用于组合视觉关系的增强异常对比学习预测推理

李成泰、何雨婷、任剑锋，*IEEE* 高级会员，白瑞斌，*IEEE* 高级会员，赵一田，*IEEE* 会员，于恒，*IEEE* 高级会员，蒋旭东，*IEEE* 会士，

arXiv:2603.01125v1 [cs.CV] 1 Mar 2026

摘要——尽管简单类比的视觉推理已受到广泛关注，但由于组合视觉关系（CVR）更加复杂，相关研究仍相对有限。为解决CVR任务，我们提出采用增强异常对比学习的预测推理（PR-A^{2}CL），即给定遵循相同组合规则的另外三张图像，识别出一张异常图像。为应对大量组合规则建模所带来的挑战，我们设计了增强异常对比学习，通过最大化正常实例之间的相似度，同时最小化正常实例与异常离群样本之间的相似度，提炼出具有判别性和泛化能力的特征。更重要的是，我们引入一种用于基于规则的推理的预测与验证范式，其中一系列预测异常推理模块（PARB）迭代利用四张图像中三张图像的特征，预测剩余一张图像的特征。在随后的验证阶段，PARB逐步定位由底层规则导致的具体差异。在SVRT、CVR和MC^{2}R数据集上的实验结果表明，PR-A^{2}CL的性能显著优于最先进的推理模型。

索引词—**抽象视觉推理、对比学习、知识表示、组合视觉推理。**

I. 引言

VISUAL 随着深度神经网络的发展，识别取得了显著进展 [1]–[3]。近年来，研究重点逐渐从物体和属性的视觉识别 [4]–[6] 转向对从图像和视频中感知到的高层概念进行视觉类比推理 [7], [8]。视觉推理大致可分为以下几个领域 [7]: 视觉问答 [9]–[10],视觉常识推理 [11]–[13],物理推理 [14], [15], 数学视觉推理 [16], [17]以及抽象视觉推理 [18]–[21]。在这些任务中，抽象视觉推理（AVR）持续受到广泛关注，因为尽管取得了显著进展，它仍是智能体面临的关键瓶颈 [22], ，尤其是在

用于计算机视觉 [23]和自然语- 言处理 [24]的大规模模型。通常，大语言模型（LLM）表现出强大的泛化能力，在各种自然语言处理 [24]任务中取得了最先进的成果

本研究得到宁波市科学技术局资助，项目编号为 2023Z138、2023Z237、2024Z110 和 2024Z124。C. Li、Y. He、J. Ren、R. Bai 和 H. Yu 任职于宁波诺丁汉大学计算机科学学院（中国宁波 315100）。C. Li、Y. Zhao 任职于中国科学院宁波材料技术与工程研究所生物医学工程研究所。X. Jiang 任职于新加坡南洋理工大学电气与电子工程学院，地址为 Nanyang Link，邮编 639798。通讯作者：任剑锋，*jianfeng.ren@nottingham.edu.cn*；*Yi-tian Zhao*，*yitian.zhao@nimte.ac.cn*，

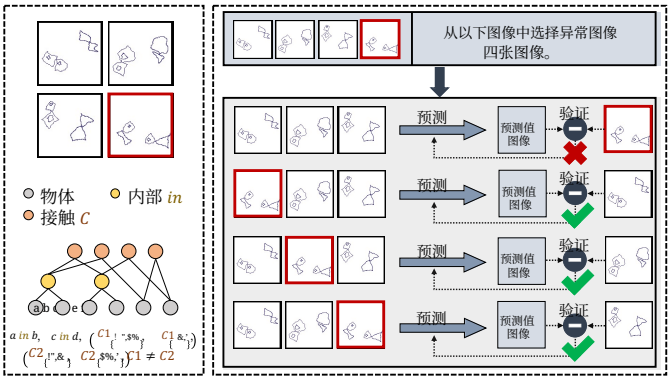


图1。左侧：CVR数据集中的一个组合规则样例。右侧：从四张图像中选出异常项，可转化为四个预测与验证问题。

以及多模态任务 [25], [26]。但大语言模型在抽象视觉推理任务 [27],上的表现缺乏竞争力，暴露出其推理能力的关键局限。因此，抽象视觉推理研究对于增强现代推理系统并大幅提升其抽象推理能力至关重要。

抽象视觉推理研究在瑞文渐进矩阵（RPMs） [28]–[31], 例如等任务上取得了显著进展；在RAVEN数据集 [18], [32], [33], 上，采用SCL [29], 、HCV- ARR [19], 和 PredRNet [20]的模型，其推理准确率已超越人类水平。然而，这些数据集通常只涉及数量有限的属性和规则，例如，RAVEN数据集 [18], [32], [33]仅包含五种属性和四条简单规则。近期推出的组合视觉关系（CVR）数据集 [34]通过要求模型从三张图像共有的组合规则中，找出略有偏离的一张图像，来评估抽象视觉推理能力。如图1所示，组合规则包含多层次的属性关系，与RAVEN数据集 [18], [32], [33]中的简单规则形成对比。

在这项工作中，我们专注于解决 CVR任务 [34], [35],——这是 AVR 中一个颇具挑战性的子类。AVR 涵盖了广泛的关系推理问题，而 CVR 则特别要求针对多个同时存在的规则及其组合交互进行推理。具体而言，CVR任务面临两项主要挑战：1）理解组合规则远比理解简单规则困难，因为这需要整合多个基本属性、建模它们之间的交互，并进行高层次抽象 [34]。2）组合规则的潜在无限空间，使推理模型在测试时遇到未见过的规则组合时，其泛化能力面临严峻挑战。

为应对这些挑战，我们提出了**P**redictive**R**easoning with **A**ugmented **A**nomaly **C**ontrastive**L**earning

(PR-A^{2}CL)。该方法整合了一个视觉感知模块、基于预训练 ResNet-50 [34]，构建的**A**ugmented **A**nomaly**C**ontrastive **L**earning（**A^{2}CL**），用于提取鲁棒且符合规则的特征；同时还包含一个**P**redictive **A**nomaly**R**easoning **M**odule（**PARM**），用于执行迭代式规则推理与验证。具体而言，所提出的 A^{2}CL 通过增强不同增强视图之间符合规则的特征对齐，提升泛化能力。它促进不同增强方式下正常样本之间形成较高的特征亲和性，以捕捉其内在特征，同时降低正常样本与异常实例之间的特征相似度，从而提高可分性。通过将离群样本与正常样本进行对比，并在增强视图之间施加一致性约束，A^{2}CL 学得了判别特征，使其在多种场景下具有很强的泛化能力。

为处理复杂的组合规则，我们提出了一种预测与验证（PAV）范式。如图1所示，我们首先将从四张图像中识别异常项的任务转化为四个预测与验证问题。每个子问题都使用所提出的**预**测**异**常**推**理**模**块（**PARB**），根据其他三张图像的特征预测目标图像的特征。其基本原理是，无法根据三张正常图像可靠地预测异常项的特征；而只要有效抑制异常项，就能根据另外两张图像准确推断正常图像的特征。随后，将预测值与目标进行验证，并将预测误差反向传播，以更新预测网络。收敛后，PARB能够捕捉正常图像之间的视觉关系，从而识别异常项。此外，我们将多个PARB堆叠起来，以便对组合规则进行层次化推理。初始PARB主要捕捉基础属性层面的关系，例如大小相同或位置相同；更深层则逐步将这些关系整合为更高阶的组合例如“大小相同，但形状和空间布局不同”。这种架构体现了CVR规则的多层结构，并使模型能够通过迭代优化推断抽象的组合模式。

我们的贡献总结如下。1）我们的 PR-A^{2}CL 通过提取鲁棒视觉特征，并针对组合规则进行抽象类比推理，攻克了组合视觉关系推理所面临的挑战。2）A^{2}CL 通过将离群样本与正常图像进行对比来提取判别特征，同时保持不同增强视图间的特征一致性。3）PARB 嵌入了一种预测—验证机制：利用其他三幅图像预测目标图像的特征，并通过将预测结果与目标进行核验来更新自身。4）在 SVRT [36],CVR [34]和 MC^{2}R [37]数据集上，PR-A^{2}CL 的表现显著优于最先进的视觉推理模型。

II. 相关工作

A. 抽象视觉推理

AVR 考察在不同场景中对抽象视觉关系进行类比推理的能力 [7]。这需要将相关知识应用于陌生情境，且被广泛用于评估人类智能 [7]。在文献中，AVR 涵盖多种任务，例如，Raven 渐进矩阵 [28]–[31], 寻找异同任务 [37], [38], 机器数字推理 [17] 和关系游戏 [39]。

在 AVR 任务中，瑞文渐进矩阵受到了广泛关注 [28]–[31]。这类任务需要提取视觉属性，分析行列关系，并从候选答案中推断缺失的图像。例如，Barrett 等人 [40]利用视觉特征对上下文—上下文关系和上下文—多选关系进行编码，并反复使用关系网络推断面板间的关系。Wu 等人 [29] 开发了散射组合学习器（SCL），从对象与属性网络中提取特征，并将其用于抽象推理。He 等人[19] 提出了层次化 ConViT，以捕获多层级特征，并结合基于注意力的关系推理器来推断潜在关系。PredRNet 使用浅层卷积神经网络提取视觉特征，并将八幅上下文图像直接映射到缺失图像，以推断抽象规则 [20]。Malkinski 等人 [41] 开发了一个用于单选式抽象视觉推理的统一模型，其中包含编码器—解码器推理框架，以及一个用于在不同推理任务间实现自适应特征组织的结构感知动态层。尽管 AVR 模型近期取得了进展 [28]–[31], ，其推理能力仍主要局限于简单的抽象规则，因为许多瑞文渐进矩阵基准都是根据这类规则构建的 [18], [32], [33]。

