GET: Unlocking the Multi-modal Potential of CLIP for Generalized Category Discovery

王恩光 [1], 彭志茂 [1], 解正源 [1], 杨飞 [2,1]*, 刘夏雷 [2,1], 程明明 [2,1]
[1]VCIP, CS, 南开大学 [2]NKIARI, Shenzhen Futian
{enguangwang,zhimao796,xiezhengyuan}@mail.nankai.edu.cn
{feiyang,xialei,cmm}@nankai.edu.cn

Abstract

广义类别发现（Generalized Category Discovery, GCD）旨在面对包含已知类和未知类的无标签数据集时，既能准确识别出新类别，又能正确分类已知类别。当前的 GCD 方法仅依赖单一的视觉模态信息，导致对视觉相似类别的分类效果较差。而文本信息作为一种不同的模态，能够提供互补的判别信息，这促使我们将文本信息引入 GCD 任务中。然而，由于无标签数据缺乏类别名称，直接利用文本信息变得不切实际。为了解决这一挑战，本文提出了一种文本嵌入合成器（Text Embedding Synthesizer, TES），用于生成无标签样本的伪文本嵌入。具体而言，TES 利用了 CLIP 模型能够生成对齐的视觉-语言特征这一特性，将视觉嵌入转换为 CLIP 文本编码器所需的 token，从而生成伪文本嵌入。此外，我们采用双分支框架，通过不同模态分支的联合学习与实例一致性约束，使视觉信息与语义信息相互增强，促进了视觉与文本知识的交互与融合。我们的方法充分挖掘了 CLIP 的多模态潜力，在所有 GCD 基准测试中显著优于基线方法，达到了新的最先进性能。我们的代码已公开：https://github.com/enguangW/GET。

1. 引言

深度神经网络在大量标注数据上训练后展现出强大的视觉识别能力 [23]。尽管这一进展令人鼓舞，但传统的“封闭集假设”（close-set assumption）严重限制了模型在实际应用场景中的部署。近年来，新类别发现（Novel Class Discovery, NCD）[15] 被提出，旨在利用从已知标签数据中学到的知识，对未标注数据中的未知类别进行分类。作为对 NCD 的现实扩展，广义类别发现（Generalized Category Discovery, GCD）[41] 假设未标注数据既包含已知类也包含未知类，而不仅仅是像 NCD 中那样仅包含未知类。该任务要求模型在正确分类未标注数据中已知类别的同时，准确发现其中的未知类别，从而打破传统封闭集的限制，使 GCD 成为一项具有挑战性且意义重大的任务。

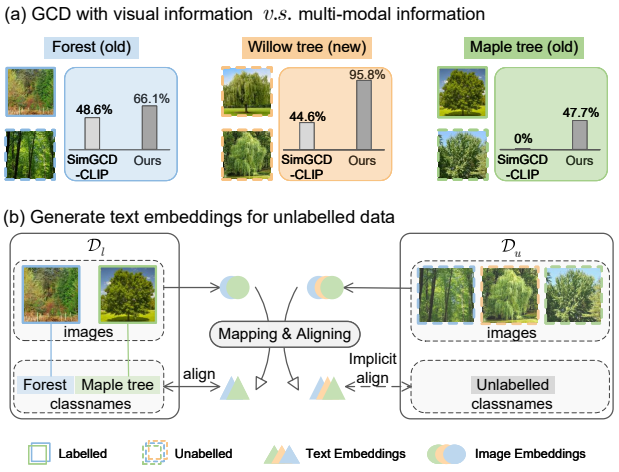

图 1. 我们方法的动机。(a) 当前的 GCD 方法 [45] 仅依赖单一的视觉特征，导致对视觉上相似类别的分类效果较差；我们的方法引入了文本信息，提升了模型的判别能力。(b) 我们提出的方法在实现模态对齐的同时，将图像嵌入映射到文本嵌入空间。

*通讯作者.

先前的 GCD 方法 [35, 41, 45, 48, 51] 通常采用基于 DINO [5] 预训练的 Vision Transformer（ViT）作为主干网络，期望模型具备良好的初始判别能力，从而便于在训练数据上进行微调。尽管这些方法取得了不错的结果，但由单一视觉主干提取的特征在面对视觉相似类别时仍面临挑战，例如所有细粒度数据集中的类别以及一些通用数据集的超类子集。如 Fig. 1(a) 所示，将参数化基线方法 [45] 的主干网络替换为强大的 CLIP [36] 视觉编码器后，仍然难以泛化某些视觉概念，导致结果不理想。受到文本模态可以提供互补判别信息这一观点的启发，我们决定将文本信息引入 GCD 任务中，以弥补视觉概念判别能力的不足。然而，在 GCD 中，由于未标注数据缺乏类别名称，使用 CLIP 的文本编码器变得不切实际，从而限制了其在 GCD 任务中多模态潜力的发挥。

为了解决这一具有挑战性的问题，本文提出了一种基于生成的方法，用于为未标注数据生成伪文本嵌入，我们将该方法命名为 GET（GEnerate pseudo Text embeddings）。具体而言，我们首先引入一个基于 CLIP 视觉-语言对齐特性的模块——文本嵌入合成器（Text Embedding Synthesizer, TES），以生成可靠且模态对齐的伪文本特征。如 Fig. 1(b) 所示，TES 学习一种映射关系，将图像嵌入转换为文本嵌入。具体来说，TES 将视觉嵌入转换为文本编码器所需的 token，从而无需原始文本输入。为了缩小生成的伪文本嵌入与真实文本嵌入之间的差距，TES 从对应于已标注数据的真实文本嵌入中提炼知识。此外，TES 还对同一实例的文本和图像表示进行对齐，增强了语言与视觉之间的一致性，同时防止模型对已知类别的过拟合。这种训练方式使 TES 相当于一个仅需视觉输入、经过特殊微调的文本编码器。从另一个角度来看，我们的 TES 可被视为执行了一个图像描述生成任务（image captioning task）[31]。

为了在 GCD 任务中有效利用这些多模态特征，我们进一步提出了一种双分支多模态联合训练策略，并设计了跨模态实例一致性目标函数。其中一条分支专注于视觉信息，另一条分支则通过文本信息进行补充。通过在 GCD 任务上的联合学习，视觉与语义信息相互增强。此外，我们的跨模态实例一致性目标函数强制要求每个实例在视觉模态与文本模态中与由已标注样本构建的锚点具有相同的关系，促进了视觉与文本嵌入空间的交互与对齐。在 TES 生成的文本嵌入支持下，结合合理的双分支训练策略，多模态特征能够修正分类超平面，在提升判别能力的同时缓解类别偏差问题。

