解剖广义类别发现：自我解构下的多重共识

Luyao Tang[1,2,5*], Kunze Huang[1,2,*], Chaoqi Chen[3,†], Yuxuan Yuan[1], Chenxin Li[4]
Xiaotong Tu[1,2], Xinghao Ding[1,2] and Yue Huang[1,2]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University [2] School of Informatics, Xiamen University
[3] Shenzhen University [4] The Chinese University of Hong Kong [5] The University of Hong Kong
{ lytang, kzhuang } @stu.xmu.edu.cn, cqchen1994@gmail.com, yhuang2010@xmu.edu.cn

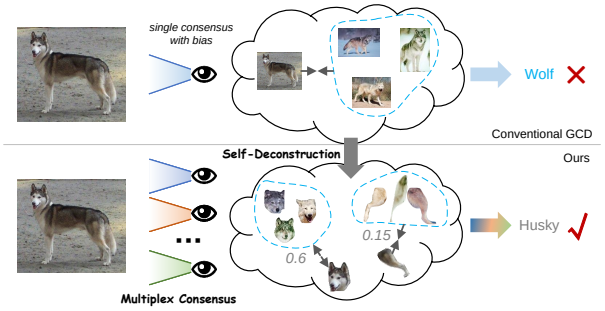


Figure 1. 广义类别发现本质上需要通过自我解构实现多重共识，拒绝图像级处理，转而支持属性感知的合成推理。

Abstract

人类的感知系统在已知和新颖类别中的推断和识别物体方面表现出色，这一能力远远超越了当前的机器学习框架。虽然广义类别发现（GCD）旨在弥合这一差距，但现有方法主要集中在优化目标函数上。我们提出了一种正交解决方案，其灵感来自于人类对新颖物体理解的认知过程：将物体分解为视觉原语，并建立跨知识比较。我们提出了 ConGCD，通过高层次语义重构来建立以原语为导向的表示，通过解构绑定类内共享属性。反映人类在视觉处理中偏好的多样性，个体可能利用主导或上下文线索，我们实现了主导和上下文一致性单元，以分别捕捉类区分模式和固有分布不变性。一个一致性调度器动态优化激活路径，最终预测通过多重一致性集成而产生。广泛的粗粒度和细粒度基准评估证明了 ConGCD 作为一种一致性感知范式的有效性。

1. 引言

将归纳和识别能力推广到未知类别的物体是人类视觉与生俱来的能力 [?]，但对于机器学习模型来说却是极具挑战性的。广义类别发现（GCD） [? ? ? ? ?] 是解决此问题的一种先进方法，它鼓励模型在经过部分标记数据训练后，对数据中的样本进行聚类和识别，这些样本包含未标记和未知类别以及已知类别。

从表示学习的角度来看，当前的 GCD 范式，包括对比学习、原型分类器和样本选择，存在一个基本的忽视：它们将图像处理为原子实体，忽视了视觉识别的组合性质。例如，CMS 对全局嵌入应用均值漂移，而 SimGCD 则从整体表示中提炼类别原型。虽然这些方法在粗粒度的区分上是有效的，但它们未能模拟人类在属性层面进行推理的能力——即将对象分解为语义原语并通过比较分析建立跨类别一致性的能力。

人类在学习新概念时可以迁移先前获得的知识 [? ? ?] 。例如，如果幼儿学会根据形状来区分“牛”和“羊”，他们可以应用相同的规则来分辨“猪”和“马”。然而，当涉及到区分“熊猫”和“熊”时，仅依赖形状可能会导致混淆，如 Fig. 1 ；这时需要关注其他视觉原始特征 [?]，例如条纹和耳朵。例如，牛津词典 [?] 将长颈鹿定义为“有着非常长的脖子、长腿和皮毛上有深色斑纹的高大动物”。这个定义涉及两个形状原始特征（长脖子，长腿）和一个纹理颜色属性（深色斑纹）。不幸的是，深度网络中的捷径学习 [?] 通常激活几个原始特征（例如大象的灰色皮肤 [?] ），导致在具有类似原始特征的类别之间（如犀牛）产生混淆。

在这项工作中，我们提出了受认知科学启发的 ConGCD，它尝试通过专注于两个方面来解决上述挑战，如 Fig. 2 所示：视觉原语的自我解构和多重共识的形成。(1) 为了分离视觉原语，我们定义了一组以原语为导向的表示，将类内共享的视觉原语以一种构建方式绑定到空间区域 [?] 。原语导向表示之间的竞争关系确保每个原语掩膜都专注于单一视觉属性。总的来说，所

* 共同贡献；† 通讯作者。

有原语导向表示可以看作是图像关键视觉原语的合成。(2) 为了多重共识的形成，我们不同于单一特征聚合，实施了主导和背景共识单元：主导单元保留高响应神经元以捕获类辨别模式，而背景单元则保留较弱的激活以编码分布不变性。动态共识调度器优化这些单元之间的激活路径，最终预测通过它们互补视角的多重集成而出现。这一过程模拟了当人类在对新的范例分类时，如何迭代地调和主导线索（例如，鸟的喙形状）与背景不变性（例如，飞行姿态） [? ?] 。

我们的主要贡献可以总结如下：

- 一种通过竞争绑定引发视觉原语的机制，能够通过拓扑竞争无监督地发现原语判别属性。
- 多重共识框架：主导单元对与类别相关的变量保留高响应的神经元，而上下文单元则对内在分布保存弱激活，由动态调度进行协调。
- 大量验证显示我们的范式与 GCD 方法的即插即用兼容性，在七个基准测试中达到最先进的性能。

