概念嵌入增强的可解释图像聚类

王 翔[1),2),3)] 刘华锋[2),3)] 景丽萍[1),2),3)] 于 剑[2),3)]
郭龙腾[4)] 杨雅君[5)]

[1)](先进轨道交通自主运行全国重点实验室(北京交通大学) 北京 100044)
[2)](交通数据挖掘与具身智能北京市重点实验室(北京交通大学) 北京 100044)
[3)](北京交通大学计算机科学与技术学院 北京 100044)
[4)](中国科学院自动化研究所 北京 100190)
[5)](天津大学智能与计算学部 天津 300350)

摘 要 作为无监督学习领域的基础性任务,聚类分析在众多数据场景中具有核心应用价值。当其与深度神经网络及大语言模型集成时,所形成的深度聚类技术展现出解析高维图像数据复杂结构的强大能力。然而,现有深度聚类方法通常采用隐式方式耦合数据的全部特征维度,以捕获非线性流形结构。这种"黑箱"特性导致模型决策难以被直观解析,进而限制了其在城市规划、医疗诊断等高风险敏感领域的应用。为应对上述挑战,本文提出一种概念嵌入增强的可解释图像聚类框架,通过跨模态语义转换机制提升聚类结果的可解释性。具体而言,本文创新地构建了一个数据自适应的文本概念生成器,能够在无监督标签缺失的条件下,自动挖掘数据集中潜在的高层语义概念;同时,设计了概念表征对齐模块和聚类决策修正模块,使模型在保持聚类性能的基础上,能够输出符合人类认知的语义解释。在六个基准图像数据集上的实验表明,所提方法不仅在聚类准确性方面表现优越,同时也显著提升了聚类结果的可解释性。

关键词 可解释聚类;图像聚类;概念瓶颈模型;自监督学习;最大编码率约束
中图法分类号 TP181 **DOI号** 10.11897/SP.J.1016.2026.00742

Interpretable Image Clustering with Concept Embedding

WANG Xiang[1),2),3)] LIU Hua-Feng[2),3)] JING Li-Ping[1),2),3)] YU Jian[2),3)]
GUO Long-Teng[4)] YANG Ya-Jun[5)]

[1)](State Key Laboratory of Advanced Rail Autonomous Operation (Beijing Jiaotong University), Beijing 100044)
[2)](Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence (Beijing Jiaotong University), Beijing 100044)
[3)](School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044)
[4)](Institute of Automation, Chinese Academy of Sciences, Beijing 100190)
[5)](College of Intelligence and Computing, Tianjin University, Tianjin 300350)

Abstract As a foundational task in the field of unsupervised learning, clustering analysis holds core application value in many data scenarios. When integrated with deep neural networks or large language models, the resulting deep clustering techniques demonstrate powerful capabilities in capturing the complex structures of high-dimensional image data. However, existing deep

收稿日期:2025-05-27;在线发布日期:2025-11-11。本课题得到国家自然科学基金(62436001,62406019,62176020)、国家重点研发计划项目(2024YFE0202900)、北京市自然科学基金(4244096)、北交大人才基金(2024XKRC075)、教育部创新团队联合基金(8091B042235)、中央高校基础研究基金(2019JBZ110)、北京交通大学轨道交通控制与安全国家重点实验室(RCS2023K006)资助。
王 翔,博士,中国计算机学会(CCF)会员,主要研究领域为机器学习、无监督聚类。E-mail: wang-xiang@bjtu.edu.cn。**刘华锋**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、概率生成模型。**景丽萍**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习。E-mail: lpjing@bjtu.edu.cn。**于 剑**,博士,教授,中国计算机学会(CCF)会士,主要研究领域为机器学习,人工智能。**郭龙腾**,副研究员,中国计算机学会(CCF)会员,主要研究兴趣包括图像分析与理解、多模态预训练模型等。**杨雅君**,博士,讲师,中国计算机学会(CCF)会员,主要研究领域为图数据管理,图挖掘。

clustering methods typically rely on implicitly coupling all feature dimensions of the data to capture nonlinear manifold structures. This “black-box” nature renders model decisions difficult to interpret intuitively, thereby limiting their applicability in high-risk and sensitive domains such as urban planning and medical diagnosis. To address these challenges, we propose a concept embedding-enhanced interpretable image clustering framework that improves the interpretability of clustering results through a cross-modal semantic transformation mechanism. Specifically, we introduce a novel data-adaptive textual concept generator that can autonomously discover high-level semantic concepts from the dataset under the condition of lacking supervised labels. Additionally, we design a concept representation alignment module and a clustering decision refinement module, enabling the model to provide human-understandable semantic explanations while maintaining clustering performance. Experiments conducted on six benchmark image datasets demonstrate that the proposed method not only achieves superior clustering accuracy but also significantly enhances the interpretability of the clustering results.

Keywords　interpretable clustering; image clustering; concept bottleneck model; self-supervised learning; maximal coding rate reduction

1　引　言

图像聚类作为数据挖掘领域的经典挑战，旨在根据图像中所蕴含的内在语义信息，将其划分到不同的类簇集合中[1-2]。面对图像数据在高维空间中呈现出的非线性和复杂结构等特性，传统依赖浅层特征的聚类方法在刻画数据潜在语义分布方面表现出明显的局限性[3-4]。近年来，深度聚类方法逐渐成为图像聚类领域的主流研究方向[5]。该类方法借助深度神经网络构建层级化特征表示，将复杂的高维视觉信息映射到低维嵌入空间，以增强聚类划分的准确性与稳定性。得益于深度神经网络强大的非线性表征能力，深度聚类在聚类精度方面显著优于传统聚类方法[6]。然而，这种性能提升的背后也隐藏着不可忽视的问题：深度神经网络的复杂建模机制易导致嵌入表示中各维度特征以高度非线性的方式相互纠缠[7]。由此产生的低维嵌入表征虽然具备优异的聚类判别性，但是却丧失了直观的可解释性[8]。在城市规划、医疗诊断等风险敏感领域，这种可解释性缺失已成为限制深度聚类方法实际应用的重要因素[9-10]。

针对上述问题，近年来可解释聚类逐渐成为研究热点[11]。与传统聚类方法主要关注聚类精度与计算效率不同[12-14]，可解释聚类强调聚类结果背后的逻辑透明性和模型决策的可解释性，其核心目的在于提供清晰且易于理解的决策证据，使得用户能够洞悉算法行为及其内在决策逻辑[15]。已有研究尝试通过引入具备解释能力的模型结构来增强聚类的可解释性[16]。例如，基于决策树的可解释聚类方法[17-19]通过预定义划分规则，在每个节点内部对样本空间进行划分，从而构建可追踪的决策路径；

基于规则的可解释聚类方法则通过挖掘最优规则集合，显式关联输入特征与类簇划分结果[20-21]。这些方法在处理结构化表格数据或者低维离散数据时表现出良好的可解释性与实用性，能够明确呈现样本归属类簇的关键决策依据[22-24]。然而，图像数据天然具有高维、非结构化的特点，其基本单元为离散像素点。这些像素点通常仅承载底层视觉信息，难以与高层语义概念之间建立直接映射关系，如图1(b)所示。因此，结构化数据场景下常用的可解释机制难以直接适用于图像聚类任务[25-26]。

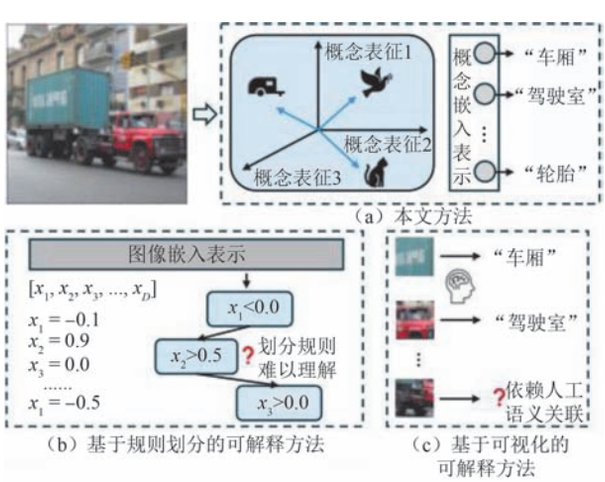


图1　IC²E与已有方法的差异

为应对图像数据的语义复杂性，当前研究主要聚焦于类簇层级的可解释证据构建[7,27]。例如，部分工作利用深度神经网络从原始图像中提取高层语义信息，构建具备认知语义的中间表示（如原型向量等）[28-29]，并通过可视化等策略为聚类结果提供全局性解释[30]。但是，这些方法大都依赖于样本整体的统计特性，难以为单个样本给出精确且针对性的解释；此外，原型表示与原始图像之间的映射过程依赖于人为设定的语义关联。如图1(c)所示，这种主观先验的引入不可避免地削弱了可解释证据的客观性与一致性[31]。为缓解这一问题，Svirsky等人[9]提出一种基于双门控注意力机制的可解释聚类模型，通过动态筛选与类簇高度相关的像素特征子集，在样本和类簇层级生成的可视化解释证据。然而，受限于像素特征与高层语义之间天然存在的抽象层次差异，该方法生成的解释证据仍然难以直接转化为人类可理解的语义解释[32]。

为应对上述挑战，本文提出一种概念嵌入增强的可解释图像聚类方法（Interpretable Image Clustering with Concept Embedding，IC^{2}E）。该方法受概念瓶颈模型启发[33]，通过构建具有明确语义含义的概念嵌入表示，实现视觉特征空间与高层语义空间的对齐，从而增强聚类模型的可解释性。在本文中，“概念”指代图像或类簇所涵盖的本质语义信息，以自然语言短语的形式呈现。如图1(a)所示，所提方法在训练过程中引导模型预测输入图像是否包含特定文本概念，从而将底层视觉特征映射到语义明确的概念空间，并生成与文本概念一一对应的嵌入表示。该方法使得嵌入表示不仅保留了图像间的语义相似性结构，还使得每一维嵌入都可被人类语义解释，从而使聚类决策具备可追溯性与可分析性。例如，如果多个图像在“轮胎”、“车灯”、“运输”等概念维度上表现一致，那么聚类算法基于这些共同的语义属性将它们划归一类就变得合理。与传统概念瓶颈模型依赖于人工概念集或先验类别信息不同[33-34]，本文提出的数据自适应概念生成机制能够在类簇标签缺失的情况下自动挖掘与数据紧密相关的高层语义短语，具备更强的通用性与适应性。此外，IC^{2}E具备良好的可扩展性，可作为可插拔模块无缝集成于多种深度聚类框架中，在保持原有聚类性能的基础上增强模型的可解释能力。本文的主要贡献可以总结如下：

1）提出一种数据驱动的文本概念生成策略，能够在类别名称缺失的情况下自动构建与数据语义相关的自然语言短语；

2）构建一种通用的可解释图像聚类框架，具备良好的模型兼容性和结构可扩展性，能够与现有聚类算法协同运行。通过概念表征对齐和聚类决策修正，该框架不仅能保留原聚类模型的判别能力，还能在聚类过程中同步生成人类可理解的可解释证据；

3）在六个被广泛使用的图像聚类数据集进行了充分的实验。实验结果表明，该方法能够在保证聚类精度的同时显著提升划分结果的可解释性。

2 相关工作

本节系统回顾了当前深度聚类和可解释聚类领域的主流方法。

2.1 深度聚类

面对高维图像数据呈现出的非线性分布与复杂结构特性，传统基于浅层特征的聚类方法往往有效建模数据的潜在语义分布，逐渐暴露出明显的性能瓶颈[1]。近年来，随着深度学习技术的迅猛发展，深度神经网络逐渐被引入聚类任务中，以提升特征表达与结构建模能力，深度聚类逐渐成为图像聚类研究的主流方向[2]。通过联合优化特征表示与聚类目标，该类方法能够从原始图像中提取具有判别性的低维嵌入表示，构建更具语义一致性的嵌入空间，从而提升聚类划分的准确性与稳定性。