总结，我们的主要贡献如下：

- 为了解决未标注数据无法使用文本编码器的问题，我们提出了 TES 模块，将视觉嵌入转换为 CLIP 文本编码器所需的 token，从而生成伪文本嵌入。
- 在所提出的双分支框架中，通过引入跨模态实例一致性目标函数，不同模态的信息相互增强，生成更具判别能力的分类原型。
- 我们的方法在多个基准数据集上取得了最先进的性能表现，为 GCD 任务提供了一种多模态范式。

2. 相关工作

新类别发现（NCD） NCD 的思想可以追溯到 KCL [18]，其中由相似性预测网络生成的成对相似性用于指导聚类重建，为跨任务和跨领域的迁移学习提供了一种有效的实现方式。早期的方法通常基于两个目标：在有标签数据上进行预训练，在未标注数据上进行聚类。RS [16] 在有标签和无标签数据上同时进行自监督预训练，缓解了模型对已知类别的偏好。同时，RS 还提出了通过秩统计（rank statistics）进行知识迁移的方法，这一策略在后续研究中被广泛采用 [49]。Zhao 等人 [49] 提出了一种具有双排序统计机制的双分支学习框架，并通过互信息蒸馏交换信息，在某种程度上与我们的方法类似。不同之处在于，我们的两个分支分别关注语义信息和视觉信息，而 [49] 中则侧重于局部与全局特征。为了简化 NCD 方法的设计，UNO [13] 提出使用统一的交叉熵损失，并结合多视角 SwAV [4] 的交换预测策略来优化任务，建立了一种新的范式。

广义类别发现（GCD） 近年来，GCD [41] 将 NCD 扩展到了一个更贴近现实的任务设定中，即未标注数据同时来自于已知类和未知类。GCD [41] 使用预训练的视觉 Transformer [11] 提供初始的视觉表示，并通过对有标签数据以及全部数据进行监督和自监督对比学习来微调主干网络。在模型学习到具有判别能力的特征表示后，通过约束有标签样本的正确聚类，采用半监督的 k-means 方法进行分类。作为一个新兴且现实意义重大的研究方向，GCD 正逐渐引起学术界的关注。PromptCAL [48] 提出了一种两阶段框架，旨在解决由

负样本误配导致的类别冲突问题，同时增强了模型在下游数据集上的适应能力。SimGCD [45] 引入了一种参数化分类方法，在缓解 GCD 聚类带来的计算开销的同时，取得了显著的性能提升。具体而言，SimGCD 在 GCD 框架上增加了一个分类器，并联合使用了自蒸馏与监督训练策略。μGCD [43] 通过引入 Clevr4 数据集，对现有方法中的类别偏倚问题进行了分析，并采用“均值教师”（mean-teacher）技术及更高效的训练策略，实现了显著的性能提升。CLIP-GCD [33] 从大规模文本语料库中挖掘文本描述，利用 CLIP 的文本编码器，并将视觉特征与文本特征简单拼接用于分类。相比之下，我们的方法专注于数据本身，无需引入额外语料库。最近，TextGCD [52] 从多个基准数据集中收集大量文本标签，并借助大语言模型（LLMs）增强这些标签，构建了一个“视觉词典”，并基于视觉词典与视觉特征之间的相似性为每个样本生成文本描述。与这些方法不同，我们的 GET 方法利用 CLIP 将多模态信息引入任务中，而无需依赖任何额外数据库或大型语言模型。

视觉-语言预训练 视觉-语言预训练（Vision-Language Pre-training, VLP） [6, 7, 9, 12, 14, 25] 旨在通过在大规模图文对数据上训练一个大模型，使其在经过微调后能够在多种下游视觉-语言任务中表现出色。一些研究 [8, 24, 27, 28, 39] 通过融合方法建模图像与文本之间的交互，在各类图文任务中取得了良好的性能提升。然而，融合方法需要对所有图文对进行编码，导致在图文检索任务中的推理速度较慢。因此，一些研究 [19, 36] 提出分别对图像和文本进行独立编码，并通过对比学习将图像嵌入和文本嵌入投影到统一的嵌入空间中。CLIP [36] 在大规模图文对数据上采用对比训练策略，最小化对应图文对之间的距离，同时最大化非对应图文对之间的距离。CLIP 所展现出的强大泛化能力和多模态特性促使我们将其引入到 GCD 任务中。

3. 预备知识

3.1. 问题公式化

在 GCD 任务中，训练数据 $\mathcal{D}$ 被划分为一个有标签数据集 $\mathcal{D}_l = \left\{(\boldsymbol{x}_i^l, y_i^l)\right\}_{i=1}^N \in \mathcal{X} \times \mathcal{Y}_l$ 和一个未标注数据集 $\mathcal{D}_u = \{(\boldsymbol{x}_i^u, y_i^u)\}_{i=1}^M \in \mathcal{X} \times \mathcal{Y}_u$，其中 $\mathcal{Y}_l$ 和 $\mathcal{Y}_u$ 分别表示标签空间，并且满足 $\mathcal{Y}_l \subset \mathcal{Y}_u$，且 $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$。$|\mathcal{Y}_l|$ 和 $|\mathcal{Y}_u|$ 分别表示有标签样本和未标注样本的类别数量。按照文献 [41, 45] 中的设定，我们假设新类别的数量 $|\mathcal{Y}_u \backslash \mathcal{Y}_l|$ 是已知的，或者可以通过一些现成的方法进行估计 [15, 41]。GCD 的目标是借助有标签样本，对未标注样本进行正确的聚类。

3.2. 参数化 GCD 方法（SimGCD）

本文采用 SimGCD [45] 提出的参数化方式来解决 GCD 问题。该方法通过 DINO 式的自蒸馏形式训练一个统一的原型分类头，用于对所有新/旧类别进行分类。具体而言，它包括两种类型的损失函数：表征学习损失和参数化分类损失。对于表征学习部分，它在所有有标签数据上进行监督表示学习 [20] $\mathcal{L}_{scon}$，并在全部训练数据上进行自监督对比学习 $\mathcal{L}_{con}$，对应的损失函数如下：

$$\mathcal{L}_{scon} = -\frac{1}{|B_l|} \sum_{i \in B_l} \frac{1}{|\mathcal{N}_i|} \sum_{q \in \mathcal{N}_i} \log \frac{\exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_q{}'}{\tau_{sc}}\right)}{\sum_{n \neq i}^{n \in B_l} \exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_n{}'}{\tau_{sc}}\right)}, \tag{1}$$

$$\mathcal{L}_{con} = -\frac{1}{|B|} \sum_{i \in B} \log \frac{\exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_i{}'}{\tau_c}\right)}{\sum_{n \neq i}^{n \in B} \exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_n{}'}{\tau_c}\right)}, \tag{2}$$