组合性在数学 [42], 逻辑推理 [43], 和视觉推理 [44], 领域已得到广泛研究，但在高层次组合式抽象推理中的探索仍然不足。为弥补这一空白，研究者引入了 CVR 数据集 [34] ，通过灵活组合视觉属性来评估组合性理解。在这项工作中，我们致力于解决组合式视觉推理面临的挑战。

B. 对比学习

对比学习（CL）通过约束同一样本不同增强视图之间的表征一致性，同时提升不同样本之间的可判别性 [45]。CL 研究主要聚焦于两个方向。1）开发训练机制，以增强骨干网络的泛化能力 [46], [47]。SimCLR [46] 约束同一样本的增强视图之间保持相似，同时最大化其与其他样本之间的差异，并采用可学习非线性变换来增强特征表示。Chen 等人 [47], 提出的 MoCo 采用队列来增加负样本，并使用动量编码器来维持队列中的特征一致性。2）设计损失函数，以更好地区分相似度各异的样本 [48]–[50]。InfoNCE 损失 [48] 遵循二元交叉熵（BCE）损失的结构，并引入一个超参数来调节对困难负样本的惩罚。Robinson 等人 [49] 构建了可调样本分布，以优先处理困难负样本。Chuang 等人 [50] 开发了鲁棒 InfoNCE（RINCE）损失，通过施加对称性来增强学习到的特征的鲁棒性。近期，一致性学习 [51], [52]通过对齐不同模态、

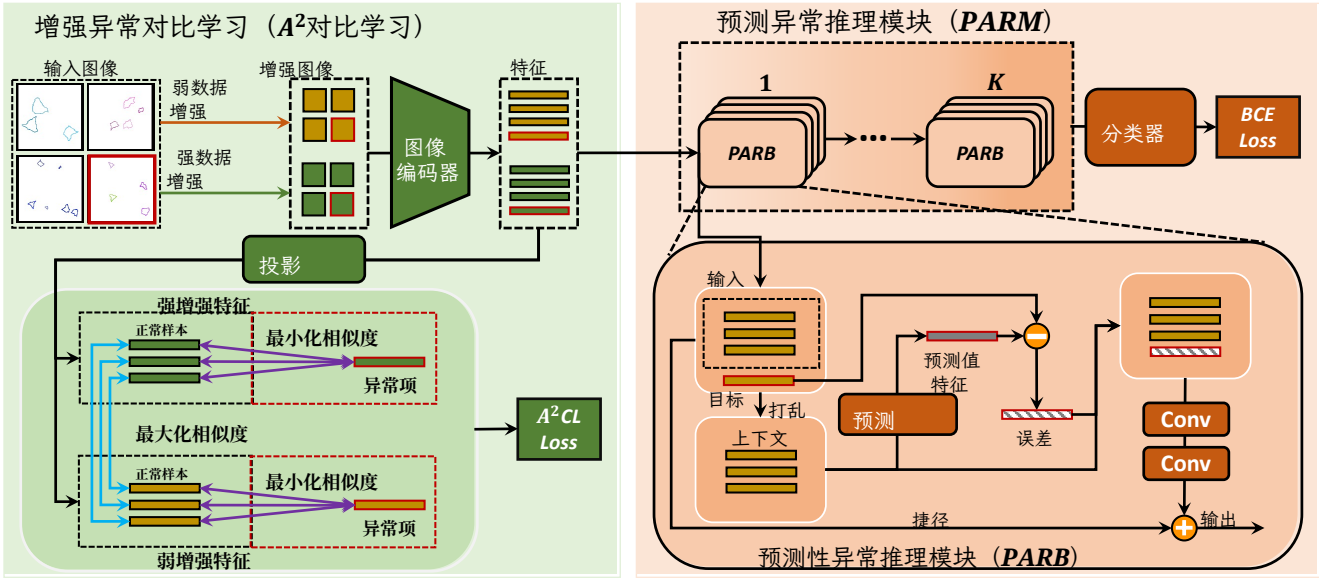


图 2. 所提出的 PR-A²CL 由两个模块组成：1）采用增强异常对比学习（A²CL）的感知模块，生成弱增强视图和强增强视图，并通过 ResNet-50 对其进行编码，从而提升特征的可判别性和泛化能力。2）由堆叠的 PARB 构成的预测异常推理模块（PARM）。每个模块块都采用预测—验证机制，根据上下文特征预测目标，并将预测结果与目标进行验证，以推断潜在组合规则。

视图或数据源之间的表征或预测，提升模型鲁棒性。在这项工作中，我们研究将对比学习用于视觉特征提取，使模型能够获取更高层次的抽象特征，从而具备更强的泛化能力并提升推理能力。

III. 所提出的 PR-A²CL

A. 所提出的 *PR-A²CL* 概述

如图 1 所示，CVR任务 [34]旨在从四张图像 $\{\boldsymbol{X}_i \in \mathbb{R}^{H\times W\times C}\}_{i=1}^{4}$中识别出异常项。三张图像是根据一组属性上的同一组合规则生成的，而异常项则是根据同一组属性上的一个略有不同的规则生成的。 H、 W和 C 分别表示图像的高度、宽度和通道数。形式化地，

$$\hat{\boldsymbol{y}} = \mathcal{F}_\Theta(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4; \Theta), \tag{1}$$

其中，$\hat{\boldsymbol{y}} \in \mathbb{R}^4$ 表示预测分数， Θ 表示模型参数， $\mathcal{F}_\Theta$ 是映射函数。

为应对组合视觉推理中的核心挑战，我们提出PR-A² CL，它由两个互补模块组成，如图2所示。首先，为解决测试时对未见过的规则组合及外观变化的泛化挑战，我们引入了带有增强异常对比学习（A²CL）的感知模块。该方法通过在各数据增强之间对齐表征，同时分离违反规则的离群样本，提取鲁棒且与规则一致的视觉特征。我们采用两种数据增强策略来提升模型的泛化能力：弱数据增强（WDA）采用旋转和色调调整等常见图像变换来丰富样本并扩展特征空间，而强数据增强（SDA）则通过对弱增强图像施加掩蔽来进一步提升鲁棒性，且不会从根本上改变其固有属性或语义规则。继 [34],我们采用ResNet-50 [1]作为图像编码器。A²CL模块旨在最小化离群样本与正常图像之间的相似性，同时最大化各增强视图之间的一致性，从而获得在多样场景中具有强泛化能力的高度判别特征。

其次，为了应对组合规则推理方面的挑战，预测异常推理模块（PARM）采用了堆叠式预测与验证架构。具体而言，PAV范式将从四张图像中选出异常项的任务转化为四个独立的预测与验证过程。每个过程利用三张图像预测目标图像，将预测结果与真实标签进行比较，并通过最小化预测误差来更新网络参数，从而隐式学习潜在的组合规则。与仅预测并验证分类标签 [29], [40], 的以往模型不同，该PAV范式被嵌入专门的推理块中，以实现更深入的规则抽象。通过一系列PARB模拟迭代式的人类推理过程，即，较早的层捕捉基本关系，并通过后续PARB逐步将其组合成更高层次的抽象结构。这种层次化细化使模型能够逐步编码组合规则，从而有助于理解日益复杂的规则结构。

虽然 PR-A²CL 专为 CVR任务 [34], 设计，但其模块化架构使其适用范围更广。如第 IV-B 节中对 SVRT数据集 [36] 的实验，以及第 IV-D 节中对 MC²R 数据集 [37] 的实验所示，该模型能够有效泛化到多上下文组合推理任务。

B. 使用 *A²CL* 进行视觉感知

人类认知可以类比为一个聚类过程：即使无法明确识别某个实体，人类也能

将其与具有相似特征的其他实体联系起来。将对比学习融入视觉推理框架，能够通过强调与底层规则相关的属性来增强特征表示，从而提升判别能力和泛化能力。

1）弱/强数据增强：弱数据增强包括随机旋转（90°、180°和 270°）、随机色调调整，以及随机水平或垂直平移 [53]。以概率 p_w应用时，它通过混合弱增强图像与原始数据来增加样本多样性。

强数据增强通过将图像划分为大小均匀的块，并对每个块内的局部区域进行随机掩蔽来实现。SDA促使模型从信息稀疏的输入中学习，从而促进更高层次的特征抽象，并增强对图像变化的鲁棒性。为保持底层语义结构的一致性，会对组合推理样本中的四张图像应用相同的数据增强。

2）增强异常对比学习：其目标是学习在各种场景下都具有强泛化能力的判别特征。形式上，令 $[\boldsymbol{X}_1^w, \boldsymbol{X}_2^w, \boldsymbol{X}_3^w, \boldsymbol{O}^w]$ 表示弱增强图像，$[\boldsymbol{X}_1^s, \boldsymbol{X}_2^s, \boldsymbol{X}_3^s, \boldsymbol{O}^s]$ 表示与之对应的强增强图像，其中 $\boldsymbol{O}^w$ 和 $\boldsymbol{O}^s$ 表示离群样本，$\boldsymbol{X}_i^w$ 和 $\boldsymbol{X}_i^s$表示增强后的正常样本。按照 [46],, 增强图像由图像编码器处理以提取特征表示，随后通过投影层映射到适用于对比学习的嵌入空间，

$$[\boldsymbol{H}_1^w, \boldsymbol{H}_2^w, \boldsymbol{H}_3^w, \boldsymbol{U}^w] = \mathcal{G}\left(\mathcal{F}([\boldsymbol{X}_1^w, \boldsymbol{X}_2^w, \boldsymbol{X}_3^w, \boldsymbol{O}^w])\right), \quad (2)$$

$$[\boldsymbol{H}_1^s, \boldsymbol{H}_2^s, \boldsymbol{H}_3^s, \boldsymbol{U}^s] = \mathcal{G}\left(\mathcal{F}([\boldsymbol{X}_1^s, \boldsymbol{X}_2^s, \boldsymbol{X}_3^s, \boldsymbol{O}^s])\right), \quad (3)$$