早期的深度聚类方法多采用堆叠自编码器结构，并使用重构损失作为辅助监督信号[35]。这类方法通常将自编码器与传统聚类算法（如k-means[36]或高斯混合模型[37]）相结合，实现端到端的嵌入表示学习与类簇分配。然而，这类方法常受到网络结构的限制：自编码器主要优化重构能力，难以确保相似样本在嵌入空间中有一致的低维表示，导致聚类性能受限[38]。随着无监督学习的迅猛发展，对比学习策略被引入聚类任务，一系列基于实例判别的对比聚类方法在图像聚类任务中展现出显著的性能提升[3]。这类方法侧重于利用原始图像与其增强版本之间的一致性信息进行训练，通过最大化同一样本不同增强视图下的表示一致性，同时最小化不同样本间表示的相似性，实现判别性表征学习[39]。然而，在缺乏显示监督信息引导的情况下，模型往往将每个图像样本视为独立类别，并在训练过程中强行推远其他所有样本的特征表示。这种过度分离的优化策略可能导致类内样本分布分散、类间样本边界模糊，从而引发次优的类簇划分[40]。

为进一步提升聚类性能，部分研究提出两阶段的深度聚类策略。考虑到特征表示在无监督聚类中的关键作用，这类方法通常使用预训练的特征提取网络获取初始嵌入表示，再基于邻域一致性[41]或伪标记自迭代机制[42]等策略进一步优化聚类结构。这种训练流程在增强聚类灵活性的同时，也提升了模型对复杂数据分布的适应能力。与此同时，大规模预训练模型的发展极大拓展了表示学习的潜力，为聚类有效性的进一步提升提供了新机遇[43]。例如，TAC方法[5]引入多模态大型模型CLIP[44]，利用图像与文本之间的语义交互信息对聚类边界进行细化，从而提升聚类性能。然而，尽管这些方法在聚类精度上取得了显著进展，但它们大都侧重于优化嵌入空间的判别性，忽视了模型决策与聚类划分的可解释性，导致模型决策过程呈现出“黑盒化”趋势。在实际应用中，特别是医疗、城市规划等对结果透明性与可控性要求较高的领域，缺乏可解释性的聚类结果不仅削弱了用户信任，也限制了模型的部署可行性。

2.2　可解释聚类

可解释聚类方法旨在揭示聚类结果的划分依据，使得用户能够理解模型的决策逻辑，从而提升模型的透明度与可信性。这在城市规划、医疗诊断等风险敏感领域尤为重要，因为错误或不透明的聚类输出可能引发严重后果[45]。在图像聚类任务中，可解释性的需求更加迫切。一方面，图像数据本身具有高维度、非结构化和特征语义模糊等特性，聚类模型通常依赖复杂的非线性变换生成潜在特征表示，这使得类簇划分背后的逻辑不再显性可见；另一方面，无监督聚类过程缺乏明确的标签监督，仅依赖样本之间的相似性关系进行划分，其结果天然具备不确定性，使得解释性成为用户理解和信任聚类输出的桥梁[5,16]。

早期的可解释聚类方法多采用决策树模型对聚类结果进行建模与解释[7,19]。通过预设特征阈值逐层划分数据，这些方法将根节点至叶节点的路径作为聚类划分的决策依据[17]。通常，树的深度被视为可解释性的量化指标之一：树越浅，意味着决策路径越短，结构越简洁，进而提升了推理过程的清晰性与可读性。然而，随着数据维度与结构复杂性的提升，树的深度和分支数目往往迅速增长，容易导致推理路径冗长，从而严重削弱了解释的可读性与实用性[25]。与决策树利用树形层次结构逐层划分数据空间不同，部分研究尝试通过规则挖掘形成“If-Then”的聚类决策规则集[20-21]，以增强解释的灵活性与可读性。这类方法通常以启发式搜索或优化策略生成紧凑的规则集，每条规则的设计目标是最大程度覆盖目标类簇样本，同时尽可能规避对非目标类簇的误覆盖[46]。当输入数据的特征具有清晰的物理含义或符合人类认知先验时，这类方法在解释的结构简洁性和语义清晰度方面表现出显著优势。然而，在处理图像等高维非结构化数据时，这些方法面临显著挑战。图像像素或感知特征通常缺乏明确语义，难以直接转化为人类可理解的划分依据。尽管部分研究尝试通过构建低维嵌入空间以缓解高维特征带来的解释困难[47-48]，但嵌入特征之间高度耦合以及语义表达的模糊性仍使得类簇划分的逻辑难以被清晰阐释。

为提升图像聚类的可解释性，一些研究将聚类任务与深度可解释框架相结合，尝试借助概念建模与可视化机制对聚类结果进行事后解释[15]。例如，Peng等人[8]借助深度神经网络学习表征类簇语义的概念原型，并通过原型可视化间接揭示聚类划分的决策依据[30]。尽管这些方法在类簇层面具有一定的语义概括能力，但由于原型通常在全局层面进行优化，它们难以为具体样本提供个性化解释。此外，这些方法通常需要将学习到的抽象概念原型映射回原始像素空间，该过程存在显著的主观性与不确定性[31]。为增强解释的个体差异性，部分研究尝试融合原型建模与层级解释结构[28-29]，但它们大多依赖标签信息，难以直接应用于无监督聚类任务。何等人[48]提出的模式挖掘方法虽然能有效分析离散序列数据中的结构模式，但其难以直接推广至高维连续图像数据。为此，Salles等人[49]提出一种自适应门控机制，能够在聚类过程中动态选择最具代表性的样本特征；Svirsky[9]提出双门控注意力机制，通过联合建模类簇层面的全局注意力与样本层面的局部注意力，使模型能够自动筛选出与类簇划分最相关的像素区域或特征维度，构建多粒度的可解释结构。这些方法在提升解释的个性化方面取得一定进展，但其解释结果仍主要停留在像素或中间特征层面，缺乏与高层语义概念的直接映射，难以满足人类用户对“语义清晰解释”的认知需求。

3　概念嵌入增强的可解释图像聚类

本节首先对所提方法的整体模型框架进行了简要概述，随后深入阐述了各模块的设计与实现机

制。通过将人类可理解的语义信息融入嵌入表示空间，本文所提方法能够有效增强聚类结果的可解释性。

3.1 整体框架

令$D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$表示包含$N$个样本的图像数据集，其中$\boldsymbol{x}_i\in\mathbb{R}^{H\times W\times 3}$表示第$i$个图像样本，$(H\times W)$为图像空间分辨率。$y_i\in\{1,2,\cdots,K\}$代表第$i$个图像对应的真实类别标签，在聚类模型训练阶段通常是不可用的。本文的目标是构建一个结果可解释的聚类模型$f_\theta(\boldsymbol{x}_i)$，该模型能够同时输出当前样本的聚类划分$\hat{y}_i\in\{1,2,\cdots,K\}$以及一组人类可理解的可解释证据$\mathcal{T}_i=\{t_i^1,t_i^2,\cdots,t_i^{M_i}\}$。其中，$t_i^j$代表第$j$个与样本$\boldsymbol{x}_i$相关的自然语言概念短语，如“人工湖”或者“金属车身”等[50]，而M_i表示模型为样本x_i输出的概念集合大小。相较于传统的后处理可视化方法，本文通过构建文本短语到概念表示的映射机制，将概念表征内嵌于深度神经网络中，使得可解释性成为聚类决策过程的内在组成部分[51-52]。这一机制借鉴了概念瓶颈模型[33,53]的设计思想，但又缓解了其对先验标签的强依赖问题，从而更适用于无监督图像聚类场景。

整体模型框架如图2所示，输入样本$\boldsymbol{x}_i$首先通过特征提取器，然后经由概念编码模块映射至概念空间，生成概念表征向量$h(\boldsymbol{x}_i)=\hat{\boldsymbol{o}}_i\in\mathbb{R}^M$，其中$M$表示概念空间的大小。该向量的每一维特征都对应于一个具体的语义概念，可直接映射到自然语言短语，具备清晰的可解释性。随后，概念表征向量$\hat{\boldsymbol{o}}_i$被送入聚类划分模块，经过稀疏投影层得到对应的聚类结果$g(\hat{\boldsymbol{o}}_i)=\hat{y}_i$。为了保证模型能生成有意义的概念解释，本文引入二值概念指示向量$\boldsymbol{o}_i\in\{0,1\}^M$，其中每个维度的特征值用于指示当前样本$\boldsymbol{x}_i$是否包含特定概念。在有监督场景中，可以直接通过对齐预测概念$\hat{\boldsymbol{o}}_i$与真实概念标签$\boldsymbol{o}_i$引导模型完成训练[54]；但在无监督场景中，真实概念标签$\boldsymbol{o}_i$往往难以获得。为此，本文设计了一种数据驱动的文本概念挖掘与嵌入机制，可在无需类别标签的情况下，从语义特征中自动提取可解释概念。在此基础上，本文提出了一种概念表征对齐和聚类边界修正策略，使得最终模型不仅具备优良的聚类性能，还能输出与人类认知一致的语义解释。在接下来的小节中，本文将依次介绍概念生成模块、概念对齐模块以及聚类决策修正策略的具体实现。

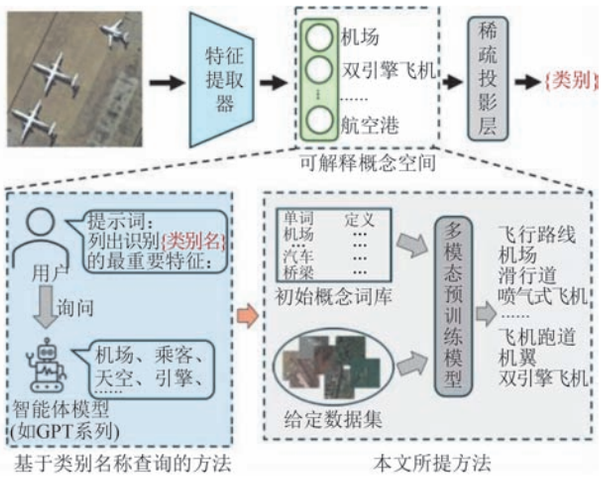


图2 整体模型框架示意图

3.2 概念生成模块

给定图像数据集$\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$，已有方法尝试利用人工标注或者预训练语言模型（如GPT系列）生成文本概念描述[33-34,53]。如图2所示，它们通过构建提示词来询问大模型，以获得与给定类簇相匹配的文本概念描述。尽管这类方法在解释性表达方面具备一定优势，但其普遍依赖预定义的类别标签或属性集合[54]，生成结果高度依赖于查询模型外部先验的覆盖范围，难以适应真实应用中类别标签未知的无监督聚类场景。

为应对上述挑战，本文提出一种数据驱动的概念词库自动构建方法。该方法可在无需人工标注的条件下，结合聚类结果与图文多模态特征相似度，自动生成与语义结构相一致的描述性短语。整体流程如图3所示，主要包含以下三个步骤：

（1）初始概念文本库构建

首先，基于现有的开放词汇资源（如WordNet词库[55]），本文构建了一个涵盖146348个基础英文名词的初始概念集合。随后，为提升文本概念的语义纯度，本文引入基于词向量的语义相似度过滤机制[5]，采用余弦相似度去除过于相似的同义词与潜在歧义词，从而缓解语义重叠与计算资源浪费[31,53]，最终得到包含73197个名词的候选词库，如图3(a)所示。

