Textual Knowledge Matters: Cross-Modality Co-Teaching for Generalized Visual Class Discovery

Haiyang Zheng[1⋆], Nan Pu[1⋆], Wenjing Li[2,3†], Nicu Sebe[1], and Zhun Zhong[2,3†]

[1]University of Trento [2]Hefei University of Technology [3]University of Nottingham

Abstract. In this paper, we study the problem of Generalized Category Discovery (GCD), which aims to cluster unlabeled data from both known and unknown categories using the knowledge of labeled data from known categories. Current GCD methods rely on only visual cues, which however neglect the multi-modality perceptive nature of human cognitive processes in discovering novel visual categories. To address this, we propose a two-phase TextGCD framework to accomplish multi-modality GCD by exploiting powerful Visual-Language Models. TextGCD mainly includes a retrieval-based text generation (RTG) phase and a cross-modality co-teaching (CCT) phase. First, RTG constructs a visual lexicon using category tags from diverse datasets and attributes from Large Language Models, generating descriptive texts for images in a retrieval manner. Second, CCT leverages disparities between textual and visual modalities to foster mutual learning, thereby enhancing visual GCD. In addition, we design an adaptive class aligning strategy to ensure the alignment of category perceptions between modalities as well as a soft-voting mechanism to integrate multi-modality cues. Experiments on eight datasets show the large superiority of our approach over state-of-the-art methods. Notably, our approach outperforms the best competitor, by 7.7% and 10.8% in All accuracy on ImageNet-1k and CUB, respectively. Code is available at https://github.com/HaiyangZheng/TextGCD.

Keywords: Generalized Category Discovery · Cross-Modality Co-Teaching

1 Introduction

Despite the remarkable advancements of deep learning in visual recognition, a notable criticism is that the models, once trained, show a significant limitation in recognizing novel classes not encountered during the supervised training phase. Drawing inspiration from the innate human capacity to seamlessly acquire new knowledge with reference to previously assimilated information, Generalized Category Discovery (GCD) [35] is proposed to leverage the knowledge of labeled data from known categories to automatically cluster unlabeled data that belong to both known and unknown categories. While current GCD methods [29, 35, 38, 44, 45] have demonstrated considerable success utilizing advanced

⋆ Equal contribution. † Corresponding author.

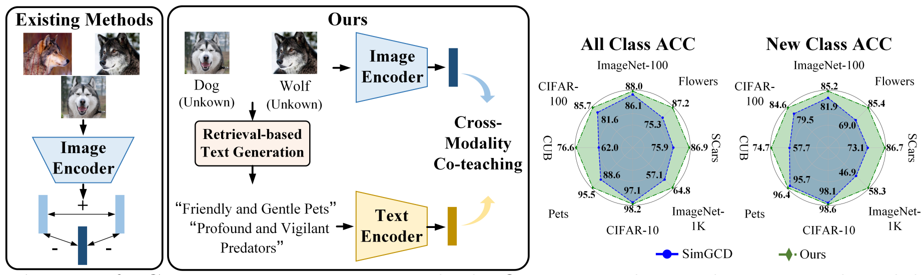

Fig. 1: Left: Comparison to existing methods. Our approach introduces textual modality information (*e.g.*, "Friendly and Gentle Pets" for Husky dogs, "Profound and Vigilant Predators" for wolves) into the framework and proposes cross-modal co-teaching for accurate generalized category discovery. Right: Performance comparison with SOTA.

large-scale visual models (*e.g.*, ViT [9]), as shown in Fig. 1, they predominantly focus on visual cues. In contrast, human cognitive processes for identifying novel visual categories usually incorporate multiple modalities [34], such as encompassing visual, auditory, and textual elements in recognizing a subject. In the light of this, unlike existing GCD methods [29,35,37,38,44,45] that rely on only visual modality, we propose to exploit both visual and textual cues for GCD.

In this paper, we advocate the utilization of large-scale pre-trained Visual-Language Models (VLMs) (*i.e.* CLIP [30]) to inject rich textual information into GCD. However, VLMs require predetermined, informative textual descriptors (*e.g.*, class names) for matching or recognizing images. This poses a significant challenge in GCD involving unlabeled data, particularly for unknown categories that lack predefined class names. Hence, how to provide relevant textual cues for unlabeled data is of a key in facilitating GCD with textual information.

