高或样本相对稀缺的关键应用领域中，展现出广阔的应用前景。

总体而言，原型学习已经逐渐成为构建可信和高效人工智能的重要途径之一。它在提升知识表示与处理能力、增强模型可解释性和鲁棒性，以及优化小样本与跨域学习方面具有显著优势，为人机协作、知识驱动推理和跨领域应用提供了重要基础。随着多模态大模型、生成式人工智能和可持续学习的不断发展，未来原型学习有望在推动人工智能系统的透明化、可信化和高效化方面发挥更加关键的作用，并为人工智能的理论突破与实践应用提供新的范式与支撑。

为了系统地阐述原型学习的研究现状与技术全貌，图 1 构建了一个原型学习方法分类框架。

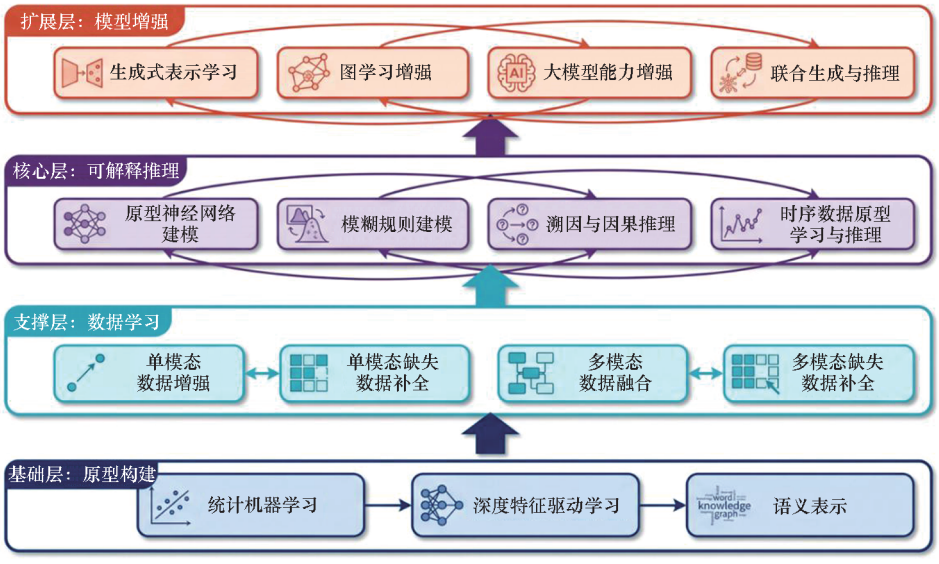

图 1　原型学习方法分类框架

Fig. 1　Classification framework of prototypical learning methods

该分类遵循“基础—支撑—核心—扩展”的自下而上层级逻辑:基础层聚焦于原型的构建范式,涵盖从传统统计机器学习到深度特征驱动学习及语义表示的演进;支撑层关注数据学习质量,探讨利用原型进行单模态增强与多模态融合补全的机制;核心层致力于解决“黑箱”问题,论述基于原型的可解释推理与因果建模方法;扩展层则面向前沿应用,展示了原型在生成式表示学习、图学习增强及大模型能力增强中的最新进展。后续章节将严格遵循这一逻辑架构,对上述关键技术进行逐层剖析。

1 原型学习的基本概念

原型的概念最早起源于认知心理学领域。20世纪70年代,认知心理学家 Rosch[1] 提出了著名的“原型范畴理论”(prototype categorization theory, PCT),其研究发现,人类在认知过程中并非依据严格的充分必要条件来划分范畴,而是依赖“家族相似性”(family resemblance, FR)机制,即通过比较对象与该范畴中最具代表性的典型实例(即原型)之间的相似程度来判断其归属。这种认知机制表明,范畴内部成员具有典型的中心-边缘结构,越接近原型的成员越能代表该范畴。这一发现为人工智能中的知识表示提供了重要的认知科学依据,即通过构建特征空间中的“中心”或“锚点”来高效、可解释地表征复杂数据分布。

在人工智能领域,原型被抽象为特征空间中能够刻画特定类别或任务核心语义特征的向量表示。不同于传统判别模型依赖复杂的非线性超平面来划分决策边界,原型学习的核心思想是学习一个度量嵌入空间,通过计算样本与类中心(即原型)之间的相似性或距离来实现分类决策。这种“以点代类”的机制赋予了模型天然的结构化知识表示能力,使模型能够以直观的典型实例为桥梁,解释样本与类别之间的隶属关系。

从数学角度而言,设给定数据 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$,其中 $\boldsymbol{x}_i\in\boldsymbol{X}$ 为输入样本,$y_i\in\{1,2,\cdots,K\}$ 为类别标签。原型可以被理解为该数据集的一系列代表,使其能够充分反映数据的结构特点。原型学习通常也被称为数据/示例选择、锚点学习、子集选择、样例选择和核心集构造等[2]。定义一个由参数 θ 构成的非线性嵌入映射函数 $f_\theta:\boldsymbol{X}\to\mathbb{R}^M$(可以为深度神经网络、聚类算法或其他机器学习模型),将样本映射至高维特征空间。在该空间内,类别 k 的原型 $\boldsymbol{c}_k$ 定义为该类支持集样本特征的中心,其数学表达式通常表示为:

$$\boldsymbol{c}_k=\frac{1}{|\boldsymbol{S}_k|}\sum_{(\boldsymbol{x}_i,y_i)\in\boldsymbol{S}_k}f_\theta(\boldsymbol{x}_i) \tag{1}$$

其中,$\boldsymbol{S}_k$ 为类别 k 的支持样本矩阵。对于待查询样本 $\boldsymbol{x}$,其属于类别 k 的预测概率通常基于距离度量 $d(\cdot)$ 通过 Softmax 函数建模为:

$$P(y=k\mid\boldsymbol{x})=\frac{\exp(-d(f_\theta(\boldsymbol{x}_i,\boldsymbol{c}_k)))}{\sum_{k'=1}^{K}\exp(-d(f_\theta(\boldsymbol{x}_i,\boldsymbol{c}_{k'})))} \tag{2}$$

综上所述,原型学习通过显式地维护特征空间中的类中心,将复杂的非线性分类问题转化为嵌入空间内的度量对齐问题。这种建模范式具有极强的归纳偏置,即假设同一类别的样本在特征空间中应呈现围绕原型聚拢的紧凑分布。由于原型本身承载了类别的全局特征信息,该范式在处理数据缺失、噪声干扰以及小样本迁移学习等任务时,能够通过稳定的原型锚点提供鲁棒的知识支撑,从而有效缓解了深度学习等机器学习方法对大规模高质量标注数据的过度依赖[1-2]。

2 原型构建的主要方法

原型学习伴随人工智能技术的发展大致经历了三个阶段,也具体反映在如何构建原型上。在统计机器学习时期,侧重于几何空间的聚类划分;在深度特征驱动时期,随着深度神经网络的兴起,实现了端到端的特征与原型联合优化,解决了高维感知的语义对齐问题;在生成式大模型时期,原型逐渐演变为非参数化的知识锚点,赋能生成式模型的可控生成与大语言模型(large language model, LLM)的知识检索,如图2所示。

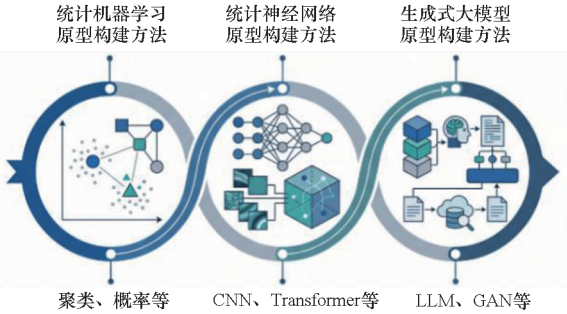

图2 原型学习方法发展脉络

Fig. 2 Development trend of prototypical learning methods

2.1 基于统计机器学习的原型构建方法

基于统计机器学习的原型构建方法主要尝试利用数据自身的样本分布特点,获得能够近似表示样本空间结构的若干原型,以实现聚合小样本信息、过滤噪声,形成稳定的数据表示。因此,统计机器学习中的众多聚类算法被用于原型构建。这些聚类算法按照能否直接生成原型可进一步分

为两类。其中,可直接生成原型的方法不仅包括 K 均值(K-Means)聚类算法[3]、K 中心点(K-Medoids)聚类算法[4]、K 原型(K-Prototypes)聚类算法[5]、亲和传播(affinity propagation, AP)算法[6]、模糊 C 均值[7](fuzzy C-means, FCM)这一类最终可以输出一系列数值原型的聚类方法(每个数值原型代表一个簇),同时涵盖了高斯混合模型[8](Gaussian mixture model, GMM)这一类最终能够输出一系列样本概率分布的聚类方法(每个概率分布代表一个簇)。不直接生成原型的方法一般是先获得数据样本的一个划分,形成一系列数据子集,原型则可以通过各个数据子集内部数据的加权平均获得。这一类方法主要包括基于图的聚类(如谱聚类[9]、社区聚类[10])、层次聚类[11]和基于密度的聚类(如有噪声基于密度的聚类[12](density-based spatial clustering of applications with noise, DBSCAN)方法和识别聚类结构的排序点[13](ordering points to identify the clustering structure, OPTICS)方法)。

综上所述,基于不同类型聚类算法及概率生成式模型的原型学习方法的核心逻辑在于通过度量数据样本空间内的相似度或概率关联,选择性地保留原始样本或生成虚拟表征,从而实现利用极少数代表性样本对全局数据分布的高效抽象。这一过程本质上是将高维、冗余且含噪的数据流形压缩至低维、精炼且具有物理或几何意义的“骨干”结构中,已成为该领域的重要研究方向。

从理论价值来看,此类方法为缓解“维度灾难”提供了坚实的统计学基础。通过原型构建,可以将复杂的非参数化分布转化为可参数化或半参数化的表达形式,极大地降低了后续学习任务的计算复杂度。同时,原型作为样本空间的“锚点”,为理解大规模数据集的拓扑结构提供了直观的数学解释[14]。

从应用前景来看,基于统计机器学习的原型构建在大规模数据分析、异常检测及联邦学习(federated learning, FL)等领域具有广阔的空间。在实时数据处理中,利用原型替代原始样本可以显著降低存储压力并提升查询效率;在隐私保护场景下,原型作为抽象后的聚合特征,能够实现在不泄露个体隐私的前提下进行跨机构的知识共享与协同建模。

2.2　深度特征驱动的原型构建方法

基于卷积神经网络(convolutional neural network, CNN)、Transformer 等深度特征,动态学习原型可以反映类别特征和样本局部结构。

近些年来,随着深度学习在特征提取、模式识别等方面的突出表现,越来越多的研究将原型学习应用于深度学习,以进一步提升深度学习模型的精度和解释能力。依据原型学习的作用层级,这类深度特征驱动的原型学习方法包括两类。一类是将原型学习应用于原始特征空间,随后基于原型构建神经网络。这类方法本质上依赖的仍是浅层特征。该类研究相对较少,例如何克磊等[15]主要采用聚类算法获取包的特征,进而在包的层面构建卷积神经网络。

另一类是将原型学习作用于深度学习网络中的特征图层(多层神经网络的隐藏层所产生的中间结果),形成深度特征的原型表示,并通过与原始样本的结合,提升深度学习的模型性能。这类研究是深度特征驱动的原型学习的主要方向。例如,杨雨龙等[16]在神经网络映射的特征空间上学习一个图注意力模型,将此模型应用于特征空间中,可以充分利用原型的信息来修正特征。李娇等[17]通过捕获每个标签的视觉原型表示,构建标签视觉原型字典,充分发挥视觉特征信息与图像分类任务的适配性。还有一些工作考虑到类别的动态变化,通过动态扩展原型中的特征,增强了模型在废旧家电识别过程中对新旧样本的适应性。Wang 等[18]提出了一种原型集成方法,用于在增量学习(incremental learning, IL)中检测新类别。该方法通过减小 CNN 所提取的特征在类内的差异并增大类间的差异,从而增强对新类别检测的鲁棒性。Snell 等[19]在文献中提出了一种原型网络,该网络能够构建一个度量空间,在这个空间里,通过计算特征与每个类别的原型之间的距离来进行分类。这种方法在小样本学习领域取得了良好的效果。

2.3　面向语义表示的原型构建方法

将原型嵌入向量空间,捕捉类别间语义结构及层次关系,实现可解释知识表示。结合全局原型与局部子原型,构建多尺度、多粒度表示。

多尺度与多层级语义原型构建方法是当前语义表示研究领域的核心议题,旨在通过整合不同粒度、不同层级的特征信息,提升语义原型的表达能力和泛化性能。例如,Tang 等[20]在类别原型 Transformer(category prototype transformer, CPT)中提出了一种基于类别原型的特征重构机制,通过将低分辨率特征图中的像素特征替换为其对应的类别原型,有效缓解了内类特征方差问题。此外,不少基于聚类分析的模糊规则模型[21-22]在构建过程中,会把原型的识别与模型参数的优化联

合起来,得到不同尺度(粒度)的原型,从而提升原型的代表性。Zhao 等[23]提出的无训练开放词汇语义分割框架引入了 LLM 引导的支持图像生成和粗细聚类策略,利用 LLM 生成具有不同属性描述的支持图像,并通过层级聚类获取鲁棒的部件级原型。Shen 等[24]针对遥感图像少样本语义分割提出的自适应自支持原型学习网络,通过自适应超原型表示(hyper-prototype representation, HPR)和自支持匹配(self-support matching, SSM)机制实现了层级化原型的动态构建与匹配。