2. 相关工作

2.1. 广义类别发现

在广义类别发现（GCD）[? ?] 中，目标是使用来自已知类别的有限标注数据集将未标记样本分类为已知和未知类别，这反映了一种现实的图像识别方法。最初，GCD 由 [?] 提出，采用对比和半监督学习方法，在推理过程中严重依赖聚类。因此，SimGCD，一种强大的基线方法，替换聚类为分类器，通过一种在 SSL [?] 中显示出显著效果的伪标签策略进行训练。为了解决这一限制，后续工作旨在利用潜在的跨实例关系来改善特征表示 [?] 。提出了一种类 EM 的框架，以交替进行对比学习和类别数估计 [?] 。最近的工作提出使用参数化分类器以避免预测偏差 [?] 。然而，之前的研究 [? ?] 关注于聚类中的类别标记，忽视了样本间相似的视觉属性可能导致的类别边界混淆。

2.2. 以对象为中心的学习

在对象中心学习（OCL）的不断发展过程中，Slot Attention 建立了一个重要的里程碑 [?] 。这种机制促成了进一步的研究，例如 [?] ，它将 OCL 扩展到视频中，利用视频序列中的时间一致性和动态性来更好地区分对象与背景。此外，为了解决单视图 OCL 中的稳定性问题，[?] 引入了模块以引导槽远离背景噪声并朝向一致的对象表示。[?] 探索了自监督，提供了首个在真实世界视频中的无监督多对象分割。而适应复杂真实世界环境的 OCL 挑战则由 [? ?] 解决，弥合了实验室与真实世界场景之间的差距。OCL 目前集中在对象级别的分离 [?] ，我们将其扩展到更细粒度的视觉属性发现水平，并使模型本质上具备视觉属性语义的归纳能力。

3. 预备知识

3.1. 视觉原语

一个样本 $\boldsymbol{x}$ 的生成过程实际上是一个视觉原语集合 $s^1, \cdots, s^K$ （$s^{1:K}$ ）与 $s^i \in \mathcal{S}$ 的组合过程，其中 $\mathcal{S}$ 是有限集 [?] 。考虑到 $\boldsymbol{x}$ 和 $s^{1:K}$ 的联合分布，由于观察和标注上的偏差以及与真实世界分布的偏离，一个样本的标签可以依赖于一个或多个原语。数据生成过程可以表达为

$$\begin{aligned} &\boldsymbol{z} \sim p(\boldsymbol{z}), \quad s^i \sim p(s^i|\boldsymbol{z}), \quad \boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{z}), \\ &p(s^{1:K}, \boldsymbol{x}) = p(s^{1:K}) \int p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}|s^{1:K}) d\boldsymbol{z}, \end{aligned} \tag{1}$$

，其中 $\boldsymbol{z}$ 是在潜在空间中的一个因子。我们认为，如果给定原始数据的不同边缘分布存在但它们共享相同的条件生成过程，则类别之间将存在分布转移，即 $p_{\text{class-n}}(s^{1:K}) \neq p_{\text{class-n}'}(s^{1:K})$ 。然而，由于在分布之间共享生成模型，某些原始数据子集 $\tilde{\mathcal{A}}$ 在类别之间高度相似，我们有 $p_{\text{class-n}}(\tilde{s}^{1:\tilde{K}}, \boldsymbol{x}) = p_{\text{class-n}'}(\tilde{s}^{1:\tilde{K}}) \int p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}|\tilde{s}^{1:\tilde{K}}) d\boldsymbol{z}$ ，其中 $\tilde{s}^i \in \tilde{\mathcal{S}}$ 和 $\tilde{\mathcal{S}} \subseteq \mathcal{S}$ 。类别间原始数据的相似性对于区分类别是灾难性的。因此，我们旨在找到一个互补集 $\mathcal{S} \setminus \tilde{\mathcal{S}}$ ，即使类别之间的分布差异最大化的区别性视觉原始数据，作为对类别间分离贡献最大的因子。

3.2. GCD 问题定义

我们考虑每个数据集的一个标记子集，记作 $\mathcal{D}_l = \{(\boldsymbol{x}_i^l, y_i^l)\} \subset \mathcal{X} \times \mathcal{Y}_l$ 。此外，还有一个未标记的子集 $\mathcal{D}_u = \{(\boldsymbol{x}_i^u, y_i^u)\} \subset \mathcal{X} \times \mathcal{Y}_u$ 。在 $\mathcal{D}_l$ 中，只存在已知类别，即 $\mathcal{Y}_l = \mathcal{C}_{known}$ 。另一方面，$\mathcal{D}_u$ 包含已知和新颖类别，即 $\mathcal{Y}_u = \mathcal{C}_{known} \cup \mathcal{C}_{novel}$ 。模型的任务是在 $\mathcal{D}_u$ 中对已知和新颖类别进行聚类。可以提前确定新颖类别的数量，表示为 N_{novel} [? ? ?] 。函数 $f(\cdot)$ 和 $g(\cdot)$ 分别作为特征提取器和投影头。特征 $\boldsymbol{h}_i = f(\boldsymbol{x}_i)$ 和投影嵌入 $\boldsymbol{z}_i = g(\boldsymbol{h}_i)$ 都要进行 L-2 归一化。