To this end, we introduce the TextGCD framework, comprising a retrieval-based text generation (RTG) phase and a cross-modality co-teaching (CCT) phase, to leverage the mutual benefit of visual and textual cues for solving GCD. In particular, RTG constructs a visual lexicon encompassing category tags from diverse datasets and attributes gleaned from Large Language Models (LLMs), which offer categorical insights. Subsequently, the most pertinent tags and attributes are selected from the lexicon to craft descriptive texts for all images through an offline procedure. While the augmentation of textual information for unlabeled images via RTG is beneficial, the effective integration of this textual information to enhance visual category discovery remains an essential challenge.

To solve the challenge, we introduce CCT to leverage the inherent disparities between textual and visual modalities, harnessing these distinctions to facilitate joint learning, which is important for co-teaching [12]. To this end, we devise a cross-modal co-teaching training strategy that fosters mutual learning and collective enhancement between text and image models. Additionally, to ensure the effective alignment of category perceptions between text and image models and to facilitate comprehensive learning from both modalities, we introduce a warm-up stage and a class-aligning stage before engaging in cross-modality co-teaching. Finally, a soft-voting mechanism is deployed to amalgamate insights from both modalities, aiming to enhance the accuracy of category determination. In short, introducing the textual modality and exchanging insights between modalities have significantly improved the precision of visual category identification, particularly for previously unseen categories. We make the following contributions:

- We identify the limitations of existing GCD methods that rely on only visual cues and introduce additional textual information through a customized RTG based on large-scale VLMs.
- We propose a co-teaching strategy between textual and visual modalities, along with inter-modal information fusion, to fully exploit the strengths of different modalities in category discovery.
- Comprehensive experiments on eight datasets demonstrate the effectiveness of the proposed method. Notably, compared to the leading competitor in terms of All accuracy, our approach achieves an increase of 7.7% on ImageNet-1k and 10.8% on CUB. The source code will be publicly available.

2 Related Works

Generalized Category Discovery (GCD) aims to accurately classify an unlabeled set containing both known and unknown categories, based on another dataset labeled with only known categories. It is an extension of novel category discovery [10, 13, 14, 32, 46, 47] where the unlabeled set only contains unknown classes. Vaze et al. [35] first proposed optimizing image feature similarities through supervised and self-supervised contrastive learning to address GCD. Building on this, SimGCD [38] designed a parametric classification baseline for GCD. For better image representations, DCCL [29] proposed a novel approach to contrastive learning at both the concept and instance levels. Similarly focusing on image representation, PromptCAL [44] introduced a Contrastive Affinity Learning approach with auxiliary visual prompts designed to amplify the semantic discriminative power of the pre-trained backbone. Furthermore, SPTNet [37] proposed iteratively implementing model finetuning and prompt learning, resulting in clearer boundaries between different semantic categories. Unlike existing methods that primarily focus on visual cues, we design a retrieval-based approach to introduce text cues from LLM for unknown categories. Recently, CLIP-GCD [27] concatenated image features with textual features obtained from a Knowledge Database and categorized them using a clustering method. Unlike CLIP-GCD, which merely concatenates text and visual features from a frozen backbone, we leverage the differences between textual and visual models to establish a dynamic co-teaching scheme.
Visual-Language Models (VLMs) are designed to map images and text into a unified embedding space, facilitating cross-modal alignment. Among the prominent works, CLIP [30] employs contrastive representation learning with extensive image-text pairs, showcasing remarkable zero-shot transfer capabilities across a variety of downstream tasks. Additionally, LENS [2] leverages VLMs as visual reasoning modules, integrating them with Large Language Models (LLMs) for diverse visual applications. As VLMs effectively bridge the visual and textual

modalities, enhancing image classification tasks with knowledge from LLMs has been extensively explored [24,26,41]. However, these works use VLMs to enhance image classification in scenarios where the names of all classes are predefined for VLMs. In this paper, we exploit VLMs to address the GCD, where the unlabeled data do not have exploitable textual information for VLMs. We utilize a retrieval-based text generation method to enable the GCD to benefit from VLMs.

Co-teaching, which stemmed from the Co-training approach [3], was originally proposed as a learning paradigm to address the issue of label noise, [12]. This strategy involves training two peer networks that select low-loss instances from a noisy mini-batch to train each other. Yuan et al. [43] established three symmetric peer proxies with pseudo-label-driven co-teaching to address the Offline model-based optimization task. Yang et al. [40] were the first to introduce the co-teaching strategy into the Person re-identification task, designing an asymmetric co-teaching framework. Similarly, Roy et al. [31] pioneered co-teaching between classifiers of a dual classifier head to tackle the multi-target domain task. Unlike these works, our focus is on applying co-teaching to the GCD task. Furthermore, rather than employing two peer networks within the same modality or learning from self-discrepancy, we initiate co-teaching between the image model and the text model, leveraging the disparities between modalities to foster mutual enhancement.