局部与全局语义原型构建方法是语义表示研究中的重要分支,其核心在于如何有效整合局部特征与全局知识,以形成具有判别性和泛化能力的原型表示。近年来,相关研究针对不同应用场景(如联邦学习、增量学习、弱监督学习等)提出了多样化的解决方案,既关注局部数据的独特性,又强调全局语义的一致性与共享性。例如,Duan 等[25]提出的多标签原型视觉空间搜索(multi-label prototype visual spatial search, MuP-VSS)方法通过全局嵌入学习和原型嵌入模块(prototype embedding module, PEM)实现了局部语义信息与全局上下文的有效结合。Zhou 等[26]提出联邦语义锚点学习(federated semantic anchor learning, FedSA)框架,将原型生成与局部表示学习解耦,引入简单有效的语义锚点作为原型,通过锚点基正则化(含边缘增强对比学习)和锚点基分类器校准,实现原型的类内紧凑性和类间可分性,同时确保决策边界的一致性。

动态演化语义原型构建方法通过设计自适应更新机制,使原型能够响应数据分布变化和概念演化,显著提升了模型的泛化能力和鲁棒性。首先,在增量学习场景中,Zhu 等[27]提出的自适应原型重放方法通过自适应偏差补偿策略动态更新存储的原型,匹配旧类在增量学习过程中的表示漂移。其次,探索层级化动态原型学习,借鉴 Zhao 等[23]的部件级原型设计,结合动态演化机制,可以实现从局部部件到全局概念的层级化动态调整,提升复杂场景下的语义理解能力。

2.4 小结

原型构建已从浅层统计聚类(适用于低维结构化数据)向深度特征驱动的端到端学习(适用于高维感知任务)演进,并呈现融合知识图谱等先验语义的趋势;深度原型尽管显著提升了判别力,但仍面临“原型坍塌”与物理含义缺失的挑战。未来的核心在于构建兼具数据拟合能力与显式语义解释性的“神经-符号”双重原型。

3 基于原型的数据增强与融合方法

深度学习模型对大规模标注数据的过度依赖,是人工智能高效落地的主要瓶颈之一。尤其在军事、医疗等长尾场景下,获取充足样本成本极高。本节探讨原型如何作为一种强先验知识,通过挖掘数据的自相关性(self-correlation)与互相关性(cross-correlation),实现高效的小样本学习与低资源适应。同时,利用原型纠正数据中的噪声与偏差,也为模型的鲁棒性提供了可信的数据基础。以下将分别从单模态数据增强与多模态数据融合两个层面展开综述。

3.1 单模态数据增强方法

单模态数据增强旨在不依赖外部多模态信息辅助的情况下,充分挖掘数据自身的内在拓扑结构,利用原型表示作为稳定的语义锚点,系统性地解决原始数据中存在的样本稀缺、分布偏差及噪声干扰等难题。该类方法超越了传统图像处理中简单的几何变换或像素级扰动,专注于在深层特征空间内,通过样本流形扩充、噪声过滤抑制以及特征重构增强等手段,对数据分布进行重构或优化。其核心逻辑在于认为原型代表了类别的本质特征,通过围绕原型进行操作,可以在保证语义不发生漂移的前提下,最大化地提升模型在复杂场景下的泛化能力与鲁棒性。

在样本生成与扩充方面,核心目标是解决小样本或长尾分布导致的特征空间稀疏与不连续问题,防止模型因决策边界过拟合而失效。利用原型作为类别的高维几何中心,不仅可以引导新样本的生成方向,还能通过分布校准技术“幻觉”出未见过的特征变化,从而修补特征流形的断裂带。Yang 等[28]针对小样本分类中样本不足导致的分布估计偏差问题,提出了一种基于原型的分布校准(distribution calibration)策略。该策略基于一个强假设:基类和新类虽然语义不同,但共享相似的方差统计特性(即类内变化模式相似)。通过将基类丰富的协方差矩阵迁移并施加到新类原型上,有效地“膨胀”了原本坍缩的新类特征分布,使得生成的增强样本在统计学特性上更符合真实数据的离散分布,从而在特征空间模拟出“从少到多”的演变过程。在生成式模型方面,Xu 等[29]提出了一种基于原型引导的代表性样本生成方法,该方法不再进行盲目的随机噪声扰动,而是通过在特征空间中精确定位类原型,并沿着语义变化最丰富的方向(如主成分方向)进行线性或非线性插值。这种方法规避了在低密度区域生成无

效样本的风险,生成了既具备高度多样性又保持严格类别一致性的高质量样本,有效填充了类间空白,缓解了分类边界模糊的问题。此外,随着生成式 AI 的发展,最新的研究开始结合扩散模型(diffusion models, DM)。Redekop 等[30]提出了一种原型引导的扩散生成框架,创新性地将类别原型作为条件嵌入到反向去噪过程中。利用原型的强语义约束,该方法能引导模型从纯噪声中恢复出具有特定类别特征的结构,确保合成数据严格围绕原型分布,避免了传统生成方法中常见的语义漂移或模式崩塌,特别是在病理图像等极度缺乏标注数据且对生成质量要求极高的少样本领域,仅凭少量样本便实现了优异的性能表现。

在噪声抑制与鲁棒性提升方面,主要利用原型的“类内紧凑性”与“类间可分性”来过滤标签噪声或抑制异常值。在实际的非理想数据采集环境中,离群点往往会拉偏类中心导致模型过拟合,而原型作为基于全局样本计算的统计量,具有天然的抗干扰稳定性,是理想的噪声过滤器。Zhi 等[31]针对域适应任务中常见的噪声环境,提出了一种基于类别原型的混淆对校正(confusing pair correction, CPC)方法。该方法利用原型作为稳定的度量标准,在特征空间中动态评估样本与所属类中心的距离可信度,精准识别并修正那些位于类别边界处、极易引发误判的“困难样本对”。通过重新加权或特征修正,该方法有效区分了对模型训练有价值的“困难样本”与具有破坏性的“有害噪声”,从而在保留关键判别信息的同时显著提升了分类边界的清晰度与模型的抗噪能力。

在特征增强与表达优化方面,主要针对单一样本特征表达能力贫乏或存在分布偏移的问题,通过“注入”原型包含的先验信息或“校正”特征的相对位置来增强判别力。钱菲等[32]提出了一种基于原型解耦的特征增强方法,引入了正交分解的思想,利用原型将复杂的图像特征解耦为不同的语义成分(如代表身份的身份原型、代表篡改特征的伪造痕迹原型)。通过这种结构化的特征分解与重组,不仅丰富了样本局部的细粒度语义信息,还通过强制不同原型之间的正交性,防止了模型学习到虚假的捷径特征,从而增强了特征在深度伪造检测等复杂任务中的可解释性与泛化性。赵红等[33]则指出,直接提取的样本特征往往因数据质量问题而偏离真实的流形分布,因此设计了一种自适应原型特征类矫正机制。该机制利用原型在特征空间中的分布特性构建“引力场”,根据样本与原型的相关性动态调整样本特征的位置,实质上是将边缘化或偏离中心的特征强行“拉”向原型方向。通过这种特征层面的校正操作,极大地增强了类内特征的紧凑度并扩大了类间距离,使模型能够学习到更本质、更具区分度的类别表达。

为更直观地展示不同单模态增强策略的技术特性,表 1 选取了文中论述的六种代表性方法,分别从样本生成与扩充、噪声抑制与鲁棒性及特征增强与优化三个维度进行策略分类,并对比了这些策略在数据类型、增强效果及计算复杂度等方面的差异。

表 1　基于原型的单模态数据增强方法特性对比

Tab. 1　Comparison of prototype-based single-modal data augmentation methods

增强策略	数据类型	增强效果	计算复杂度
样本生成与扩充	稀疏特征	插值填充	低
	图像	去噪声	高
噪声抑制与鲁棒性	含噪标签	修正误判	低
	偏离样本	增强紧凑度	中
特征增强与优化	长尾分布	丰富特征分布	低
	复杂特征	正交分离内容	中

3.2　单模态缺失数据补全方法

单模态缺失数据补全旨在不依赖外部模态信息辅助的情况下,充分挖掘数据自身的内在拓扑结构与自相关性。与利用异构信息互补的多模态方法不同,此类方法的核心在于利用原型表示作为数据流形上的稳定“语义锚点”。它假设数据分布具有局部紧凑性,通过度量观测样本与原型之间的几何或概率关联,将缺失值的推断转化为基于局部流形结构的特征重构过程。这种“内省式”的修复机制,使得模型能够利用未缺失特征的上下文信息,在单一视图内部恢复信息的完整性。

针对表格(矩阵)型数据,原型通常被用作划分数据局部子空间的基准,通过“分而治之”的策略实现填补。在这类数据中,利用原型作为聚类中心可以将复杂的高维数据空间划分为若干相对简单的局部线性或非线性子结构,进而建立精准的映射关系。例如,文献[34]通过引入基于原型的信息粒度的概念,利用聚类和模糊规则原型构建模型方法,实现了在数据不完整和缺失数据填补时对系统输入输出关系的鲁棒覆盖。研究[35]引入了“可控原型”概念,模型将补全过程

解耦为“原型检索”和“细节生成”两个阶段,通过将提取点云的几何结构原型作为先验,驱动生成式模型在极高缺失率下重建出多样且精确的三维形状。在文本模态缺失的行人重识别任务中,Gong 等[36]通过构建跨模态原型图,利用图像端的原型知识补全了缺失的文本语义特征,并证明了原型引导在跨模态补全中的可靠性。

面对更加非平稳或动态变化的时间序列数据,单一的静态原型往往不足以覆盖样本随时间演化的复杂规律,因此研究者将高相关性的邻居节点或序列片段视为“参考原型”并进行填补。在这种视角下,原型不再是固定的点,而是表现为动态聚合的上下文信息,旨在捕捉时空依赖性。高杨等[37]针对缺失环境下的多元时间序列异常检测问题,利用传感器物理拓扑或信号相关性将节点构建为图结构,将空间上相邻的节点视为互为支撑的“空间原型”,通过图消息传递(message passing)机制聚合邻域信息来推断并填补当前的缺失值。为解决非平衡采样时间序列中忽视序列间信息的问题,文献[38]提出了原型循环补全模型(prototype recurrent imputation model, PRIME),通过原型记忆模块、双向门控循环单元(gated recurrent unit,GRU)和精细化调整模块整合序列内与序列间特征,显著提升了对缺失值的补全精度。

针对图数据的特征缺失,Tu 等[39]提出了一种端到端的属性缺失图聚类框架,通过将网络学习到的聚类原型作为提示(prompts)来引导缺失节点特征的插补,实现了数据补全与图聚类任务的相互促进。针对少样本场景下的分布偏移问题,Zhang 等[40]提出了一种原型补全网络(ProtoComNet),该方法通过引入物体属性等原始知识作为先验,对样本稀疏导致的残缺类别原型进行精准补偿,显著提升了模型对长尾分布数据的泛化能力。

原型学习通过提取数据的宏观结构先验,能够有效地弥补单模态数据在局部或全局上的信息缺失,有助于实现更具解释性和鲁棒性的数据补全与表征学习。

3.3 多模态数据融合方法

多模态数据(如图像、文本、语音及各类传感器数据)通常分布于不同的特征空间,在联合建模与融合过程中普遍面临表示异构、跨模态对齐不确定以及信息冗余等挑战。通过引入原型学习机制,将原型作为跨模态对齐与特征融合的桥梁,可以在统一的潜在语义空间中实现不同模态之间的对齐、比较与融合,从而支持视觉、文本、语音等多源信息以及多视图数据的联合表征。

基于对比学习的原型融合方法将原型作为锚点或记忆单元,引入对比学习思想,拉近同一样本在不同模态的表示与共享原型的距离,同时推远不同样本视图的距离。例如,Zhou 等[41]通过使用跨模态对比学习,提出聚类原型一致性正则化项,缩短完整原型与局部模态共享表示之间的距离,鼓励局部模态特定表示接近同一类的完整原型,并远离不同类的原型。文献[42]提出了一种基于对比学习的多视图核函数,通过在联合语义空间中显式建模视图间的互补性与差异性,实现兼容传统核理论的多视图核生成方法,并将其应用于多视图聚类以显著提升性能。

此外,针对分布式存储的多视图数据,文献[43]提出了一种通信高效的联邦多视图聚类框架,通过以伪标签和质心矩阵近似数据表示并结合隐式线性核建模样本相似性,在显著降低通信与计算开销的同时,实现了隐私保护下的大规模多视图聚类性能提升。研究[44]将联邦学习机制系统引入多视图模糊聚类,通过在联邦优化框架下进行共识原型学习与原型通信,提出联邦多视图模糊 C 均值共识原型聚类方法,在隐私保护条件下实现了高效且性能优越的多视图聚类。

基于注意力与推理的原型融合方法将原型视为可推理的概念单元,通过注意力机制动态地计算样本与不同原型的关联度,并基于此关联度从多模态信息中检索、加权和融合信息。Ni 等[45]在用语义增强模块处理文本特征以获得文本原型的基础上,基于注意力机制或加权聚合思想,提出了多模态原型增强模块,以融合视觉和文本原型。Guo 等[46]引入了可学习的多视图原型,并通过视图引导的注意力模块来增强文本特征。Zhang 等[47]借助对比语言 - 图像预训练[48](contrastive language-image pre-training, CLIP)进行特征提取,并通过使用跨模态注意力和残差学习动态融合多模态原型,进而提出了一种多模态原型融合神经网络,取得了良好且鲁棒的图像分类性能。

基于生成式网络的原型融合方法利用生成式模型,如生成对抗网络(generative adversarial network, GAN)、变分自编码器(variational autoencoder, VAE),来生成或增强原型,实现模态或视图间的特征对齐和融合。Li 等[49]针对不同模态间数据互补性未充分利用这一问题,通过记忆库形式构建了样本级原型,并基于生成对抗网络获得模态级原型,并进行整合,发挥两者互补

性。Cheng 等[50]设计了一个跨模态条件重建模块,在训练阶段通过重建掩码图像和报告来交换不同模态的信息,同时构建了一个基于句子的原型记忆库,使网络能够专注于低层次的局部视觉特征和高层次的临床语言特征,并使用非自回归生成范式来重建非顺序报告。