4. 方法论

4.1. 视觉基元的自我解构

视觉原始元素的迭代发现。指的是人在收到视觉信号后，将物体的视觉特征与大脑中的现有知识进行比较并进行循环搜索的过程 [?] ，我们将视觉原始元素的发现定义为一个迭代过程。这起源于槽注意力 [?] 。我们预定义了以原始元素为中心的表示（可学习的嵌入），$\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathbb{R}^{K \times D_s}$ 。K 表示视觉原始元素的数量，D_s 表示维度数量。

对于预训练编码器的输出 $\boldsymbol{z} = g(f(\boldsymbol{x})) \in \mathbb{R}^{N \times D_z}$ ，实现了人类感知过程的注意力机制 [?] 。表示 $\mathcal{K}_\beta$ 、$\mathcal{Q}_\gamma$ 和 $\mathcal{V}_\phi$ 为 $\boldsymbol{z}$ 的键、查询和值的投影网络，参数分别为 β 、γ 和 ϕ 。dis 过程的 v 象 p 元件 vpdis(·) 和注意力函数 attn(·) 可以定义为

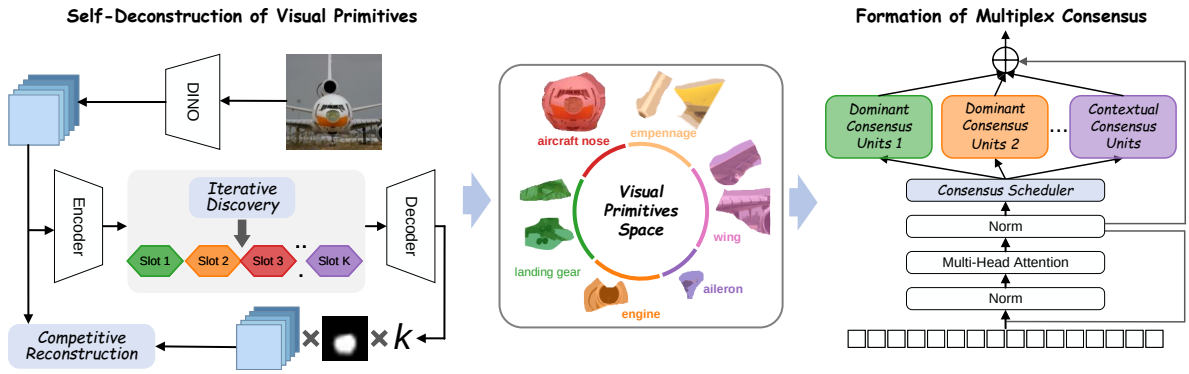


Figure 2. 所提出的 ConGCD 的概述：（1）通过竞争重建迭代地发现视觉原语；（2）每个视觉原语由带有原语遮罩的主要共识单元处理，而固有分布由具有整体表示的上下文共识单元调度器处理。每个单元的权重由共识调度器决定。

$$\begin{aligned} \text{vpdis}(\boldsymbol{A}, \boldsymbol{v}) &= \boldsymbol{A}^T \boldsymbol{v}, \quad A_{ij} = \frac{\text{attn}(\boldsymbol{q}, \boldsymbol{k})_{ij}}{\sum_{l=1}^{K} \text{attn}(\boldsymbol{q}, \boldsymbol{k})_{lj}}, \\ \text{attn}(\boldsymbol{q}, \boldsymbol{k}) &= \frac{e^{M_{ij}}}{\sum_{l=1}^{N} e^{M_{il}}}, \quad \boldsymbol{M} = \frac{\boldsymbol{k}\boldsymbol{q}^T}{\sqrt{D_s}}, \end{aligned} \tag{2}$$

其中 $\boldsymbol{A}$ 是注意矩阵。$\boldsymbol{k} = \mathcal{K}_\beta(\boldsymbol{z}) \in \mathbb{R}^{N \times D_s}$ 和 $\boldsymbol{v} = \mathcal{V}_\phi(\boldsymbol{z}) \in \mathbb{R}^{N \times D_s}$ 分别是键和值向量。为了确保视觉原语与特征嵌入之间的注意关系的绑定，查询是视觉原语 $\boldsymbol{s}$ 的一个函数，正式表达式是 $\boldsymbol{q} = \mathcal{Q}_\gamma(\boldsymbol{s}) \in \mathbb{R}^{K \times D_s}$ 。

类似于思维和推理的过程 [?]，视觉原语在 T 注意迭代过程中被迭代细化。具体来说，在迭代 t 中，查询由 $\hat{\boldsymbol{q}}^t = \mathcal{Q}_\gamma(\boldsymbol{s}^t)$ 给出。应用门控循环单元 (GRU) [? ?] ，记为 $\mathcal{G}_\theta$ 。视觉原语的更新过程可以总结为 $\boldsymbol{s}^{t+1} = \text{vpdis}\left(\text{attn}\left(\hat{\boldsymbol{q}}^t, \boldsymbol{k}\right), \boldsymbol{v}\right)$ 。

给定第 $t = T$ 次迭代中的 $\boldsymbol{x}$ ，视觉基元的后验分布为

在 $t = 0$ 次迭代初始化时，我们从高斯分布 [?] 中抽取 s ，其中 $\delta(\cdot)$ 表示 Dirac delta 分布 [?] ，以使 $\boldsymbol{s}^0$ 的分布表示相对均匀的视觉原始，防止在视觉原始发现过程中产生偏差。