其中，$\boldsymbol{H}_i^w$、$\boldsymbol{H}_i^s$、$\boldsymbol{U}^w$和 $\boldsymbol{U}^s$表示投影后的特征，$\mathcal{F}$ 表示图像编码器，$\mathcal{G}$ 表示投影层。对比学习通常受益于生成多样化的增强视图，并最大化这些视图所得到的表征之间的相似度 [54]。在我们的框架中，WDA 生成多样化的数据视图，而 SDA 增强数据鲁棒性。模型采用对比学习策略，使弱增强和强增强正常样本的特征对齐，同时排斥异常样本的特征，从而形成一个特征空间：符合相同组合规则的样本聚集成紧密的簇，以确保语义一致性；而违反规则一致性的异常则被有效地推离。因此，我们按如下方式最大化弱增强和强增强正常样本之间的相似度，

$$\alpha_i = \mathcal{F}_S(\boldsymbol{H}_i^w, \boldsymbol{H}_i^s) = \frac{\boldsymbol{H}_i^{wT}\boldsymbol{H}_i^s}{|\boldsymbol{H}_i^w||\boldsymbol{H}_i^s|}, \quad (4)$$

其中，$\mathcal{F}_S$ 是余弦相似度。此处不考虑离群样本之间的相似度，因为离群样本的性质各不相同。通过最大化 α_i，促使弱增强视图和强增强视图所得到的特征表示保持较高的相互相似度。

为增强判别能力，我们最小化正常样本与异常项样本之间的相似度，如下所示：

$$\beta_i^w = \mathcal{F}_S(\boldsymbol{H}_i^w, \boldsymbol{U}^w), \quad \beta_i^s = \mathcal{F}_S(\boldsymbol{H}_i^s, \boldsymbol{U}^s). \quad (5)$$

综合所有相似度，A^{2}CL 损失定义如下：

$$\mathcal{L}_C = -\frac{1}{N_I}\sum_{i=1}^{N_I}\log\frac{\exp(\alpha_i)}{\exp(\beta_i^w)+\exp(\beta_i^s)}, \quad (6)$$

其中 N_I 为正常图像的数量。通过最小化 $\mathcal{L}_C$，所学表征促进类内紧凑性并增强类间分离。这显著提升了模型区分正常样本与异常样本的判别能力，同时也增强了特征在不同视图间的泛化能力。

A^{2}对比学习通过最大化弱增强样本与强增强样本之间的相似度，同时最小化正常样本与异常项之间的相似度以分离异常，从而在各增强视图间强制实现高层结构一致性。然而，它并不显式建模组合规则结构，而是侧重于提取具有判别性且稳健的特征，以支持后续推理。基础关系规则向高阶结构的整合与组合由PARB通过迭代的预测与验证操作实现，详见下一小节。

C. 预测异常推理模块

认知科学领域的既有研究表明，人类推理依赖于假设生成、检验与修正的迭代循环 [22], [55],例如，Lake 等人将概念学习刻画为通过反馈不断精炼抽象假设的迭代过程 [55]。Zador等人进一步主张，下一代人工智能系统应整合类似的迭代与校正机制，以模拟类人推理能力 [22]。受认知科学原理启发，我们的模型采用迭代式预测与验证架构，以模拟人类对复杂关系结构的推理。该设计通过堆叠的PARB实现，其中每一层通过最小化预测误差逐步精炼关系规则。这种逐步精炼不仅提升了规则组合的准确率，也更贴近驱动人类推理的反馈机制。

1) 预测与验证范式：从四张图像中选出一个异常项构成一项微型聚类任务：三张正常图像在共享的组合规则下形成一个连贯的聚类，而剩余的一张则为异常项。然而，这一表述带来两大挑战：1）仅有三个正常样本和一个异常项，实例数量有限，难以可靠地估计聚类结构；2）尽管遵循相同的组合规则，三张正常图像仍可能表现出相当大的视觉多样性。因此，仅依赖特征级相似性不足以捕捉潜在的规则一致性并确保语义对齐。为应对这些挑战，我们将微型聚类任务重构到预测与验证范式中。令 $[\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4]$ 表示四张图像的特征。我们构建四个 PAV 实例，每个实例以一个特征 $\boldsymbol{F}_t$ 为目标、其余特征为上下文来对其进行预测，

$$\hat{\boldsymbol{F}}_t = \mathcal{F}_P(\{\boldsymbol{F}_i\}_{i\neq t}), \quad (7)$$

其中 $\mathcal{F}_P$ 为预测函数。随后，我们将ˆ预测 $\boldsymbol{F}_t$ 与目标进行验证，如下，

$$\tilde{\boldsymbol{F}}_t = \boldsymbol{F}_t - \hat{\boldsymbol{F}}_t, \tag{8}$$

˜其中 $\boldsymbol{F}_t$ 为预测误差。根据这样一种直觉识别异常项：特征偏离正常图像所共有的组合规则，因此无法根据其余三张图像准确预测。相比之下，抑制异常项的影响后，正常图像可以根据其余正常图像可靠地预测出来。预测网络通过根据上下文特征预测目标特征，习得内在的关系模式。通过迭代优化，模型不断减小预测误差并持续优化其参数。因此，对于符合一致组合规则的正常图像，模型能够生成准确的预测；而对于违反这些模式的异常样本，则会产生较大的误差。因此，预测误差与语义异常标签高度相关。潜在的组合规则在迭代优化过程中得到隐式学习和强化。值得注意的是，我们的方法并非根据其余图像进行显式的像素级重建，而是在结构化特征空间中推断语义关系并确保组合一致性。

除了式（7）中所述的预测（该预测使用另外三幅图像作为上下文）外，我们还引入了一种替代预测方案，仅依赖其余图像中的两幅。这种设计通过排除可能不一致或造成干扰的来源来减少上下文噪声，从而促进更专注、更稳健的关系推理。此外，它还增加了训练期间上下文组合的多样性。我们将这两种预测策略整合到模型中，以提升整体性能。

2）预测异常推理模块：不失一般性，我们展示一个PARB，它根据其他三个特征预测目标 $\boldsymbol{F}_4$ 。为确保顺序不变性，首先对 $\boldsymbol{F}_1$、 $\boldsymbol{F}_2$ 和 $\boldsymbol{F}_3$ 进行打乱。随后，推理网络 [20] 根据打乱后的上下文预测 $\boldsymbol{F}_4$ ：$\hat{}\boldsymbol{F}_4 = \mathcal{F}_P(\mathcal{F}_\pi(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3))$，其中 $\mathcal{F}_\pi$ 表示置换函数。通过最小化 $\tilde{}$ $\hat{}$ 预测误差 $\boldsymbol{F}_4 = \boldsymbol{F}_4 - \boldsymbol{F}_4$来优化网络 $\mathcal{F}_P$ ，使其能够捕捉正常图像之间的潜在关系，并在 $\boldsymbol{F}_4$ 不是异常项时部分学习组合规则。随后，我们将 $\boldsymbol{F}_4$ 与上下文特征拼接，形成$\tilde{\boldsymbol{Y}}_4^0 = [\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \tilde{\boldsymbol{F}}_4]$，再通过两个卷积层 $\mathcal{F}_C$处理，以建模特征间交互，$\hat{}\,\boldsymbol{Y}_4^0 = \mathcal{F}_C(\tilde{\boldsymbol{Y}}_4^0)$。此外，还引入残差捷径 [1]，以利用原始特征 $\boldsymbol{Y}^0 = [\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4]$,

$$\boldsymbol{Y}_4^1 = \hat{\boldsymbol{Y}}_4^0 + \boldsymbol{Y}_4^0. \tag{9}$$

另外三个 PARB 将分别选择 $\boldsymbol{F}_1$、 $\boldsymbol{F}_2$ 和 $\boldsymbol{F}_3$ 作为目标，以获得 $\boldsymbol{Y}_1^1$、 $\boldsymbol{Y}_2^1$ 和 $\boldsymbol{Y}_3^1$。

3）层次化 *PARB*：为对复杂的组合规则进行建模，PARB 按层次堆叠 K 次，每个 PARB 将前一个 PARB 的输出作为输入，

$$\boldsymbol{Y}_i^j = \mathcal{F}_{\text{PARB}}^j(\boldsymbol{Y}_i^{j-1}), \quad i = 1, 2, 3, 4, \tag{10}$$

算法 1PR-A^{2}CL 的训练流程

输入： 训练样本

输出： PR-A^{2}CL 模型

1：初始化 PR-A^{2}CL 模型

2：**对于** $i = 1$ **到** 训练迭代 T **执行**

3: // 视觉感知

4: 生成弱增强图像 $\boldsymbol{H}_i^w$, $\boldsymbol{U}^w$ 和强增强图像 $\boldsymbol{H}_i^s$、 $\boldsymbol{U}^s$;

5: 生成特征并将其映射到适合如公式（2）和公式（3）所示进行对比学习;