基于图的原型融合方法显式地构建图结构,即以节点为原型或样本、边为关系,并利用图神经网络、矩阵优化等方式进行消息传递,从而更新和融合原型。Shi 等[51]在构建原型图的基础上,进行谱嵌入以获得实矩阵,随后使用谱旋转以获得指标矩阵,并提出了交替优化策略进行模型求解。Yang 等[52]融合了共识二进制编码、代码压缩、有符号原型图和基于原型的聚类分配等技术,捕捉数据的底层结构,提高了数据融合性能、计算效率。Yu 等[53]为每个模态引入一组不同数量的原型,从而灵活地提取属于每个模态自身的数据特征,由于生成的图具有不同规模,更能全面地描述数据的整体相似性。

基于对齐的原型融合方法利用原型来对齐不同模态的潜在分布,或通过原型来引导跨模态的生成过程,实现原型空间模态间的语义一致性。Le 等[54]利用完整的数据原型,在模态共享层面通过跨模态正则化以及在模态特定层面通过跨模态对比机制来提供多样化的知识,同时引入了跨模态对齐,为模态特定特征提供正则化,从而增强了整体性能。Xie 等[55]引入了原型发现和继承机制,在初始特征编码期间建立可靠的跨模态对齐,提取并汇总多个邻域语义原型,以促进开放词汇识别。

3.4　多模态缺失数据补全方法

与单模态方法侧重于数据内部的自修复不同,多模态缺失数据补全的核心在于利用模态间的互相关性来弥补单一视角的盲区。在这一场景下,原型不再仅仅是单一视图的聚类中心,而是被视为连接异构特征空间的“语义桥梁”。该类方法旨在建立跨模态的原型映射与对齐机制,使模型能够将完整模态的丰富语义信息“翻译”并迁移至缺失模态的表达中。通过这种“交互式”的补偿策略,原型学习在潜在语义空间中实现了跨视角的知识共享,从而有效解决信息缺失带来的语义鸿沟问题。

在多视图聚类任务中,研究[56]通过学习共享或视图特定的语义原型,将原型作为跨视角对齐的中介结构,使缺失视角样本能够借助原型映射获得间接表达。该研究通过原型匹配、一致性约束或对比学习机制,减小不同视角之间的分布差异,并在潜在空间中实现缺失视角信息的隐式补全。进一步地,有研究[57]将原型学习与图结构建模及注意力机制结合,通过刻画样本—原型及原型—原型之间的关系,实现跨视角语义信息的结构化传播,从而在高缺失率条件下仍保持稳定的聚类性能。除集中式学习场景外,原型驱动的不完整多视图建模也被拓展至隐私约束环境中。相关研究[58]在联邦学习框架下引入模糊原型与隶属度建模,通过在本地视图上学习原型相关统计信息并进行跨方协同优化,实现不完整视角条件下的聚类补全与决策一致性,为原型学习在分布式多模态场景中的应用提供了新的范式。

在多模态下游任务中,原型学习常与知识蒸馏(knowledge distillation)策略相结合,用于缓解模态缺失对模型判别能力的影响。文献[59]通过从完整模态教师模型中提取原型级语义知识,并将其映射至缺失模态条件下的学生模型,使学生模型在推理阶段能够借助原型对齐获得接近完整模态的语义表达。该类方法强调将原型作为跨模态共享语义锚点,在补全缺失模态信息的同时,有效提升模型的鲁棒性与泛化能力。

针对医学影像等对模态依赖性较强的应用场景,研究[60-61]关注模态缺失率不均衡对学习过程的影响,并在原型或语义层面引入偏好感知与多层次蒸馏机制,以动态调节不同模态在训练过程中的贡献权重,避免模型过度依赖高可用模态,从而提升对稀缺模态的补全效果与整体分割性能。

基于原型学习的多模态缺失数据补全方法通过显式建模高层语义结构与跨模态对齐关系,为缺失模态样本提供了一种具有可解释性的间接表达途径。相较于传统基于数据重建的补全策略,该类方法更加关注语义一致性与结构约束,在多种任务中展现出良好的鲁棒性与扩展潜力。

3.5　小结

通过对比可以发现,基于原型的数据增强本质是利用先验分布约束特征空间的修复与生成:单模态方法侧重挖掘数据的自相关性与流形结构,计算高效但受限于信息熵上限;而多模态方法利用跨模态原型的互相关性与语义对齐,虽然计算复杂度随模态数量增加,但能通过异构信息的互补显著提升复杂场景下的鲁棒性,是解决信息缺失的关键路径。

4　基于原型的可解释建模与推理方法

将“黑箱”神经网络转化为可信的透明系统,

是人工智能应用于高风险决策领域的关键前提。与传统端到端模型仅输出预测结果不同,基于原型的推理方法致力于构建基于案例或基于规则的显式决策路径。通过度量测试样本与原型的相似度,模型能够提供人类可理解的决策依据,从而显著提升系统透明度与用户信任感。

4.1 原型神经网络建模与可解释推理方法

原型神经网络作为可解释人工智能的重要实现途径,通过将类别表示为原型向量并在度量空间中进行相似度比较,为模型决策提供了直观的推理依据。该类方法不仅在小样本学习中表现出色,更重要的是其内在的可解释性机制,使得决策过程对用户透明可理解。Snell 等[19]提出的原型网络是小样本学习中的里程碑工作,其核心思想是将每个类别表示为支持集中样本的特征均值(原型),通过计算查询样本与各类别原型之间的欧氏距离进行分类决策。该方法简洁有效,且决策依据可直接通过距离可视化展示,为后续可解释原型学习研究奠定了基础。Allen 等[62]进一步扩展了这一框架,提出了无限混合原型(infinite mixture prototypes, IMP),通过为每个类别动态分配多个聚类中心,有效提升了模型处理复杂、多模态数据分布的能力。在图像分类任务中,原型网络通过可视化原型与测试样本之间的特征相似性,为用户提供了直观的决策解释。例如,在细粒度鸟类分类中,模型可以展示测试图像与各类别典型羽毛图案原型的匹配程度,使分类结果更具说服力[63]。为提升原型网络的可解释性,研究者提出了多种增强机制。Chen 等[64]开发的可解释原型部分网络(interpretable prototype part network, ProtoPNet)将原型嵌入卷积网络的中间层,并通过反向投影技术将原型可视化显示为训练图像中的典型局部模式。该方法不仅提供了"此图像看起来像那类原型"的直观解释,还通过优化原型与训练样本的对应关系增强了表示的语义一致性。Liu 等[65]提出的分层组合网络(hierarchical compositional network, HCN)通过学习共享语义空间中的属性原型,构建了一个可解释模型,其以"上级类别 + 属性原型"的线性组合来明确定义每个类别,从而模拟了人类依据层次化标准进行分类的决策过程。

原型注意力机制通过聚焦输入样本的关键区域进一步增强了解释能力。Yu 等[66]提出了一个基于全局与局部自适应原型网络(global-local adaptive prototypical network, GLAP),它通过注意力机制增强特征表示,并利用多尺度类别原型进行元学习度量,以综合解决小样本学习中数据依赖、特征表示弱和复杂任务性能差三大问题。Wang 等[67]提出了一种基于 Transformer 的原型搜索网络(transformer-based prototype search network, TPSN),它通过编码器整合图像区域信息,并利用解码器为每个类别生成多个自适应原型,以替代传统单一的原型,从而更合理地表达对象特征及其背景依赖关系,最终在小样本语义分割任务上取得了领先的性能。Hong 等[68]提出一个名为 ProtoryNet 的可解释深度神经网络,它引入了"原型轨迹"的新概念,通过捕捉文本序列中每个句子与最相似原型之间的动态关联模式来进行分类,从而实现了与人类分析文本相似的高质量、细粒度的模型解释。针对分布式存储数据,研究[69]提出了一种联邦原型学习(federated prototype learning, FedProto)框架,通过以类别原型替代梯度进行跨客户端知识聚合与对齐,有效缓解异构联邦学习中的优化失配问题,并在理论与实验层面验证了其收敛性与性能优势。随后,Yang 等[70]在联邦学习中引入多层原型对比学习(prototype contrastive learning, PCL)与原型引导的软标签协同机制,以原型与软标签替代模型参数进行知识传递,从而缓解数据异构带来的偏置并提升通信效率与模型性能。

在医疗影像分析领域,原型网络通过展示输入图像与病理原型之间的相似性,为医生提供可信的诊断依据。Wang 等[71]设计了可解释医学图像分类框架 MProtoNet,其通过引入多模态注意力机制(利用医学报告辅助)和位置嵌入模块,有效利用文本信息和图像空间位置来优化原型的学习与激活,从而在提升疾病分类准确率的同时增强了模型的可解释性。类似地,在皮肤病变诊断中,原型模型通过展示病变图像与恶性/良性原型模式的匹配程度,提供了直观的诊断解释[72]。针对电力变压器故障数据稀缺的问题,Deng 等[73]提出了一种基于高斯原型网络(Gaussian prototype network, GPN)的小样本诊断方法,通过将嵌入网络与距离度量有机结合,即使在少量样本下也能实现准确预测。论文[74]提出"一致性得分"与"稳定性得分"两项指标,定量评估原型网络的可解释性,并设计了特征对齐与分数聚合模块以显著提升原型的解释质量与鲁棒性。研究[75]在乳腺 X 光与结直肠息肉数据集上验证了原型网络在医学图像分析中的可解释性,发现原型与临床关注区域具有视觉一致性,并探讨了将其融入临床决策的可行路径与挑战。

4.2　基于原型的模糊规则建模与推理方法

在基于原型的模糊规则建模与推理研究中,原型作为构建可解释模糊规则的核心载体,其生成、优化与集成方法得到了广泛探索。Li 等[76]提出了通过粒子群优化(particle swarm optimization, PSO)算法重新定位原型并分配信息粒度的模糊规则模型,实现从数据中学习内部结构信息并构建更鲁棒的粒度模糊模型。针对规则的可读性与效率问题,Niu 等[77]从粒度约简的视角出发,提出了一种基于模糊概念格子的增量式模糊规则分类方法(fuzzy rule-based classification method based on incremental fuzzy concept lattices, FRCM),有效简化了复杂问题并实现了高效的规则学习。Hu 等[78]则设计了面向多分类的模糊规则分类器,并通过联邦梯度优化机制实现了在分散数据下的隐私保护建模,展现了原型与规则模型在分布式环境下的应用潜力。

为了应对模糊规则系统的复杂性并优化其结构,Mamaghani 与 Pedrycz[79]提出了一种结构优化方法,通过粒子群优化算法分别对规则前件和后件进行高效安排,实现了模型复杂度的有效管理。Tsekouras[80]引入了一种结合多维尺度和非线性约束优化的新颖方法,通过近似模糊集间的相似性度量来简化和合并相似规则,从而在保持高精度的同时获得简单透明的模型结构[80]。对于高维问题,Zhao 等[81]提出了一种多目标进化分层模糊回归系统(multi-objective evolutionary hierarchical fuzzy regression system, MOEHFRS),通过灵活构建拓扑结构来交换和组合子模糊系统,在保证精度的同时显著减少了规则总数。

将模糊系统与深度学习结合是另一个重要趋势。Wang 等[82]提出了一种基于改进 WM 的深度模糊规则分类系统(deep fuzzy rule-based classification system based on an improved Wang-Mendel, DFRBCS),通过分层特征变换在模型可解释性和预测精度之间取得了良好平衡。Gu 等[83]针对零阶演化智能系统,提出了一种多目标进化优化方法,通过同时优化训练误差和类内方差来获取最优原型,从而在避免过拟合的同时提升了分类性能。

在动态与增量学习方面,Li 等[84]提出了增量模糊模型的概念,利用模糊规则来补偿全局简单模型产生的误差,并通过增强型模糊 C 均值进行设计,适用于在线学习环境。Rudnik 等[85]则提出了一种从增量数据集中生成有序模糊规则的方法,能够记录不确定性及其变化,为动态不确定环境下的决策提供了支持。Škrjanc 等[86]的综述系统总结了演化模糊与神经 - 模糊方法在聚类、回归和分类中的在线、实时应用,为增量式原型模糊系统的发展提供了框架性指导。

在规则精简与可解释性评估方面,Bai 等[87]提出了一种基于多核学习的规则约简(multi-kernel learning based rule reduction, MKLRR)方法,通过将多个模糊集合并并映射到高维空间,在保持建模精度的同时大幅减少了规则数量。Pontoizeau 等[88]开发了名为模糊推理编译器(fuzzy inference compiler, FuzzIC)的工具,通过计算多种互补性指标来系统评估模糊规则库的可解释性,并通过用户实证研究揭示了哪些指标对实际应用中的可解释性提升最为关键。

最后,在应用层面,该类模型在医疗、金融、安全等领域展现出强大价值。Czabanski 等[89]提出了一种基于模糊聚类和进化策略的规则库精简方法,并将其成功应用于基于胎心宫缩信号的胎儿状态评估,在保持高分类性能的同时增强了模型的实用性。Almseidin[90]提出了一种结合模糊规则插值(fuzzy rule interpolation, FRI)和深度神经网络的混合入侵检测方法,利用置信度驱动融合机制,有效处理了稀疏规则问题并提升了系统对各类攻击的检测能力与适应性。

4.3　基于原型的溯因与因果推理方法

基于原型的溯因与因果推理方法将原型表示与因果推断结合,为机器学习模型提供了深层次的解释能力。该方法通过将原型作为因果推理的参考点和干预基准,不仅能够解释“是什么”,更能回答“为什么”的问题,推动了可解释人工智能从相关性分析向因果推理的深化发展。基于原型的溯因与因果推理方法的核心在于利用原型作为因果机制的载体或干预的基准,从而将统计关联提升为可解释的因果陈述。因果推断的理论为该方法提供了坚实基础,如 Peters 等[91]的著作为从观测和干预数据中学习因果模型提供了系统的理论基础与算法框架,强调了从数据中推断因果关系的原则与方法。在实际应用中,反事实推理(counterfactual reasoning)是实现因果解释的关键手段。Duong 等[92]提出了一个原型驱动的反事实解释(prototype-driven counterfactual explanation, ProCE)框架,通过保留特征间的潜在因果关系来生成既合理又可实现的反事实样本,为分类模型的决策提供了因果性解释。同样地,Shao 等[93]提出了基于因果干预的反事实解释(causal intervention-based counterfactual

explanation, CUBE)方法,通过因果干预建模反事实生成过程,并利用因果导向器捕获数据分布中的因果关系,从而生成高质量且符合因果律的解释。在小样本学习场景中,结合因果视角能有效缓解由混淆因素导致的伪相关问题。Lin 等[94]从因果角度重新审视了小样本学习,将现有度量方法解释为前门调整的具体形式,并据此提出了考虑样本间关系和表示多样性的新方法,以学习更稳定的因果关系。