（2）初始类簇图像库构建

基于已有或者预训练聚类模型的划分结果，本文在图像嵌入空间中计算每个类簇的聚类中心；随后，从每个类簇中选取距离聚类中心最近的前L个样本作为该簇的样本代表，构建类簇图像集合，用于后续的跨模态语义匹配，如图3(b)所示。其中，参

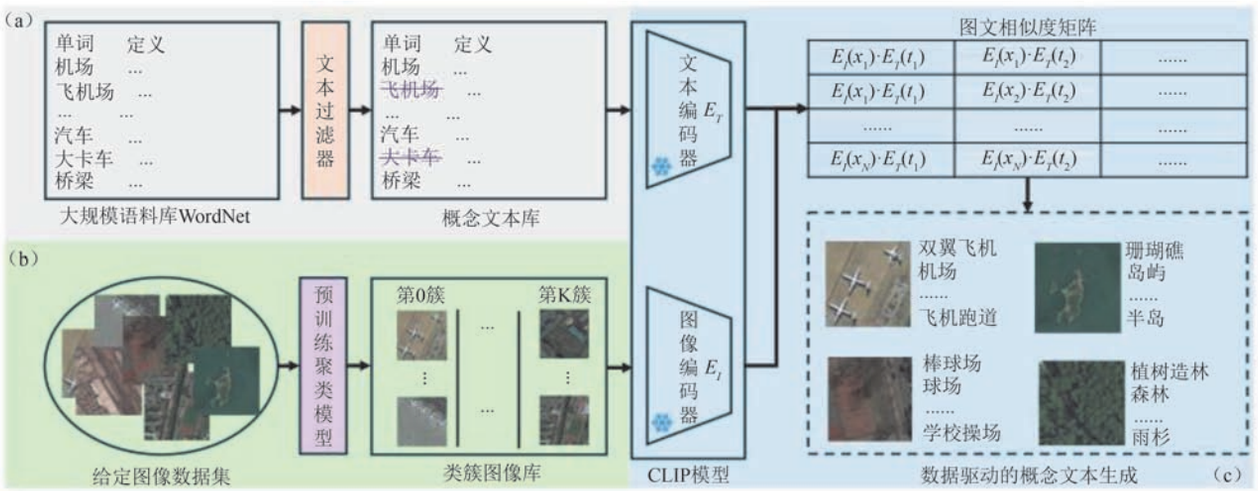


图3　数据驱动的概念词库自动构建方法

数L用于控制每簇代表样本的数目，在本文中固定为100。

(3) 数据驱动的概念词库构建

基于前述构建的文本库和图像库，本文采用预训练的多模态大模型CLIP[44]，将图像和文本编码到统一的语义嵌入空间中，并计算两者之间的跨模态相似度矩阵。对于每张图像，本文选取Top-5相似度最高的概念短语作为其初始语义标注；随后，统计每个类簇中所有样本的语义标注频次，并选取出现频率最高的前V个概念词作为该簇的语义概念标注。最终，本文构建的类簇概念集合为$\{t_l\}_{l=1}^{M}$，其中$M=K\times V$，K代表类簇个数，V为每簇保留的概念数目，如图3(c)所示。

通过上述三个步骤，本节所提方法能够为给定数据集自动生成语义明确的可解释性标注。同时，借助近邻筛选与高频过滤机制，该方法能够在一定程度上确保所生成的概念标签与实际类簇语义信息之间的一致性。在表1中，本文展示了该方法为不同数据集上生成概念词库的时间成本，验证了其良好的可扩展性。此外，该方法可进一步与基于大语言模型的描述生成方法[34-53]相结合，从而构建更加细粒度的类簇语义描述。

表1　数据集统计

数据集	图像数量	图像尺寸	类别数
CIFAR10	50,000	$3\times32\times32$	10
STL10	5,000	$3\times96\times96$	10
ImageNet-10	13,000	$3\times224\times224$	10
ImageNet-100	130,000	$3\times224\times224$	100
DTD	5,640	$3\times224\times224$	47
RESISC45	31,500	$3\times224\times224$	45

3.3　概念对齐模块

给定图像数据集$\{\boldsymbol{x}_i\}_{i=1}^{N}$和第3.2节中构建的概念集合$\mathcal{T}=\{t_l\}_{l=1}^{M}$，本节旨在挖掘图像样本$\boldsymbol{x}_i$中潜藏的高层语义概念，并将样本映射到概念嵌入空间，使得图像嵌入表征$\hat{\boldsymbol{o}}_i\in\mathbb{R}^M$的每一特征维度都对应于特定语义概念，从而提升后续聚类划分的可解释性。其中，M为第3.2节所提方法为当前数据集生成的文本概念数目。

为了实现上述目标，本文借助CLIP[44]模型的图像编码器E_I和文本编码器E_T，分别对图像$\boldsymbol{x}_i$和文本概念t_l进行编码，并基于余弦相似度构建概念代理标签矩阵$\boldsymbol{S}\in\mathbb{R}^{N\times M}$，矩阵元素定义如下：

$$S_{il}=\left\langle E_I(\boldsymbol{x}_i),E_T(t_l)\right\rangle \tag{1}$$

其中，$\langle\cdot,\cdot\rangle$代表向量内积运算，在不引起歧义的情况下将其简记为$S_{il}=E_I(\boldsymbol{x}_i)\cdot E_T(t_l)$。该矩阵可视为伪监督概念标签，用以度量样本与类簇概念之间的语义关联程度[34,53]，为模型学习语义一致的概念嵌入提供引导。具体损失函数定义如下：

$$Loss_s=-\sum_{i=1}^{N}\frac{\boldsymbol{S}_{i,:}\cdot\hat{\boldsymbol{o}}_i}{\left|\boldsymbol{S}_{i,:}\right|_2\cdot\hat{\boldsymbol{o}}_i} \tag{2}$$

其中，$\boldsymbol{S}_{i,:}$代表第i个样本对应的概念代理向量。通过最小化该损失函数，模型能够最大化嵌入表征$\hat{\boldsymbol{o}}_i$和目标代理$\boldsymbol{S}_{i,:}$之间的相似性，从而提升嵌入空间的可解释性，为后续聚类任务提供语义基础。

在此基础上，为进一步提升嵌入表征的可解释性与简洁性，本文引入稀疏化策略，对原始概念相似度矩阵$\boldsymbol{S}$进行筛选，仅保留每个样本对应的Top-$V_{/2}$高置信度的概念关联维度，从而构建稀疏化概念代理矩阵$\tilde{\boldsymbol{S}}$。具体来说，矩阵元素定义为：

$$\tilde{S}_{il}=\begin{cases}S_{il}, & \text{如果 } l\in D_{\mathrm{Top}-V/2}\\ 0, & \text{否则}\end{cases} \tag{3}$$

其中，$D_{\mathrm{Top}-V/2}$表示针对样本$\boldsymbol{x}_i$的相似度得分中排名前$V/2$的维度集合，V为第3.2节中所设定的高频概念保留参数。该稀疏策略在保留关键语义的同时有效控制了用于解释的维度数目，有助于减轻用户对可解释证据的理解负担[23-25]。

此外，在第3.2节构建概念集合的过程中，本文观察到不同类簇所对应的文本概念之间存在一定程度的语义重叠。为增强不同类簇样本之间可解释证据的独特性，进一步提升类间判别能力，本文引入最大编码率约束损失。该损失旨在最大化类间语义表达能力，同时抑制类内语义冗余，其表达式如下：

$$Loss_m=\exp\left(1-\frac{R\left(\hat{O},\boldsymbol{\Pi};\varepsilon\right)}{R\left(\hat{O};\varepsilon\right)}\right) \tag{4}$$

其中，$R\left(\hat{O};\varepsilon\right)=\frac{1}{2}\log\det\left(\boldsymbol{I}+\frac{M}{N\varepsilon^2}\hat{O}\hat{O}^{\top}\right)$，$M$和$N$分别代表嵌入表示的特征维度和参与训练的样本数目。$\boldsymbol{I}\in\mathbb{R}^{N\times N}$为单位对角矩阵。$\varepsilon$为超参数，在本文中固定为0.05。$R\left(\hat{O},\boldsymbol{\Pi};\varepsilon\right)=\sum_{j=1}^{K}\frac{tr\left(\boldsymbol{\Pi}_j\right)}{2N}\log\det\left(\boldsymbol{I}+\frac{M}{tr\left(\boldsymbol{\Pi}_j\right)\varepsilon^2}\hat{O}\boldsymbol{\Pi}_j\hat{O}^{\top}\right)$，$O=\left[\hat{\boldsymbol{o}}_1,\hat{\boldsymbol{o}}_2,\cdots,\hat{\boldsymbol{o}}_N\right]\in\mathbb{R}^{M\times N}$表示概念嵌入表征矩阵，$\boldsymbol{\Pi}=\left\{\boldsymbol{\Pi}_j\in\mathbb{R}^{N\times N}\right\}_{j=1}^{K}$是一组对角矩阵，其对角线元素表示这$N$个样本在$K$个类簇中的隶属关系，可以通过预训练模型的聚类划分构建得到。最小化该损失可以促使不同语义类别的可解释证据在特征空间形成正交分布，从而增强聚类划分结果的可解释性和判别能力[43]。

3.4 概念-数据双驱动聚类

理想情况下，若视觉特征能够在文本语义空间中找到明确对应，则其聚类划分结果应表现出较强的判别能力[5]。因此，本文进一步将第3.3节中获得的概念嵌入表征$\hat{\boldsymbol{o}}_i$投影至类簇概率空间，以实现可解释性与聚类性能的统一优化，如图4(b)所示。