其中 $\mathcal{N}_i$ 表示在一个 batch 中与 $\boldsymbol{x}_i$ 具有相同语义标签的其他图像索引，B_l 是 mini-batch B 中的有标签子集，τ_{sc} 和 τ_c 是温度系数。对于图像编码器生成的两个视图 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_i'$ 的视觉嵌入 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_i'$，使用一个 MLP 层 $g(\cdot)$ 将其映射为高维嵌入 $\boldsymbol{h}_i = g(\boldsymbol{z}_i)$ 和 $\boldsymbol{h}_i' = g(\boldsymbol{z}_i')$。对于参数化分类部分，所有有标签数据通过交叉熵损失 $\mathcal{L}_{\text{cls}}^s$ 进行训练，所有训练数据则通过自蒸馏损失 $\mathcal{L}_{\text{cls}}^u$ 进行训练：

$$\mathcal{L}_{\text{cls}}^s = \frac{1}{|B_l|} \sum_{i \in B^l} \mathcal{H}\left(y_i, \sigma(\boldsymbol{p}_i, \tau_s)\right), \tag{3}$$

$$\mathcal{L}_{\text{cls}}^u = \frac{1}{|B|} \sum_{i \in B} \mathcal{H}\left(\sigma(\boldsymbol{p}_i{}', \tau_t), \sigma(\boldsymbol{p}_i, \tau_s)\right), \tag{4}$$

其中 $\sigma(\cdot)$ 表示 softmax 函数，$\boldsymbol{p}_i$ 和 $\boldsymbol{p}_i{}'$ 分别是两个视图 x_i 和 x_i' 在原型分类器上的输出，τ_s 是温度参数，τ_t 是更锐化的版本。$\mathcal{H}(\cdot)$ 表示交叉熵函数，y_i 是 x_i 对应的真实标签，$\sigma(\boldsymbol{p}_i{}', \tau_t)$ 是 x_i 的软伪标签。

此外，SimGCD 还引入了一个最大平均熵正则化项 $H(\overline{\boldsymbol{p}})$ 来防止模型陷入平凡解，其中 $H(\cdot)$ 是预测结果的熵 [37]，$\overline{\boldsymbol{p}} = \frac{1}{2|B|} \sum_{i \in B} \left(\sigma(\boldsymbol{p}_i{}', \tau_s) + \sigma(\boldsymbol{p}_i, \tau_s)\right)$ 表示一个 batch 的平均 softmax 概率。通过上述损失函数和正则化项联合训练模型，SimGCD 取得了显著的性能提升，然而由于仅依赖单一视觉模态信息，在视觉相似类别上的表现仍存在不足。

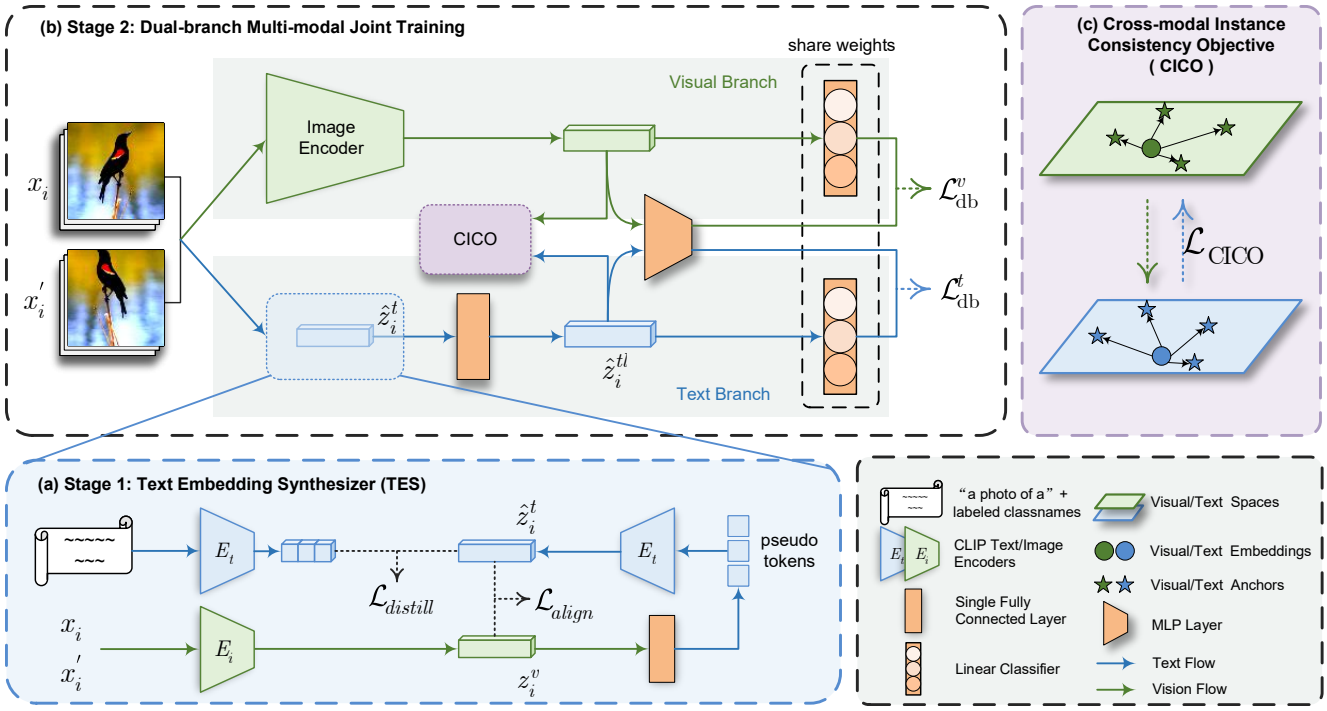

图 2. 我们 GET 框架的总体结构。(a) 在第一阶段，我们引入了一个文本嵌入合成器，为未标注数据生成伪文本嵌入。TES 学习一个线性映射，将图像特征转换为文本编码器的输入 token。所生成的伪文本嵌入将在第二阶段用于联合训练。(b) 我们在第二阶段提出了一种双分支多模态联合训练框架，并引入了跨模态实例一致性目标函数。两个分支采用相同的参数化训练策略 [45]，分别关注文本与视觉信息。(c) 我们的跨模态实例一致性目标函数使得视觉与文本信息能够相互交换并互相促进。