3 Method

Task Configuration: In GCD, we are given a labeled dataset, denoted as $\mathcal{D}_L = \{(\mathbf{x}_i, y_i^l)\}_{i=1}^M \subseteq \mathcal{X} \times \mathcal{Y}_L$, and an unlabeled dataset, denoted as $\mathcal{D}_U = \{(\mathbf{x}_i, y_i^u)\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}_U$, where M and N indicate the number of samples in the $\mathcal{D}_L$ and $\mathcal{D}_U$, respectively. $\mathcal{Y}_L$ and $\mathcal{Y}_U$ indicate the label spaces for labeled and unlabeled datasets, respectively, where $\mathcal{Y}_L \subseteq \mathcal{Y}_U$. That is, $\mathcal{D}_U$ contains data from unknown categories, and $\mathcal{Y}_U$ is unavailable. The objective of GCD is to leverage the prior knowledge from known categories within $\mathcal{D}_L$ to classify samples in $\mathcal{D}_U$ effectively. Following [38], it is assumed that the number of classes in $\mathcal{D}_U$, represented by $K = |\mathcal{Y}_U|$, is predetermined.

3.1 Overview

The framework of the proposed TextGCD, depicted in Fig. 2, consists of the Retrieval-based Text Generation (RTG) phase and Cross-modal Co-Teaching (CCT) phase. In RTG, we initially develop a visual lexicon comprising a broad spectrum of tags, along with attributes obtained from LLMs. Subsequently, we extract representations of images and the visual lexicon with an auxiliary VLM model and generate textual category information for each image in a retrieval manner. Building upon this, we design textual and visual parametric classifiers and propose the CCT phase to foster mutual learning and collective progress between text and image models.

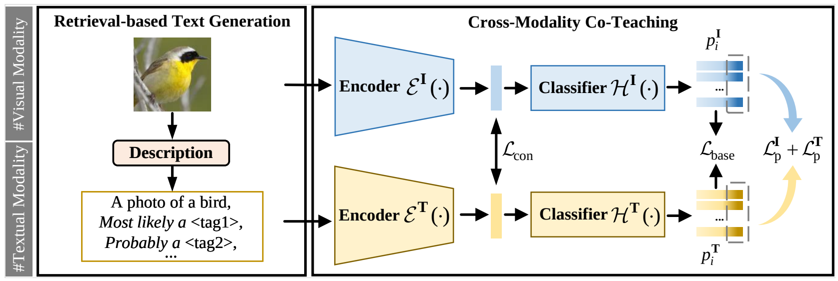

Fig. 2: The TextGCD framework comprises two main phases: Retrieval-based Text Generation (RTG) and Cross-modality Co-Teaching (CCT). In the RTG phase, descriptions for each sample are generated using a visual lexicon. The CCT phase involves developing a two-stream parametric model that leverages the interaction of visual and textual modalities for enhanced mutual progress. The gray dashed box on the right illustrates the text and image models independently selecting high-confidence samples with pseudo labels for the co-teaching process.

3.2 Baseline

We follow the SimGCD [38] to build our parametric learning baseline, which consists of a supervised loss for labeled data and an unsupervised loss for both labeled and unlabeled data.

Specifically, the supervised loss utilizes the conventional cross-entropy loss:

$$\mathcal{L}_{\text{sup}} = \frac{1}{|B^l|} \sum_{i \in B^l} \ell(\boldsymbol{y}_i, \boldsymbol{p}_i), \tag{1}$$

where B^l indicates the mini-batch for labeled data, ℓ is the cross-entropy loss, and $\boldsymbol{p}_i = \sigma(\mathcal{H}(\mathcal{E}(\mathbf{x}_i))/\tau)$ is the predicted probabilities of input $\mathbf{x}_i$. Here, $\mathcal{E}$ represents the backbone encoder, and $\mathcal{H}$ indicates the parametric classifier. $\sigma(\cdot)$ denotes the softmax function, and τ is a temperature parameter set to τ_s.