6: 按照公式（4）、（5）和（6）计算 A^{2}CL 损失 $\mathcal{L}_C$ 。

7: // 预测推理

8: **对于** $k = 1$ **到** 推理迭代次数 K **执行**

9: **对于** $t = 1$ **到** 图像数量 N **执行**

10: 按照公式（7）使用上下文特征预测 $\boldsymbol{F}_t$ ；

11: 计算预测误差如下 $\tilde{\boldsymbol{F}}_t = \boldsymbol{F}_t - \hat{\boldsymbol{F}}_t$;

12: 按照式（**??**）利用特征交互;

13: 利用残差结构融入原始特征集，如式（9）所示。

14: **结束循环**

15: **结束循环**

16: 使用公式（12）计算BCE损失 $\mathcal{L}_{\text{BCE}}$ ；

17: 计算总损失 $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_C$;

18: 通过反向传播更新 PR-A^{2}CL 模型。

19: </style}结束循环

其中 $\mathcal{F}^j$ PARB 表示第 j 个 PARB 的运算，且$\boldsymbol{Y}_i^0 = \boldsymbol{Y}^0$。这种迭代推理机制通过执行预测与验证的循环，密切模拟人类认知过程，从而逐步完善推断出的抽象规则 [20]。尽管表达能力强大，PARM 在处理涉及细微视觉变换或相互冲突的重叠规则的组合规则时仍可能遇到困难。在这种情况下，预测误差可能无法有效区分正常样本图像与异常图像，从而可能无法检测出离群样本，详见第IV-F节。

经过两个全连接层 $\mathcal{F}_{\text{FC}}$ 后，为目标 $\boldsymbol{F}_i$生成离群点预测分数，

$$\hat{y}_i = \mathcal{F}_{\text{FC}}(\boldsymbol{Y}_i^K), \quad i = 1, 2, 3, 4. \tag{11}$$

最后，采用二元交叉熵（BCE）损失，

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[\sigma(y_i)\log\sigma(\hat{y_i}) + (1-\sigma(y_i))\log(1-\sigma(\hat{y_i}))], \tag{12}$$

其中， σ 表示 Sigmoid函数， $N = 4$ 表示 CVR面板中的图像数量， y_i 是真实标签的独热编码；即，如果第 i 张图像是异常项，则为 $y_i = 1$ ；如果是正常样本，则为 $y_i = 0$ 。PR-A^{2}对比学习的总损失定义为:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_C, \tag{13}$$

其中 λ 是 $\mathcal{L}_C$的加权因子。我们的模型以端到端方式进行训练，网络通过同时使用对比损失 $\mathcal{L}_C$ 和BCE损失 $\mathcal{L}_{\text{BCE}}$ 进行更新。算法1总结了PR-A^{2}CL的训练流程。

IV. 实验结果

A. 实验设置

三个组合数据集用于评估

.SVRT [36]: 一个用于分析基本组合规则的经典基准测试。它包含 23 项任务，每项任务均由两条不同的规则定义。每项任务都要求通过二元分类，从视觉示例中识别潜在规则。这些规则由七种基本属性构成：形状、位置、大小、旋转、翻转、内部和接触。然而，原始任务形式依赖严格的规则匹配，这简化了问题，并限制了对推理能力的评估。为解决这一问题，我们参照 CVR 数据集 [34]，将任务重新定义为四选一问题，大幅提高了难度，并确保与当代基准测试保持一致。

CVR [34]: 一个基于组合抽象规则的合成基准测试，由九种基本属性构成。其中包括六种对象属性：形状、位置、大小、颜色、旋转和翻转；以及三种关系属性：数量、内部和接触。这些属性构成九条基本规则，并组合形成组合规则。该数据集包含 103 条独特规则：20 条基于单条基本规则，65 条由两条基本规则组合而成，另有 9 条包含两条以上规则。

MC^{2}R [37]: 一个具有挑战性的抽象视觉推理基准测试。该基准旨在评估模型在复杂组合规则中识别细微差异的能力，而非依赖表面的外观匹配。每个样本由五幅图像组成：三幅符合相同规则，另外两幅离群样本仅有细微偏差。规则由九种属性类型构成：六种对象级属性（形状、位置、大小、颜色、旋转、翻转）和三种关系属性（数量、内部、接触）。这些属性通过逻辑（与、或、非）、关系（=、>、<）和算术（+、-）运算符，以树状结构进行层次化组合。该数据集包含 103 个不同任务。三个数据集 SVRT [36], 、CVR [34], 和 MC^{2}R [37],均包含 10,000 个训练样本、500 个验证样本和每个任务 1,000 个测试样本。与 CVR 相比，MC^{2}R 提供了更高的属性多样性和多答案格式，大幅增加了难度并减少了捷径学习。

我们比较以下最先进模型。

WReN [40] 利用关系网络对特征间关系进行建模，以解决瑞文渐进矩阵问题。本文将其改造用于解决 CVR 问题。

SCL [29] 由用于对象提取的对象网络、用于属性编码的属性网络，以及用于推断潜在关系的关系网络组成。在 **SCL-ResNet-18** [29],中，原始编码器被 ResNet-18 取代。

PredRNet [20], 是一种最先进的瑞文渐进矩阵模型，现已改造用于 CVR。它采用由四个 ResBlock 组成的编码器进行视觉特征提取，并使用预测推理网络对上下文与答案关系进行建模。

SCAR [41]通过其结构感知动态层实现动态适应性，该层根据视觉任务结构调整计算过程和权重。它代表了用于多样化抽象推理任务的最先进统一模型。

表 I 在与 [34]相同的设置下，SVRT 上的最先进方法比较[36] 。我们的方法在所有训练样本规模下，始终显著优于所有基线方法。

模型	20	50	100	200	500	1000	曲线下面积	10000
ViT-Small [2]	24.7	24.8	24.9	26.6	27.6	29.8	26.4	40.0
IN-ViT-Small [2]	25.6	25.1	26.2	26.1	28.7	33.0	27.5	92.4
SCAR [41]	25.8	25.9	27.0	28.3	31.0	32.9	28.5	50.3
WReN [40]	26.0	28.2	27.9	29.7	33.0	39.7	30.8	70.3
SCL [29]	25.8	25.1	25.9	26.6	31.9	55.0	31.7	73.4
ResNet-50 [1]	29.8	31.5	32.4	33.4	41.6	54.0	37.1	96.0
SCL-ResNet-18 [29]	28.2	31.8	35.8	41.6	59.5	68.2	38.2	95.8
PredRNet [20]	31.2	32.2	33.6	37.3	44.2	51.3	38.3	96.2
SSL-ViT-Small [47]	41.7	43.2	49.4	55.5	62.7	67.1	53.3	85.3
IN-ResNet-50 [1]	36.6	40.6	40.4	49.8	72.1	83.0	53.8	93.0
CLIP-ResNet-50 [57]	34.2	35.5	44.8	49.3	81.3	89.3	55.7	97.6
CLIP-ViT-Base [57]	31.0	39.2	42.1	55.0	88.1	92.1	57.9	98.6
SSL-ResNet-50 [47]	57.5	62.2	65.0	70.0	77.1	83.5	69.2	97.7
R^{3}PCL [37]	64.8	71.9	77.2	80.7	87.5	93.5	79.3	98.5
DBCR [35]	65.2	73.4	79.5	83.1	89.8	94.9	81.0	98.8
所提方法	**68.5**	**78.5**	**85.9**	**92.1**	**96.9**	**98.2**	**86.7**	**99.4**

R^{3}PCL [37] 在图像三元组之间执行残差回归推理，并融入伪标签对比学习，以增强特征判别性和泛化能力。**DBCR [35]** 采用双分支架构，利用层次回归对簇内关系进行建模，并通过对比注意力捕捉正常样本与异常图像之间的簇间差异。

ResNet-50 [1]和 **ViT-Small** [2] 已广泛用于视觉识别。按照 [34], ，我们通过添加两个全连接层作为推理器，将它们选为基准方法。此外，使用 ImageNet 数据集预训练的**IN-ResNet-50** [1]和**IN-ViT-Small** [2] [56], 、使用 MoCo-v3 [47] 并以 103 万个样本进行预训练的**SSL-ResNet-50** [47] 和 **SSL-ViT-Small** [47] ，以及采用 CLIP 的视觉编码器的 **CLIP-ResNet-50** [57] 和 **CLIP-ViT-Base** [57] [57] 也被选作对比方法。

PARB 的默认数量为 $K = 3$, $\mathcal{L}_C$ 的加权因子 λ 为 0.1。第 IV-E 节详细介绍了这两个参数的消融实验。输入图像分辨率为 128×128 像素。训练轮数最多设为 100 轮；若连续 20 轮没有提升，则提前停止。使用 Adam优化器，初始学习率为 0.0001，权重衰减为 1×10^{-4}，批量大小为 64。

B. SVRT数据集上的比较

我们在 SVRT数据集上评估了对比方法，并按照 [34]报告了每个任务使用 20、50、100、200、500、1k 和 10k 个训练样本时的结果。根据表 I，我们观察到以下几点。1) PR-A^{2}对比学习在所有实验条件下均持续优于所有基线方法，表明其能够有效提升特征判别和泛化能力，同时也验证了所提出 PARM 模块在抽象推理方面的优势。2) PR-A^{2}对比学习相较于第二优方法 DBCR [35], 取得了显著的性能提升：在每个任务的训练集规模分别为 20、50、100、200、500、1k 和 10k 个样本时，优势幅度分别达到 3.3%、5.1%、6.4%、9.0%、7.1%、3.3% 和 0.6%。这些稳定的增益表明，PR-A^{2}对比学习能够保持