4. 方法

本文中，我们提出了 GET，通过引入多模态范式来解决 GCD 任务。如 Fig. 2 所示，GET 主要包含两个阶段。在第一阶段，我们学习一个文本嵌入合成器（Text Embedding Synthesizer, TES，在 Sec. 4.1 中介绍），为每个样本生成伪文本嵌入。在第二阶段，我们引入一种双分支多模态联合训练策略，并结合跨模态实例一致性约束（详见 Sec. 4.2），以充分利用多模态特征。

4.1. 文本嵌入合成器

未标注数据缺乏自然语言形式的类别名称，这使得将文本信息引入 GCD 任务变得困难。在本文中，我们尝试从特征层面出发，为每张图像生成与视觉嵌入对齐的伪文本嵌入。

受 BARON [46] 的启发，该方法将边界框内的嵌入视为句子中的词嵌入，用于解决开放词汇检测任务。我们提出了一个文本嵌入合成器（Text Embedding Synthesizer, TES）。具体而言，我们的 TES 利用 CLIP 能够生成对齐的视觉-语言特征这一特性，将视觉嵌入转换为 CLIP 文本编码器所需的 token，从而为每个样本生成伪文本嵌入。TES 的结构如 Fig. 2 (a) 所示。对于 mini-batch 中的每张图像 $\boldsymbol{x}_i$，我们使用 CLIP 的图像编码器获取其视觉嵌入 $\boldsymbol{z}_i^v$。接着通过一个全连接层 l 将视觉嵌入映射为伪 token，并作为 CLIP 文本编码器的输入，从而生成对应的伪文本嵌入 $\hat{\boldsymbol{z}}_i^t$。

TES 的目标包括对所有样本的对齐损失（align loss）和对有标签样本的蒸馏损失（distill loss）。为了使生成的伪文本嵌入 $\hat{\boldsymbol{z}}_i^t$ 与其对应的视觉特征 $\boldsymbol{z}_i^v$ 对齐，我们的对齐损失利用了 CLIP 编码器的模态对齐特性，在拉近正确图文嵌入对的同时，推开错误匹配的图文对。对齐损失包含两个对称部分 $\mathcal{L}_{align}^v$ 和 $\mathcal{L}_{align}^t$，其具体计算如下：

$$\mathcal{L}_{align}^v = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp\left(\boldsymbol{z}_i^{v\top}\hat{\boldsymbol{z}}_i^t/\tau_a\right)}{\sum_{j\in B}\exp\left(\boldsymbol{z}_i^{v\top}\hat{\boldsymbol{z}}_j^t/\tau_a\right)}, \quad (5)$$

$$\mathcal{L}_{align}^t = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp\left(\hat{\boldsymbol{z}}_i^{t\top}\boldsymbol{z}_i^v/\tau_a\right)}{\sum_{j\in B}\exp\left(\hat{\boldsymbol{z}}_i^{t\top}\boldsymbol{z}_j^v/\tau_a\right)}, \quad (6)$$

其中 $\hat{\boldsymbol{z}}_i^t$ 和 $\boldsymbol{z}_i^v$ 是经过 ℓ_2-归一化的，τ_a 是温度参数。因

此，对齐损失为：$\mathcal{L}_{align} = \mathcal{L}^v_{align} + \mathcal{L}^t_{align}$。为了确保我们生成的伪文本特征位于与真实文本特征相同的嵌入空间中并保持一致性，我们引入了蒸馏损失 $\mathcal{L}_{distill}$：

$$\mathcal{L}_{distill} = -\frac{1}{|B_l|}\sum_{i\in B_l}\log\frac{\exp\left(\hat{\boldsymbol{z}}_i^{t\top}T(n_i)\right)}{\sum_{j=0}^{|\mathcal{Y}_l|}\mathbb{1}_{[j\neq n_i]}\exp\left(\hat{\boldsymbol{z}}_i^{t\top}T(j)\right)} + \frac{1}{|B_l|}\sum_{i\in B_l}\left(\hat{\boldsymbol{z}}_i^t - T(n_i)\right)^2, \tag{7}$$

其中 $T \in |\mathcal{Y}_l| \times dim$ 表示 $|\mathcal{Y}_l|$ 个语义类别的真实文本嵌入，$n_i \in \mathcal{Y}_l$ 表示 $\boldsymbol{x}_i$ 在所有已知类别名称中的索引，$T(j)$ 表示所有已知类别中第 j 个类别的真实文本嵌入，$\mathbb{1}_{[\cdot]}$ 是指示函数。T 中的每个向量由文本编码器生成，使用的提示词为 “a photo of a {CLS}”，其中 {CLS} 表示对应类别名称。

我们文本嵌入合成器的整体目标为：$\mathcal{L}_{TES} = \mathcal{L}_{align} + \mathcal{L}_{distill}$. 其中，蒸馏损失用于引导网络输出的伪文本嵌入向真实语义空间靠拢，并使模型适应数据集的分布；而对齐损失则防止模型对已知类别的过拟合，并强制视觉与文本模态之间的一致性。此外，我们还为 TES 引入了多视角策略。具体而言，在一个 mini-batch 中，我们对同一图像的两个不同视图 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_i'$ 分别计算对齐损失 $\mathcal{L}_{align}$ 和蒸馏损失 $\mathcal{L}_{distill}$。这一策略进一步隐式增强了我们 TES 训练过程中的实例判别能力 [47]，使得同一张有标签图像的不同视图能够生成相同的伪文本嵌入。所生成的伪文本嵌入 $\hat{\boldsymbol{z}}_i^t$ 将用于第二阶段的联合训练。

4.2. 双分支多模态联合训练

直观上，引入多模态信息可以对 GCD 任务产生积极影响。文本信息可作为视觉信息的有效补充，提升模型的判别能力。然而，如何在 GCD 任务中有效利用视觉与文本信息，并充分发挥它们各自的作用，仍是一个具有挑战性的问题。本文提出了一种双分支架构，如 Fig. 2 (b) 所示，分别关注语义和视觉信息。我们对每个分支采用相同的参数化训练策略（见 Sec. 3.2），促使模型对同类别的视觉与文本特征具有对齐且互补的判别能力。此外，我们引入了一个跨模态实例一致性损失，约束样本在视觉与文本空间中的实例关系，使两个分支能够相互学习。我们用 v 表示视觉概念，t 表示文本概念。