通过竞争性重建进行解构。$\hat{\boldsymbol{z}} = \Phi_d(\boldsymbol{s}_k)$ 视觉原语的发现过程是无监督的，我们通过重建的代理任务来完成它。在 $\boldsymbol{s}^T$ 经过上述迭代发现后，它被作为解码器 [?] ，即 $\Phi_d(\cdot)$ ，输入为一个通用的轻量级多层感知器 (MLP)。$\Phi_d(\cdot)$ 空间广播 K 视觉原语到与 $\hat{\boldsymbol{z}}_k = \Phi_d(\boldsymbol{s}_k) \in \mathbb{R}^{N \times (D_z+1)}$ 相同的空间维度。重建的 $\hat{\boldsymbol{z}}_k$ 由 $D_z + 1$ 通道组成。经过通道划分，我们获得重建后的特征嵌入 $\hat{\boldsymbol{z}}_k \in \mathbb{R}^{N \times D_s}$ 和激活区域 $\alpha_k \in \mathbb{R}^{N \times 1}$ 。为了将视觉原语嵌入与图像中具有相同语义的区域绑定，我们在 α_k 上使用 softmax 对 K 个视觉原语进行归一化，以引入视觉原语之间的竞争关系，使得某个原语与某些像素相关联。

$$\hat{\boldsymbol{z}} = \sum_{k=1}^{K} \hat{\boldsymbol{z}}_k \odot \boldsymbol{m}_k, \quad \boldsymbol{m}_k = \underset{k}{\text{softmax}}(\alpha_k). \tag{3}$$

视觉原语的迭代发现和解耦过程的优化目标很简单

$$\mathcal{L}_{\text{rec}} = \|\hat{\boldsymbol{z}} - \boldsymbol{z}\|^2, \quad \hat{\boldsymbol{z}} = \Phi_d(\boldsymbol{s}). \tag{4}$$

这种自监督解构通过竞争绑定建立了基本的视觉原语，使机器能够模仿人类的属性级别感知。通过将对象拆解成可组合的因素，我们为模型赋予了细粒度比较的内在解释能力，同时保持对新类别结构的适应性。

4.2. 多重共识的形成

模型需要理解视觉原语。然而，单纯依赖主要特征激活 [?] 会形成自我强化的偏差，降低对未知类别的泛化能力，如 Tab. 4 中所示，强激活的神经元编码与类别无关的结构原语，而弱激活的神经元则保留跨类别的上下文模式 [?] 。

这种对偶性（如在 Fig. 3 所示）通过两条互补的途径推动我们的共识形成：（1）主导共识单元通过神经竞争层传输高激活视觉原语来提炼类别区分知识，防止原语过度承诺于已知类别。（2）上下文共识单元不断同化弱激活的嵌入以建模超越类别的关系。最终的多路架构通过共识调度器实现自我稳定，其中原语的特异性和上下文的普遍性共同进化。

主导共识单元。在网络前向过程中的第 i 个块的输出 $\boldsymbol{z}^i \in \mathbb{R}^{N \times D_z^i}$ 中，为了与视觉原语的像素级关系保持一致，我们在空间维度上选择性地保留标记级神经元激活响应的前 r （%）。具体而言，我们自定义了一个从大到小排序的函数 sort(·) ，用于在过滤神经元激活值后生成 $\boldsymbol{z}^i_{attr}$ 。此外，对于每个共识单元，其注意力应集中在单一视觉原语上，以确保单元的专业性（类似于人类的视觉偏好，一些人依靠头部来区分动物，而另一些人依靠尾巴）。我们直接使用 $\boldsymbol{m}_k$ 来表示视觉原语区域的激活，以限制第 k 个单元的输入。每个主导共识单元的前向过程是

	Method	CUB-200			FGVC-Aircraft			Stanford-Cars			Average		
		All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	GCD [?]	51.3	56.6	48.7	45.0	41.1	46.9	39.0	57.6	29.9	45.1	51.8	41.8
	GPC [?]	52.0	55.5	47.5	43.3	40.7	44.8	38.2	58.9	27.4	44.5	51.7	39.9
	XCon [?]	52.1	54.3	51.0	47.7	44.4	49.4	40.5	58.8	31.7	46.8	52.5	44.0
	PromptCAL [?]	62.9	64.4	62.1	52.2	52.2	52.3	50.2	70.1	40.6	55.1	62.2	51.7
	AMEND [?]	64.9	75.6	59.6	52.8	61.8	48.3	56.4	73.3	48.2	58.0	70.2	52.0
	μ GCD [?]	65.7	68.0	64.6	53.8	55.4	53.0	56.5	68.1	50.9	58.7	63.8	56.2
	CMS [?]	68.2	76.5	64.0	56.0	63.4	52.3	56.9	76.1	47.6	60.4	72.0	54.6
	InfoSieve [?]	69.4	77.9	65.2	56.3	63.7	52.5	55.7	74.8	46.4	60.5	72.1	54.7
	SimGCD [?]	60.3	65.6	57.7	54.2	59.1	51.8	53.8	71.9	45.0	56.1	65.5	51.5
	+ ConGCD	61.6	65.1	59.5	55.0	58.4	53.1	54.5	72.2	47.8	57.1	65.2	53.5
	LegoGCD [?]	63.8	71.9	59.8	55.0	61.5	51.7	57.3	75.7	48.4	58.7	69.7	53.3
	+ ConGCD	65.3	72.6	61.7	55.6	61.2	54.0	58.1	75.9	49.5	59.7	69.9	55.1
	SPTNet [?]	65.8	68.8	65.1	59.3	61.8	58.1	59.0	79.2	49.3	61.4	69.9	57.5
	+ ConGCD	68.1	68.5	67.8	59.7	61.3	59.2	59.1	79.0	49.8	62.3	69.6	58.9
	SelEx [?]	78.7	81.3	77.5	60.9	70.3	56.2	57.0	77.4	47.2	65.5	76.3	60.3
	+ ConGCD	81.7	80.4	82.4	62.5	70.2	58.7	57.5	77.5	47.9	67.3	76.0	63.0
	Avg. △	+2.03	-0.25	+2.83	+0.85	-0.40	+1.80	+0.53	+0.10	+1.28	+1.18	-0.18	+1.98
DINOv2	GCD [?]	71.9	71.2	72.3	55.4	47.9	59.2	65.7	67.8	64.7	64.3	62.3	65.4
	SimGCD[?]	71.5	78.1	68.3	63.9	69.9	60.9	71.5	81.9	66.6	69.0	76.6	65.3
	μ GCD [?]	74.0	75.9	73.1	66.3	68.7	65.1	76.1	91.0	68.9	72.1	78.5	69.0
	SelEx[?]	86.0	86.5	85.7	82.1	84.8	80.7	80.5	91.8	75.0	82.9	87.7	80.5
	+ ConGCD	86.3	87.4	85.8	81.7	83.3	81.0	79.8	93.1	73.3	82.6	87.9	80.1