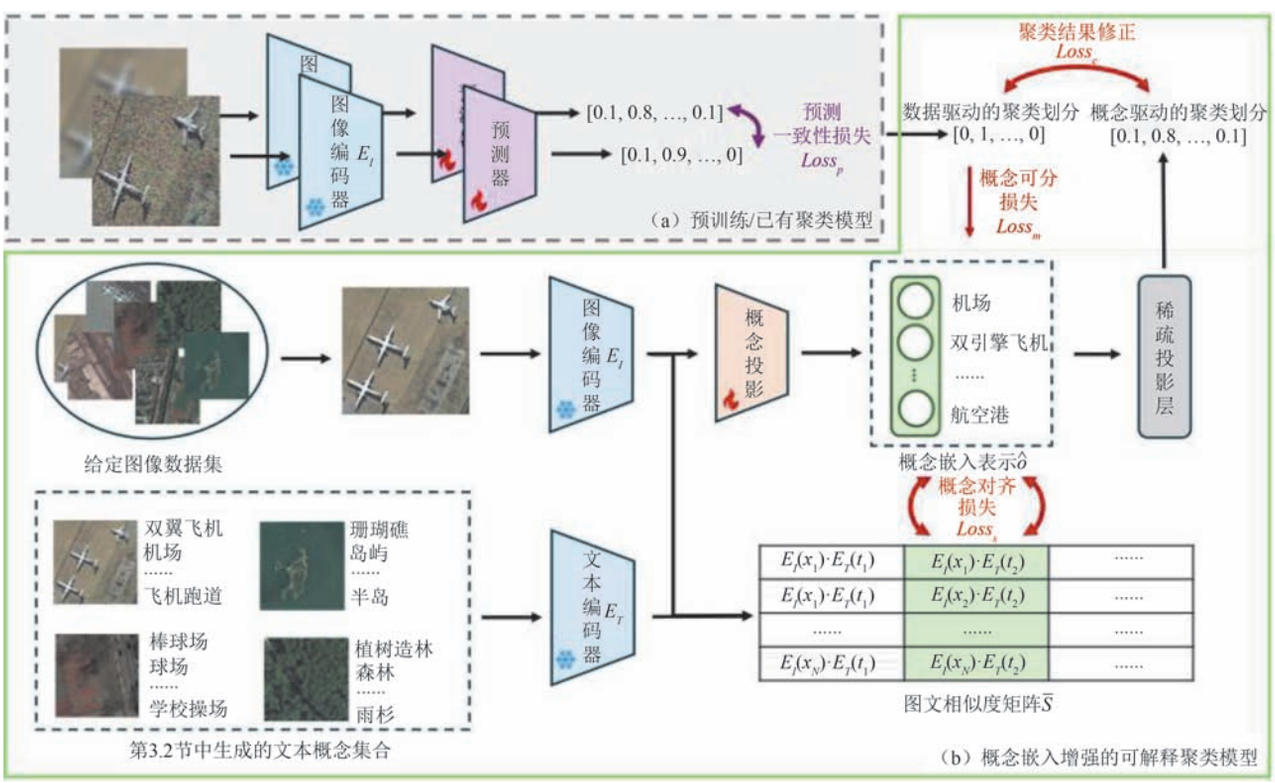


图4 概念嵌入增强的可解释聚类模型示意图

具体而言，采用单层神经网络作为投影层，可以得到类簇概率分配：

$$\boldsymbol{p}_i=\hat{\boldsymbol{o}}_i^{T}\boldsymbol{W} \tag{5}$$

其中，$\boldsymbol{p}_i$表示样本$\boldsymbol{x}_i$在K个簇上的响应值，矩阵$\boldsymbol{W}\in\mathbb{R}^{M\times K}$为可学习的权重矩阵，其第$k$列对应于第$k$个语义簇的原型表示。然而，若权重矩阵$\boldsymbol{W}$过

于稠密，则可能引入噪声概念，进而影响模型可解释性[54,56]。为缓解此问题，本文提出权重稀疏性诱导策略，通过非线性变换抑制非判别性概念的影响：

$$\boldsymbol{p}_i=\log\left(\left(\hat{\boldsymbol{o}}_i^T\boldsymbol{W}\right)^2+1\right) \tag{6}$$

通过修改投影层输出，模型可以在训练过程中逐渐减少不相关概念的权重。具体来说，本文限制矩阵$\boldsymbol{W}\geqslant 0$，以确保每个概念的贡献具有物理可解释性。同时，平方操作$(\cdot)^2$强化显著概念对簇响应的主导作用。在训练阶段，模型使用soft max函数对向量$\boldsymbol{p}_i$进行归一化处理，以输出类簇分布概率；在推理阶段则直接采用线性投影$\boldsymbol{p}_i=\hat{\boldsymbol{o}}_i^T\boldsymbol{W}$进行聚类预测，以提升聚类结果的可解释性。

算法1.　IC^{2}E聚类算法

输入：数据集$\{\boldsymbol{x}_i\}_{i=1}^N$，概念词库$\{t_l\}_{l=1}^M$，类簇个数$K$，参数冻结的CLIP编码器$E_I(\cdot)$和$E_T(\cdot)$，预训练模型$G(\cdot)$，权重参数$\alpha$；

输出：聚类划分$\{\hat{y}_i\}_{i=1}^N$，可解释证据$\mathcal{T}_i=\{t_i^1,t_i^2,\cdots,t_i^{M_i}\}$；

1. 利用文本编码器$E_T(\cdot)$提取文本概念的嵌入表示：
$$\boldsymbol{h}_l=E_T(t_l);$$
2. 对数据集中的每一个样本$\boldsymbol{x}_i$：
3. 利用预训练模型$G(\cdot)$得到类簇隶属度$\boldsymbol{q}_i$；
4. 利用图像编码器$E_I(\cdot)$提取图像嵌入表示：
$$\boldsymbol{h}_i=E_I(\boldsymbol{x}_i);$$
5. 经过概念投影模块$f(\cdot)$得到概念嵌入表征：
$$\hat{\boldsymbol{o}}_i=f(\boldsymbol{h}_i);$$
6. 依据公式(6)得到类簇概率分配向量$\boldsymbol{p}_i$；
7. 依据公式(1)和公式(2)计算概念代理信号$\tilde{\boldsymbol{S}}_{i,:}$；
8. 最小化损失函数：
$$Loss=Loss_s+Loss_m+\alpha*Loss_c;$$
9. 得到聚类划分$\hat{y}_i$和可解释证据$\mathcal{T}_i=\{t_i^1,t_i^2,\cdots,t_i^{M_i}\}$；

为增强聚类过程中的语义一致性，本文借鉴了对比学习中的一致性损失[41]，利用数据驱动的聚类划分结果引导概念嵌入增强的可解释模型进行训练。具体损失函数如下：

$$Loss_c=-\log\sum_{i=1}^N\boldsymbol{p}_i^T\boldsymbol{q}_i \tag{7}$$

其中，$\boldsymbol{p}_i$为可解释模型的聚类划分，$\boldsymbol{q}_i$为预训练聚类模型的预测输出。整体训练流程如算法1所示。

在本文中，预训练模型可通过预测一致性损失优化得到，如图4(a)所示。具体来说，给定图像数据集$\{\boldsymbol{x}_i\}_{i=1}^N$，在预训练阶段，本文使用基于数据增强驱动的一致性约束损失：

$$Loss_p=-\sum_{i=1}^N\left(\boldsymbol{q}_i^1*\log\boldsymbol{q}_i^2\right)-\log\sum_{i=1}^N\boldsymbol{q}_i^{1\,T}\boldsymbol{q}_i^2 \tag{8}$$

其中，$\boldsymbol{q}_i^1$和$\boldsymbol{q}_i^2$分别是图像$\boldsymbol{x}_i$的不同增强样本的预测输出。通过确保来自同一图像的不同增强样本之间享有相近的预测，预训练模型能够逐步学习数据的潜在结构。考虑到聚类性能与表示质量息息相关，本文冻结CLIP模型中的图像编码器，并在此基础上引入可训练的聚类投影层，将样本映射至与预设聚类数K对齐的潜空间。所得预测向量$\boldsymbol{q}_i\in\mathbb{R}^K$经过soft max归一化后，其第$k$维响应值可解释为样本属于第$k$个语义簇的置信度概率。同时，为防止模型聚焦于少数簇类别，本文附加熵正则项以鼓励模型进行均匀的聚类分配。在实际使用过程中，预训练模型可以替换为现有的无监督聚类方法，从而简化训练流程，并增强已有模型结果的可解释性。

4　实验与分析

本节首先介绍所使用的数据集与实验设置，随后对实验结果进行详细分析，并探讨模型各组成模块的作用与贡献。

4.1　数据集

为了充分评估所提方法的有效性，本文在六个具有代表性的数据集上进行了实验验证，并表1中总结了所使用数据集的基本统计信息。

具体而言，本文首先选取了三个被广泛使用的图像聚类基准数据集：CIFAR10[57]，STL10[58]和ImageNet-10[59]。这些数据集涵盖不同图像尺寸和语义类别，能够有效衡量模型在通用图像聚类任务上的性能表现。之后，为进一步评估模型在可解释性与复杂场景适应性方面的能力，本文引入了三个具有挑战性的图像数据集：图像分类数据集ImageNet-100，从ImageNet中随机抽取100个类别构成；纹理识别数据集DTD[60]，包含多种具备语义标签的自然纹理图像；遥感图像场景分类数据集RESISC45[61]，涵盖了多种地理环境和复杂背景。图5展示了部分图像样例及本文方法在多个数据集上自动生成的语义概念描述。从图中可以观察到，所提出的IC^{2}E方法倾向于捕获图像的主体结构或场景特征，生成较为粗粒度的语义标签，如“卡车”、“蜂窝”、“公路”等。这些标签能够从整体上概括图像的核心语义，与人类对类别的认知高度一致。相比之下，基于类别名称查询的大型语言模型方法所

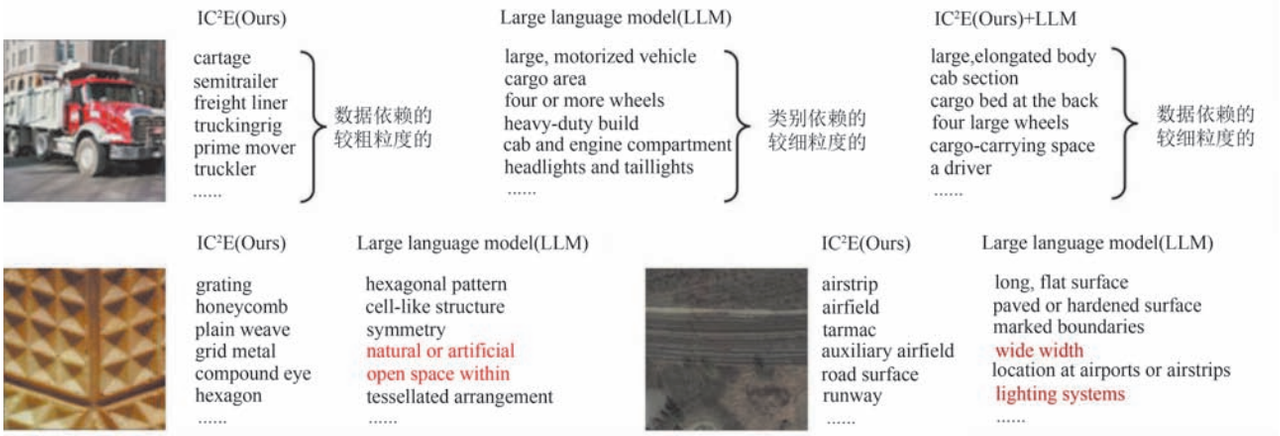


图5 针对不同数据集生成的文本概念

生成的文本描述通常更加细粒度，能够挖掘如“车灯”、“发动机”等局部概念。

产生这一差异的原因主要有两点：首先，所选用的预训练多模态模型CLIP在图像编码过程中更关注图像的主要目标区域，因此在计算图文相似度时更倾向于响应图像的全局或主体语义。这种机制虽然有利于捕捉类簇语义，但可能忽略了样本局部细节。其次，在使用语言模型生成图像描述时，生成结果往往受到提示词(prompt)的影响，例如“列出识别汽车最重要的特征”等。这类提示会促使大模型围绕已有的类别名称进行细粒度扩展，从而生成结构合理但受限于先验概念空间的描述。尽管这种带有弱监督引导的生成策略在图像识别任务中更符合人类认知方式，但其高度依赖预设类别标签或手工设计的语义查询模板，生成的文本描述容易出现与实际图像不一致或噪声描述，如图5中的红色字体标注的示例。

相比之下，本文提出的IC²E方法利用聚类结构和图文相似度信息，能够以无监督的方式挖掘与图像内容更紧密相关的语义概念。该方法不仅提升了模型对特定领域数据的适应性，也降低了对先验知识(如类别标签、提示模板等)的依赖。在实际应用中，本文发现，通过在聚类语义空间中生成粗粒度概念标签，IC²E可为大模型生成的更加细粒度的描述提供语义约束或候选补充，从而在多种下游任务中展现更强的可拓展性与实用性。此外，本文所使用的初始文本概念库在所有数据集上保持一致，仅需预先构建一次；在初始类簇图像库和最终概念词库的生成过程中，模型参数保持冻结状态，避免了额外的反向传播计算，因而整体计算开销较低。表2列出了在六个实验数据集上构建最终概念词库所耗费的时间，从表中可以看出，本文提出的概念词库构建策略具备良好的可扩展性，能够部署到较大规模的数据集上。