视觉分支 视觉分支的目标包含表示学习部分和参数化分类部分。给定图像编码器生成的图像 $\boldsymbol{x}_i$ 的视觉嵌入 $\boldsymbol{z}_i^v$，我们使用一个 MLP 层 $g(\cdot)$ 将其映射为高维嵌入 $\boldsymbol{h}_i^v = g(\boldsymbol{z}_i^v)$。同时，我们使用原型分类器 $\eta(\cdot)$ 生成分类概率分布 $\boldsymbol{p}_i^v = \eta(\boldsymbol{z}_i^v)$。只需将 Eq. (1) 和 Eq. (2) 中的所有高维嵌入 $\boldsymbol{h}$（为简洁起见省略下标）替换为其对应的视觉分支版本 $\boldsymbol{h}^v$，即可得到监督对比损失 $\mathcal{L}^v_{scon}$ 和自监督对比损失 $\mathcal{L}^v_{ucon}$。整体表示学习损失通过 λ 进行平衡，表达如下：

$$\mathcal{L}^v_{\text{rep}} = (1-\lambda)\mathcal{L}^v_{ucon} + \lambda\mathcal{L}^v_{scon}. \tag{8}$$

对于参数化分类部分，只需将 Eq. (3) 和 Eq. (4) 中的 $\boldsymbol{p}_i$ 和 $\boldsymbol{p}_i'$ 替换为 $\boldsymbol{p}_i^v$ 和 $\boldsymbol{p}_i^{v\prime}$，即可得到交叉熵损失 $\mathcal{L}^s_{\text{cls-v}}$ 和自蒸馏损失 $\mathcal{L}^u_{\text{cls-v}}$。因此，分类损失为：$\mathcal{L}^v_{\text{cls}} = (1-\lambda)\mathcal{L}^u_{\text{cls-v}} + \lambda\mathcal{L}^s_{\text{cls-v}}$.

视觉分支的整体目标如下：

$$\mathcal{L}^v_{\text{db}} = \mathcal{L}^v_{\text{rep}} + \mathcal{L}^v_{\text{cls}}. \tag{9}$$

文本分支 我们的文本分支简单地采用了与视觉分支相同的训练策略。具体而言，给定由 TES 生成的文本嵌入 $\hat{\boldsymbol{z}}_i^t$，我们首先将其输入一个全连接层，以获得可学习的文本嵌入 $\hat{\boldsymbol{z}}_i^{tl}$ 并调整其维度。只需将表示学习目标 $\mathcal{L}^v_{\text{rep}}$ 中的 $\boldsymbol{h}_i^v$ 替换为 $\boldsymbol{h}_i^t = g(\hat{\boldsymbol{z}}_i^{tl})$，并将分类部分 $\mathcal{L}^v_{\text{cls}}$ 中的 $\boldsymbol{p}_i^v$ 替换为 $\boldsymbol{p}_i^t = \eta(\hat{\boldsymbol{z}}_i^{tl})$，即可得到对应的文本目标 $\mathcal{L}^t_{\text{rep}}$ 和 $\mathcal{L}^t_{\text{cls}}$。换句话说，将视觉概念指示符 v 替换为文本概念指示符 tl，即可得到文本分支的目标函数。因此，文本分支的整体目标可以形式化为：$\mathcal{L}^t_{\text{db}} = \mathcal{L}^t_{\text{rep}} + \mathcal{L}^t_{\text{cls}}$.

为了缓解旧类与新类之间的偏差，我们将均值-熵正则化：[1] 扩展为多模态均值熵正则化：$H_{mm} = H(\overline{p}_{mm}, \overline{p}_{mm})$，其中：$\overline{p}_{mm} = \frac{1}{2|B|}\sum_{i\in B}\left(\sigma(\boldsymbol{p}_i^v, \tau_s) + \sigma(\boldsymbol{p}_i^t, \tau_s)\right)$. 通过这种方式，每个原型在不同模态下的预测概率被约束为一致，防止模型陷入平凡解。

跨模态实例一致性目标 为了使两个分支能够相互学习并鼓励两种模态之间达成一致，我们提出了一种跨模态实例一致性目标 (Cross-modal Instance Consistency Objective, CICO)，如 Fig. 2 (c) 所示。我们的 CICO 形式上与 [49] 中的互信息蒸馏相同，但我们提炼的是两个分支之间的实例一致性。对于每个 mini-batch B，我们选择其包含 K 个类别的有标签子集 B_l 作为锚样本，并分别计算 K 个类别的视觉原型和文本原型，记

作视觉锚点 $\mathcal{P}_v$ 和文本锚点 $\mathcal{P}_t$。我们定义视觉分支和文本分支中的实例关系如下：

$$\begin{aligned} s_i^v &= \sigma(\boldsymbol{z}_i^{v\top}\mathcal{P}_v)\,, \\ s_i^t &= \sigma(\hat{\boldsymbol{z}}_i^{tl\top}\mathcal{P}_t)\,. \end{aligned} \tag{10}$$

因此，CICO 的形式可以写作：

$$\mathcal{L}_{\text{CICO}} = \frac{1}{2|B|}\sum_{i\in B}\left(D_{KL}(s_i^t||s_i^v) + D_{KL}(s_i^v||s_i^t)\right)\,, \tag{11}$$

其中 D_{KL} 表示 Kullback-Leibler 散度。对两个模态之间的实例关系进行互信息蒸馏，使得视觉流与文本流能够相互交换并从中受益，从而使两个分支成为彼此互补的判别辅助器。

我们方法的整体优化目标为：

$$\mathcal{L}_{\text{Dual}} = \mathcal{L}_{\text{db}}^v + \mathcal{L}_{\text{db}}^t - \epsilon H_{mm} + \lambda_c \mathcal{L}_{\text{CICO}}\,. \tag{12}$$

由于信息通过 CICO 在不同模态间交换并注入到视觉主干中，我们在推理阶段使用最后一个 epoch 的视觉分支进行预测。

5. 实验

5.1. 实验设置

数据集 我们在多个基准数据集上评估了我们的方法，包括三个通用图像分类数据集（即 CIFAR 10/100 [22] 和 ImageNet-100 [10]），来自 Semantic Shift Benchmark [42] 的三个细粒度数据集（即 CUB [44]、Stanford Cars [21] 和 FGVC-Aircraft [30]），以及三个具有挑战性的数据集（即 Herbarium 19 [40]、ImageNet-R [17] 和 ImageNet-1K [10]）。我们首次将 ImageNet-R 引入 GCD 任务。该数据集包含来自 200 个 ImageNet 类别的多种变体图像，从而对“数据来自同一领域”的 GCD 假设提出了挑战。具体的数据划分信息请参见 Supp（补充材料）。