在要求高可靠性的医疗领域,因果推理与原型学习的结合尤为重要。Zhang 等[95]针对医学图像中的异质性问题,提出了混合原型校正因果推理(mixed prototype correction-based causal inference, MPCCI)方法,通过前门调整因果框架和混合原型校正模块进行因果干预,有效缓解了未知混杂因素对诊断模型的影响。Prosperi 等[96]强调了在精准医疗中从预测模型转向因果干预模型的必要性,并讨论了目标试验、可移植性等关键概念,为医疗领域的因果推理提供了指导框架。Sanchez 等[97]进一步探讨了如何将因果机器学习整合到临床决策支持系统中,并指出因果表示学习等方向是解决医疗高维数据等挑战的潜在路径。

在时序因果效应估计方面,Grecov 等[98]提出了一种基于全局循环神经网络的反事实预测方法,通过同时建模处理组和对照组的时间序列,能够更精确地分离和评估政策干预的因果效应。此外,从更基础的认知层面看,Stukker 与 Sanders[99]的研究揭示了因果关系在语言连接词中的原型结构,表明人类对因果关系的认知本身具有基于原型的范畴化特点,这为构建更符合人类认知习惯的可解释因果模型提供了语言学启示。

因果发现是从数据中推断因果图结构的关键步骤,为基于原型的因果推理提供了先验知识。Zanga 等[100]对因果发现领域进行了全面综述,系统性地概述了从数据中恢复因果图以识别和估计因果效应的各类算法与实用工具。针对时序这一特殊且普遍的数据结构,Gong 等[101]对时序因果发现进行了系统梳理,涵盖了多元时间序列和事件序列两大类别,为在动态系统中应用因果推理提供了重要指导。在此基础之上,Pfister 等[102]提出了面向时序数据的因果预测方法,即使在没有明确环境信息的情况下,也能利用数据的顺序性来推断瞬时因果关系,增强了对时间序列中因果关系的识别能力。为了高效地从非平稳时间序列中发现动态变化的因果关系,Pan 等[103]提出了 EffCause 算法,通过优化滑动窗口内的因果检测,在保证准确性的同时大幅提升了计算效率。

在具体应用层面,基于因果与原型的方法在故障诊断、医疗影像和自然语言处理等多个领域展现出强大解释力和可靠性。Luo 等[104]提出了因果时序图注意力网络(causal temporal graph attention network, CTGAN),通过因果推断构建化工过程故障传播图,并结合注意力机制定位关键变量,实现了高性能与高可解释性的故障诊断。Castro 等[105]强调了因果推理对于解决医学影像中数据稀缺和分布不匹配等挑战的重要性,为建立图像与其标注间可靠的因果关系提供了指导。在视频问答任务中,Zang 等[106]从因果表征的视角出发,提出了一个新颖的推理框架,通过捕捉与问题语义因果相关的视觉特征,削弱了局部语言语义的偏差,增强了模型的泛化能力。

将因果干预机制与深度学习模型结合,是提升模型鲁棒性和可解释性的有效途径。Huang 等[107]提出了基于因果干预的多头注意力网络(causal intervention-based multi-head attention network, CaIMA),通过探索多区域注意力与诊断结果之间的内在因果关系,鼓励网络学习对医学影像诊断更有用的注意力图。Chen 等[108]则引入了一种稀疏时序逻辑网络(sparse temporal logic network, STLN),将神经元概念化为逻辑命题,通过时序逻辑语言为轴承故障诊断决策提供形式化、可解释的说明。在药物发现这一强因果性领域,Michoel 与 Zhang[109]指出因果推断是减少认知偏差、改进决策的关键,并综述了其在药物研发价值链中的应用。Zhou 等[110]更进一步提出了一种基于因果发现的多目标结构式药物设计方法,通过构建性质间的因果图来合理分解联合分布,从而引导生成同时满足多个目标要求的候选分子。

最后,从理论层面审视可解释性本身,Marconato 等[111]提出了一个基于因果表示学习的人类可解释表征学习(human-interpretable representation learning, HRL)数学框架,通过显式地将人类利益相关者建模为外部观察者,形式化了机器表征与人类概念词汇之间的“对齐”概念,为构建真正可被人理解的解释奠定了理论基础。Chou 等[112]的系统性综述则指出,当前多数与模型无关的反事实解释算法缺乏因果理论形式化基础,因而难以向人类决策者提供真正的因果解释能力,这为未来研究指明了方向。

4.4　时序数据的原型学习与推理方法

时序数据的原型学习与推理方法通过将复杂的时间动态抽象为具有代表性的原型,为理解序列数据提供了可解释的框架。在表示学习层面,Cai 等[113]提出了因果导向的表示学习预测器(causal-oriented representation learning predictor, CReP),通过将原始空间分解为与目标变量因果相关、效应相关及非因果的三个正交潜在因子,从统一视角实现了多步预测与因果发现。Chen 等[114]则针对现实世界中普遍存在的非可逆生成过程,提出了非可逆生成过程下的因果表征(causal representation under non-invertible generation, CaRiNG)方法,在理论可识别的保证下学习时序数据的因果表示,有效提升了时序理解与推理能力。这些工作印证了 Schölkopf 等[115]的论断,即因果表示学习对于解决机器学习中的迁移与泛化等核心问题至关重要。

在具体应用任务中,原型作为跨模态对齐和特征学习的枢纽发挥了关键作用。Wei 等[116]针对连续手语识别任务,提出了用于连续手语识别的跨模态自适应原型学习(cross-modal adaptive prototype learning for continuous sign language recognition, CAP-SLR)模型,通过自适应地融合视觉特征、词汇原型和文本特征,有效缓解了数据冗余和标注稀疏问题,提升了识别精度。对于自监督时序表示学习,Wei 等[117]提出了排序邻域与类别原型对比学习(ranking neighborhood and class prototype contrastive learning, RESEAL)框架,通过利用相邻时间戳的相似性排序信息并结合类别原型对比学习,缓解了采样偏差问题,在分类、异常检测和预测等多个任务上学习到了泛化性强的表示。在联邦学习场景下,Fang 等[118]提出了对比原型引导的联邦学习(federated contrastive prototype guided learning, FedCPG)方法,利用全局原型生成器和分层对比学习策略,有效应对旋转机械故障诊断中时空域偏移带来的挑战。

动态时间规整(dynamic time warping, DTW)作为时序相似性度量的经典方法,与原型学习结合能有效捕捉时间形变下的模式。Iwana 与 Uchida[119]提出将 DTW 匹配过程中产生的局部距离特征作为一种新颖的输入,与原始数据结合,共同送入多模态融合网络进行分类,探索了不同原型选择方法的影响。Shi 等[120]在地震数据任务中,将稀疏表示与平滑形状 DTW 结合,通过修改原型对齐算法,有效减少了波形失真。原型能够有效概括时间序列的全局与局部特征。此外,针对存在外部干扰的复杂预测场景,研究[121]提出了条件因果表示(conditional causal representation, CCR)模型,通过提取与干扰相关的因果表示并学习受外部干扰的因果机制,显著提升了预测模型的精度、鲁棒性和泛化能力。

在视频行为理解领域,Li 等[122]提出了 PCL 框架,通过原型学习显式地发现标注帧与未标注帧之间的类别关系,并利用对比学习拉近同类原型、推远异类原型,为点监督时序动作检测生成了高质量的伪标签。对于多元时间序列分类,Liu 等[123]提出的时序动态图神经网络(temporal dynamic graph neural network, TodyNet)模型能够提取潜在的时空依赖关系,并通过动态图捕捉时间槽之间的关联,其引入的时序图池化层有效获得了全局的图级表示(可视为一种复杂的原型),显著提升了分类性能。

在视频理解领域,原型学习被广泛应用于弱监督和少样本场景下的时序动作定位与分割。Luo 等[124]针对弱监督时序动作定位中的定位不完整和背景干扰问题,提出了自适应原型学习(adaptive prototype learning, APL)方法,该方法利用自适应 Transformer 网络提取视频级的自适应原型,并结合基于最优传输的协同训练策略,实现了鲁棒的定位。Tang 等[125]则针对少样本视频目标分割,提出了整体原型注意力网络(holistic prototype attention network, HPAN),通过原型图注意力模块和双向原型注意力模块,从所有前景特征中生成局部原型并利用其内部相关性来增强整体原型的表示,实现了支持 - 查询语义一致性和帧内时序一致性。

对于少样本动作识别,研究者们探索通过双原型和自校准机制来增强原型代表性。An 等[126]提出了无监督的原型自混合校准(unsupervised prototype self-hybrid calibration, UPSHC)方法,通过混合注意力网络、双自适应对比学习机制和无监督原型自校准模块,以无监督的方式利用未标注查询样本来优化目标原型,无须额外训练即可提升性能。Zhang 等[127]针对小样本动作识别任务,提出了一个集成高效时空建模与跨模态语义的新框架,其中文本增强原型模块通过在多层次融合文本和视觉特征来增强原型表示,提高了原型的可判别性和泛化性。

层次化原型学习通过在不同粒度上构建和优化原型,能够更精细地捕捉数据的语义结构。Gao 等[128]提出了一种用于不平衡时序推荐的层次类别增强推荐(hierarchical category-enhanced

recommendation, HCRec)算法,该算法根据层次类别解耦变体和不变体信息,然后使用不变体原型来降低特殊行为的影响。Gu[129]提出了一种新颖的自训练层次化原型方法用于半监督分类,该方法从标记样本中识别多个粒度级别的有意义的原型,并自组织成一个高度透明的、多层级的识别模型。Zheng 等[130]提出了带有层次原型对比学习的谣言检测框架,通过对比学习构建一组动态更新的层次原型,以鼓励捕捉谣言内部的层次化语义结构。在少样本关系分类中,Bi 等[131]提出了一种新颖的多尺度层次化原型学习方法,在集合间、类间和类内三个层面捕捉关系交互信息,增强模型对全局语义信息的理解,帮助其区分类间的细微差异。Gao 等[132]提出了一种用于少样本关系三元组提取的层次化原型优化方法,通过提示学习将关系标签信息合并到文本中,并引入层次化对比学习来分别改善实体和关系原型之间的度量空间。Zhang 等[133]提出了层次化原型网络用于持续图表示学习,该网络以原型的形式提取不同层次的抽象知识来表示不断扩展的图,从而在遇到新任务时,只有相关的原子特征提取器和每层的原型会被激活和优化,而其他部分保持不变以维持对现有节点的性能。

在异常检测方面,原型能有效表征正常模式,从而识别偏离这些模式的异常点。Cai 等[134]提出了基于原型的模糊粗糙集方法,通过基于可分离属性选择的原型学习来执行异常检测,有效消除了异常点之间关系的影响。Li 等[135]提出了一种面向原型的多元时间序列无监督异常检测方法,将多个时间序列视为一组原型上的分布,这些原型被提取出来代表多样化的正常模式,并利用元学习得到的可迁移原型,使模型对新时间序列具有高适应能力。Liu 等[136]提出了基于逆向(反转)的异常检测(inversion-based anomaly detection, InvAD)统一框架,通过采用具有数学保证的信息保持可逆神经网络,将原始复杂信号分解为分布内特征和分布外特征,从而能够同时检测分布内和分布外异常。Huang 等[137]提出了一种用于视频异常检测的原型引导和动态感知长距离帧预测范式,采用原型引导的动态匹配网络来减弱模型对异常的泛化能力,并通过动态原型匹配机制探索时序上下文。

在脑机接口等专业领域,原型学习也展现出其价值。Han 等[138]提出了一种空 - 谱与时序双原型网络,通过双原型学习来优化特征空间分布和训练过程,从而提高了模型在小样本运动想象数据集上的泛化能力。此外,Gama 等[139]关于概念漂移适应的综述,为在动态变化的时序环境中进行自适应学习(包括原型模型的更新与演化)提供了重要的理论基础和方法论指导。

4.5 原型学习的可解释性评估

随着原型推理模型从简单的特征匹配向复杂的因果与时序逻辑演进,如何构建一套客观、系统的可解释性评估体系已成为衡量模型“可信度”的关键。早期的评估主要依赖定性可视化,即通过热力图或图像块裁剪,人工检查学习到的原型是否与人类认知的语义概念(如鸟的头部、车的轮胎)在视觉上保持一致。然而,这种方法存在主观性强且难以量化的问题。近年来,研究者开始引入定量评估指标,主要分为以下四类:

1)语义一致性与稳定性评估。针对原型神经网络和时序原型,核心在于评估原型是否在不同样本间稳定指代同一语义概念。Huang 等[74]提出了“一致性分数”,通过量化原型在类内样本激活区域的重叠度来评估其语义纯度,并利用“稳定性分数”衡量输入受到微小扰动时解释结果的鲁棒性,从而解决了传统可视化评估的主观性问题。

2)推理忠实度评估。针对模糊规则与因果推理,重点在于验证模型输出的解释是否真实反映其内部决策逻辑。Nauta 等[140]提出的基于图块的直观原型网络(patch-based intuitive prototypes network, PIP-Net)引入了分布外(out-of-distribution, OOD)检测作为忠实度指标,证明了高质量的原型系统应具备拒判能力,即当样本特征无法被任何原型解释时,模型应输出低置信度以示拒判,而非强行匹配。