表2 不同数据集生成概念词库耗时

数据集	CIFAR10	STL10	ImageNet-10
时间	39秒	14秒	51秒
数据集	ImageNet-100	DTD	RESISC45
时间	7分36秒	1分13秒	39秒

4.2 实验设置

本文提出的可解释聚类方法(Interpretable Image Clustering with Concept Embedding，IC²E)基于Python语言和Pytorch框架[62]实现，所有实验均在配备Nvidia Geforce RTX 2080 12 GB的服务器上进行。对于使用的多模态框架CLIP模型[44]，本文统一选用预训练的VIT-B/32作为主干网络。鉴于CLIP模型本身已经具备较强的跨模态语义表示能力，其参数在整个训练过程中保持冻结，仅对后续新增模块进行训练。概念投影模块被设计为由5层全连接层构成的前馈神经网络，其输出维度根据文本概念的数目M动态调整。稀疏投影层由单层全连接网络构成，负责将概念嵌入表征映射至类簇概率空间，其网络权重在训练过程中受到非负约束以增强可解释性。模型使用Adam优化器进行参数更新，初始学习率设置为1e-3，每个批次包含512张图像，损失权重参数α设置为0.1。受益于CLIP提供的高质量表示能力，模型整体训练轮数设置为50轮即可获得相对稳定的聚类性能。为了缓解噪音信息的干扰，增强语义近邻的置信度，在3.2节中提出的近邻控制参数L和V分别设置为100和20，并在所有数据集上保持一致，以验证方法的鲁棒性与通用

性。对于聚类有效性评估，本文采用三种主流的聚类性能评价指标：聚类准确率（Accuracy，ACC）、归一化互信息（Normalized Mutual Information，NMI）以及调整兰德指数（Adjusted Rand Index，ARI），这些指标的数值越高，表示聚类性能越好。

对于模型的可解释性，本文主要使用F1值进行评估，并引入有效概念数目[23-24,53]（Number of Effective Concepts，NCE）度量解释的简洁性。为确保实验结果的稳定性与统计可靠性，所有实验均重复进行5次，并报告平均结果。

4.3　对比方法

为验证所提方法IC²E的有效性，本文在六个广泛使用的数据集上进行了系统评估，并将其与当前主流的深度聚类算法进行了对比。对比方法主要包括以下三类：

（1）基于ResNet等卷积网络的深度聚类方法

该类方法普遍采用ResNet-18或ResNet-34作为图像特征提取器，并结合各类辅助任务或结构约束提升聚类性能，代表性算法有CC[39]，IDFD[63]，SRL[40]，SCAN[41]和SPICE[42]等；

（2）基于预训练多模态大模型的深度聚类方法

该类方法多采用CLIP等大规模视觉语言模型的视觉编码器作为主干网络，并冻结编码器参数。在此基础上附加聚类模块以获取最终聚类结果。这类方法在多个任务中取得显著提升，代表性算法包括：SIC[12]，TAC[5]，CPP[43]和DXMC[64]等。为公平比较，本文保持一致的CLIP模型参数，并按照原文代码在部分数据集上重新训练；

（3）基线方法

为更全面地评估所提模型的性能与优势，本文还引入了以下两种基线模型进行对比：CLIP（K-means）代表直接对CLIP视觉编码器的输出进行K-means聚类，不使用任何额外的训练策略；Pre-model代表仅利用本文提出的预训练策略进行聚类预测，若无特殊表述，则后续实验均基于该预训练模型进行。

之后，为进一步验证IC²E的可扩展性与通用性，本文将预训练模型替换为现有聚类模型，如TAC[5]和DXMC[64]，构建了两种模型变体IC²E（TAC）和IC²E（DXMC），用于检验所提方法与现有聚类模型的适配能力。括号内容在不引起歧义的情况下省略。

4.4　模型有效性分析

如表3所示，本文首先在3个广泛使用的数据集上评估了所提方法的聚类性能，并与现有深度聚类方法进行了系统对比。实验结果表明，早期基于轻量级主干网络（如ResNet-18或ResNet-34）的聚类方法，如CC[39]，IDFD[63]，SRL[40]等，在整体聚类精度上偏低，尤其在图像分辨率较高或语义分布复杂的数据集上，其性能差距更加明显。尽管SCAN[41]和SPICE[42]等方法通过引入预训练机制提升了网络的特征表达能力，但由于其模型结构相对简单，仍难以充分捕获高层语义信息，导致最终聚类效果未能达到理想水平。

表3　在三个广泛使用数据集上的聚类性能（%）

方法	CIFAR10			STL10			ImageNet-10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CC	79.0	70.5	63.7	85.0	76.4	72.6	89.3	85.9	82.2
IDFD	81.5	71.1	66.3	75.6	64.3	57.5	95.4	89.8	90.1
SRL	90.3	83.4	81.7	81.8	72.4	68.2	95.9	90.2	91.2
SCAN	88.3	79.7	77.2	80.9	69.8	64.6	/	/	/
SPICE	83.8	73.4	70.5	90.8	81.7	81.2	96.9	92.7	93.3
SIC	92.6	84.7	84.4	98.1	95.3	95.9	98.2	97.0	96.1
TAC	91.9	83.3	83.1	98.2	95.5	96.1	99.2	**98.5**	98.3
DXMC	92.8	84.7	84.9	98.3	95.8	96.7	99.4	98.2	98.6
CLIP（K-means）	73.9	68.5	59.3	94.5	92.2	89.5	98.2	96.9	96.1
Pre-model	92.9	85.0	85.1	97.9	95.0	95.4	99.1	97.5	98.0
IC²E（Pre-model）	**93.0**	**85.0**	**85.1**	97.5	94.6	94.5	99.2	97.7	98.2
IC²E（TAC）	91.5	82.4	82.2	**98.4**	**96.0**	**96.7**	99.3	98.1	98.4
IC²E（DXMC）	92.0	83.2	83.3	98.3	95.7	96.4	**99.4**	98.1	**98.6**

与上述方法相比，近年来兴起的基于预训练多模态大模型的方法在多个数据集上展现出显著的性能优势。例如，直接对CLIP模型的视觉编码器输出进行K-means聚类（表3中CLIP（K-means）），

在STL10和ImageNet-10数据集上已经取得了相当优异的聚类精度。进一步地，SIC[12]，TAC[5]等方法在冻结CLIP编码器的基础上，设计了可训练的聚类投影层，并结合各类辅助任务或结构约束对模型进行微调，显著提升了模型性能，验证了强表征模型在聚类任务中的有效性。然而，这些方法通常依赖外部先验信息（如文本相似度或标签近邻）引导训练过程。相比之下，本文所提出的预训练方法仅采用数据增强策略构建一致性约束，无需引入外部知识，在三个数据集上取得了与当前最优方法相当甚至更优的聚类精度。同时，本文所设计的两种模型变体IC²E（TAC）和IC²E（DXMC）进一步验证了本方法的高度灵活性与可扩展性，其可无缝集成于现有聚类模型中以提升整体精度。

为进一步验证所提方法在真实场景中的适应能力，本文在图像分类数据集ImageNet-100、纹理分类数据集DTD和遥感图像场景分类数据集RESISC45上展开了对比实验，实验结果如表4所示。考虑到基于传统卷积网络的深度聚类方法在精度上明显弱于基于预训练大模型的方法，本文主要选取两个具有代表性的模型进行比较。从结果中可以观察到，本文方法在ImageNet-100数据集上的表现低于TAC[5]方法，主要原因在于该数据集包含较多类别，单一的一致性约束在此类场景中可能难以有效实现簇间样本的充分分离；而TAC[5]所采用的实例判别损失在增强簇间区分性方面具有更明显优势。尽管如此，IC²E方法具备良好的兼容性，能够与TAC[5]集成使用，从而实现性能的进一步提升。

表4 在三个更具挑战性数据集上的聚类性能（%）

方法	ImageNet100		DTD		RESISC45	
	ACC	NMI	ACC	NMI	ACC	NMI
CC	/	/	25.0	35.3	29.1	40.4
IDFD	/	/	36.8	47.3	70.2	74.1
CLIP	56.0	70.1	45.4	55.8	66.4	72.5
CPP	57.3	70.6	50.2	57.7	70.1	73.8
TAC	73.7	79.5	50.1	57.6	67.2	73.0
DXMC	60.8	77.6	44.9	56.2	66.0	73.1
IC²E	64.2	76.2	**51.5**	**62.0**	**82.6**	**84.5**
IC²E（TAC）	**75.0**	**80.4**	51.3	58.9	67.2	72.9
IC²E（DXMC）	64.6	78.2	47.6	58.2	66.2	73.1

在DTD与RESISC45这两个具有背景噪音显著的数据集上，本文方法表现出明显的优越性。这主要得益于数据增强策略的引入，其在一定程度上提升了模型对背景干扰的鲁棒性，使其在复杂视觉环境中仍能保持相对稳定的聚类结果。总体而言，本文方法及其变体在所有数据集上均取得了优于现有方法的聚类精度，充分验证了其在多样化场景中的有效性与泛化能力。

4.5 模型可解释性分析

为深入理解模型在样本层面上的聚类行为，本文从语义概念层面对聚类结果进行了可视化分析。如图6所示，本文展示了IC²E方法为多个典型样本生成的可解释证据。可以观察到，模型能够自动提取与人类语义认知一致的关键概念。例如，在“卡车”类别中，模型学习到“司机”、“交通运输”、“橡胶轮胎”等与交通场景密切相关的描述；在“跑道”场景中，识别出了“柏油碎石路面”等高层语义属性；而在纹理数据集中，对于“蜂窝状”图像，模型也能准确提取“蜂窝”、“晶格”、“规则结构”等特征性描述。这些结果表明，IC²E不仅在聚类精度上具备优势，同时能够以自然语言的形式清晰、准确地解释聚类结果，从而显著提升了聚类模型的可解释性。

进一步地，图7展示了同一类簇中不同样本对应的语义解释结果。其中，绿色边框代表模型正确划分的样本，而红色边框则代表错误划分的样本。从图中可以看出，对于同一类簇的不同样本，模型能够根据各自图像内容输出相近但略有差异的可解释证据，展现出一定的个体区分能力。此外，通过对错误划分样本的可解释性结果进行分析，可以为模型失误提供有价值的诊断信息。以被错误划分到“卡车”类簇的样本为例，模型从图像中提取到“驾驶室”，“矩形集装箱”等语义信息。尽管这些概念在语义上与“卡车”密切相关，但该图像实际属于“轿车”类别。这种误判可能源于模型对局部结构的过度关注，而忽略了整体图像语义。通过语义标签的可视化，能够清晰揭示模型判断的依据，为进一步引入注意力机制或多尺度语义融合策略提供了启示。在图8中，本文展示了预训练模型与IC²E方法最终用于聚类的有效概念数目。从图中可以观察到，在保证聚类精度的前提下，IC²E方法所使用的有效语义维度显著少于预训练模型输出的原始特征维度，从而在显著提升可解释性的同时，实现了对冗余特征的有效压缩，进一步确保了解释结果的简洁性与结构精炼性。