评估与实现细节 遵循文献 [41, 45] 中的标准评估协议，我们使用聚类准确率（ACC）来评估模型性能。我们采用 CLIP [36] 预训练的 ViT-B/16 [11] 作为图像编码器和文本编码器。在第一阶段，我们训练一个全连接层。在第二阶段，我们移除了图像编码器中的投影层，使得输出特征维度为 768。其他实现细节及伪代码请参见 Supp（补充材料）。

5.2. 与最先进方法的比较

在本节中，我们将 GET 与多种当前最先进的方法进行比较。GCD 和 SimGCD 分别提供了非参数和参数化的范式。为了公平比较，我们将它们的主干网络替换为 CLIP，分别记作 GCD-CLIP 和 SimGCD-CLIP。

在细粒度与通用数据集上的评估 如 Tab. 1 所示，我们的方法在三个细粒度数据集上均取得了显著且稳定的性能提升。具体而言，在 CUB、Stanford Cars 和 Aircraft 数据集的 ‘All’ 类别上，我们的方法分别超越 SimGCD-CLIP 5.3%、8.5% 和 4.6%。在细粒度数据集中，不同类别的视觉概念具有高度相似性，仅依赖视觉信息进行分类极具挑战性。然而，文本信息能够提供额外的判别信息。因此，我们的 GET 通过文本与视觉信息流之间的相互增强，显著提升了分类准确率。在 Tab. 1 中，我们也展示了方法在三个通用图像分类数据集上的性能表现。由于 CIFAR 数据集的分辨率较低以及模型偏倚（CLIP 在 CIFAR100 上本身表现较差，其零样本性能仅为 68.7），所提方法在新类别上的结果相比 DINO 主干网络略有不足。然而，在 CLIP 本身判别能力受限的前提下，我们的方法在 CIFAR10 的“旧类”上仍提升了 0.4%，在 CIFAR100 的“新类”上提升了 2.2%，优于 SimGCD-CLIP。对于 ImageNet-100 数据集，SimGCD-CLIP 在 ‘All’ 类别上已达到 90.8% 的高度饱和性能，进一步提升面临较大挑战。然而，借助引入的多模态信息，我们的 GET 将性能上限提升至令人印象深刻的 91.7%。

在更具挑战性数据集上的评估 如 Tab. 2 所示，GET 在 Herb19 和 ImageNet-1K 数据集的 ‘All’ 类别和 ‘New’ 类别上均优于所有其他方法。特别地，在 Herb19 和 ImageNet-1K 的 ‘New’ 类别上，我们的方法分别取得了 1.4% 和 1.7% 的显著提升。此外，GCD 和 SimGCD 在 ImageNet-R 数据集上使用 DINO 主干网络时表现欠佳，突显了 DINO 在面对多领域图像时发现新类别的困难。尽管同一类别下的图像可能存在多个领域差异，其对应的文本信息仍保持一致。我们的方法有效融合了文本信息，在 ‘All’ 类别和 ‘Old’ 类别上分别比当前最先进方法提升了 3.2% 和 6.0%。值得注意的是，由于同一类别内文本信息的一致性，我们的文本分支在 ImageNet-R 和 ImageNet-1K 的 ‘All’ 类别上分别取得了 62.6% 和 63.5% 的准确率，表现出色。

Method	CUB			Stanford Cars			FGVC-Aircraft			CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k-means [29]	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RS+ [16]	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [13]	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	68.6	98.3	53.8	69.5	80.6	47.2	70.3	95.0	57.9
ORCA [2]	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1	81.8	86.2	79.6	69.0	77.4	52.0	73.5	92.6	63.9
GCD [41]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
GPC [50]	55.4	58.2	53.1	42.8	59.2	32.8	46.3	42.5	47.9	92.2	98.2	89.1	77.9	85.0	63.0	76.9	94.3	71.0
DCCL [34]	63.5	60.8	64.9	43.1	55.7	36.2	-	-	-	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
PromptCAL [48]	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	97.9	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3
SimGCD [45]	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9
μGCD [43]	65.7	68.0	64.6	56.5	68.1	50.9	53.8	55.4	53.0	-	-	-	-	-	-	-	-	-
LegoGCD [3]	63.8	71.9	59.8	57.3	75.7	48.4	55.0	61.5	51.7	97.1	94.3	98.5	81.8	81.4	82.5	86.3	94.5	82.1
GCD-CLIP	57.6	65.2	53.8	65.1	75.9	59.8	45.3	44.4	45.8	94.0	97.3	92.3	74.8	79.8	64.6	75.8	87.3	70.0
SimGCD-CLIP	71.7	76.5	69.4	70.0	83.4	63.5	54.3	58.4	52.2	97.0	94.2	98.4	81.1	85.0	73.3	90.8	95.5	88.5
GET (Ours)	77.0	78.1	76.4	78.5	86.8	74.5	58.9	59.6	58.5	97.2	94.6	98.5	82.1	85.5	75.5	91.7	95.7	89.7

表 1. 细粒度与通用数据集上的实验结果（%）最优结果以**加粗**形式标出。

Method	Herbarium 19			ImageNet-1K			ImageNet-R		
	All	Old	New	All	Old	New	All	Old	New
k-means [29]	13.0	12.2	13.4	-	-	-	-	-	-
RS+ [16]	27.9	55.8	12.8	-	-	-	-	-	-
UNO+ [13]	28.3	53.7	14.7	-	-	-	-	-	-
ORCA [2]	20.9	30.9	15.5	-	-	-	-	-	-
μGCD [43]	45.8	61.9	37.2	-	-	-	-	-	-
LegoGCD [3]	45.1	57.4	38.4	62.4	79.5	53.8	-	-	-
GCD [41]	35.4	51.0	27.0	52.5	72.5	42.2	32.5	58.0	18.9
SimGCD [45]	44.0	58.0	36.4	57.1	77.3	46.9	29.5	48.6	19.4
GCD-CLIP	37.3	51.9	29.5	55.0	65.0	50.0	44.3	79.0	25.8
SimGCD-CLIP	48.9	64.7	40.3	61.0	73.1	54.9	54.9	72.8	45.3
GET (Ours)	49.7	64.5	41.7	62.4	74.0	56.6	58.1	78.8	47.0