3)人类可模拟性评估。这是衡量人机互信的核心指标。Hase 等[141]定义了“模拟测试”范式,即让不知晓模型参数的人类测试者仅根据原型提供的解释(如相似案例、推理路径)来预测模型的输出结果。如果人类预测的准确率接近模型实际准确率,则说明该原型系统的推理逻辑与人类认知高度对齐。

4)不确定性量化评估。针对高风险决策场景,评估原型能否作为不确定性的度量基准至关重要。Zelenka 等[142]的研究表明,通过计算测试样本与属性原型的距离分布,可以构建可解释的“认知不确定性”指标。该指标不仅用于评估模型的置信度,还能有效识别数据中的异常值与分布偏移,是评价系统安全性的重要维度。

4.6　小结

本节总结了原型推理从“静态特征匹配”向“动态逻辑归因”的演进:原型神经网络与模糊规则建模主要通过几何距离与模糊隶属度处理分类边界的不确定性,提供了直观的“案例”与“规则”解释,计算相对高效;溯因与因果推理则进一步突破相关性局限,通过显式建模原型间的因果链条实现“知其所以然”的逻辑溯源,代表了高可信决策的研究方向;而时序原型推理将这一机制扩展至动态数据流,能够有效解决时间维度的动态建模问题。

5　基于原型的生成式模型构建方法

在人工智能生成内容(artificial intelligence generated content, AIGC)与大模型时代,如何兼顾生成的质量与逻辑的可信性,同时降低模型微调的高效算力成本,是当前研究的热点。原型学习在此展现出新的生命力。通过在表示空间中引入类级或语义级“原型”,可以为生成式模型提供一种紧凑而可解释的知识结构,从而在少样本场景下更好地对齐语义,在复杂关系建模中增强结构一致性,并在推理过程中提供更直观的解释支撑。利用原型辅助大模型推理,不仅有望解决大模型幻觉问题、提升可信度,还可以避免昂贵的全参数训练,实现高效的知识更新。

5.1　原型引导的生成式表示学习

生成式模型旨在学习数据分布并生成新样本,但其在小样本场景下面临模式崩溃、语义模糊和域偏移等挑战。原型学习通过提供清晰、强语义的类中心先验,为各类生成式模型提供了有效的结构约束和条件引导,显著提升了生成过程的可控性、语义一致性和跨域泛化能力。

变分自编码器作为经典的生成式模型,依赖高斯先验学习数据分布,但这种先验在类别可分性和可解释性上存在不足。引入原型学习后,VAE 的潜在空间得以结构化约束,原型向量不仅作为类级中心提升了语义可分性,还为生成过程提供了更强的监督信号,使得模型在小样本条件下具备更好的聚类性、解耦性与可解释性。围绕这一思路,已有方法大体可以分为两类。一类是通过确定性原型直接约束潜在空间分布,如原型变分自编码器[143](prototypical variational autoencoder, ProtoVAE)在潜在空间中引入类级原型向量,使得生成样本更具语义一致性;变分原型编码器[144](variational prototyping-encoder, VPE)将真实样本向对应的原型样本进行转换作为元任务,利用原型样本提供的强监督信号,使同类真实样本特征在隐空间围绕原型聚集。另一类则是通过概率建模增强原型表示的灵活性,如高斯过程变分自编码器[145](Gaussian process variational autoencoder, GP-VAE)通过构建以原型为中心的高斯混合分布,实现多样化且可区分的特征采样,从而提升小样本下原型表示的鲁棒性。

生成对抗网络通过对抗训练提升样本的真实性与多样性,但在小样本场景下容易出现模式崩溃和语义不可控的问题。引入原型学习后,GAN 框架能够借助类级原型作为条件或约束,使生成过程具备更强的语义一致性和可控性。原型生成对抗网络[146](prototype generative adversarial network, ProtoGAN)将类原型输入生成器,以确保生成样本围绕原型特征分布,同时在判别器中引入原型距离的度量,使判别器不仅区分真假,还能强化生成样本与类原型之间的语义关联,从而输出更具代表性和多样性的结果。

扩散模型作为新一代生成式模型,通过迭代去噪过程生成样本。其尽管生成质量卓越,但在小样本和条件生成任务中,仍需有效的语义控制机制。原型思想被引入扩散过程,作为引导生成轨迹的“语义锚点”。基于任务引导扩散的原型网络[147](prototypical networks by task-guided diffusion, ProtoDiff)提供任务特定的“过拟合原型”作为监督信号,引导扩散过程从随机噪声中生成更具判别力的任务相关原型,从而提升小样本分类的准确性与泛化能力。结合原型学习的无分类器扩散引导[148](classifier-free diffusion guidance with prototype learning, ProtoDiffusion)方法使用原型学习为扩散模型提供了更具代表性的类条件嵌入,使扩散过程从一开始就获得更稳定的引导,从而加快训练收敛速度并提升生成质量。解耦原型判别学习[149](decoupled prototype discriminative learning, DPDL)方法构建结构化的多高斯分布目标空间,引导正常样本向紧凑原型聚集、异常样本远离,从而增强扩散建模的判别性与对未知异常的泛化能力。原型学习通过提供强语义、结构化的先验引导,增强了扩散模型在生成过程中的语义可控性、训练稳定性以及对下游任务的判别性泛化能力。

生成式零样本学习(generative zero-shot learning, GZSL)旨在利用类别属性或语义表示生成未见类样本,但静态语义原型与真实视觉分布

之间存在显著的域偏移,导致生成样本难以兼顾语义一致性与视觉真实性。为此,近年来大量研究引入动态语义原型,使模型在生成过程中能够自适应调整语义表示,从而缩小语义与视觉之间的鸿沟并提升跨域泛化能力。Chen 等[150]通过迭代更新语义原型,使其逐步贴近视觉分布,缓解了固定原型的失配问题;Wang 等[151]借助语义引导与置信约束提升类判别性;Jiang 等[152]与 Yu 等[153]利用跨模态对齐与任务模拟增强泛化能力;Chen 等[154]与 Wu 等[155]则通过语义对齐、自监督结构引导缓解类别混淆与域偏移。在此基础上,Hou 等[156]进一步结合视觉统计先验与语义更新机制,在生成过程中动态修正语义原型,从而显著提升生成样本的真实性与稳定性。动态语义原型方法通过演化更新与跨模态交互,解决了静态语义原型的域偏移问题,使生成样本在语义一致性与视觉真实性之间取得平衡。

5.2 原型增强的图学习方法

在图学习中,仅依赖消息传递难以充分捕捉复杂语义与关系。原型学习以类中心或语义代表为核心,与图神经网络和异构图等结合后,能够在更高层次上增强表示能力,既可作为中介提高模型可解释性,又能在少样本、不平衡与开放集场景中提供稳健的度量基准,还可在异构图中抽取跨域概念,提升知识迁移与对比学习效果。

在图神经网络的应用中,可解释性问题一直备受关注。传统 GNN 的预测往往依赖高维嵌入,难以直观揭示模型的决策依据。原型学习通过在潜在空间中学习一组可复用的“语义原型”,使模型能够以类比的方式进行推理,并在预测时显式地衡量输入子图或节点与这些原型的相似度。Zhang 等[157]和 Dai 等[158]均通过将输入图与潜在空间中的类原型进行相似性比较,分别实现节点/图级别的分类和自解释;Zheng 等[159]借助可学习的聚类中心向量,显式建模图结构中的社团分布,并通过分布一致性约束增强表示稳定性。这类方法使原型同时承担判别基准和语义可视化双重功能。Shin 等[160]利用原型发现每类中具有代表性的子图模式,揭示模型依赖特征;Seo 等[161]进一步结合信息瓶颈机制,用原型识别对预测最关键的子结构,在保持性能的同时提升解释的简洁性和可辨性。在上述模型中,原型既充当了类别判别的基准,又为用户提供了可视化、可解释的语义支撑。

在图学习任务中,训练样本稀少、类别分布不平衡,以及测试中出现未知类别(开放集)的情形广泛存在。原型学习通过建立每个类别的“类中心”或代表向量/区域,来提供一种度量基础,从而在上述挑战场景中提升模型的泛化性和判别稳定性。Zhang 等[162]构建每类的内部与边界原型以应对开集识别中的类间混淆与类内差异问题;Zhu 等[163]通过原型与标签传播的联合优化,迭代增强小样本分类的准确性与鲁棒性。此外,原型方法也被成功扩展至更复杂的少样本图结构预测任务,如 Li 等[164]将谓词类别分解为多个基于主客体的可组合原型,并通过自适应聚合有效建模高类内方差,提升了小样本场景图生成的泛化能力。

在异构图和知识图谱中,不同类型实体与多种关系使得语义结构复杂、类别分布与类型约束多样。将原型学习引入这类结构中,可用于抽象类型约束或关系簇,或作为对比目标,提高表示的语义一致性与区分性,并在跨模态、跨结构,以及医疗健康、推荐系统等领域中提升下游任务性能与泛化能力。Wang 等[165]为每类节点引入可学习的原型表示,并借助超边正则化机制引导节点嵌入向原型靠拢,从而提升超图对异质信息网络中噪声的鲁棒性与可解释性。Zhang 等[166]进一步扩展了这一思路,通过为每种节点和边类型生成代表性原型,有效捕捉语义和结构差异,增强对异构程序中不同类型实体的区分能力。面对关系语义的多重性与零样本泛化需求,Li 等[167]通过动态生成多个关系原型,并利用注意力机制聚合与查询相关的参考实体,以适应不同实体对间的多义性。

5.3 基于原型的大模型能力增强方法

随着 LLM 在多任务、多场景中广泛应用,LLM 在解释性不足、少样本泛化能力有限以及知识迁移效率不高等方面的挑战逐渐凸显。原型学习为此提供了可解释且结构化的增强路径。通过在语义空间中引入类中心或原型表示,模型能够更好地对齐跨模态知识、抽象迁移共享特征,并在少样本或开放环境下保持鲁棒性。

在大模型适配中,原型常作为类别或任务的语义中心,用于引导模型学习。它既可嵌入上下文提示中,充当提示锚点,也可在零/少样本场景中构造向量原型,提升泛化与解释性。此外,将原型作为微调过程中的对齐或正则目标,能够增强跨任务迁移的稳定性和在下游任务中的适配性。Pan 等[168]将支持集中相同关系的实体对聚合成关系原型,通过比较查询实体对与原型之间的相似度进行分类,有效缓解了训练与测试阶段关系

语义不一致的问题。Wen[169]构建了可解释的类原型向量,使语言模型能够以“最近原型”的方式进行决策,在保持高分类精度的同时增强了可解释性。Li 等[170]融合 LLM 提供的语义先验与视觉特征,构建出兼具类别共性和判别性的目标原型,从而显著改善小样本分割的泛化能力与匹配精度。Wang 等[171]则利用大模型生成的特征先验与少量样本数据,构建出与训练无关的类原型,实现了零样本和小样本场景下的高效分类。针对微调过程中的适应性问题,Guo 等[172]通过为每类构建多个可学习的聚类原型,增强了小样本下对大语言模型的微调效果,能够更好地拟合复杂分类边界并缓解灾难性遗忘。

在从大规模语义模型向小型学生模型蒸馏的过程中,仅靠模仿教师的输出概率难以充分迁移知识。借助原型表示可以提取教师模型在中间层或局部结构上的通用特征,并通过对齐这些类别/任务原型,使学生模型能够在任务或标签空间异质、标注稀少或领域切换时保持性能。原型知识蒸馏[61](prototype knowledge distillation, ProtoKD)通过建模教师模型的多模态类内与类间特征变化,将其作为结构化知识蒸馏到单模态学生模型中,显著增强了在缺失模态下的分割鲁棒性。在联邦学习背景下,Wu 等[173]通过聚合本地类原型并基于奇异值分解(singular value decomposition, SVD)构建具有判别性与泛化性的全局原型,引导各客户端模型进行知识蒸馏,从而提升联邦模型在未知域上的泛化能力;Lyu 等[174]和 Zhang 等[175]利用类别原型改善非独立同分布数据下的学习效果,前者通过原型辅助过滤低质量样本并正则化训练过程,提升聚合知识与通信效率;后者通过原型相似度蒸馏对齐本地与全局特征空间,缓解数据异构问题。此外,Wang 等[176]将少量支持样本提炼为类别原型,用于与查询特征匹配,实现对新类别的快速识别与定位。

在多模态大语言模型(multi-modal large language model, MLLM)的少样本适配中,采用多原型或聚类中心能够更有效地表征类内多模态差异,并通过原型对齐或对比学习机制增强跨模态一致性。典型方法包括借助多原型聚类微调增强开放词汇识别能力,引入原型感知模块辅助细粒度匹配,以及利用原型表征作为跨模态对齐锚点等。Liu 等[177]通过构建跨模态锚点,对齐视觉与文本原型的分布并最大化其互信息,实现源域到目标域的无监督跨模态检索适配。Yang 等[178]通过构建每类点云特征的支持原型,与 3D MLLM 生成的伪标签原型进行相似度匹配,筛选高质量伪标签并指导自适应填充。Chen 等[179]通过提取多模态原型初始化分类器,替代粗糙类别名称,实现大规模词汇下的精准物体识别。

5.4　原型驱动的生成与推理联合框架

在生成与推理任务中引入原型,能够为大模型提供结构化的先验,在数据有限或质量较低的场景下提升模型的可控性与解释性。一方面,原型作为语义中心可与生成式模型结合,用于数据补全、样本合成与跨模态生成,以缓解生成式模型训练中的灾难性遗忘;另一方面,原型也可以作为推理锚点嵌入到生成式推理流程中,帮助模型在复杂任务中保持语义一致性、减少幻觉并提升可解释性。