For unsupervised loss, we use the predicted probabilities of an augmented counterpart $\mathbf{x}'_i$ as the supervision to calculate the classification loss for the original input $\mathbf{x}_i$, which is formulated as:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{|B|} \sum_{i \in B} \ell(\boldsymbol{q}'_i, \boldsymbol{p}_i) - \varepsilon H(\overline{\boldsymbol{p}}), \tag{2}$$

where $\boldsymbol{q}'_i$ is the predicted probabilities of $\mathbf{x}'_i$ using a sharper temperature value τ_t. We also include a mean-entropy maximization regularizer $H(\overline{\boldsymbol{p}})$ [1] for the unsupervised objective. $H(\overline{\boldsymbol{p}}) = -\sum_k \overline{\boldsymbol{p}}^{(k)} \log \overline{\boldsymbol{p}}^{(k)}$, where $\overline{\boldsymbol{p}} = \frac{1}{2|B|} \sum_{i \in B} (\boldsymbol{p}_i + \boldsymbol{p}'_i)$. $\boldsymbol{p}_i$ and $\boldsymbol{p}'_i$ are the probabilities of $\mathbf{x}_i$ and $\mathbf{x}'_i$, respectively, which use the same temperature of τ_u. B indicates the mini-batch for both labeled and unlabeled data. The hyperparameter ε aligns with the configuration used in SimGCD [38].

The classifier is jointly trained with supervised loss and unsupervised loss, formulated as:

$$\mathcal{L}_{\text{base}} = \lambda \cdot \mathcal{L}_{\text{sup}} + (1 - \lambda) \cdot \mathcal{L}_{\text{unsup}}, \tag{3}$$

where λ serves as the balancing factor.

3.3 Retrieval-based Text Generation

The existing foundational VLMs and LLMs, trained on massive amounts of data, have demonstrated remarkable capabilities in aligning visual images with corresponding textual descriptions and generating detailed object descriptions. Drawing inspiration from these advancements, we introduce the Retrieval-based Text Generation (RTG) phase for GCD. This process involves two stages: constructing a comprehensive visual lexicon and retrieving probable tags from this lexicon to generate textual information for images.

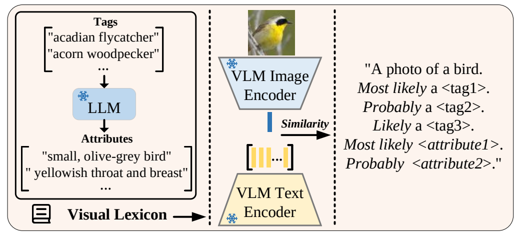

Fig. 3: Schema of retrieval-based text generation.

Building the Visual Lexicon. To construct a visual lexicon, we aggregate a diverse and extensive collection of tags from established benchmarks in semantic segmentation, object detection and classification. These benchmarks are selected for their comprehensive coverage of the visual categories in the world. Additionally, we utilize a large language model (*e.g.*, GPT3 [5]) to further describe each tag by using the prompt of "What are useful features for distinguishing a {tag} in a photo?", which can augment the visual lexicon with the attribute of "{tag} which has the {feature}".

Tag Retrieval. Given the visual lexicon, comprising tags and corresponding attributes, we utilize an auxiliary VLM (*e.g.*, CLIP [30]) to encode it, thereby creating a textual feature bank. For each image, we use the auxiliary model to generate its visual feature, which is then compared against the entries in the textual feature bank using cosine similarity. Subsequently, we identify the most relevant n_t tags and n_a attributes exhibiting the highest similarity to the image. We then concatenate the textual descriptors of these tags and attributes based on the similarity ranking. This process yields the categorical descriptive text $\mathbf{t}_i$ for each image $\mathbf{x}_i$. The workflow of the RTG phase is delineated in Fig. 3.

Discussion. In this paper, we aim to harness the capabilities of the existing foundational models to deliver comprehensive and accurate textual information for GCD. To this end, we employ GPT-3 [5] as the LLM to obtain attributes and the ViT-H-based CLIP [30] as the auxiliary VLM to identify tags and attributes. We also attempt to use a smaller auxiliary VLM, *e.g.,* ViT-B-based CLIP, to generate the descriptive text. However, it demonstrated limited efficacy for fine-grained classification because ViT-B-based CLIP tends to focus on more general Tags and Attributes in the Visual Lexicon. Using Flower102 for instance, ViT-B-based CLIP tends to give high similarity to broader Tags like "Plant & Flower". Compared to existing methods that only use visual cues for GCD, we explore the possibility of introducing the textual information to facilitate the GCD task by using the freely available, public foundational models, which would be a new trend in the community. In addition, our approach enjoys several advantages. Firstly, the RTG is an offline phase, where we only need to process each sample once, avoiding the need for excessive computational overhead. Secondly, given the generated descriptive text, we only need to train the GCD model with a smaller model (ViT-B-based CLIP), achieving significantly higher results than existing methods. Third, the auxiliary VLM can be replaced with other advanced foundational models, such as FLIP [21] and CoCa [42]. Consequently, the effectiveness of our method is likely to be enhanced in line with the advancements in these foundational models.