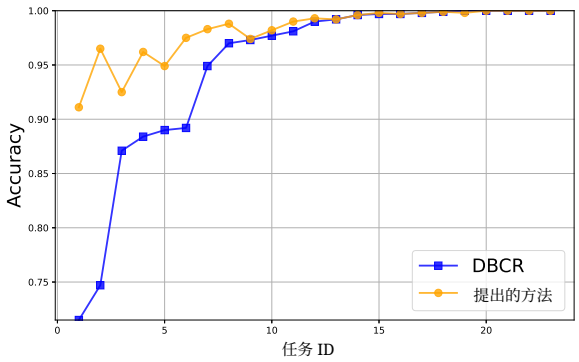


图 3. 使用每个任务 1k 个样本时与 DBCR [35] 在 SVRT [36] 上的比较。PR-A^{2}CL 在大多数任务上表现出色，而 DBCR 在若干任务上表现不佳。

相对于强竞争方法的稳健优势。3) PR-A^{2}对比学习在每个任务使用 1k 个训练样本时达到 98.2% 的准确率，使用 10k 个样本时达到 99.4%，这表明充足的数据使其能够准确建模并推理 SVRT数据集中几乎所有的组合规则。4) 即使在低数据条件下，PR-A^{2}对比学习仍能保持出色的性能：使用 100 个样本和 200 个样本时，准确率分别达到 85.9 % 和 92.1%，相较于 DBCR 分别高出 6.4% 和 9.0%。这些结果凸显了该模型的鲁棒性和泛化能力，证明了其适用于有限监督条件下的抽象推理。

为评估任务级别的性能，我们按照每个任务 1k 个训练样本的设置，对 PR-A^{2}CL 和 DBCR [35]进行逐任务准确率的详细比较，如图 3 所示。任务按 DBCR 的准确率升序排列。PR-A^{2}CL 在每项任务上的准确率均超过 90%，凸显了其在 SVRT数据集上的稳健推理能力。相比之下，DBCR 仅在部分任务上的准确率超过 90%，且在两项任务上低于 80%。我们推测，PR-A^{2}CL 得益于通过其预测—验证机制持续进行的自我精炼，因此能够稳定地处理复杂规则，而 DBCR 缺乏这种能力。

*C. CVR*数据集上的比较

遵循 [34]，的实验方案，我们在 CVR数据集 [34] 上使用每个任务 20、50、100、200、500、1k 和 10k 个样本，将 PR-A^{2}CL 与先进方法进行比较。实验采用两种配置：在所有任务上联合训练的统一模型，以及针对每个任务单独训练的模型。结果见表 II。大多数基线结果取自 [34], [35]。对于 PredR-Net [20], ，我们调整了官方实现以适用于 CVR；必要时重新运行了部分单独训练的模型。

1）联合训练一个统一模型：如表 II 所示，在联合训练设置下可以得出以下观察结果。1）在所有样本数量下，PR-A^{2}CL 的表现始终显著优于所有基线方法，证明了其在增强特征判别能力、泛化能力以及 PARM 捕获潜在组合规则的能力方面的有效性。2）与排名第二的方法 DBCR [35], 相比，PR-A^{2}CL 的性能提升随着每个任务的训练样本数从 20 增至 50、100 和 200，分别从 0.9% 增至 1.3%、2.3% 和 3.0%。这表明，随着数据增多，A^{2}CL 能更好地捕捉视觉线索并提升推理能力。3）当每个任务使用 500、1k 和 10k 个样本时，相较于 DBCR 的优势分别降至 1.1%、1.4% 和 0.7%。我们推测，充足的数据（例如，10k 个样本）使模型能够较好地逼近底层分布，从而缩小性能差距。然而，为每个任务收集 10k 个样本往往不切实际。因此，以往工作 [34] 主要使用 1k 个样本进行比较，在这种情况下，PR-A^{2}CL 超过 DBCR 1.2%。4）在训练样本有限时，许多方法的表现接近随机水平（25% 的准确率），而 SSL-ResNet-50 [34] （经 MoCo-V3 预训练）、R^{3}PCL、[37],DBCR 以及我们的方法则取得了显著更好的表现，凸显了 A^{2}CL 在泛化方面的优势。

2）为每个任务单独训练一个模型：如表 II 所示，在该设置下，PR-A^{2}对比学习在不同样本数量上均持续超越所有基线方法。值得注意的是，在每个任务仅使用 20 或 50 个样本时，其性能较次优方法 DBCR [35]，提升了 1.9%，这一改进幅度大于联合训练配置中所观察到的提升。在联合训练中，混合来自多个任务的样本会增加每个任务的有效训练数据量。例如，同时接触颜色任务和尺寸任务可以隐式地提升颜色与尺寸等组合任务上的性能。相比之下，单独训练在每个任务上使用的样本显著更少。尽管如此，我们的方法利用 A^{2}对比学习模块生成增强数据视图，从而促进更高效的推理，并在极低样本情形（20 或 50 个样本）下相较联合训练取得更大的性能增益。最后，我们注意到，在单一任务上单独训练的模型通常不如联合训练的模型，这主要是由于样本稀缺。因此，本文其余部分聚焦于联合训练的结果。

如表 II 最后两列所示，我们比较了各模型的时间复杂度。PR-A^{2}CL 仅带来适度的计算开销，同时在所有设置下都保持了较高的效率和出色的准确率。此外，PR-A^{2}CL 的参数效率较高，仅包含 27.8M 个参数，与基于 ResNet-50 的模型相当，但远少于 DBCR [35],，实现了良好的准确率与效率的平衡。

3）各组合规则的比较：图 4 比较了 PR-A^{2}CL 和 DBCR [35] 在每个任务 1k 个样本条件下的逐规则推理准确率；对于包含多个组合的规则对，结果取其平均值。主要观察结果包括：1）PR-A^{2}CL 在大多数规则组合上的表现优于 DBCR，尤其是在接触+翻转和位置+翻转等具有挑战性的组合中，这表明它在建模具有结构依赖的多属性组合方面能力更强。2）在对角线（基本规则）上，PR-A^{2}CL 在所有类别中都达到了近乎完美的准确率，而 DBCR 在形状、旋转和翻转等抽象规则上的表现则明显下降，这表明其编码高阶关系抽象的能力有限。3）最显著的提升出现在翻转和旋转等语义敏感任务中，这些任务需要理解空间结构和全局

表 II 在 CVR[34] 联合（左）和单独（右）训练设置下的最先进方法比较。PR-A^{2}CL 在所有样本数量下均优于所有基线方法，同时计算成本可接受。

模型	联合/单独训练使用不同每个任务样本数时的准确率（%）								参数（M）	十亿次浮点运算
	20	50	100	200	500	1000	AUC	10000		
SCAR [41]	26.2/27.1	26.2/27.4	26.6/27.6	27.1/28.0	27.7/28.0	29.4/32.3	27.4/28.4	42.7/41.4	0.5	0.7
ViT-Small [2]	27.3/28.6	27.8/30.1	28.0/30.9	28.1/31.9	29.9/33.8	31.4/35.1	28.7/31.7	58.7/42.5	22.6	11.2
WReN [40]	26.8/30.0	27.6/32.0	28.5/32.9	30.1/34.1	36.4/36.3	42.3/39.0	32.0/34.1	64.5/52.5	1.5	0.3
SCL [29]	25.8/26.9	25.8/30.0	28.3/30.3	34.1/30.0	43.2/31.4	46.2/33.4	33.9/30.3	56.9/53.2	0.2	0.1
IN-ViT-Small [2]	27.9/31.7	28.2/33.6	28.6/35.3	30.0/37.8	35.6/40.3	47.2/43.6	32.9/37.0	72.4/60.5	22.6	11.2
PredRNet [20]	26.9/28.9	30.0/32.6	31.5/35.3	36.2/39.8	45.9/45.3	54.5/49.0	37.5/38.5	92.2/68.7	8.4	5.8
ResNet-50 [1]	27.5/28.0	28.2/31.1	29.9/32.5	33.9/34.0	52.1/38.7	59.2/44.8	38.4/34.9	93.7/84.0	28.0	10.8
SSL-ViT-Small [47]	39.3/46.7	39.5/51.6	40.8/54.8	44.1/57.5	53.3/62.0	60.7/65.5	46.3/56.4	81.6/76.1	22.6	11.2
SCL-ResNet-18 [29]	26.4/31.4	28.4/37.3	31.6/37.8	40.7/39.6	51.4/42.7	64.0/48.3	40.4/39.5	78.9/72.8	11.5	4.8
IN-ResNet-50 [1]	32.0/33.9	35.1/36.9	39.0/42.8	43.8/48.3	57.7/55.3	69.5/62.5	46.2/46.6	85.1/82.0	28.0	10.8
CLIP-ResNet-50 [57]	28.7/34.0	32.0/37.9	40.8/43.2	46.9/57.1	59.7/57.9	74.4/60.8	47.1/48.5	82.7/75.1	28.0	10.8
CLIP-ViT-Base [57]	31.1/33.9	37.4/45.7	43.9/49.4	56.0/53.7	68.9/54.8	78.8/56.1	52.7/48.9	88.9/74.6	22.6	11.2
SSL-ResNet-50 [47]	44.3/40.5	50.3/47.3	55.3/52.9	59.5/56.8	68.9/61.9	79.2/67.7	59.6/54.5	93.1/90.8	28.0	10.8
R^{3}PCL [37]	44.8/40.9	52.1/49.2	57.6/54.1	63.0/60.1	76.4/71.1	88.9/86.2	63.8/60.3	95.7/93.9	29.6	12.7
DBCR [35]	46.5/45.3	53.0/51.8	60.0/59.1	70.2/66.9	83.4/81.2	90.4/89.5	67.3/65.6	96.4/94.9	73.1	11.3
所提方法	**47.4/47.2**	**54.3/53.7**	**62.3/61.0**	**73.2/69.7**	**84.5/82.1**	**91.8/91.0**	**68.9/67.5**	**97.1/95.9**	27.8	13.3