在原型条件化生成范式中,类别或模态的原型被用作生成器的条件信号或对齐锚点,以在样本合成、缺失数据补全和跨模态生成中注入语义约束,以实现高效的增量学习。Zhu 等[180]通过保存并增广旧类别的类代表原型,在特征空间中维护旧类别的决策边界,有效减少增量学习中的灾难性遗忘问题。Shi 等[181]通过存储旧类原型并与新类特征做双向插值“回忆”,即时合成大量旧类增强特征,起到无须保存旧样本就能持续扩增训练数据的作用。Barsellotti 等[182]提出了一种基于扩散增强原型的生成方法。该方法通过将文本语义与视觉空间精准对齐,能够在缺乏真实标注的情况下扩充特征的语义多样性,进而实现数据增强。与前面几种方法不同,Tagade 等[183]为每类优化一个原型图像,作为数据无关的全局解释,并通过高频抑制与随机变换等正则化手段实现生成增强。

在原型辅助的推理框架中,原型不仅可以作为生成式模型的条件信号,还可以作为推理过程中的语义约束,帮助模型在复杂任务中保持一致性并避免语义漂移。通过将原型嵌入到推理模型中,可以显著提高推理的可解释性和准确性,尤其是在面对多模态信息融合和长序列推理任务时。Ren 等[184]先为每类合成初始原型,再以高置信度查询样本加权更新原型,从而同时完成数据增强与分类推理。Zhang 等[185]通过 SVD 生成任务感知原型完成鲁棒匹配,再以超像素原型为节点进行图推理,引导查询特征重编码。Li 等[186]通过动态原型生成模块跨帧一致原型,并以原型交叉引导解码器进行推理,从而显式挖掘相邻帧的共享与私有特征,实现端到端视频行人去重计数。

5.5 小结

本节揭示了在生成式 AI 时代,原型角色正从判别基准转变为生成控制锚点与非参数化知识库:在生成式模型中,离散原型有效缓解了模式坍塌并提升了样本的语义可控性;而在大模型应用中,基于检索的原型增强(如检索增强生成(retrieval-augmented generation, RAG))以低成本解决了幻觉与知识过时问题,这种"参数化推理 + 非参数化记忆"的混合范式成为构建高效可信系统的核心策略。

6 原型学习的优势与局限性

原型学习作为一种连接符号主义与连接主义的桥梁,在当前的人工智能研究中展现出独特的两面性。系统地审视其优势边界与应用局限,对于指导未来的模型设计至关重要。

在优势方面,原型学习的核心价值不仅在于其显式的可解释性,更体现在对复杂数据环境的适应能力上。首先,在不确定性量化方面,原型提供了一种天然的概率度量基准。通过计算样本与属性原型的距离,可以直观地估计预测的不确定性,这对于高风险领域的决策至关重要[142]。其次,在开放环境适应性方面,原型学习打破了封闭世界的假设。利用原型作为紧凑的分布描述符,能够有效界定已知类的边界,从而在开放集识别中精准捕捉未见过的异常类别,展现出优于传统判别模型的灵活性[149]。

然而,随着应用场景向高维、动态及大规模拓展,当前原型学习的研究也展现出一些局限性,主要体现在以下三个方面:一是高维空间的分布估计失效。在复杂的高维特征空间中,传统的基于简单统计量(如均值)的原型往往难以准确刻画精细的特征分布[187]。基于统计高斯分布的原型在高维空间中无法准确估计复杂的检测特征分布,导致新旧任务的知识重叠与混淆,这种"维度灾难"限制了静态原型在高维增量学习中的表现[188]。二是动态环境下的概念漂移。在流式数据处理中,数据分布往往随时间发生非平稳变化。现有的原型方法大多假设分布静态,当发生概念漂移时,固定的原型无法自适应地跟踪数据流的统计特性变化,导致模型难以区分真实的语义漂移与噪声干扰,从而引发性能衰退[189]。三是大规模检索的计算与存储瓶颈。在 RAG 与大语言模型应用中,为了覆盖长尾知识,往往需要构建极大规模的原型库。研究[190]中分析了基于万亿级 Token 构建数据存储的挑战,指出随着数据存储规模的指数级增长,其索引构建、维护及实时检索的计算成本急剧上升,成为原型检索模式在大规模预训练模型中高效部署的关键瓶颈。

7 总结与展望

综上所述,原型学习作为衔接底层特征表达与高层语义逻辑的桥式范式,在构建可信、透明且高效的智能系统中发挥着愈发关键的作用。基于原型的生成式模型构建方法已在少样本学习、跨模态生成及推理可解释性等多个前沿领域取得显著进展;通过引入原型作为语义锚点与类别中心,不仅有效抑制了深度学习模型的"黑箱"效应,更在提升模型可控性、透明度及泛化能力方面展现出独特优势。

然而,面对日益复杂的现实应用需求,原型学习仍面临诸多亟待破解的关键挑战。首先,在处理海量异构数据时,跨模态语义对齐的精度不高与模态失衡问题依然严峻;其次,在非平稳时序流数据环境下,原型表示难以有效应对概念漂移引发的动态适应性难题;此外,如何在大模型时代实现高效的大小模型知识协同,以及在复杂语义空间中避免高维特征的表示塌陷,亦是当前研究的攻坚焦点。展望未来,原型学习将从单一分类范式向动态、协同且可持续演进的智能架构跨越。其研究重点将聚焦于四个前沿方向。一是多模态深度联觉:构建跨模态共享语义锚点,实现在信息缺失或噪声环境下的稳健跨模态补全与对齐。二是动态自适应演化:引入动态拓扑更新机制,赋予原型随任务环境变化自适应生长与更迭的能力,以适应时序非平稳分布。三是端云一体化协同:探索原型驱动的轻量化知识蒸馏范式,实现大模型先验知识向边缘端的高效迁移与反馈。四是长效可持续学习:完善基于原型的增量记忆保护框架,在无须回溯原始数据的条件下抑制灾难性遗忘,构建具备自主进化能力的知识系统。

总之,通过深化原型在生成式建模、时序推理及跨域迁移中的核心枢纽作用,原型学习将不仅在众多垂直领域发挥关键价值,更将为构建透明、可信、高效、人机协同的通用人工智能(artificial general intelligence, AGI)提供坚实的理论基础与技术支撑。

参考文献(References)

[1] ROSCH E. Cognitive representations of semantic categories[J]. Journal of Experimental Psychology: General, 1975, 104(3): 192 – 233.

[2] 张幸幸, 朱振峰, 赵亚威, 等. 机器学习中原型学习研究

进展[J]. 软件学报, 2022, 33(10): 3732 – 3753.
ZHANG X X, ZHU Z F, ZHAO Y W, et al. Prototype learning in machine learning: a literature review[J]. Journal of Software, 2022, 33(10): 3732 – 3753. (in Chinese)

[3] IKOTUN A M, EZUGWU A E, ABUALIGAH L, et al. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences, 2023, 622: 178 – 210.

[4] HEIDARI J, DANESHPOUR N, ZANGENEH A. A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers [J]. Pattern Recognition, 2024, 155: 110639.

[5] KUO R J, WU C Y, KUO T. An ensemble method with a hybrid of genetic algorithm and K-prototypes algorithm for mixed data classification [J]. Computers & Industrial Engineering, 2024, 190: 110066.

[6] CHAKRABORTY S, MALI K, MITRA S. Affinity propagation in semi-supervised segmentation: a biomedical application[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(10): 6023 – 6032.

[7] SHEN Y H, PEDRYCZ W, CHEN Y, et al. Hyperplane division in fuzzy C-means: clustering big data[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(11): 3032 – 3046.

[8] GU C Z, LU X Q, ZHANG C. Example-based color transfer with Gaussian mixture modeling[J]. Pattern Recognition, 2022, 129: 108716.

[9] CAI H M, HU Y, QI F, et al. Deep tensor spectral clustering network via ensemble of multiple affinity tensors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 5080 – 5091.

[10] YU X T, YANG Y F, WANG A B, et al. Adapt-Infomap: face clustering with adaptive graph refinement in infomap[J]. Pattern Recognition, 2023, 143: 109792.

[11] 郭昕刚, 王佳, 程超. 层次聚类算法和基于图的分割算法相融合的图像分割算法[J]. 国防科技大学学报, 2022, 44(3): 194 – 200.
GUO X G, WANG J, CHENG C. Image segmentation algorithm combining hierarchical clustering algorithm and graph-based segmentation algorithm[J]. Journal of National University of Defense Technology, 2022, 44(3): 194 – 200. (in Chinese)

[12] 吴超凡, 黄鹤, 贾睿, 等. 基于改进 DBSCAN 算法的道路障碍物点云聚类[J]. 南京大学学报(自然科学), 2025, 61(5): 738 – 751.
WU C F, HUANG H, JIA R, et al. Point cloud clustering of road obstacles based on improved DBSCAN algorithm[J]. Journal of Nanjing University (Natural Science), 2025, 61(5): 738 – 751. (in Chinese)

[13] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering points to identify the clustering structure[J]. ACM SIGMOD Record, 1999, 28(2): 49 – 60.

[14] NIE F P, XUE J J, YU W Z, et al. Fast clustering with anchor guidance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 1898 – 1912.

[15] 何克磊, 史颖欢, 高阳, 等. 一种基于原型学习的多示例卷积神经网络[J]. 计算机学报, 2017, 40(6): 1265 – 1274.
HE K L, SHI Y H, GAO Y, et al. A prototype learning based multi-instance convolutional neural network [J]. Chinese Journal of Computers, 2017, 40(6): 1265 – 1274. (in Chinese)

[16] 杨雨龙, 郭田德, 韩丛英. 基于原型学习改进的伪标签半监督学习算法[J]. 中国科学院大学学报, 2021, 38(6): 841 – 851.
YANG Y L, GUO T D, HAN C Y. Improving pseudo-labeling semi-supervised learning based on prototype learning[J]. Journal of University of Chinese Academy of Sciences, 2021, 38(6): 841 – 851. (in Chinese)

[17] 李娇, 范浩东, 洪旭东, 等. 基于标签视觉原型学习的多标签图像分类[J]. 计算机工程, 2026, 52(4): 229 – 238.
LI J, FAN H D, HONG X D, et al. Multi-label image classification based on label visual prototype learning[J]. Computer Engineering, 2026, 52(4): 229 – 238. (in Chinese)

[18] WANG Z Y, KONG Z L, CHANGRA S, et al. Robust high dimensional stream classification with novel class detection[C]//Proceedings of 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019: 1418 – 1429.

[19] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning [C]//Proceedings of the 31st International Conference on Neural Information Processing System, 2017: 4080 – 4090.

[20] TANG Q, LIU C J, LIU F G, et al. Rethinking feature reconstruction via category prototype in semantic segmentation[J]. IEEE Transactions on Image Processing, 2025, 34: 1036 – 1047.

[21] HU X C, PEDRYCZ W, WANG X M. Granular fuzzy rule-based models: a study in a comprehensive evaluation and construction of fuzzy models[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(5): 1342 – 1355.

[22] SHEN Y H, PEDRYCZ W. Collaborative fuzzy clustering algorithm: some refinements [J]. International Journal of Approximate Reasoning, 2017, 86: 41 – 61.

[23] ZHAO X P, SHENG D M, TAN Z T, et al. Training-free open-vocabulary semantic segmentation via diverse prototype construction and sub-region matching[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(10): 10474 – 10482.

[24] SHEN W H, MA A L, WANG J J, et al. Adaptive self-supporting prototype learning for remote sensing few-shot semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5634116.

[25] DUAN S S, YANG X, WANG N N. Multi-label prototype visual spatial search for weakly supervised semantic segmentation [C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 30241 – 30250.

[26] ZHOU Y B, QU X M, YOU C L, et al. FedSA: a unified representation learning via semantic anchors for prototype-based federated learning [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39 (21): 23009 – 23017.

[27] ZHU G L, WU D Y, GAO C X, et al. Adaptive prototype replay for class incremental semantic segmentation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(10): 10932 – 10940.

[28] YANG S, WU S H, LIU T L, et al. Bridging the gap between few-shot and many-shot learning via distribution calibration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9830 – 9843.

[29] XU J Y, LE H. Generating representative samples for few-shot classification [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 8993 – 9003.

[30] REDEKOP E, PLEASURE M, IVEZIC V, et al. Prototype-guided diffusion for digital pathology: achieving foundation model performance with minimal clinical data [C]// Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025: 5187 – 5195.

[31] ZHI C R, ZHUO J B, WANG S H. Confusing pair correction based on category prototype for domain adaptation under noisy environments[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(15): 17060 – 17068.

[32] 钱菲, 李威, 陈鹏, 等. 基于原型解耦的虚假人脸检测和生成方法溯源[J]. 计算机科学与探索, 2025, 19(2): 490 – 501.
QIAN F, LI W, CHEN P, et al. Face forgery detection and attribution via prototype disentanglement [J]. Journal of Frontiers of Computer Science and Technology, 2025, 19(2): 490 – 501. (in Chinese)

[33] 赵红, 钟杨清, 金杰, 等. 基于自适应原型特征类矫正的小样本学习方法[J]. 自动化学报, 2025, 51(2): 475 – 484.
ZHAO H, ZHONG Y Q, JIN J, et al. Few-shot learning based on class rectification via adaptive prototype features[J]. Acta Automatica Sinica, 2025, 51(2): 475 – 484. (in Chinese)

[34] HU X C, SHEN Y H, PEDRYCZ W, et al. Granular fuzzy rule-based modeling with incomplete data representation[J]. IEEE Transactions on Cybernetics, 2022, 52(7): 6420 – 6433.

[35] YU X M, WANG Y B, ZHOU J, et al. ProtoComp: diverse point cloud completion with controllable prototype [C]// Proceedings of Computer Vision – ECCV 2024, 2024: 270 – 286.

[36] GONG T T, DU G D, WANG J S, et al. Prototype-guided cross-modal completion and alignment for incomplete text-based person re-identification[C]//Proceedings of the 31st ACM International Conference on Multimedia, 2023: 5253 – 5261.