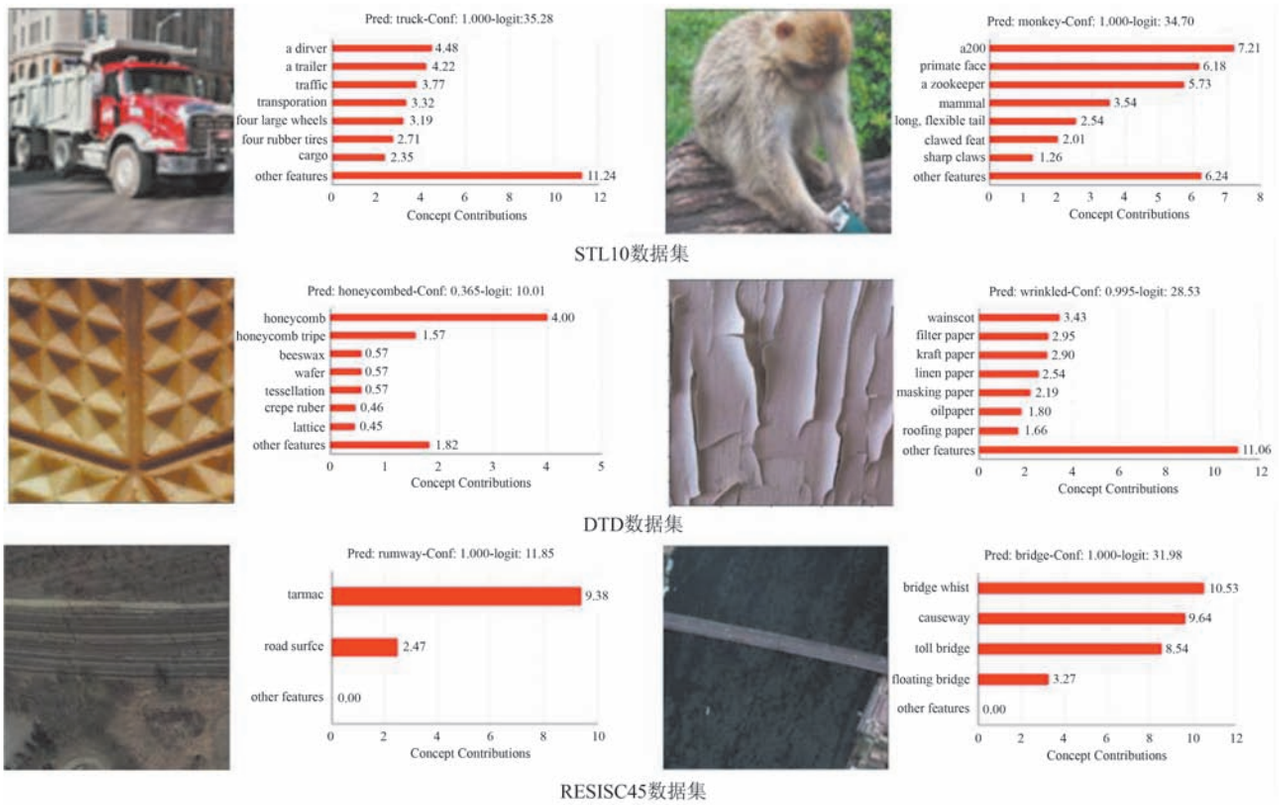


图6　不同类簇样本对应的可解释证据

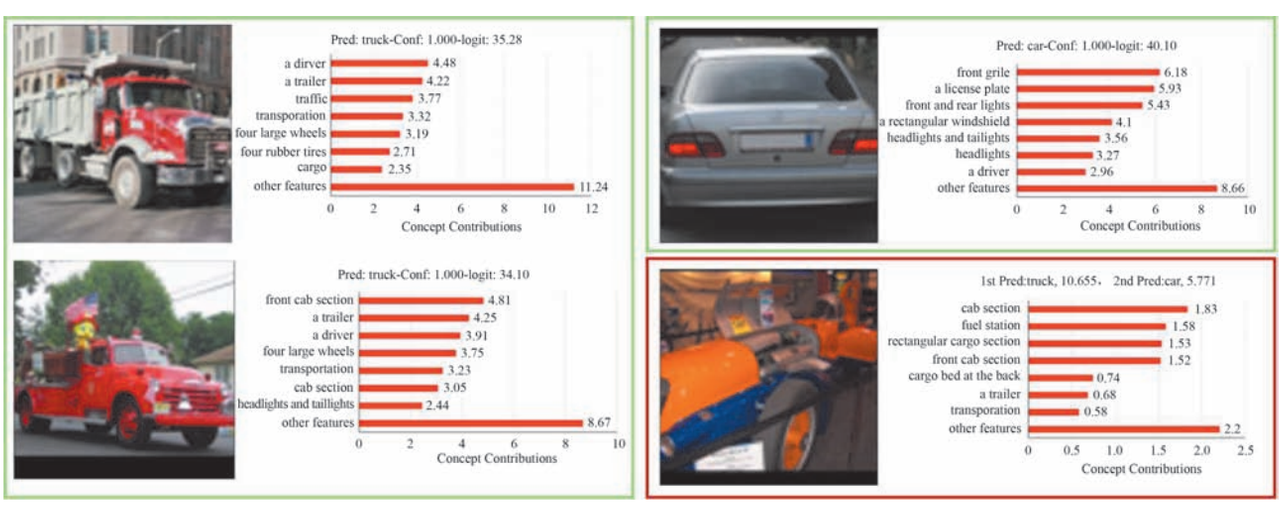


图7　相同类簇样本对应的可解释证据(绿色边框代表正确划分,红色边框代表错误划分)

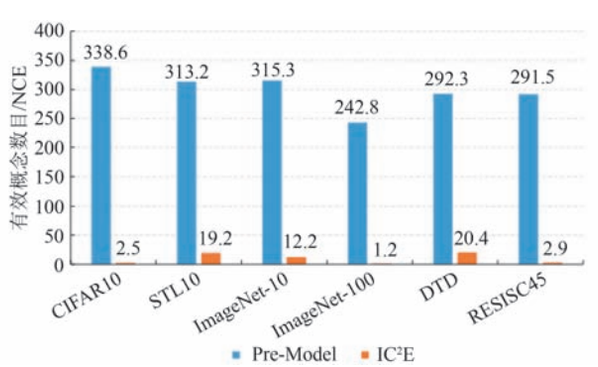


图8　不同模型最终用于聚类的有效特征数目

4.6　消融实验

为系统评估所提方法中各关键模块对聚类性能与语义可解释性的贡献,本文设计并实施了一系列消融实验,从聚类精度与解释质量两个维度展开,验证各模块在整体模型性能中的作用。

具体来说,在表5中,本文展示了在引入概念对齐模块与聚类修正模块前后,模型聚类性能的变化。可以观察到,引入概念对齐模块后,模型在部分数据集上的聚类精度略有下降。这是因为该模块在

优化过程中引入了对语义空间结构的约束，使得模型在保持聚类结构一致性的同时，更倾向于实现嵌入表征的语义对齐，从而在一定程度上削弱了特征表示的判别能力。然而，这种结构性约束为模型带来了显著的可解释性提升。为进一步量化各模块对可解释性的贡献，本文基于第3.3节中构建的概念代理矩阵$\tilde{S}$，使用F1分数评估概念表征的语义对齐质量。具体而言，本文将模型生成的概念表征与代理矩阵进行0/1二值化匹配，并计算二者间的F1分数作为解释性能的度量。在表5中，概念对齐模块在各数据集上均显著提升了F1分数，表明其有效增强了语义标签与视觉表征之间的一致性；而聚类修正模块则主要负责提升聚类精度，对可解释性贡献相对有限。此外，本文还对比了仅使用预训练模型（Pre-model）与最终可解释模型（IC²E）的整体性能表现。实验结果表明，所提出的可解释聚类方法基本保持了预训练模型在各数据集上的聚类精度，同时显著增强了语义可解释性，验证了本文方法在聚类有效性与可解释性之间实现了较好的平衡。最后，图9展示了逐步去除概念对齐模块中各个组件对模型可解释性的影响。结果表明，这些组件均在不同程度上均有助于提升可解释证据的质量，进一步佐证了其设计的合理性与有效性。

表5 不同模块对模型性能的影响（%）

方法	STL10			DTD		
	ACC	NMI	F1	ACC	NMI	F1
w/o cluster	49.1	46.7	86.0	18.5	25.0	**98.2**
w/o concept	**97.9**	**94.9**	2.3	49.7	61.6	1.6
IC²E	97.5	94.6	**86.2**	**51.5**	**62.0**	97.5
Pre-model	**97.9**	**95.0**	/	50.0	**62.4**	/
IC²E	97.5	94.6	/	**51.5**	62.0	/

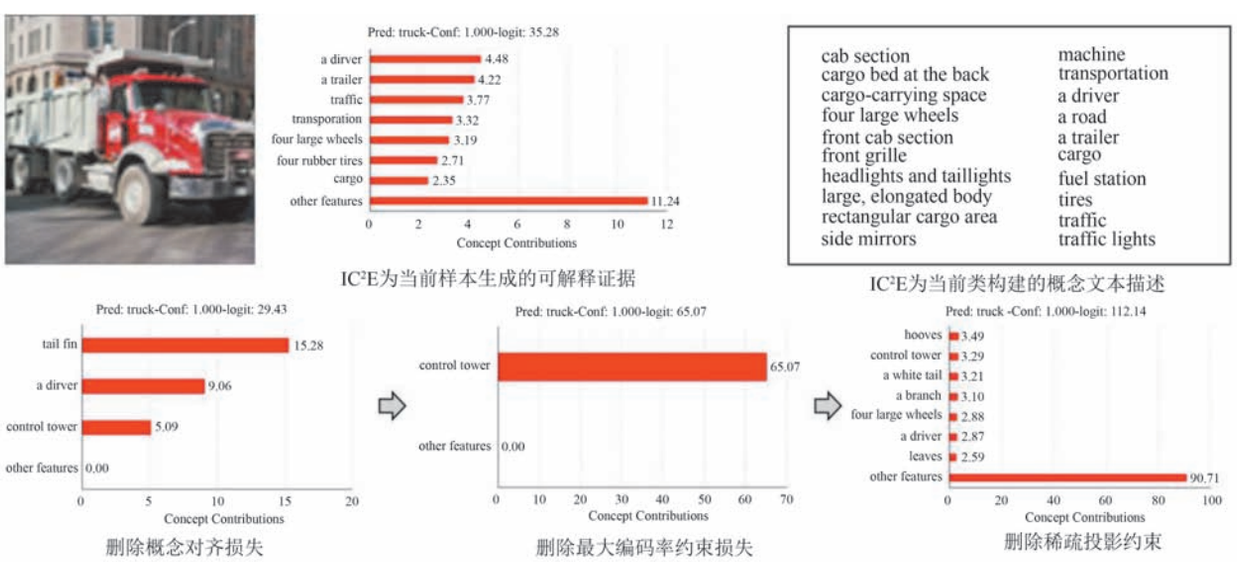


图9 去除各项约束对模型可解释性的影响

4.7 参数分析

在所提出的可解释深度聚类模型IC²E中，主要的超参数为损失项权重系数α，用于平衡聚类有效性与语义可解释性之间的关系，如算法1所示。为深入分析该权重系数对模型性能的具体影响，本文在STL10数据集上进行了相关实验，评估模型在不同损失系数取值下模型性能的变化。从图10中可以看出，随着α取值的增大，模型的聚类精度（ACC、NMI、ARI）整体保持稳定，而可解释性评价指标F1分数则呈现明显下降趋势。这是由于损失系数α增大时，概念对齐损失在总损失中的比重降低，其优化效果减弱，从而削弱了概念表征的语义一致性与可解释性。因此，综合考虑聚类有效性与可解释性之间的平衡，本文将α统一设置为0.1，该值在多个评估指标上均展现出较为稳定且优良的性能。