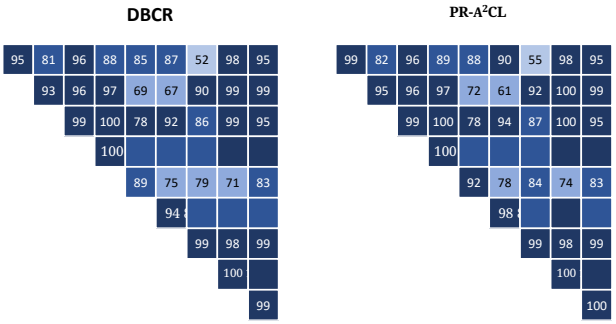


图 4. 特定组合规则的推理准确率对比。对于几乎所有规则，PR-A^{2}CL 的表现都优于 DBCR [35]。

表III 人类与 PR-A^{2}CL 在 CVR[34] 上的比较：每项任务使用20/1K个样本，在9条基础规则和36条组合规则上进行评估。

	人类	PR-A^{2}CL ，20	PR-A^{2}CL ，1k
基础	97.6	71.8	99.3
组合	74.0	39.2	89.1
颜色与位置	57.5	26.2	98.3
大小与接触	60.0	61.5	95.4
接触与旋转	30.0	65.1	94.0
颜色与翻转	45.0	27.4	88.5
计数与旋转	45.0	28.5	83.5

上下文。PR-A^{2}CL 在这些方面的优势凸显了其通过对比学习和迭代预测推理提取高层语义并对复杂视觉关系进行推理的有效性。

4) 与人类智能的比较：表 III 对比了 CVR数据集上45条抽象规则的 人类基线与 PR-A^{2}CL 的表现，采用两种训练方案：标准设置下每项任务使用1k个样本，如 [34],所示；少样本设置下每项任务仅使用20个样本，以便与人类学习者进行公平比较。虽然在少样本条件下，PR-A^{2}CL 的表现不及人类，但每项任务使用1k个样本时，其表现超过了人类。

在标准的每任务1k样本设置下，人类与PR-A^{2}CL在单属性基本规则上均表现良好，但多规则组合仍然困难。PR-A2CL将基本规则上的准确率从97.6%提升至99.3%，并将组合任务上的准确率从74.0%提升至89.1%。它在位置、翻转和旋转等与其他属性相结合的空间语义整合上取得了显著提升。尽管人类在示例有限（每次试验仅四张图像）时表现吃力，PR-A^{2}CL却凭借其预测—验证机制表现出色。

在少样本条件下，该模型在组合规则上的准确率急剧下降至39.2%，显著低于74.0%的人类基线。差距在颜色与位置、颜色与翻转以及计数与旋转等复杂组合中最大，这些任务需要将视觉属性解耦并整合为抽象结构。在监督极少的情况下，PR-A^{2}CL无法学习到稳健的表征，尤其是对于细微或罕见的属性，这凸显了少样本抽象推理的固有挑战。

D. 在 *MC^{2}R* 数据集上的比较

我们将表 IV 中的主要观察结果总结如下：1）在 MC^{2}R 数据集的所有实验设置下，所提方法始终取得最高性能，展现了其处理复杂多上下文组合推理任务的强大能力。2）与排名第二的方法 DBCR [35], 相比，我们的方法在所有数据条件下都持续取得提升。值得注意的是，在每个任务仅有 20、50 和 100 个训练样本的低数据场景中，所提模型的准确率分别达到 22.7%、32.1% 和 39.8%，分别领先 DBCR 1.3、2.0 和 1.6 个百分点。考虑到 MC^{2}R 数据集本身具有更高的组合复杂度和规则多样性，这些提升尤为显著。3）即使在规模更大的

表 IV 不同训练样本规模下 MC^{2}R [37] 的最先进方法比较。

模型	20	50	100	200	500	1,000	10,000
ViT-Small [2]	10.1	10.1	10.4	10.5	11.1	11.7	14.1
SCL [29]	10.1	10.0	10.1	10.2	10.5	12.5	30.4
MRNet [58]	10.5	10.7	10.9	11.2	11.7	13.5	31.6
WReN [40]	10.6	10.5	11.0	11.0	11.6	12.6	37.6
SSL-ViT-small [47]	15.9	16.6	17.1	17.6	18.7	21.7	38.9
SCAR [41]	10.2	10.1	10.2	10.8	11.5	15.9	39.0
PredRNet [20]	11.4	12.6	13.1	15.6	23.3	41.2	59.3
ResNet-50 [1]	10.6	10.5	10.7	13.1	17.1	28.1	75.3
SSL-ResNet-50 [47]	19.4	22.7	24.9	27.4	37.7	44.9	82.3
R^{3}PCL [37]	19.8	27.8	35.1	45.5	59.6	72.9	87.9
DBCR [35]	21.4	30.1	38.2	48.7	60.8	76.0	89.3
所提方法	**22.7**	**32.1**	**39.8**	**50.3**	**63.1**	**77.4**	**90.4**

表 V CVR [34] 上主要组件的消融研究。

A^{2}CL	PARM	200	500	1000
✘	✘	59.5	68.9	79.2
✔	✘	67.5	79.2	88.9
✘	✔	66.5	77.8	85.4
✔	✔	**73.2**	**84.5**	**91.8**

表 VI 数据增强策略消融。

	200	500	1000
无对比学习	66.5	77.8	85.4
仅使用WDA	70.8	82.2	89.1
仅使用SDA	67.9	79.4	87.2
A^{2}CL	**73.2**	**84.5**	**91.8**

包含 1k 和 10k 个样本的训练集上，所提方法仍保持很强的竞争力，准确率分别达到 77.4% 和 90.4%。这些结果表明，该模型不仅能在监督有限的情况下有效泛化，也能随着数据量的增加而稳健扩展。

*E. CVR*数据集上的消融研究

1）主要组件消融：为评估所提出的两个模块对性能的贡献，我们在三种数据条件下进行了消融研究：每个任务分别使用 200、500 和 1k 个训练样本。我们采用 **SSL-ResNet-50** [34] 作为基线，分别用所提出的 A^{2}CL 和 PARM 替换其原有的感知模块和推理模块。结果如表 V 所示。仅集成 A^{2}CL，便使三种数据规模下的准确率分别提升了 6.6%、9.2% 和 8.7%，证实了该模块在学习具有判别性和泛化能力的视觉表征方面的有效性。仅使用 PARM，则分别提升了 4.9%、7.1% 和 5.0%，表明预测与验证推理范式能更有效地处理复杂的组合规则。两个模块结合使用后，性能进一步提升，准确率分别达到 70.4%、83.9% 和 91.5%。这些结果验证了各模块对整体框架的独立贡献及互补作用。

表VII 关键参数消融研究 K 和 λ。

K	1	2	3	4
准确率（%）	91.2	91.4	**91.8**	91.1
λ	0.02	0.05	0.10	0.20
准确率（%）	91.5	91.4	**91.8**	91.4

2）数据增强策略消融：为验证 A^{2}CL 的有效性，我们进行了消融研究，对比以下配置：无对比学习（NoCL）、仅 WDA、仅 SDA，以及所提出的 A^{2}CL。如表VI所示，在所有样本数量下，A^{2}CL 均取得最高的推理准确率：每类使用 200、500 和 1k 个训练样本时，准确率分别为 73.2%、84.5% 和 91.8%。相比之下，不使用对比学习会导致性能显著下降，准确率仅为 66.5%、77.8% 和 85.4%。仅采用弱或强数据增强虽然优于无对比学习的基线，但效果仍不及 A^{2}CL。这些发现表明，对强弱增强视图进行对比，能够增强模型在特征空间中捕捉规则一致且扰动不变的语义表征的能力，从而证实了 A^{2}CL 的有效性。

*3) PARB*数量的影响：我们使用每个任务 1k 个训练样本，研究推理模块中 PARB 数量（K）的影响，结果汇总于表VII。结果表明，将 K 从 1 增加到 3，可使模型性能从 91.2% 提升至 91.8%，这说明采用多个 PARB 进行层次化推理有助于捕捉复杂的组合规则。然而，将K 增加到 4 会导致性能略微下降至 91.1%，这可能是由于过多堆叠 PARB 导致过拟合。当模型适应了训练数据中的噪声和特异性，而非学习具有泛化能力的推理模式时，就会出现这种情况，导致模型关注虚假细节，而不是稳健抽象。因此，我们默认设置 $K = 3$ 。

4）加权因子的影响 λ： PR-A^{2}CL采用两个损失函数：A^{2}CL损失 $\mathcal{L}C$，用于增强视觉特征的判别能力和泛化能力；以及BCE损失 $\mathcal{L}_{\text{BCE}}$。表VII展示了针对加权因子 λ的消融研究。鉴于 $\mathcal{L}_{\text{BCE}}$ 直接提供来自标签的监督信号，而 $\mathcal{L}_C$ 作为特征增强的辅助损失，我们在（0。02，0。20）范围内评估λ 。如表VII所示，模型在此区间内的性能保持稳定，并在 $\lambda = 0.1$处取得最高准确率91.8%。因此，我们将该值设为默认值。