Table 1. 细粒度语义变化基准的结果。ConGCD 提供了出色的跨方案和跨模型兼容性。

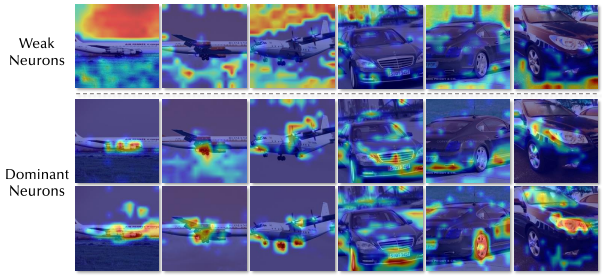


Figure 3. 主导和弱势神经元分别捕获与类别相关的变量和固有分布。

$$\boldsymbol{z}^{ik}_{Dom} := \begin{cases} \boldsymbol{z}^i \odot \boldsymbol{m}_k, \text{ if } \boldsymbol{z}^i[:,:] \geq \text{Top-R}_k(\text{sort}(\boldsymbol{z}^i[:,:])) \\ 0, \text{ otherwise }, \end{cases} \tag{5}$$

在 Top-$\text{R}_\text{k}(\cdot)$ 中 $R_k = N \times r_k$ ，对于用于捕捉类相关变量的视觉基本调度器，它们需要感知高度激活的神经元，并提供有助于类之间分离的特征。

上下文共识单元。由于数据集的固有分布，仅仅最大化提取与类别相关的视觉原件会导致模型在识别新类别时遇到困难。因此，我们提出了基于弱神经元的上下文共识单元。它接收完整的 $\boldsymbol{z}^i$ 作为输入，并感知对象的所有视觉原件。我们分解了 $\boldsymbol{z}^i$ 中的低激活区域

$$\boldsymbol{z}^i_{Con} := \begin{cases} \boldsymbol{z}^i, \text{ if } \boldsymbol{z}^i[:,:] \leq \text{Top-R}(\text{sort}(\boldsymbol{z}^i[:,:])) \\ 0, \text{ otherwise }, \end{cases} \tag{6}$$

类似于 Eq. (5) ，但在弱神经元上激活了对类别独立特征的支配共识单元，旨在从特征嵌入本身中提取每个样本的共同信息。这防止模型过度关注已知类别的视觉原始元素，并有效感知新类别。

一致调度器。自我解构过程在 ViT 块中表现为通过双重激活极化的 [?] ：虽然多头自注意力（MHA）[?] 捕捉跨补丁依赖关系，分解出的主导/情境一致路径表现出互补的神经元激活模式。我们将 FFN 重新表述为由激活分离控制的并行一致单元：

在这里，$\boldsymbol{W}$ 和 $\boldsymbol{W}^{1/2}_{\text{FFN}_k}$ 是可学习参数。$\phi(\cdot)$ 是一个非线性激活函数。原始调度器在 K 主导单元和 M 上下文单元中密集操作，并为每个单元分配非零概率。

由于单位之间的偏好不一致，在检测显性/上下文激活神经元时确定比率 r_k （%）变得具有挑战性。我们通过一个激活分离分配器（ASR）来解决这个问题，该分配器从 $\boldsymbol{z}$ 动态生成单位特定的分离比率，从而为每个样本的激活分解提供自适应边界确定。调度器采用一个新的 R^* 来使用 Top-$\text{R}^*_\text{k}(\cdot)$ 筛选神经元：

其中 W_{ASR} 是可学习的，η 控制缩放，N 表示补丁数量。这允许样本自适应地调整到多样的视觉原语，同时保持类间激活平衡。一般来说，给定 $\boldsymbol{z}$ 作为输入，更新过程为

其中 LN(·) 表示层归一化，z_{out} 是输出并被发送到下一个模块。多路复用共识框架通过共生协调解决了 GCD 中的根本矛盾，主导单元在强调类别边界的同时，语境单元保持开放世界的连续性，这反映了人类视觉对焦点属性和环境背景的双重处理方式。

5. 实验

通过全面的实验验证和分析，我们旨在回答以下问题。（1）ConGCD 能否提高 GCD 的准确性？（2）ConGCD 能否在现有的 GCD 方案中即插即用？（3）ConGCD 性能提升的来源是什么？（4）ConGCD 对超参数变化是否具有鲁棒性?