表 2. 更具挑战性数据集上的实验结果（%）

	TES	Dual-branch	CICO	Stanford Cars			CIFAR100		
				All	Old	New	All	Old	New
(1)	✗	✗	✗	70.0	83.4	63.5	81.1	85.0	73.3
(2)	✓	✓	✗	76.2	85.3	71.7	81.0	85.3	72.3
(3)	✓	✓	✓	78.5	86.8	74.5	82.1	85.5	75.5

表 3. 不同组件的消融实验

5.3. 消融实验与分析

各组件有效性分析 为了评估不同组件的有效性，我们在 SCars 和 CIFAR100 数据集上进行了消融实验，结果如 Tab. 3 所示。将 (2) 与 (1) 对比可以看出，在引入 TES 生成的文本特征后，SCars 的 'All' 类别性能提升了 6.2%，CIFAR100 的 'Old' 类别性能提升了 0.3%。进一步将 (3) 与 (1) 对比可以看出，CICO 使得两个分支能够相互交换信息并互相促进，在 SCars 的 'New' 类别上带来了 11% 的显著提升，在 CIFAR100 的 'New' 类别上提升了 2.2%。

不同融合方法的对比实验 在 Tab. 4 中，我们将所提出的双分支策略与其它多模态融合方法进行比较，包括拼接（concatenation）和均值（mean）融合方式。尽管这些方法由于引入了多模态信息可能带来性能提升，但我们验证了双分支联合学习的有效性，因为它促使模型对同类别的视觉与文本特征具有互补且对齐的判别能力，从而生成更具判别性的多模态原型。

TES 的有效性验证 为了验证 TES 的有效性，我们在 CUB 数据集上进行了实验，将 TES 生成的文本嵌入替换为通过 Text-Retrieval、VQA 和 Captioning 得到的文本嵌入进行对比。对于 Text-Retrieval，我们

	Dual-branch	Concat	Mean	Stanford Cars			CIFAR100		
				All	Old	New	All	Old	New
(1)	✗	✓	✗	68.9	79.1	64.0	79.9	85.5	68.7
(2)	✗	✗	✓	72.0	85.0	65.6	81.1	84.3	74.8
(3)	✓	✗	✗	78.5	86.8	74.5	82.1	85.5	75.5

表 4. 不同融合方法的对比实验

Method		Total Params	All	Old	New
Baseline	SimGCD-CLIP	92.2M	71.7	76.5	69.4
Text-Retrieval	WordNet	155.8M	69.8	77.1	66.2
	CC3M	155.8M	72.3	79.1	68.9
VQA	BLIP (ViT-L)	625.8M	67.1	74.3	63.5
	BLIP-2 (opt2.7b)	3.9B	71.3	73.5	70.2
Captioning	BLIP (ViT-L)	625.8M	40.5	54.6	33.4
	BLIP-2 (opt2.7b)	3.9B	42.6	56.1	35.8
Feature-Generation	TES (Ours)	165.1M	77.0	78.1	76.4
	TES w/o $\mathcal{L}_{align}$	165.1M	74.7	76.9	73.5
	TES w/o $\mathcal{L}_{distill}$	165.1M	75.3	77.5	74.2

表 5. 不同伪文本嵌入方法的对比实验

基于图像与文本嵌入之间的余弦相似度，从两个语料库（WordNet [32] 和 CC3M [38]）中检索每张图像最相似的文本。我们使用 BLIP [25] 和 BLIP-2 [26] 进行 Captioning，并使用问题“图片中的鸟叫什么名字？”进行 VQA。如 Tab. 5 所示，由于细粒度图像之间具有高度视觉相似性，通过 VQA 检索或生成的类别名称往往不够准确，导致性能受限。同时，描述类方法（Captioning）更倾向于描述对象姿态和场景信息，而非类别特异性特征，因此同一类别的样本可能生成差异较大的描述，显著影响了类别发现的效果。相比之下，我们的方法在参数量适中的情况下取得了最佳性能。关于 TES 的更多实验（包括结构设计、特征分布与灵活性分析）请参见 Supp（补充材料）。

使用不同提示的实验结果 在我们的方法中，我们使用了一个简单的提示词：“a photo of a {CLS}”。如 Tab. 6 所示，我们还探索了以下几种不同的提示方式：(1) “a photo of a {CLS}”，(2) “a photo of a {CLS}, which is a type of bird/car”，(3) 使用大语言模型（GPT4o-mini）生成的 {CLS} 描述文本，(4) 对 (1) 至 (3) 的文本特征进行平均。实验结果表明，我们所采用的提示词虽然简单，但已经取得了良好的效果；而更精细设计的提示词可以进一步提升模型性能。

Prompts	CUB			Stanford Cars		
	All	Old	New	All	Old	New
(1)	77.0	78.1	76.4	78.5	86.8	74.5
(2)	76.3	78.2	75.4	78.5	88.2	73.8
(3)	76.8	78.7	75.8	78.6	90.4	72.9
(4)	78.3	77.6	78.7	79.1	88.8	74.3

表 6. 使用不同提示的实验结果

Methods	NEV			TV-100		
	All	Old	New	All	Old	New
CLIP(zero-shot)	10.7	-	-	1.93	-	-
SimGCD	54.7	88.0	38.0	35.2	50.3	29.2
SimGCD-CLIP	79.1	96.7	70.3	55.7	75.8	47.8
GET (Ours)	85.3	96.0	80.0	57.1	77.3	49.2

表 7. 在 MEV 和 TV-100 数据集上的实验结果

关于在 GCD 中使用 CLIP 的讨论 GCD 的一个核心目标是发现新类别，而这一能力高度依赖主干模型提供的初始特征判别能力。由于 CLIP 具有强大的泛化能力，它能够编码更具判别性的特征，因此将 CLIP 引入 GCD 任务是一个自然的想法。一个值得关注的问题是：CLIP 是否在训练过程中已经见过 GCD 任务中的未知类别或类别名称。为此，我们从以下三个方面讨论在 GCD 中使用 CLIP 的意义。**方法论意义：**尽管 CLIP 在大规模数据集上进行了预训练，并可能与某些类别存在重叠，但其知识是隐式的、非结构化的。要有效地将其用于 GCD 任务，仍需设计新的方法，尤其是在如何利用文本编码器处理未标注数据方面提出了挑战。我们的方法在性能上优于 SimGCD-CLIP，验证了这一方法论价值。**前瞻性意义：**为了评估类别发现方法在 CLIP 未见过的场景下的表现能力，我们构建了一个包含 2023 年推出的新车型的小规模细粒度数据集（NEV）。此外，我们还在 NEV 和 TV-100 [53]（一个 CLIP 预训练未接触过的电视剧人物数据集）上进行实验，结果表明即使面对 CLIP 未曾见过的类别，利用文本模态对于有效的类别发现仍然至关重要。这为探索 CLIP 泛化能力在未来真正新颖类别的 GCD 任务中的应用提供了前瞻视角。**实际意义：**探索 CLIP 在现实场景中的潜力具有重要意义。因此，我们在 SSSupp 中展示了其在医学图像和超细粒度数据集上的实验结果。本研究为利用 CLIP 解决具有挑战性的 GCD 应用奠定了基础。