[37] 高杨, 王新宇, 贺达, 等. 面向缺失多元时间序列的图神经网络异常检测算法[J]. 国防科技大学学报, 2025, 47(3): 32 – 40.
GAO Y, WANG X Y, HE D, et al. Anomaly detection algorithm based on graph neural network for missing multivariate time series[J]. Journal of National University of Defense Technology, 2025, 47(3): 32 – 40. (in Chinese)

[38] YU Z H, CHU X, MA L T, et al. Imputation with inter-series information from prototypes for healthcare time series[EB/OL]. (2024 – 01 – 14) [2026 – 01 – 08]. https://arxiv.org/abs/2401.07249.

[39] TU W X, GUAN R X, ZHOU S H, et al. Attribute-missing graph clustering network [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(14): 15392 – 15401.

[40] ZHANG B Q, LI X T, YE Y M, et al. Prototype completion for few-shot learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12250 – 12268.

[41] ZHOU Z A, YANG B, HUANG W K, et al. Unbiased prototype consistency learning for multi-modal and multi-task object re-identification [C]//Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025), 2025: 1 – 21.

[42] LIU J Y, LIU X W, YANG Y X, et al. Contrastive multi-view kernel learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9552 – 9566.

[43] LIU J Y, LIU X W, WANG S Q, et al. Communication-efficient federated multi-view clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026, 48(1): 17 – 32.

[44] HU X C, QIN J D, SHEN Y H, et al. An efficient federated multiview fuzzy C-means clustering method [J]. IEEE Transactions on Fuzzy Systems, 2024, 32(4): 1886 – 1899.

[45] NI X Z, LIU Y, WEN H, et al. Multimodal prototype-enhanced network for few-shot action recognition [C]// Proceedings of 2024 International Conference on Multimedia Retrieval, 2024: 1 – 10.

[46] GUO Z, TANG Y W, ZHANG R, et al. ViewRefer: grasp the multi-view knowledge for 3D visual grounding [C]// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 15326 – 15337.

[47] ZHANG X S, CHEN D, QIN Y B. Multimodal prototype fusion network for paper-cut image classification [J]. npj Heritage Science, 2025, 13(1): 462.

[48] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 8748 – 8763.

[49] LI Y N, QI T Y, MA Z Q, et al. Seeking a hierarchical prototype for multimodal gesture recognition [J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 198 – 209.

[50] CHENG P J, LIN L, LYU J Y, et al. PRIOR: prototype representation joint learning from medical images and reports[C]// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 21304 – 21314.

[51] SHI S J, NIE F P, WANG R, et al. Fast multi-view clustering via prototype graph[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 443 – 455.

[52] YANG G P, YANG S S, YANG Y Y, et al. SPGMVC: multiview clustering via partitioning the signed prototype graph[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(5): 8036 – 8048.

[53] YU S J, DONG Z B, WANG S W, et al. Simple yet effective incomplete multi-view clustering: similarity-level imputation and intra-view hybrid-group prototype construction [C]// Proceedings of the Thirteenth International Conference on Learning Representations, ICLR 2025, 2025: 51543 – 51569.

[54] LE H Q, THWAL C M, QIAO Y, et al. Cross-modal prototype based multimodal federated learning under severely missing modality [J]. Information Fusion, 2025, 122:

103219.

[55] XIE J N, ZHENG X L, ZHENG L. Prototype-aware multimodal alignment for open-vocabulary visual grounding[EB/OL]. (2025 - 09 - 08) [2026 - 01 - 08]. https://arxiv.org/abs/2509.06291.

[56] YUAN H L, SUN Y, ZHOU F, et al. Prototype matching learning for incomplete multi-view clustering [J]. IEEE Transactions on Image Processing, 2025, 34: 828 - 841.

[57] DU Y M, WANG Y, WANG Z Y, et al. PGFormer: a prototype-graph transformer for incomplete multiview clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2026, 37(3): 1163 - 1175.

[58] LI Y, HU X C, YU S J, et al. A vertical federated multiview fuzzy clustering method for incomplete data [J]. IEEE Transactions on Fuzzy Systems, 2025, 33(5): 1510 - 1524.

[59] LI M C, YANG D K, ZHAO X, et al. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 12458 - 12468.

[60] SHI J J, SHANG C Z, SUN Z B, et al. PASSION: towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates[C]//Proceedings of the 32nd ACM International Conference on Multimedia, 2024: 456 - 465.

[61] WANG S, YAN Z P, ZHANG D A, et al. Prototype knowledge distillation for medical segmentation with missing modality[C]//Proceedings of ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023: 1 - 5.

[62] ALLEN K, SHELHAMER E, SHIN H, et al. Infinite mixture prototypes for few-shot learning[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 232 - 241.

[63] HUANG H W, WU Z K, LI W B, et al. Local descriptor-based multi-prototype network for few-shot learning [J]. Pattern Recognition, 2021, 116: 107935.

[64] CHEN C F, LI O, TAO C F, et al. This looks like that: deep learning for interpretable image recognition [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 8930 - 8941.

[65] LIU H M, WANG R P, SHAN S G, et al. What is a tabby? Interpretable model decisions by learning attribute-based classification criteria [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1791 - 1807.

[66] YU Q H, WANG Y. Global and local attention-based multiscale prototypical network for few-shot learning [J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2026, 10(1): 468 - 479.

[67] WANG W J, DUAN L J, EN Q, et al. TPSN: transformer-based multi-prototype search network for few-shot semantic segmentation [J]. Computers and Electrical Engineering, 2022, 103: 108326.

[68] HONG D, WANG T, BAEK S. ProtoryNet: interpretable text classification via prototype trajectories[J]. Journal of Machine Learning Research, 2023, 24(1): 12344 - 12382.

[69] TAN Y, LONG G D, LIU L, et al. FedProto: federated prototype learning across heterogeneous clients [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(8): 8432 - 8440.

[70] YANG W X, HU X C, ZHU X B, et al. FedMPS: federated learning in a synergy of multi-level prototype-based contrastive learning and soft label generation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2026, 37(2): 781 - 794.

[71] WANG G C, LI J B, TIAN C, et al. A novel multimodal prototype network for interpretable medical image classification[C]// Proceedings of 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2023: 2577 - 2583.

[72] LIU Y, JAIN A, ENG C, et al. A deep learning system for differential diagnosis of skin diseases[J]. Nature Medicine, 2020, 26(6): 900 - 908.

[73] DENG W H, XIONG W, LU Z Y, et al. Few-shot power transformers fault diagnosis based on Gaussian prototype network [J]. International Journal of Electrical Power & Energy Systems, 2024, 160: 110146.

[74] HUANG Q H, XUE M Q, HUANG W Q, et al. Evaluation and improvement of interpretability for self-explainable part-prototype networks [C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 2011 - 2020.

[75] GORT P C, CLAESSENS C H B, DE WITH P H N, et al. Evaluating the interpretability of prototype networks for medical image analysis[C]//Proceedings of Medical Imaging 2025: Image Processing, 2025.

[76] LI Y, CHEN C, HU X C, et al. Fuzzy rule-based models: a design with prototype relocation and granular generalization[J]. Information Sciences, 2021, 562: 155 - 179.

[77] NIU J J, CHEN D G, LI J H, et al. Fuzzy rule-based classification method for incremental rule learning[J]. IEEE Transactions on Fuzzy Systems, 2022, 30(9): 3748 - 3761.

[78] HU X C, ZHU X B, YANG L, et al. A design of fuzzy rule-based classifier for multiclass classification and its realization in horizontal federated learning[J]. IEEE Transactions on Fuzzy Systems, 2024, 32(9): 5098 - 5108.

[79] MAMAGHANI A S, PEDRYCZ W. Structural optimization of fuzzy rule-based models: towards efficient complexity management[J]. Expert Systems with Applications, 2020, 152: 113362.

[80] TSEKOURAS G E. Fuzzy rule base simplification using multidimensional scaling and constrained optimization [J]. Fuzzy Sets and Systems, 2016, 297: 46 - 72.

[81] ZHAO T, ZHU Y, XIE X P. Topology structure optimization of evolutionary hierarchical fuzzy systems[J]. Expert Systems with Applications, 2024, 238(Part B): 121857.

[82] WANG Y G, LIU H R, JIA W J, et al. Deep fuzzy rule-based classification system with improved Wang-Mendel method[J]. IEEE Transactions on Fuzzy Systems, 2022, 30(8): 2957 - 2970.

[83] GU X W, LI M Q, SHEN L, et al. Multiobjective evolutionary optimization for prototype-based fuzzy classifiers[J]. IEEE Transactions on Fuzzy Systems, 2023, 31(5): 1703 - 1715.

[84] LI J B, PEDRYCZ W, WANG X M. A rule-based development of incremental models[J]. International Journal of Approximate Reasoning, 2015, 64: 20 - 38.

[85] RUDNIK K, CHWASTYK A, PISZ I, et al. Ordered fuzzy rules generation based on incremental dataset [C]//

Proceedings of 2021 IEEE International Conference on Fuzzy Systems, 2021: 1 – 7.

[86] ŠKRJANC I, IGLESIAS J A, SANCHIS A, et al. Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: a survey[J]. Information Sciences, 2019, 490: 344 – 368.

[87] BAI Y X, LU X J. Multiple kernel learning-based rule reduction method for fuzzy modeling[J]. Fuzzy Sets and Systems, 2023, 465: 108534.

[88] PONTOIZEAU T, CAILLIÈRE R, LESOT M J, et al. Assessing the interpretability of fuzzy rule bases[C]// Proceedings of 2025 IEEE International Conference on Fuzzy Systems (FUZZ), 2025: 1 – 6.

[89] CZABANSKI R, JEZEWSKI M, LESKI J, et al. Refining the rule base of fuzzy classifier to support the evaluation of fetal condition[J]. Applied Soft Computing, 2023, 147: 110790.

[90] ALMSEIDIN M. Confidence-driven hybrid intrusion detection: combining fuzzy rule interpolation and deep learning[J]. Journal of Network and Systems Management, 2026, 34(1): 12.

[91] PETERS J, JANZING D, SCHÖLKOPF B. Elements of causal inference: foundations and learning algorithms[M]. Cambridge: the MIT Press, 2017.

[92] DUONG T D, LI Q, XU G D. Causality-based counterfactual explanation for classification models[J]. Knowledge-Based Systems, 2024, 300: 112200.

[93] SHAO X Y, WANG H Z, CHEN X, et al. CUBE: causal intervention-based counterfactual explanation for prediction models[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(6): 2416 – 2429.

[94] LIN G L, XU Y H, LAI H J, et al. Revisiting few-shot learning from a causal perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6908 – 6919.

[95] ZHANG Y J, HUANG Z A, HONG Z L, et al. Mixed prototype correction for causal inference in medical image classification[C]//Proceedings of the 32nd ACM International Conference on Multimedia, 2024: 4377 – 4386.

[96] PROSPERI M, GUO Y, SPERRIN M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare[J]. Nature Machine Intelligence, 2020, 2(7): 369 – 375.

[97] SANCHEZ P, VOISEY J P, XIA T, et al. Causal machine learning for healthcare and precision medicine[J]. Royal Society Open Science, 2022, 9(8): 220638.

[98] GRECOV P, BANDARA K, BERGMEIR C, et al. Causal inference using global forecasting models for counterfactual prediction[C]// Proceedings of Advances in Knowledge Discovery and Data Mining, 2021: 282 – 294.

[99] STUKKER N, SANDERS T. Subjectivity and prototype structure in causal connectives: a cross-linguistic perspective[J]. Journal of Pragmatics, 2012, 44(2): 169 – 190.

[100] ZANGA A, OZKIRIMLI E, STELLA F. A survey on causal discovery: theory and practice[J]. International Journal of Approximate Reasoning, 2022, 151: 101 – 129.

[101] GONG C, ZHANG C Z, YAO D, et al. Causal discovery from temporal data: an overview and new perspectives[J]. ACM Computing Surveys, 2024, 57(4): 100.

[102] PFISTER N, BÜHLMANN P, PETERS J. Invariant causal prediction for sequential data[J]. Journal of the American Statistical Association, 2019, 114(527): 1264 – 1276.

[103] PAN Y C, ZHANG Y F, JIANG X R, et al. EffCause: discover dynamic causal relationships efficiently from time-series[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(5): 1 – 21.

[104] LUO J J, JIN Z H, JIN H P, et al. Causal temporal graph attention network for fault diagnosis of chemical processes[J]. Chinese Journal of Chemical Engineering, 2024, 70: 20 – 32.

[105] CASTRO D C, WALKER I, GLOCKER B. Causality matters in medical imaging[J]. Nature Communications, 2020, 11(1): 3673.

[106] ZANG C Q, WANG H Q, PEI M T, et al. Discovering the real association: multimodal causal reasoning in video question answering[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 19027 – 19036.

[107] HUANG S S, WANG L, LIAO J, et al. Multi-attentional causal intervention networks for medical image diagnosis[J]. Knowledge-Based Systems, 2024, 299: 111993.

[108] CHEN G, DONG G M. Temporal logic inference for interpretable fault diagnosis of bearings via sparse and structured neural attention[J]. ISA Transactions, 2025, 158: 256 – 271.

[109] MICHOEL T, ZHANG J D. Causal inference in drug discovery and development[J]. Drug Discovery Today, 2023, 28(10): 103737.

[110] ZHOU J Y, ZHAO D W, QIAN H, et al. Multi-objective structure-based drug design using causal discovery[J]. IEEE Transactions on Computational Biology and Bioinformatics, 2025, 22(4): 1789 – 1800.

[111] MARCONATO E, PASSERINI A, TESO S. Interpretability is in the mind of the beholder: a causal framework for human-interpretable representation learning[J]. Entropy, 2023, 25(12): 1574.

[112] CHOU Y L, MOREIRA C, BRUZA P, et al. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications[J]. Information Fusion, 2022, 81: 59 – 83.

[113] CAI S H, PENG H, LIU R, et al. Causal-oriented representation learning for time-series forecasting based on the spatiotemporal information transformation[J]. Communications Physics, 2025, 8(1): 242.