5.1. 设置

基准。ConGCD 在总共七个图像识别基准上进行了评估，即四个细粒度数据集，CUB-200-2011 [?] 、Stanford Cars [?] 、FGVC Aircraft [?] 、Herbarium19 [?] ，以及三个粗粒度数据集，CIFAR10、CIFAR100 [?] 和 ImageNet100 [?] 。在处理 CUB、Stanford Cars 和 FGVC Aircraft 时，为了将目标类别划分为已知和未知集合，我们遵循语义转移基准 (SSB) [?] 定义的划分。对于剩余的数据集，我们采用之前研究 [?] 的划分。在 CIFAR100 基准下，我们将 80 % 的类别指定为已知类别。对于其他基准，已知类别的比例为 50 %。我们的标记集，称为 $\mathcal{D}_l$ ，由所有基准中已知类别的 50 % 的图像组成。
评估协议。遵循现有研究 [?] ，我们使用平衡半监督 K-means [?] 来聚类整个图像集合 $\mathcal{D}$ 并计算模型在数据集 $\mathcal{D}_u$ 上的识别准确性，该数据集没有访问真实标签。匈牙利算法 [?] 被用来获得未标注聚类集和真实标签之间的最佳映射关系。基于类别原语，未标注集中的示例 $\mathcal{D}_u$ 被分为已知和新类别，以进行聚类准确性评估。此外，还报告了所有类别的准确性，因为它代表了模型的综合性能。
实现细节。我们使用在 ImageNet-1K [?] 上预训练的 DINOv1 [?] 作为图像编码器，并保持其冻结。来自 10 -th 块的特征用于视觉原语的解耦。我们将视觉原语的数量设置为 8。通过 Adam 优化器在学习率为 0.0005、批量大小为 64 的条件下训练 Slot-Attention 共 400 轮。我们构建了最后两个块的 ConGCD，遵循之前的研究 [? ?] 。在细粒度数据集中，$\lambda = 0.35$，标签平滑系数 α 设置为 0.5，而在粗粒度数据集中为 0.1。此外，我们在 ImageNet-22K [?] 上预训练的 DINOv2 [?] 上进行了上述训练。所有实验均在 RTX-4090 上进行，每个阶段的随机数种子固定，以避免数据标签泄露。

5.2. 基线

除了基于对比的方法 CMS [?] 和在 SimGCD [?] 中使用原型分类器外，我们还将 ConGCD 与其他几种现有的最先进方法进行比较。LegoGCD [?] [?] 通过结合不同组件的模块化方法在未标记的数据集中发现新的类别。SPTNet [?] 通过空间提示调优迭代优化模型参数和数据参数。SelEx [?] 通过分层伪标签的自我专长策略解决 GCD 问题。

5.3. 结果

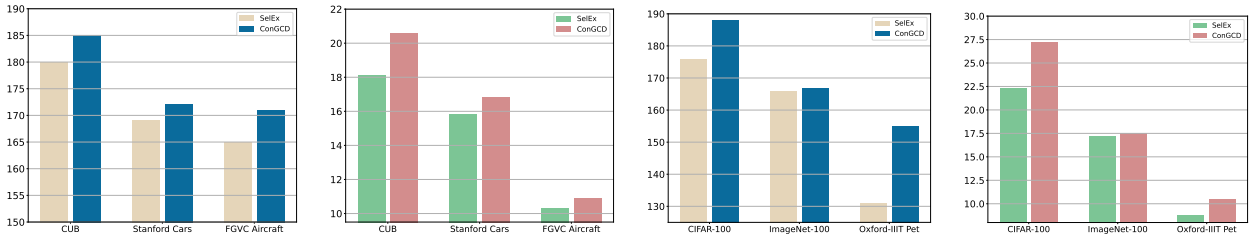
定量结果。ConGCD 在细粒度和粗粒度分类任务中表现出显著的改进，特别是在分类新类别时表现出色，同时保持对已知类别的性能。在细粒度任务中（表 1 ），ConGCD 在四个基线上平均提高了 1.18 % 的整体准确率，新类别的准确率提高了 1.98 %。在粗粒度任务中（表 2 ），ConGCD 在 CIFAR-100 上新类别的准确率提高了 0.98 %，而已知类别的准确率保持稳定或略有提高。这些结果突出了 ConGCD 在有效分类新类别的同时不影响已知类别的性能，显示了其作为一种具备强大兼容性的即插即用方法的有效性。
定性结果。(1) 上下文共识单元捕捉固有分布，而主导共识单元则关注与类别相关的变量。如 Fig. 3 所示，弱神经元在低信息区域（如背景）中激活，而主导神经元则锁定高语义差异区域（例如，车轮、车灯）以增强 GCD。(2) 为了突出 ConGCD 的性能提升，Fig. 8 中的注意力图显示 ConGCD 强调对象的主体部分（前景），不同于 SelEx 更关注背景和特定小范围区域。这使得对任务相关变量的提取更加均匀。(3) 如 Fig. 5 所示，可以观察到在同一数据集中，视觉原语具有很高的跨样本一致性。一个原语专注于具有相似语义的区域。