5.4. 定性结果分析

注意力图可视化 如 Fig. 3 所示，与 SimGCD-CLIP 相比，我们的方法额外关注了鸟类的羽毛纹理，这对于区分视觉上相似的鸟类种类至关重要。在文本信息的辅助下，我们方法中的视觉分支注意力图变得更加精细，能够聚焦于更具判别性的区域。

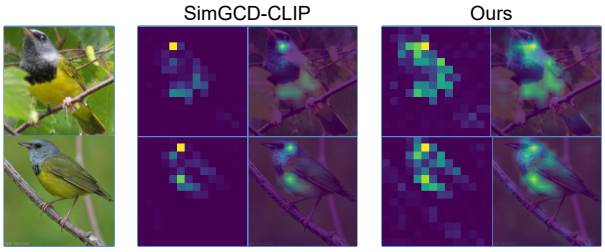

图 3. 类别 token 的注意力图可视化

t-SNE **可视化** Fig. 4 展示了在 CUB 数据集上随机选取的 20 个类别中，视觉特征与文本特征的 t-SNE 可视化结果。我们方法所提取的视觉特征和文本特征均展现出更清晰、更紧凑的聚类分布。更多可视化结果与聚类分析请参见 Supp（补充材料）。

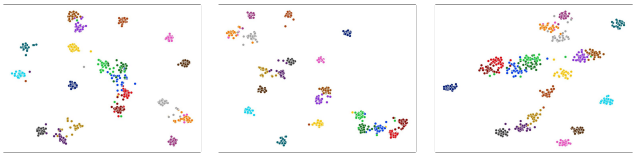

图 4. t-SNE 可视化

6. 结论

在本项工作中，我们提出利用多模态信息来解决 GCD 任务。具体而言，我们引入了一个文本嵌入合成器，用于为未标注数据生成伪文本嵌入。该模块使得使用 CLIP 的文本编码器成为可能，从而解锁了 GCD 任务中多模态信息的潜力。同时，我们采用了一种双分支训练策略，并引入跨模态实例一致性目标函数，促进了不同模态之间的协同作用与相互学习。我们的研究将 GCD 推向了一种多模态范式，并在多个基准测试中取得了优越性能，验证了所提方法的有效性。

局限性与未来工作 本方法的一个局限在于我们将视觉信息与文本信息视为同等重要。事实上，某些样本可能包含比文本信息更丰富且更具判别性的视觉信息，反之亦然。更合理的方式可能是让模型自适应地利用多模态信息，自主判断哪种模态的信息更为关键。这将是我们在未来工作中重点探索的方向。

致谢

本研究工作得到了以下项目资助：深圳市科技创新计划（JCYJ20240813114237048），国家自然科学基金（NO. 62206135，62225604），中国科协青年人才托举工程（2023QNRC001），中央高校基本科研业务费（南开大学，070-63233085），以及“科技涌江 2035”重点技术攻关计划项目（2024Z120）。本研究计算资源由南开大学超级计算中心提供支持。

参考文献

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In European Conference on Computer Vision, pages 456–473. Springer, 2022. 5

[2] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In International Conference on Learning Representations, 2022. 7

[3] Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16880–16889, 2024. 7

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems, 33:9912–9924, 2020. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021. 2

[6] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. Machine Intelligence Research, 20(1):38–56, 2023. 3

[7] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A

survey on vision-language pre-training. Machine Intelligence Research, 20(1):38–56, 2023. 3

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In European conference on computer vision, pages 104–120. Springer, 2020. 3

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences, 67(12):220101, 2024. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jua Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021. 2, 6

[12] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. arXiv preprint arXiv:2202.10936, 2022. 3

[13] Enrico Fini, Enver Sangineto, Stéphane Lathuiliere, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9284–9292, 2021. 2, 7

[14] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. Visual Intelligence, 2(1): 1–17, 2024. 3

[15] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8401–8409, 2019. 1, 3

[16] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In International Conference on Learning Representations (ICLR), 2020. 2, 7

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8340–8349, 2021. 6

[18] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In International Conference on Learning Representations, 2018. 2

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning, pages 4904–4916. PMLR, 2021. 3

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673, 2020. 3

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), 2013. 6

[22] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009. 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84–90, 2017. 1

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34:9694–9705, 2021. 3

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training

for unified vision-language understanding and generation. In ICML, 2022. 3, 8

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In ICML, 2023. 8

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019. 3

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32, 2019. 3

[29] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pages 281–297. Oakland, CA, USA, 1967. 7

[30] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013. 6

[31] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. arXiv preprint arXiv:2209.15162, 2022. 2

[32] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 1995. 8

[33] Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. Clip-gcd: Simple language guided generalized category discovery. arXiv preprint arXiv:2305.10420, 2023. 3

[34] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7579–7588, 2023. 7

[35] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7579–7588, 2023. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. 2, 3, 6

[37] Claude Elwood Shannon. A mathematical theory of communication. The Bell system technical journal, 27 (3):379–423, 1948. 3

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of ACL, 2018. 8

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019. 3

[40] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. In Workshop on Fine-Grained Visual Categorization, 2019. 6

[41] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7492–7501, 2022. 1, 2, 3, 6, 7

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In International Conference on Learning Representations, 2022. 6

[43] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. Advances in Neural Information Processing Systems 37, 2023. 3, 7

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[45] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 16590–16600, 2023. 1, 2, 3, 4, 6, 7

[46] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In CVPR, 2023. 4

[47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric

instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3733–3742, 2018. 5

[48] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3479–3488, 2023. 2, 7

[49] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In Conference on Neural Information Processing Systems (NeurIPS), 2021. 2, 5

[50] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 16623–16633, 2023. 7

[51] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. arXiv preprint arXiv:2305.06144, 2023. 2

[52] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In European Conference on Computer Vision, pages 41–58. Springer, 2024. 3

[53] Da-Wei Zhou, Zhi-Hong Qi, Han-Jia Ye, and De-Chuan Zhan. Tv100: A tv series dataset that pre-trained clip has not seen. Frontiers of Computer Science, 18(5):185349, 2024. 8