5) PARB 的 *t-SNE* 可视化：为展示迭代优化过程中特征组织的演变，我们给出了 PARB1、PARB2 和 PARB3 输出的 t-SNE 图。为确保视觉清晰度并减轻重叠，我们从总计 103 条组合规则中随机选取了 20 条。如图5所示，随着 PARB 的逐步迭代，特征表示形成了越来越紧凑且彼此分明的簇。这种渐进式分离表明，模型不断优化特征空间内的语义组织

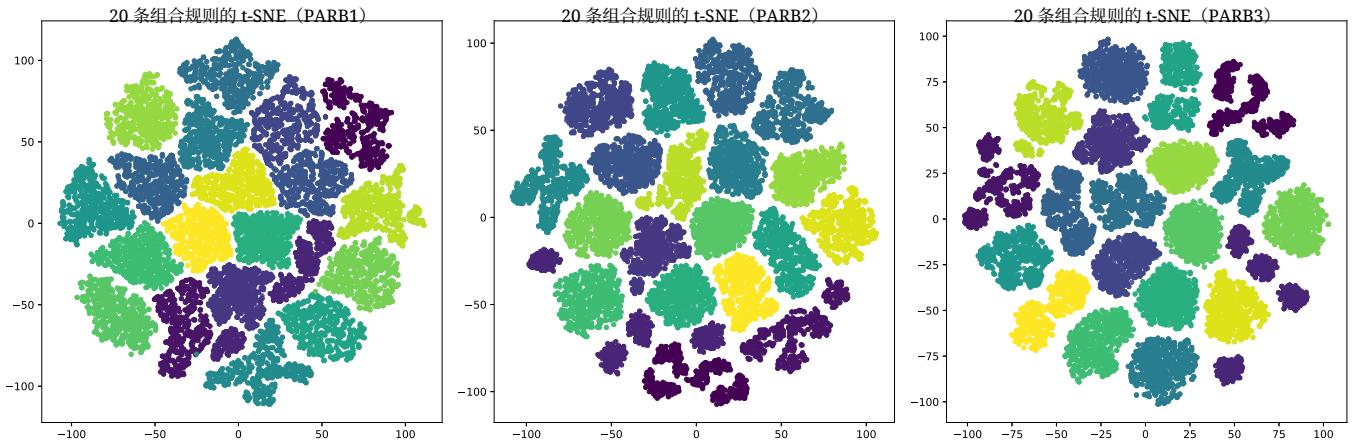


图5. t-SNE可视化　　连续的PARB中的特征表示，显示随着PARB加深，聚类紧凑度有所提高。

，从而形成更忠实地表示底层组合规则的几何配置。

*F. CVR*数据集上的失败分析

图6展示了两个失败案例。在旋转与计数任务中，DBCR [35] 未能学习组合规则，而PR-A2CL则成功完成了任务。该任务需要将物体分组成子集，并识别随机旋转物体的类别和数量，因此需要高层次的语义推理。DBCR难以应对如此复杂的组合性，而PR-A^{2}CL在抽象视觉概念上的推理能力更强。位置与翻转任务尤其具有挑战性，因为DBCR和PR-A^{2}CL都未能完成任务。这主要是由于对翻转规则的误解：旋转被视为强烈的干扰噪声，而不是组合规则的组成部分。旋转变化严重干扰了两个模型，妨碍其准确进行基于翻转的推理。值得注意的是，人类在这些任务上也面临困难，这凸显了它们本身的难度。

我们的进一步研究发现了两种主要失败模式：1）特征注意力偏差。在复杂任务场景（例如，位置与翻转）中，模型倾向于优先关注物体形状或数量等显著视觉特征，而忽略翻转方向或空间关系等更为细微但关键的属性。这种注意力失衡会导致系统性推理错误。2）规则纠缠与过拟合。模型未能解耦相关规则（例如，位置与翻转），从而过拟合于表层训练模式，而非学习抽象且可泛化的规则，尤其是在非加性交互条件下。

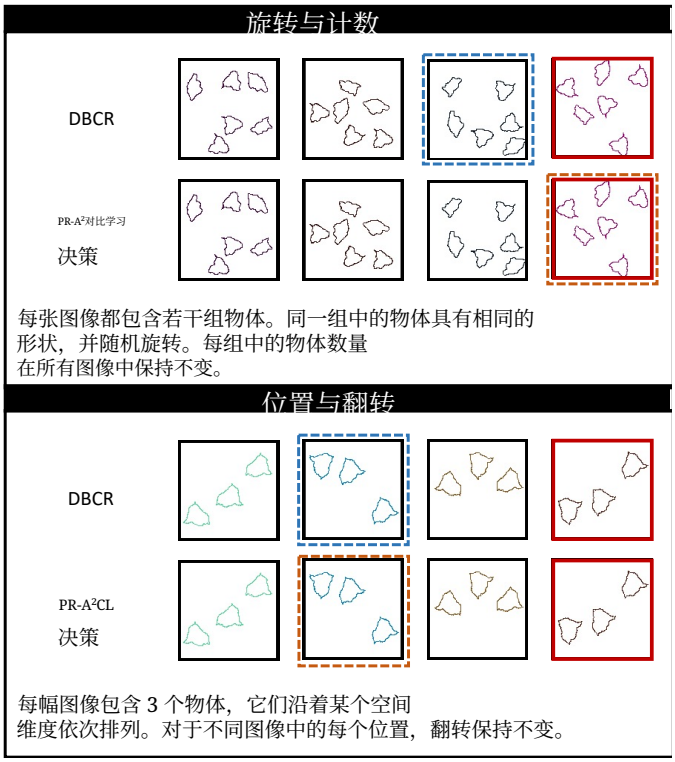


图6. CVR数据集上的两个失败案例：1）旋转与计数：DBCR [35] 未能捕捉特定的组合规则。相比之下，所提出的PR-A^{2}CL成功识别了异常项。2）位置与翻转：DBCR [35] 与PR-A^{2}CL均未能推断出底层规则，主要原因是旋转操作引入的随机噪声造成了干扰。

V. 结论与未来工作

尽管针对简单抽象规则的推理已得到广泛探索，组合式视觉推理仍是一个研究不足但至关重要的领域。本文提出预测推理增强异常对比学习（PR-A^{2}CL），一种用于组合式视觉推理的新框架。我们的方法包含一个视觉感知模块

通过增强异常对比学习进行增强，该方法利用多层次的数据增强提取判别特征，同时保持正常样本在特征空间中的相似性，并将离群样本分离。此外，预测异常推理模块采用预测—验证机制，根据其余三张图像推断每张图像。为了更好地理解组合规则，我们设计了层次化预测异常推理模块，以模拟人类的推理过程。在 SVRT、CVR 和 MC^{2}R

数据集上的大量实验表明，PR-A^{2}CL 在各种设置下始终优于现有最先进模型。

未来工作中，可通过引入规则解耦或不确定性建模来扩展 PR-A^{2}CL，从而提升其在复杂任务上的表现。结构特征解耦或可产生更具可解释性且彼此独立的视觉属性表征，而不确定性感知方法（如置信度校准或反事实增强）则有望提升模型在规则模糊或相互冲突情形下的鲁棒性。

参考文献

[1] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于 CVPR，2016 年，第 770–778 页。 [2] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等，“一张图像胜过 16×16 个词：面向大规模图像识别的 Transformer”，载于 ICLR，2020 年。[3] Y. He、C. Li、H. Yu、J. Ren、Z. Wang、H. Du 和 Y. Xia，“采用相互注意力与 U-Net去噪的双分支 StarNet：同时识别关键词和说话人”，载于 ICONIP。Springer，2024 年，第 286–300 页。[4] W. Song、X. Wang、Y. Guo、S. Li、B. Xia 和 A. Hao，“CenterFormer：一种用于无约束牙菌斑分割的新型簇中心增强 Transformer”，TMM，第 26 卷，第 10 965–10 978 页，2024 年。[5] A. Umam、C.-K. Yang、J.-H. Chuang 和 Y.-Y. Lin，“通过共同注意的超点生成与聚合进行无监督点云共部件分割”，TMM，第 26 卷，第 7775–7786 页，2024 年。[6] B. Wu、J. Xie、K. Shen、Z. Kong、J. Ren、R. Bai、R. Qu 和 L. Shen，“MG-MotionLLM：用于跨多个粒度进行运动理解与生成的统一框架”，载于 CVPR，2025 年，第 27 849–27 858 页。[7] M. Małki´nski 和 J. Ma´ndziuk，“抽象视觉推理新兴研究方向综述”，信息融合，第 91 卷，第 713–736 页，2023 年。[8] R. Pan、H. Yang、C. Li 和 J. Yang，“联合细粒度内外推理：重新审视图文检索中的语义一致性”，TMM，第 26 卷，第 4912–4925 页，2024 年。[9] J. Jiang 和 N. Zheng，“MixPHM：面向低资源视觉问答的冗余感知参数高效调优”，载于 CVPR，2023 年，第 24 203–24 213 页。[10] D. Xue、S. Qian、Q. Fang 和 C. Xu，“LININ：用于可解释视觉问答的逻辑集成神经推理网络”，TMM，第 27 卷，第 16–27 页，2025 年。[11] J. Zhu、H. Wang 和 B. He，“用于视觉常识推理的多模态结构嵌入图Transformer”，TMM，第 26 卷，第 1295–1305 页，2024 年。[12] J. Zhang、X. Wang、C. Yao、J. Ren 和 X. Jiang，“采用分组注意力和 Gamma校正门控融合的视觉语言跨域特征学习，用于提取常识知识”，载于 ACM MM，2024 年，第 4650–4659 页。[13] M. Yuan、G. Jia 和 B.-K. Bao，“用于视觉常识推理的关系推理增强网络”，TMM，第 27 卷，第 2221–2231 页，2025 年。[14] F. Baradel、N. Neverova、J. Mille、G. Mori 和 C. Wolf，“CoPhy：物理动力学的反事实学习”，载于 ICLR，2020 年。[15] K. R. Allen、K. A. Smith 和 J. B. Tenenbaum，“通过模拟支持实现快速试错学习，从而灵活使用工具并进行物理推理”，PNAS，第 117 卷，第 47 期，第 29 302–29 310 页，2020 年。[16] W.-Z. Dai、Q. Xu、Y. Yu 和 Z.-H. Zhou，“通过溯因学习架起机器学习与逻辑推理之间的桥梁”，NeurIPS，第 32 卷，2019 年。[17] C. Li、Y. Y. Tan、Y. He、J. Ren、R. Bai、Y. Zhao、H. Yu 和 X. Jiang，“DARR：用于解决机器数字推理问题的双分支算术回归推理框架”，载于 AAAI，第 39 卷，第 2 期，2025 年，第 1373–1382 页。[18] Y. Benny、N. Pekar 和 L. Wolf，“尺度局部化抽象推理”，载于 CVPR，2021 年，第 12 557–12 565 页。[19] W. He、J. Zhang、J. Ren、R. Bai 和 X. Jiang，“采用基于注意力的关系推理器的层次化ConViT，用于视觉类比推理”，载于 AAAI，第 37 卷，第 1 期，2023 年，第 22–30 页。[20] L. Yang、H. You、Z. Zhen、D. Wang、X. Wan、X. Xie 和 R.-Y. Zhang，“神经预测误差使人类标准智力测验中的视觉类比推理成为可能”，载于 ICML，2023 年，第 39 572–39 583 页。