5.4. 分析

视觉原语增加冯 · 诺依曼熵。我们从表示学习和信息量化的角度分析了 ConGCD 有效性的来源。我们采用冯 · 诺依曼熵 [?] 作为衡量表示中信息量的方法。对于 GCD 中使用的类别标记 $\boldsymbol{z}$ ，其自相关矩阵为 $\mathcal{A} \triangleq \sum_{i=1}^{N} \frac{1}{N} \boldsymbol{z}_i \boldsymbol{z}_i^{\top}$ 。冯 · 诺依曼熵（VNA）表示为 $\hat{H}(\mathcal{A}) \triangleq -\sum_j \lambda_j \log \lambda_j$ ，其中 λ 是 $\mathcal{A}$ 的特征值。如 Fig. 4 所示，与 SelEx 相比，ConGCD 的表示具有更高的 $\hat{H}(\mathcal{A})$ 和更高的 $log(\text{rank}(\mathcal{A}))$ ，这意味着信息量更大。这表明 ConGCD 通过提供更多的判别特征使 GCD 过程能够更有辨别力地分离聚类。
视觉原语的多功能性。为了验证 ConGCD 作为即插即用解决方案的普遍性，我们将 ConGCD 整合到现有的 GCD 方案中，并在 Tab. 1 和 Tab. 2 中报告了结果。显然，ConGCD 为它们提供了准确性的一致改进，这可以归因于视觉原语从对象中的无结构解耦，使得模型具有一定的拓扑关系建模能力。模型不再仅仅依赖于单一视角的图像级嵌入，而是将这种类别无关的拓扑关系应用于新类别的发现。
ConGCD 提供了准确的分布估计。GCD 的核心是尽可能准确地估计类别分布并区分不同类别。我们从两个角度评估 ConGCD 对分布估计的贡献：聚类可视化和类别数量估计。（1）我们使用 t-SNE [?] 对嵌入进行可视化，如 Fig. 6 所示。值得注意的是，ConGCD 在不修改优化目标的情况下显著增强了聚类质量。它减少了类内距离，明确了聚类边界。（2）如 Tab. 3 所示，ConGCD 与 CMS [?] 结合，实现了最准确的类别数量估计 $|\mathcal{Y}_u|$ ，误差率最低，展示了其对数据建模和特

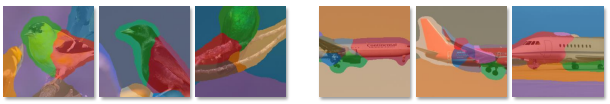
	Method	CIFAR-10			CIFAR-100			ImageNet-100			Average		
		All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	ORCA [?]	96.9	95.1	97.8	74.2	82.1	67.2	79.2	93.2	72.1	83.4	90.1	79.0
	GCD [?]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3	79.5	88.0	73.7
	GPC [?]	90.6	97.6	87.0	75.4	84.6	60.1	75.3	93.4	66.7	80.4	91.9	71.3
	XCon [?]	96.0	97.3	95.4	74.2	81.2	60.3	77.6	93.5	69.7	82.6	90.7	75.1
	PIM [?]	94.7	97.4	93.3	78.3	84.2	66.5	83.1	95.3	77.0	85.4	92.3	78.9
	PromptCAL [?]	97.9	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3	87.4	91.2	84.0
	DCCL [?]	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2	84.0	87.9	81.1
	InfoSieve [?]	94.8	97.7	93.4	78.3	82.2	70.5	80.5	93.8	73.8	84.5	91.2	79.2
	SimGCD [?]	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9	86.7	89.8	84.6
	+ ConGCD	97.3	95.1	98.4	81.3	82.5	78.9	83.5	92.8	78.6	87.4	90.1	85.3
	LegoGCD [?]	97.1	94.3	98.5	81.8	81.4	82.5	86.3	94.5	82.1	88.4	90.1	87.7
	+ ConGCD	97.2	94.8	97.9	82.2	82.1	82.3	85.8	94.4	81.7	88.4	90.4	87.3
	SPTNet [?]	97.3	95.0	98.6	81.3	84.3	75.6	85.4	93.2	81.4	88.0	90.8	85.2
	+ ConGCD	97.4	95.2	98.5	82.5	85.9	77.3	85.9	93.4	82.5	88.6	91.5	86.1
	SelEx [?]	94.1	97.7	92.2	80.0	84.8	70.4	82.3	93.9	76.5	85.4	92.1	79.7
	+ ConGCD	95.7	97.6	94.8	80.5	84.9	71.7	83.5	94.2	78.0	86.6	92.2	81.5
	Avg. △	+0.50	+0.15	+0.55	+0.83	+0.93	+0.98	+0.43	-0.03	+0.73	+0.63	+0.35	+0.75

Table 2. 在粗粒度分类基准测试上的结果。ConGCD 提供了出色的跨方案和跨模型兼容性。



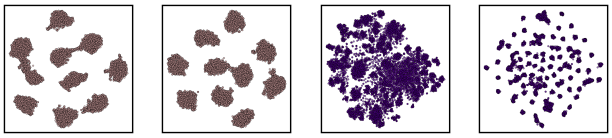
(a) $log(\text{rank}(\mathcal{A}))$ of fine-grained datasets (b) $\hat{H}(\mathcal{A})$ of fine-grained datasets (c) $log(\text{rank}(\mathcal{A}))$ of coarse-grained datasets (d) $\hat{H}(\mathcal{A})$ of coarse-grained datasets

Figure 4. （a）$log(\text{rank}(\mathcal{A}))$ 与（b）$\hat{H}(\mathcal{A})$ 的比较。为说明 99 % 的总特征值能量，所需的最大特征值的计数用作秩的替代。