对于第3.2节中使用的近邻参数L，其仅在概念生成阶段中使用，用于控制每个簇选取的代表样本数目。较大的L值可以获得更全面的簇内信息，但会增加噪声样本的干扰；较小的L值则可能导致语义信息不足。为此，本文将其固定为100，以在信息充分性与噪声抑制之间取得平衡。另一近邻参数V同时服务于概念生成与概念对齐。如表6所示，实验结果表明，V的变化对聚类精度影响并不显著，但会影响可解释概念的多样性。当V取值过大时

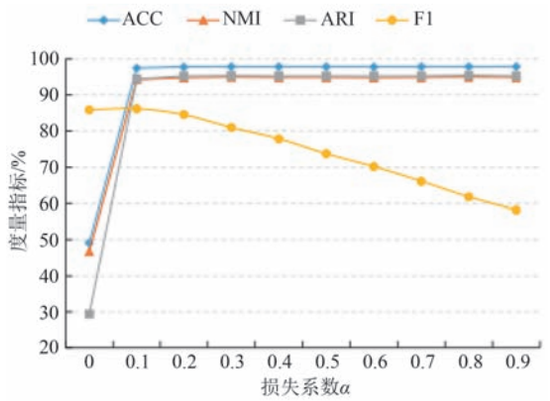


图10　STL10数据集上不同损失系数α对模型性能的影响

会增加人工理解的负担；当V取值过小时则可能遗漏重要语义。为此，本文受已有文献[65]启发，将V设定为20，在保证概念集具有充分表达能力的同时，降低用户的认知负荷，提高信息的易读性。

4.8　嵌入表示可视化

为直观分析语义概念的引入对类簇结构的影响，本文使用t-SNE对不同数据集上的嵌入表示进行降维可视化，并比较引入语义概念前后的可视化结果，如图11所示。可以观察到，在引入语义概念之后，嵌入空间中不同簇之间的分界更加清晰，簇内样本分布更为紧凑。为了更好地说明这一点，本文在模型输出的嵌入表示上直接应用k-means聚类，并观察到聚类精度（NMI）的提升。这表明语义概念引导机制不仅增强了模型在特征空间与语义空间之间的一致性，还优化了嵌入表示的可分性，从而提升了聚类的有效性与可解释性。

表6　近邻参数V在RESISC45数据集上不同取值下的模型性能

近邻V取值	4	12	20	30	50
ACC	82.5	82.6	82.6	82.7	82.6
NMI	84.5	84.4	84.5	84.6	84.5
ARI	73.6	73.6	73.8	73.9	73.8
F1	92.8	92.2	91.7	91.5	91.0
NCE	1.3	2.2	2.9	3.4	4.5

4.9　适用性分析

为进一步验证所提方法在不同领域数据集上的适用性与泛化能力，本文在保持前述所有超参数设置不变的情况下，将模型直接应用于医学图像数据集LC25000进行实验评估。该数据集包含不同类型的组织切片图像，具有较高的专业性与领域特异性。实验结果如图12所示。

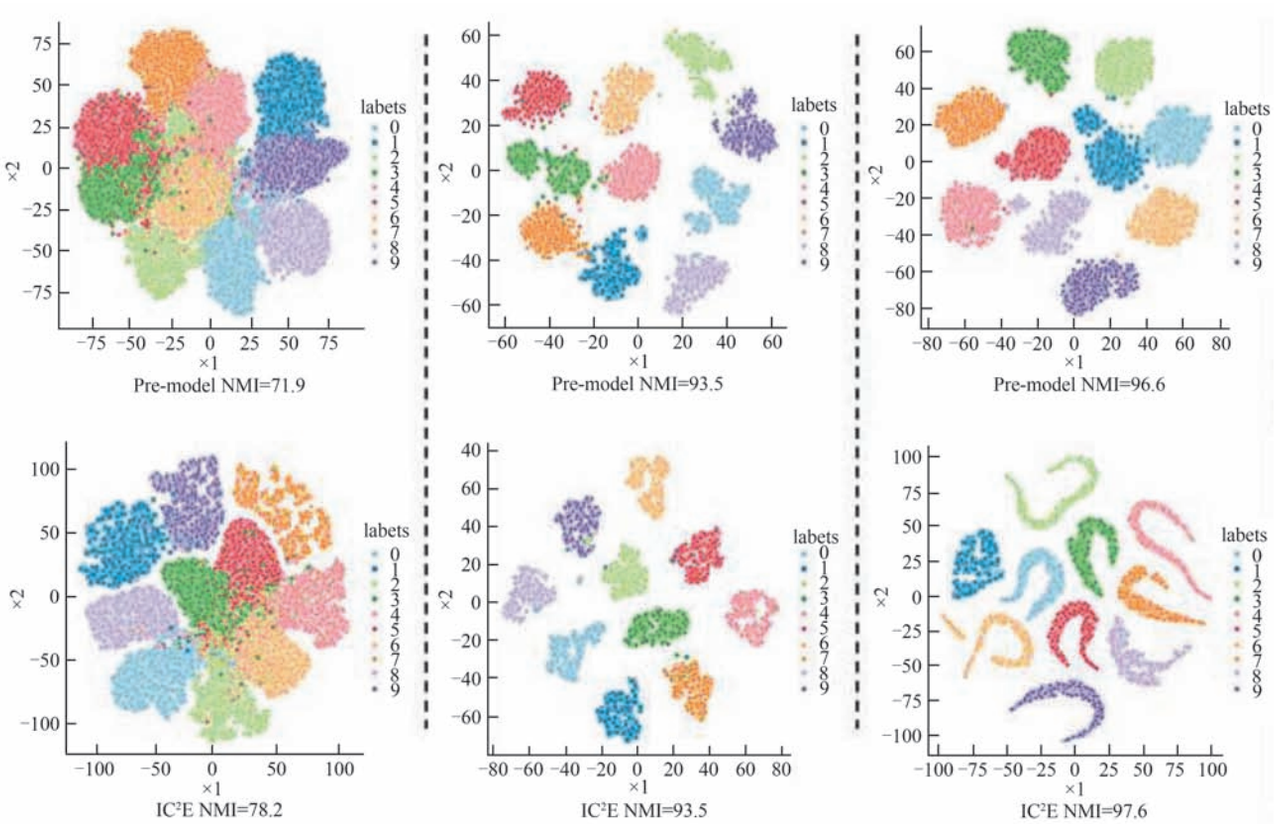
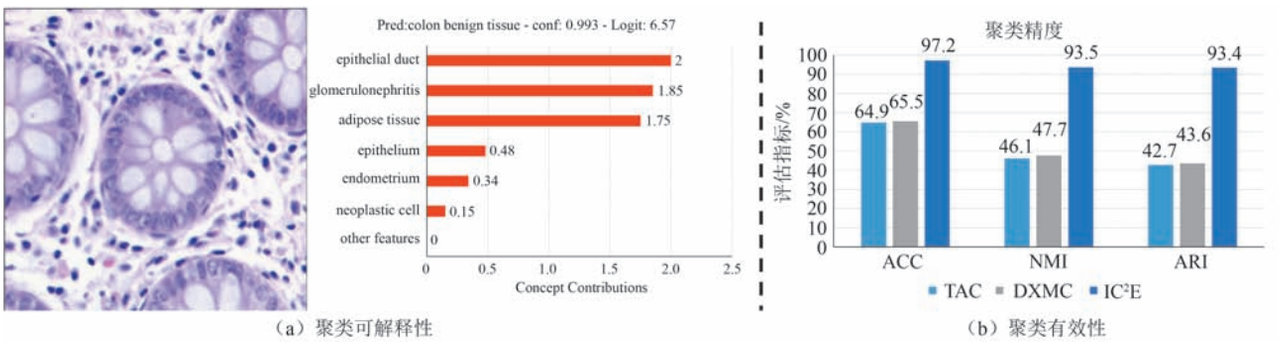


图11　不同数据集上的嵌入表示可视化（从左至右分别为CIFAR10，STL10，ImageNet-10数据集）



（a）聚类可解释性 （b）聚类有效性

图12 在LC25000数据集上的聚类精度和可解释证据

从结果可以观察到，在不进行额外参数调整或领域特定预处理的情况下，所提方法依然在LC25000数据集上取得了较为优异的聚类精度。这表明该方法在跨领域场景中具备一定的稳健性与迁移性。相比之下，TAC[5]和DXMC[64]在该数据集上的性能显著下降，其主要原因在于这两类方法依赖事先构建的文本近邻先验进行模型引导，而这些先验通常基于通用词汇库（如WordNet[55]）构建，难以准确反映医学领域的专有概念与语义结构，从而影响了聚类性能。在可解释性方面，所提方法能够正确识别主要图像类簇，但部分生成的可解释证据与实际图像内容存在一定程度的不一致。这一现象一方面源于通用词汇库在医学语义覆盖上的不足，另一方面与所使用的CLIP模型在医学领域的特征表征能力有限密切相关。尽管如此，随着领域适配与知识注入技术的不断发展，这一问题有望在未来得到有效缓解。

5 结 论

本文提出了一种概念嵌入增强的可解释图像聚类方法（Interpretable Image Clustering with Concept Embedding，IC²E）。该方法能够在类别名称缺失的情况下自动生成与给定数据集高度相关的描述性短语，并将其嵌入深度神经网络。通过提出的概念表征对齐和聚类决策修正模块，所提方法能够在完成聚类决策的同时输出人类可理解的文本化概念。本文对6个被广泛使用的数据集进行了充分的实验，并与已有聚类算法进行性能对比，证明了本文方法的有效性。

此外，本文在实验中也观察到，当前方法在解释能力方面也存在一定的局限性。所生成的文本概念在一定程度上依赖于预训练模型的语义编码能力，以及初始聚类模型在特征空间中的表征鲁棒性。同时，当面对更加复杂的语义类别时，当前方法存在一定程度的语义冲突与重叠。然而，当结合更强大的语言模型以及更加丰富且精准的先验信息时，这些问题有望被解决。

参 考 文 献

[1] Huang Huajuan, Wang Chen, Wei Xiuxi, et al. Deep image clustering: a survey. Neurocomputing, 2024, 599 (000): 128101

[2] Liu Jiaxin, Wang Dongwei, Yu Siquan, et al. A survey of image clustering: taxonomy and recent methods//Proceedings of the IEEE International Conference on Real-time Computing and Robotics. Qinghai, China, 2021: 375-380

[3] Ren Yazhou, Pu Jingyu, Yang Zhimeng, et al. Deep clustering: a comprehensive survey. IEEE transactions on neural networks and learning systems, 2024, 36(4): 5858-5878

[4] Li Chao, Liao Hongmei, Xu Xiao, et al. Robust spectral clustering based on density distribution. Chinese Journal of Computer, 2024,47(11):2645-2663
(李超，廖红梅，徐晓，等. 基于密度分布的鲁棒谱聚类算法. 计算机学报,2024,47(11):2645-2663)

[5] Li Yunfan, Hu Peng, Peng Dezhong, et al. Image clustering with external guidance//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 27890-27902

[6] Zhou Sheng, Xu Hongjia, Zheng Zhuonan, et al. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. ACM Computing Surveys, 2024, 57(3): 1-38

[7] Moshkovitz M, Dasgupta S, Rashtchian C, et al. Explainable k-means and k-medians clustering//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2020: 7055-7065

[8] Peng Xi, Li Yunfan, Tsang I W, et al. XAI beyond classification: interpretable neural clustering. Journal of Machine Learning Research, 2022, 23(6): 1-28

[9] Svirsky J, Lindenbaum O. Interpretable deep clustering for tabular data//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 47314-47330

[10] Alvarez-Garcia M, Ibar-Alonso R, Arenas-Parra M. A comprehensive framework for explainable cluster analysis. Information Sciences, 2024, 663: 120282