[21] C. Li、Y. He、J. Ren、R. Bai、Y. Zhao 和 X. Jiang，“Dsrf：一种用于解决瑞文渐进矩阵的动态可扩展推理框架”，载于 NeurIPS，2025。[22] A. Zador、S. Escola、B. Richards、B. ¨Olveczky、Y. Bengio、K. Boahen、M. Botvinick、D. Chklovskii、A. Churchland、C. Clopath 等，“通过神经人工智能推动下一代人工智能的发展”，《自然-通讯》，第 14 卷，第 1597 页，2023。[23] A. Kirillov 、 E. Mintun 、 N. Ravi 、 H. Mao 、 C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo 等，“分割一切”，载于 CVPR，2023，第 4015–4026 页。[24] E. Kasneci、K. Seßler、S. K¨uchemann、M. Bannert、D. Dementieva、F. Fischer、U. Gasser、G. Groh、S. G¨unnemann、E. H¨ullermeier 等，“善用 ChatGPT？大语言模型在教育领域的机遇与挑战”，《学习与个体差异》，第 103 卷，第 102274 页，2023。[25] L. Qu、S. Wu、H. Fei、L. Nie 和 T.-S. Chua，“LayoutLLM-T2I：从大语言模型中提取文本到图像生成的布局指导”，载于 ACM MM，2023，第 643–654 页。[26] H. Xiong、Y. Zhuge、J. Zhu、L. Zhang 和 H. Lu，“3UR-LLM：用于三维场景理解的端到端多模态大语言模型”，TMM，第 27 卷，第 2899–2911 页，2025。[27] G. Camposampiero、L. Houmard、B. Estermann、J. Mathys 和 R. Wat- tenhofer，“由语言驱动的抽象视觉推理”，载于 CVPR，2023，第 2642–2646 页。[28] W. He、J. Ren、R. Bai 和 X. Jiang，“基于瑞文渐进矩阵的两阶段规则归纳视觉推理及其在视频预测中的应用”，《模式识别》，第 160 卷，第 111151 页，2025。[29] Y. Wu 、 H. Dong 、 R. Grosse 和 J. Ba，“散射组合学习器：在类比推理中发现物体、属性和关系”，arXiv 预印本，arXiv:2007.04212，2020。[30] W. He、J. Ren 和 R. Bai，“通过形态学混合进行数据增强以解决瑞文渐进矩阵问题”，《视觉计算机》，第 40 卷，第 2457–2470 页，2024。[31] W. He、J. Ren、R. Bai 和 X. Jiang，“用于抽象视觉推理的分层感知与预测类比推理网络”，载于 ACM MM，2024，第 4841–4850 页。[32] C. Zhang、F. Gao、B. Jia、Y. Zhu 和 S.-C. Zhu，“Raven：一个用于关系与类比视觉推理的数据集”，载于 CVPR，2019，第 5317–5327 页。[33] S. Hu、Y. Ma、X. Liu、Y. Wei 和 S. Bai，“用于抽象视觉推理的分层规则感知网络”，载于 AAAI，第 35 卷，第 2 期，2021，第 1567–1574 页。[34] A. Zerroug、M. Vaishnav、J. Colin、S. Musslick 和 T. Serre，“组合式视觉推理基准测试”，NeurIPS，第 35 卷，第 29 776–29 788 页，2022。[35] C. Li、G. Su、J. Ren、R. Bai、Y. Zhao 和 X. Jiang，“DBCR：利用簇内和簇外关系进行组合推理”，载于 ICASSP，2025，第 1–5 页。[36] F. Fleuret、T. Li、C. Dubout、E. K. Wampler、S. Yantis 和 D. Geman，“在视觉分类测试中比较机器与人类”，PNAS，第 108 卷，第 43 期，第 17 621–17 625 页，2011。[37] C. Li、Y. He、J. Ren、R. Bai、Y. Zhao、H. Yu 和 X. Jiang，“利用伪标签对比学习进行回归残差推理，以揭示多种复杂组合关系”，载于 IJCAI，第 4 卷，2024，第 3466–3474 页。[38] M. Ricci、R. Cad`ene 和 T. Serre，“同异概念化：机器视觉视角”，《行为科学当前观点》，第 37 卷，55，2021 47页。– 。[39] M. Shanahan、K. Nikiforou、A. Creswell、C. Kaplanis、D. Barrett 和 M. Garnelo，“显式关系神经网络架构”，载于 ICML，2020，第 8593–8603 页。[40] D. Barrett、F. Hill、A. Santoro、A. Morcos 和 T. Lillicrap，“测量神经网络中的抽象推理能力”，载于 ICML，2018，第 511–520 页。[41] M. Małki´nski 和 J. Ma´ndziuk，“一个可自行配置的模型，解决多种抽象视觉推理问题”，载于 AAAI，第 38 卷，13，2024，14 297 14 305 no. 页码。– 。[42] D. Saxton、E. Grefenstette、F. Hill 和 P. Kohli，“分析神经模型的数学推理能力”，载于 ICLR，2018。[43] B. Lake 和 M. Baroni，“无需系统性的泛化：序列到序列循环网络的组合能力”，载于 ICML，2018，第 2873–2882 页。[44] T. Thrush、R. Jiang、M. Bartolo、A. Singh、A. Williams、D. Kiela 和 C. Ross，“Winoground：探测视觉语言模型的视觉语言组合性”，载于 CVPR，2022，第 5238–5248 页。[45] J. Son，“通过自循环一致性进行时空对应的对比学习”，载于 CVPR，2022，第 14 679–14 688 页。[46] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，“视觉表征对比学习的简单框架”，载于 ICML，2020，第 1597–1607 页。

[47] X. Chen、S. Xie 和 K. He，“自监督视觉 Transformer 训练的实证研究”，载于 *ICCV*，2021，第 9620–9629 页。[48] A. v. d. Oord、Y. Li 和 O. Vinyals，“基于对比预测编码的表征学习”，*arXiv* 预印本 *arXiv:1807.03748*，2018。[49] J. D. Robinson、C.-Y. Chuang、S. Sra 和 S. Jegelka，“利用困难负样本进行对比学习”，载于 *ICLR*，2020。[50] C.-Y. Chuang、R. D. Hjelm、X. Wang、V. Vineet、N. Joshi、A. Torralba、S. Jegelka 和 Y. Song，“针对噪声视图的鲁棒对比学习”，载于 *CVPR*，2022，第 16 670–16 681 页。[51] L. Nie、X. Song 和 T.-S. Chua，从多个社交网络中学习。Springer Nature，2022。[52] L. Nie、L. Qu、D. Meng、M. Zhang、Q. Tian 和 A. D. Bimbo，“面向搜索的微视频描述”，载于 *ACM MM*，2022，第 3234–3243 页。[53] M. Małki´nski 和 J. Ma´ndziuk，“用于抽象视觉推理的多标签对比学习”，*TNNLS*，第 35 卷，第 2 期，第 1941–1953 页，2024。[54] Y. Tian、C. Sun、B. Poole、D. Krishnan、C. Schmid 和 P. Isola，“什么样的视图适合对比学习？”，*NeurIPS*，第 33 卷，第 6827–6839 页，2020。[55] B. M. Lake、T. D. Ullman、J. B. Tenenbaum 和 S. J. Gershman，“构建能够像人一样学习和思考的机器”，行为与脑科学，第 40 卷，第 e253 页，2017。[56] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 李飞飞，“ImageNet：大型层次化图像数据库”，载于 *CVPR*，2009，第 248–255 页。[57] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，“通过自然语言监督学习可迁移的视觉模型”，载于 *ICML*，2021，第 8748–8763 页。[58] Y. Benny、N. Pekar 和 L. Wolf，“尺度局部化抽象推理”，载于 *CVPR*，2021，第 12 557–12 565 页。