[114] CHEN G Y, SHEN Y F, CHEN Z H, et al. CaRiNG: learning temporal causal representation under non-invertible generation process[C]//Proceedings of the 41st International Conference on Machine Learning, 2024: 7236 – 7259.

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

[116] WEI D, YANG X H, WENG Y Y, et al. Cross-modal adaptive prototype learning for continuous sign language recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(8): 7354 – 7367.

[117] WEI C X, YUAN J D, ZHANG Y, et al. Ranking neighborhood and class prototype contrastive learning for time series[J]. IEEE Transactions on Big Data, 2025, 11(4): 1907 – 1917.

[118] FANG L, SHI J H, QU H Y, et al. Contrastive prototype

guided federated learning for rotating machinery fault diagnosis under spatio-temporal domain shift [J]. Mechanical Systems and Signal Processing, 2025, 232: 112707.

[119] IWANA B K, UCHIDA S. Time series classification using local distance-based features in multi-modal fusion networks[J]. Pattern Recognition, 2020, 97: 107024.

[120] SHI Z Z, ZHANG Z J, ZHOU H L, et al. Inversion-based pre-stack gather flattening by exploiting temporal sparsity[J]. Digital Signal Processing, 2023, 132: 103783.

[121] FENG X Z, FAN D X, JIANG S H, et al. A causal representation learning based model for time series prediction under external interference [J]. Information Sciences, 2024, 663: 120270.

[122] LI P, CAO J C, YE X C. Prototype contrastive learning for point-supervised temporal action detection [J]. Expert Systems with Applications, 2023, 213(Part B): 118965.

[123] LIU H Y, YANG D H, LIU X Z, et al. TodyNet: temporal dynamic graph neural network for multivariate time series classification [J]. Information Sciences, 2024, 677: 120914.

[124] LUO W, REN H, ZHANG T Z, et al. Adaptive prototype learning for weakly-supervised temporal action localization[J]. IEEE Transactions on Image Processing, 2025, 34: 3154 – 3168.

[125] TANG Y, CHEN T, JIANG X R, et al. Holistic prototype attention network for few-shot video object segmentation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 6699 – 6709.

[126] AN Y Y, YI Y M, WU L, et al. Unsupervised prototype self-calibration based on hybrid attention contrastive learning for enhanced few-shot action recognition[J]. Applied Soft Computing, 2025, 168: 112558.

[127] ZHANG Q, YAN S, SHAO M W, et al. Efficient spatio-temporal modeling and text-enhanced prototype for few-shot action recognition [J]. Neurocomputing, 2025, 638: 130119.

[128] GAO X Y, MA Z Q, CUI J T, et al. Hierarchical category-enhanced prototype learning for imbalanced temporal recommendation [C]//Proceedings of the 31st ACM International Conference on Multimedia, 2023: 6181 – 6189.

[129] GU X W. A self-training hierarchical prototype-based approach for semi-supervised classification[J]. Information Sciences, 2020, 535: 204 – 224.

[130] ZHENG P, DOU Y, YAN Y Q. Sensing the diversity of rumors: rumor detection with hierarchical prototype contrastive learning [J]. Information Processing & Management, 2024, 61(6): 103832.

[131] BI H J, LIU L, CUI H, et al. Improving few-shot relation classification with multi-scale hierarchical prototype learning[J]. Neural Networks, 2026, 194: 108124.

[132] GAO C, ZHANG X, JIN Z, et al. Few-shot relational triple extraction with hierarchical prototype optimization [J]. Pattern Recognition, 2024, 156: 110779.

[133] ZHANG X K, SONG D J, TAO D C. Hierarchical prototype networks for continual graph representation learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4622 – 4636.

[134] CAI M J, QI D Y, HUANG C Q, et al. Prototype-based fuzzy rough sets for outlier detection[J]. Fuzzy Sets and Systems, 2025, 517: 109460.

[135] LI Y X, CHEN W C, CHEN B, et al. Prototype-oriented unsupervised anomaly detection for multivariate time series[C]//Proceedings of the 40th International Conference on Machine Learning, 2023: 19407 – 19424.

[136] LIU C, HE S B, LI S Z, et al. Detecting both seen and unseen anomalies in time series[J]. ACM Transactions on Knowledge Discovery from Data, 2025, 19(4): 87.

[137] HUANG C, LI Q Y, ZHANG B. Prototype-guided and dynamic-aware video anomaly detection [J]. Neural Networks, 2025, 189: 107583.

[138] HAN C, LIU C, WANG J, et al. A spatial-spectral and temporal dual prototype network for motor imagery brain-computer interface[J]. Knowledge-Based Systems, 2025, 315: 113315.

[139] GAMA J, ŽLIOBAITĖ I, BIFET A, et al. A survey on concept drift adaptation [J]. ACM Computing Surveys (CSUR), 2014, 46(4): 1 – 37.

[140] NAUTA M, SCHLÖTTERER J, VAN KEULEN M, et al. PIP-Net: patch-based intuitive prototypes for interpretable image classification[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 2744 – 2753.

[141] HASE P, BANSAL M. Evaluating explainable AI: which algorithmic explanations help users predict model behavior? [EB/OL]. (2020 – 05 – 04) [2026 – 01 – 08]. https://arxiv.org/abs/2005.01831.

[142] ZELENKA C, GÖHRING A, KAZEMPOUR D, et al. A simple and explainable method for uncertainty estimation using attribute prototype networks[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2023: 4572 – 4581.

[143] GAUTAM S, BOUBEKKI A, HANSEN S, et al. ProtoVAE: a trustworthy self-explainable prototypical variational model[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022: 17940 – 17952.

[144] KIM J, OH T H, LEE S, et al. Variational prototyping-encoder: one-shot learning with prototypical images[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 9454 – 9462.

[145] TANG W L, YANG B Q, LI X Z, et al. Prototypical variational autoencoder for few-shot 3D point cloud object detection [C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 2566 – 2579.

[146] DWIVEDI S K, GUPTA V, MITRA R, et al. ProtoGAN: towards few shot learning for action recognition [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019: 1308 – 1316.

[147] DU Y J, XIAO Z H, LIAO S C, et al. ProtoDiff: learning to learn prototypical networks by task-guided diffusion[C]//Proceedings of the 37th International Conference on Neural Information Processing System, 2023: 46304 – 46322.

[148] BAYKAL G, KARAGOZ H F, BINHURAIB T, et al. ProtoDiffusion: classifier-free diffusion guidance with

prototype learning [C]//Proceedings of the 15th Asian Conference on Machine Learning, 2024: 106 – 120.

[149] WANG F Y, ZHANG T, WANG Y Z, et al. Distribution prototype diffusion learning for open-set supervised anomaly detection[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 20416 – 20426.

[150] CHEN S M, HOU W J, HONG Z M, et al. Evolving semantic prototype improves generative zero-shot learning[C]// Proceedings of the 40th International Conference on Machine Learning, 2023: 4611 – 4622.

[151] WANG Y Y, MAO J, GUO C G, et al. Contrastive prototype-guided generation for generalized zero-shot learning[J]. Neural Networks, 2024, 176: 106324.

[152] JIANG H J, LI Z X, HU Y L, et al. Dual prototype contrastive network for generalized zero-shot learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(2): 1111 – 1122.

[153] YU Y L, JI Z, HAN J G, et al. Episode-based prototype generating network for zero-shot learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 14032 – 14041.

[154] CHEN S M, HONG Z M, YOU X G, et al. Semantics-conditioned generative zero-shot learning via feature refinement[J]. International Journal of Computer Vision, 2025, 133(7): 4504 – 4521.

[155] WU J M, ZHANG T Z, ZHA Z J, et al. Prototype-augmented self-supervised generative network for generalized zero-shot learning [J]. IEEE Transactions on Image Processing, 2024, 33: 1938 – 1951.

[156] HOU W J, CHEN S M, CHEN S H, et al. Visual-augmented dynamic semantic prototype for generative zero-shot learning [C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 23627 – 23637.

[157] ZHANG Z X, LIU Q, WANG H, et al. ProtGNN: towards self-explaining graph neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(8): 9127 – 9135.

[158] DAI E Y, WANG S H. Towards prototype-based self-explainable graph neural network[J]. ACM Transactions on Knowledge Discovery from Data, 2025, 19(2): 1 – 20.

[159] ZHENG Y M, JIA C Y. ProtoMGAE: prototype-aware masked graph auto-encoder for graph representation learning[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(6): 1 – 22.

[160] SHIN Y M, KIM S W, SHIN W Y. PAGE: prototype-based model-level explanations for graph neural networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(10): 6559 – 6576.

[161] SEO S, KIM S, PARK C. Interpretable prototype-based graph information bottleneck[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 76737 – 76748.

[162] ZHANG Q, LI X W, LU J X, et al. ROG_PL: robust open-set graph learning via region-based prototype learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(8): 9350 – 9358.

[163] ZHU H, KONIUSZ P. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement [C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 23996 – 24006.

[164] LI X C, XIAO J, CHEN G K, et al. Decomposed prototype learning for few-shot scene graph generation [J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(1): 1 – 24.

[165] WANG S, SHEN J Y, EFTHYMIOU A, et al. Prototype-enhanced hypergraph learning for heterogeneous information networks[C]// Proceedings of MultiMedia Modeling, 2024: 462 – 476.

[166] ZHANG T H, XU R, ZHANG J P, et al. DSHGT: dual-supervisors heterogeneous graph transformer: a pioneer study of using heterogeneous graph learning for detecting software vulnerabilities [J]. ACM Transactions on Software Engineering and Methodology, 2024, 33(8): 1 – 31.

[167] LI Y L, YU K, ZHANG Y H, et al. Adaptive prototype interaction network for few-shot knowledge graph completion[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(11): 15237 – 15250.

[168] PAN D H, SUN Y Y, XU B, et al. Prototype tuning: a meta-learning approach for few-shot document-level relation extraction with large language models[C]//Proceedings of Findings of the Association for Computational Linguistics: NAACL 2025, 2025: 1112 – 1128.

[169] WEN X M. Language model meets prototypes: towards interpretable text classification models through prototypical networks [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(28): 29307 – 29308.

[170] LI P F, LIU F, JIAO L C, et al. LLM knowledge-driven target prototype learning for few-shot segmentation [J]. Knowledge-Based Systems, 2025, 312: 113149.

[171] WANG P, WANG D S, ZHAO H, et al. LLM empowered prototype learning for zero and few-shot tasks on tabular data[EB/OL]. (2025 – 08 – 12) [2026 – 01 – 08]. https://arxiv.org/abs/2508.09263.

[172] GUO M H, ZHANG Y, MU T J, et al. Tuning vision-language models with multiple prototypes clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 11186 – 11199.

[173] WU A M, YU J P, WANG Y X, et al. Prototype-decomposed knowledge distillation for learning generalized federated representation [J]. IEEE Transactions on Multimedia, 2024, 26: 10991 – 11002.

[174] LYU F, TANG C, DENG Y H, et al. A prototype-based knowledge distillation framework for heterogeneous federated learning[C]//Proceedings of 2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS), 2023: 1 – 11.

[175] ZHANG C, XIE Y, CHEN T B, et al. Prototype similarity distillation for communication-efficient federated unsupervised representation learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6865 – 6876.

[176] WANG Z C, YANG B, YUE H N, et al. Fine-grained prototypes distillation for few-shot object detection [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5859 – 5866.

[177] LIU Y, CHEN Q C, ALBANIE S. Adaptive cross-modal prototypes for cross-domain visual-language retrieval[C]//

Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 14949 - 14959.

[178] YANG S Q, DING H H, JIANG X D. Generalized few-shot 3D point cloud segmentation [C]//Proceedings of 2024 IEEE International Symposium on Circuits and Systems (ISCAS), 2024: 1 - 5.

[179] CHEN Y T, YAO W H, MENG L C, et al. Comprehensive multi-modal prototypes are simple and effective classifiers for vast-vocabulary object detection [C]//Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, 2025: 2320 - 2328.

[180] ZHU F, ZHANG X Y, WANG C, et al. Prototype augmentation and self-supervision for incremental learning[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 5867 - 5876.

[181] SHI W X, YE M. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 1772 - 1781.

[182] BARSELLOTTI L, AMOROSO R, CORNIA M, et al. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 3689 - 3698.

[183] TAGADE A, RUMBELOW J. Prototype generation: robust feature visualisation for data independent interpretability[EB/OL]. (2023 - 09 - 29) [2026 - 01 - 08]. https://arxiv.org/abs/2309.17144.

[184] REN H H, LIU S, YU X L, et al. Transductive prototypical attention reasoning network for few-shot SAR target recognition [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5206813.

[185] ZHANG Y M, LI H L, GAO Y J, et al. Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation [J]. IEEE Transactions on Medical Imaging, 2024, 43(11): 4041 - 4054.

[186] LI R, LIU Y S, LI H F, et al. Prototype-guided dual-transformer reasoning for video individual counting [C]//Proceedings of the 32nd ACM International Conference on Multimedia, 2024: 10258 - 10267.

[187] HUANG W D, HU J W, XIAO J H, et al. Prototype-guided graph reasoning network for few-shot medical image segmentation[J]. IEEE Transactions on Medical Imaging, 2025, 44(2): 761 - 773.

[188] WANG Y J, CHEN L Q, ZHAO T M, et al. High-dimension prototype is a better incremental object detection learner [C]//Proceedings of International Conference on Learning Representations 2025 (ICLR 2025), 2025: 51297 - 51314.

[189] HINDER F, VAQUET V, HAMMER B. One or two things we know about concept drift: a survey on monitoring in evolving environments. Part A: detecting concept drift[J]. Frontiers in Artificial Intelligence, 2024, 7: 1330257.

[190] SHAO R L, HE J, ASAI A, et al. Scaling retrieval-based language models with a trillion-token datastore [C]//Proceedings of the 38th Conference on Neural Information Processing Systems, 2024: 91260 - 91299.