[11] Yang Haoyu, Jiao Lianmeng, Pan Quan. A survey on

interpretable clustering//Proceedings of the Chinese Control Conference. Shanghai, China, 2021: 7384-7388

[12] Cai Shaotian, Qiu Liping, Chen Xiaojun, et al. Semantic-enhanced image clustering//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(6): 6869-6878

[13] Liu Jiyuan, Liu Xinwang, Cai Zhiping, et al. On the correlation measurement of data representations. Chinese Journal of Computer, 2024,47(7):1568-1581
(刘吉元,刘新旺,蔡志平,等. 数据表示的相关性度量方法. 计算机学报,2024,47(7):1568-1581)

[14] Liang Weixuan, Liu Xinwang, Lan Long, et al. On the generalization of spectral clustering on bipartite graph. Chinese Journal of Computer, 2025,48(5): 1065-1081
(梁伟轩，刘新旺，蓝龙，等. 关于二部图谱聚类泛化性的研究. 计算机学报,2025,48(5): 1065-1081)

[15] Pan Yuangang, Yao Yinghua, Tsang I. PC-X: Profound clustering via slow exemplars//Proceedings of the Conference on Parsimony and Learning. Hong Kong, China, 2024: 1-19

[16] Hu Lianyu, Jiang Mudi, Dong Junjie, et al. Interpretable clustering: a survey. arXiv preprint, arXiv:2409.00743, 2024

[17] Gamlath B, Jia X, Polak A, et al. Nearly-tight and oblivious algorithms for explainable clustering//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021, 34: 28929-28939

[18] Hwang H, Whang S E. Xclusters: explainability-first clustering//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(7): 7962-7970

[19] Bandyapadhyay S, Fomin F V, Golovach P A, et al. How to find a good explanation for clustering. Artificial Intelligence, 2023, 322: 103948

[20] Carrizosa E, Kurishchenko K, Marín A, et al. On clustering and interpreting with rules by means of mathematical optimization. Computers & Operations Research, 2023, 154: 106180

[21] Saisubramanian S, Galhotra S, Zilberstein S. Balancing the tradeoff between clustering value and interpretability// Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Madrid, Spain, 2020: 351-357

[22] Frost N, Moshkovitz M, Rashtchian C. ExKMC: expanding explainable k-means Clustering. arXiv preprint, arXiv: 2006.02399, 2020

[23] Jiang Mudi, Hu Lianyu, He Zengyou, et al. Interpretable multi-view clustering. Pattern Recognition, 2025, 162: 111418

[24] Dong Junjie, Yang Xinyi, Jiang Mudi, et al. Interpretable sequence clustering. Information Sciences, 2025, 689: 121453

[25] Laber E, Murtinho L, Oliveira F. Shallow decision trees for explainable k-means clustering. Pattern Recognition, 2023, 137: 109239

[26] Ma Xu, Zhou Yuqian, Wang Huan, et al. Image as set of points//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023

[27] Kang Yaming, Ye Peishun, Bai Yuxin, et al. Hyperspectral image based interpretable feature clustering algorithm. Computers, Materials & Continua, 2024, 79(2):2151-2168

[28] Nauta M, Van Bree R, Seifert C. Neural prototype trees for interpretable fine-grained image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. virtual, 2021: 14933-14943

[29] Wan A, Dunlap L, Ho D, et al. NBDT: neural-backed decision tree//Proceedings of the International Conference on Learning Representations. Virtual , Austria, 2021

[30] Davidson I, Livanos M, Gourru A, et al. Explainable clustering via exemplars: Complexity and efficient approximation algorithms. arXiv preprint, arXiv:2209.09670, 2022

[31] Oikarinen T, Das S, Nguyen L, et al. Label-free concept bottleneck Models//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023

[32] Cui Yan, Liu Shuhong, Li Liuzhuozheng, et al. Ceir: concept-based explainable image representation learning. arXiv preprint, arXiv:2312.10747, 2023

[33] Koh P W, Nguyen T, Tang Y S, et al. Concept bottleneck models//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 5338-5348

[34] ang Y, Panagopoulou A, Zhou S, et al. Language in a bottle: language model guided concept bottlenecks for interpretable image classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 19187-19197

[35] Lin Yuxiu, Liu Hui, Yu Xiao, et al. A multi-view unified representation learning network for subspace clustering. Journal of Computer Research and Development, 2025, 62(5): 1248-1261
(林毓秀，刘慧，于晓，等. 面向子空间聚类的多视图统一表示学习网络. 计算机研究与发展，2025, 62(5): 1248-1261)

[36] Xie Junyuan, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis//Proceedings of the International Conference on Machine Learning. New York City, USA, 2016: 478-487

[37] Jiang Zhuxi, Zheng Yin, Tan Huachun, et al. Variational deep embedding: an unsupervised and generative approach to clustering//Proceedings of the International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 1965-1972

[38] Huang Jiabo, Gong Shaogang, Zhu Xiatian. Deep semantic clustering by partition confidence maximization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 8849-8858

[39] Li Y, Hu P, Liu Z, et al. Contrastive clustering//AAAI Conference on Artificial Intelligence. 2021, 35(10): 8547-8555

[40] Xiang Wang, JingLiping, LiuHuafeng, and Jian Yu. Structure-driven representation learning for deep clustering. ACM Transactions on Knowledge Discovery from Data, 18 (1):1-25, 2023

[41] Van Gansbeke W, Vandenhende S, Georgoulis S, et al. Scan: learning to classify images without labels//Proceedings of European Conference on Computer Vision. Glasgow, UK, 2020: 268-285

[42] Niu C, Shan H, Wang G. Spice: Semantic pseudo-labeling for image clustering. IEEE Transactions on Image Processing,

2022, 31: 7264-7278
[43] Chu T C, Tong S, Ding T, et al. Image clustering via the principle of rate reduction in the age of pretrained models// Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024
[44] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of International Conference on Machine Learning. Virtual, 2021: 8748-8763
[45] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206-215
[46] Guilbert M, Vrain C, Dao T B H. Towards explainable clustering: a constrained declarative based approach. arXiv preprint arXiv:2403.18101, 2024
[47] Hu Lianyu, Jiang Mudi, Liu Xinying, et al. Significance-based decision tree for interpretable categorical data clustering. Information Sciences, 2025, 690: 121588
[48] He Zengyou, Hu Lianyu, He Jinfeng, et al. Significance-based interpretable sequence clustering. Information Sciences, 2025, 704: 121972
[49] IsadoraSalles, PaolaMejia-Domenzain, VinitraSwamy, et al. Interpret3C: interpretable student clustering through individualized feature selection . In Proceedings of the International Conference on Artificial Intelligence in Education, Xi'an, China, 2024: 382-390
[50] Espinosa Zarlenga M, Barbiero P, Ciravegna G, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. Advances in Neural Information Processing Systems, 2022, 35: 21400-21413
[51] Liu H, Yan W, Abbeel P. Language quantized autoencoders: Towards unsupervised text-image alignment. Advances in Neural Information Processing Systems, 2023, 36: 4382-4395
[52] Ju Tianjie, Liu Gongshen, Zhang Jongsheng, et al. A review of probe interpretable methods in natural language processing. Chinese Journal of Computers, 2024, 47(4): 733-758
(鞠天杰，刘功申，张倬胜，等. 自然语言处理中的探针可解释方法综述. 计算机学报，2024, 47(4): 733-758)
[53] Srivastava D, Yan G, Weng L. Vlg-cbm: training concept bottleneck models with vision-language guidance. Advances in Neural Information Processing Systems, 2024, 37: 79057-79094
[54] He H, Zhu L, Zhang X, et al. V2C-CBM: building concept Bottlenecks with Vision-to-Concept Tokenizer. arXiv preprint arXiv:2501.04975, 2025
[55] Miller, G. A. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39-41, 1995
[56] WongEric, SanturkarShibani, and MadryAleksander. Leveraging sparse linear layers for debuggable deep networks//Proceedings of the International Conference on Machine Learning, virtual, 2021: 11205-11216
[57] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images, 2009
[58] Coates A, Ng A, Lee H. An analysis of single-layer networks in unsupervised feature learning//Proceedings of the International Conference on Artificial Intelligence and Statistics Workshop, Ft. Lauderdale, USA, 2011: 215-223
[59] Chang J, Wang L, Meng G, et al. Deep adaptive image clustering//Proceedings of IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5879-5887
[60] Cimpoi M, Maji S, Kokkinos I, et al. Describing textures in the wild//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3606-3613
[61] Cheng G, Han J, Lu X. Remote sensing image scene classification: Benchmark and state of the art. IEEE, 2017, 105(10): 1865-1883
[62] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch. 2017
[63] Tao Y, Takagi K, Nakata K. Clustering-friendly representation learning via instance discrimination and feature decorrelation// Proceedings of the International Conference on Learning Representations. Virtual, Austria, 2021
[64] Zhang H, Li Y, Huang D. Dual-level cross-modal contrastive clustering. arXiv preprint arXiv:2409.04561, 2024
[65] Miller George A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 1956, 63(2): 81

WANG Xiang, PhD. His main research interests include unsupervised clustering and fairness representation learning.

LIU Hua-Feng, Ph. D., associate professor. His main research directions are machine learning, probabilistic generative model related theories and their applications.

JING Li-Ping, Ph. D., professor. Her main research interests include machine learning and its application in the artificial intelligence field.

YU Jian, Ph. D., professor. His main research areas are artificial intelligence and machine learning.

GUO Long-Teng, Ph. D., associate researcher. His main research interests include graph data management and graph mining.

YANG Ya-Jun, Ph. D., associate professor. His main research interests include graph data management and graph mining.

Background

Image clustering aims to group images according to their underlying semantic content. Due to the nonlinear and complex structure of image data in high-dimensional spaces, traditional clustering methods based on shallow features struggle to capture intrinsic semantic patterns effectively. Recently, deep clustering has become the dominant approach by leveraging deep neural networks to learn hierarchical feature representations that project complex visual data into a low-dimensional embedding space, improving clustering accuracy and robustness. However, the nonlinear entanglement of features in deep embeddings often leads to a lack of interpretability, limiting their use in decision-critical applications.

Interpretable clustering has thus emerged as an important research direction. Instead of focusing solely on accuracy, it emphasizes transparency and explainability of clustering decisions. Approaches using decision trees or rule-based mechanisms enhance interpretability by providing explicit decision paths, but they are mainly effective for structured or low-dimensional data and do not generalize well to unstructured image data. For image clustering, recent studies have explored cluster-level semantic explanations, such as prototype representations or visualization-based interpretations. Yet, these methods rely on statistical summaries and subjective priors, which reduce the objectivity and consistency of explanations.

In this paper, we propose an Interpretable Image Clustering with Concept Embedding (IC^{2}E) framework. Inspired by the concept bottleneck model, IC^{2}E introduces semantically meaningful concept embeddings to align visual and semantic spaces, enabling each embedding dimension to be explained in natural language. A data-adaptive concept generation mechanism is further developed to automatically discover high-level semantic concepts without label supervision, enhancing both interpretability and scalability while maintaining strong clustering performance.

This work was partly supported by the National Key Research and Development Program of China under Grant (2024YFE0202900); the National Natural Science Foundation of China under Grant (62436001, 62406019, 62176020); Beijing Natural Science Foundation (4244096); the Talent Found of Beijing Jiaotong University (2024XKRC075); the Joint Foundation of the Ministry of Education for Innovation team (8091B042235); the Fundamental Research Funds for the Central Universities (2019JBZ110); and the State Key Laboratory of Rail Traffic Control and Safety (RCS2023K006).