(a) CUB-200 (b) FGVC-Aircraft

Figure 5. 视觉基本元素的可视化。在数据集中，基本元素以无监督的方式与特定语义绑定。



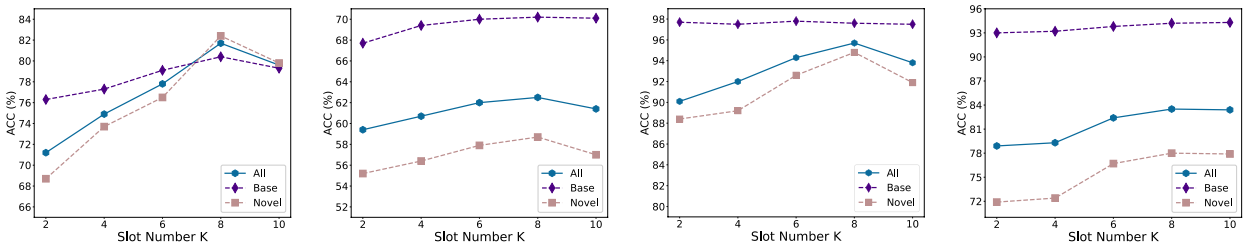
(a) CIFAR-10 with SelEx (b) CIFAR-10 with ConGCD (c) CIFAR-100 with SelEx (d) CIFAR-100 with ConGCD

Figure 6. 使用 t-SNE 对嵌入空间进行可视化 [?] 。

征空间划分的卓越能力，这对于达到 GCD 的上限至关重要。

视觉原语数量的消融研究。作为一种即插即用的解决方案，ConGCD 不需要任何超参数调优，这对于其广泛应用是一个重要的前提。唯一可能影响性能的参数是视觉原语的数量 K 。我们对细粒度和粗粒度数据集进行了消融研究，如 Fig. 7 所示。可以看出，在开放世界中，适当的 K 足以处理多个下游任务。当 K 在 7 到 10 的范围内时，ConGCD 实现了一致且稳定的性能提升，在所有数据集中，K 被设定为 8。

多层共识机制的消融研究。我们对 Sec. 4.2 中 ConGCD 的关键组件进行了消融研究，结果总结在 Tab. 4 中。专家混合（MoE）是与我们的多层共识机制最具可比性的方法，因为它直接与视觉原语集成。然而，传统的 MoE 对 GCD 任务有害，因为它仅接受图像级别的嵌



(a) CUB-200 (b) FGVC-Aircraft (c) CIFAR-10 (d) ImageNet-100

Figure 7. 视觉基元 K 数量在 (a, b) 细粒度和 (c, d) 粗粒度基准上的消融分析。

Method	CIFAR-100		ImageNet-100		CUB-200		Stanford-Cars	
	$\lvert\mathcal{Y}_u\rvert$	Err(%)	$\lvert\mathcal{Y}_u\rvert$	Err(%)	$\lvert\mathcal{Y}_u\rvert$	Err(%)	$\lvert\mathcal{Y}_u\rvert$	Err(%)
Ground truth	100	-	100	-	200	-	196	-
GCD [?]	100	0	109	9	231	15.5	230	17.3
DCCL [?]	146	46	129	29	172	9	192	0.02
PIM [?]	95	5	102	2	227	13.5	169	13.8
GPC [?]	100	0	103	3	212	6	201	0.03
CMS [?]	94	6	98	2	176	12	149	23.9
ConGCD	97	3	99	1	186	7	160	18.3

Table 3. $\lvert\mathcal{Y}_u\rvert$ 的估计数量和误差率。

Figure 8. SoTA 和 ConGCD 的注意力图可视化。

Components	CUB-200			FGVC-Aircraft			CIFAR-10			ImageNet-100		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
conventional MoE	78.2	79.1	77.3	59.4	78.1	53.6	93.6	97.1	91.8	80.4	92.5	74.9
Ours w/o Dominant Consensus Units	81.1	80.3	81.5	62.1	69.7	58.3	94.9	97.8	94.1	82.9	94.0	77.2
Ours w/o Contextual Consensus Units	80.4	79.7	81.0	61.3	70.6	57.1	94.2	97.6	92.3	81.5	93.5	75.4
Ours w/o Consensus Scheduler	79.8	79.5	80.0	60.9	70.1	56.7	94.4	97.7	92.7	81.4	92.7	77.0

Table 4. 多重共识组件的消融研究。

入，并且过拟合于已知类别的标签数据。专家们仅仅记住了已知类别的特征，无法泛化到新颖类别。相反，我们的多层共识机制提供了三个关键优势：(1) 主导共识单元处理多样化的视觉原语，能够在原语级进行感知，提高新颖类别的分类准确性。(2) 上下文共识单元捕捉已知和未知类别之间共享的内在分布，确保平衡聚类并减轻偏差。(3) 共识调度器引入了一个可学习的阈值，以优化弱神经元和主导神经元之间的区分，优化单元选择和权重分配以提高准确性。

6. 结论

当前的 GCD 方法将图像视为原子实体，忽视了人类的组合推理能力。我们提出了一种名为 ConGCD 的认知启发范式，通过两个关键创新来弥合这种差距：自我解构通过语义重构将物体分解为竞争的视觉基元，而多路复用共识动态整合了双路径，主导共识单位捕捉类别特定属性，情境共识单位则建模分布不变性。ConGCD 在粗粒度和细粒度基准上实现了有竞争力的性能，增强了开放世界识别的可解释性和泛化能力。

这项工作部分得到了中国国家自然科学基金 82172033、U19B2031、61971369、52105126、82272071 和 62271430 资助；部分得到了江淮先进技术中心的梦想基金资助；还部分得到了国家红外检测技术重点实验室开放基金的资助。