3.4 Cross-modality Co-teaching

In light of the categorical descriptive texts generated during the RTG phase, the primary challenge lies in effectively utilizing them to train an accurate GCD model. To address this, we develop two-stream parametric classifiers for visual and textual modalities. Recognizing that the intrinsic differences between the modalities naturally meet the requirements for the model disparity in co-teaching strategies [12], we propose a Cross-modal Co-Teaching (CCT) phase to realize the mutual benefits of the visual and textual classifiers. However, a notable challenge arises due to the lack of annotations during the co-teaching process: the classifiers for each modality are often misaligned. This misalignment implies that consistent class indexes across the two modalities cannot be guaranteed, posing a significant obstacle to effective co-teaching. We introduce two preliminary stages to overcome this issue: a warm-up stage and a class-aligning stage, designed to establish modality-specific classifiers while ensuring their alignment.

Basic Loss. Given the inputs from two modalities, $\mathbf{x}_i$ and $\mathbf{t}_i$, we construct a text model and an image model based on the parametric baseline in Sec. 3.2. Thus, the basic loss for our method is formulated as:

$$\mathcal{L}_{\text{base}} = \mathcal{L}^{\mathbf{I}}_{\text{base}} + \mathcal{L}^{\mathbf{T}}_{\text{base}}, \tag{4}$$

where both components are implemented by Eq. 3. The difference is that $\mathcal{L}^{\mathbf{T}}_{\text{base}}$ is calculated by the text model and textual description.

Image-Text Contrastive Learning. To strengthen the association between the image and text modalities, we further introduce a cross-modal contrastive learning between image features and textual features, which is formulated as:

$$\mathcal{L}_{\text{con}} = \frac{1}{|B|} \sum_{i \in B} -\log \frac{\exp\left(\mathbf{f}^{\mathbf{I}}_i \cdot (\mathbf{f}^{\mathbf{T}}_i)^\top / \tau_c\right)}{\sum_j \mathbb{I}_{[j \neq i]} \exp\left(\mathbf{f}^{\mathbf{I}}_i \cdot (\mathbf{f}^{\mathbf{T}}_j)^\top / \tau_c\right)}, \tag{5}$$

where τ_c is a temperature parameter. $\mathbf{f}^{\mathbf{I}}_i = \mathcal{E}^{\mathbf{I}}(\mathbf{x}_i)$ is the image feature and $\mathbf{f}^{\mathbf{T}}_i = \mathcal{E}^{\mathbf{T}}(\mathbf{t}_i)$ is the textual feature. $\mathcal{E}^{\mathbf{I}}$ and $\mathcal{E}^{\mathbf{T}}$ represent the image and text encoders, respectively. We next introduce the three stages of the proposed CCT (see Fig. 4), *i.e.,* warm-up, class-aligning, and co-teaching.

Stage I: Warm-up. The image and text models initially undergo a warm-up training stage with e_w epochs, as depicted in Fig 4(a). During this stage, both the text model and image model are trained using the basic loss (Eq. 4) and cross-modal contrastive loss (Eq. 5). This warm-up training aims to enable both models to adequately learn category knowledge from their respective modality data, thereby developing modality-specific category perceptions.

Stage II: Class-aligning. Subsequently, to rectify the class misalignment between image and text classifiers, a class-aligning stage with e_a epochs is introduced, as illustrated in Fig 4(b). This stage aims to align the classifiers of two modalities. To this end, we select high-confidence samples based on the text model and use them to guide the training of the image model. Specifically, for each category k, we select the top s samples that have the highest probabilities in the k-th category produced by the text model, formulated as $\mathrm{Top}_k^{\mathbf{T}}(s)$. In addition, for each sample i in $\mathrm{Top}_k^{\mathbf{T}}(s)$, we assign it with the hard pseudo label of $\hat{y}_i^{\mathbf{T}} = k$. Thus, besides the basic and contrastive losses, we additionally train the image model by the pseudo-labeling loss:

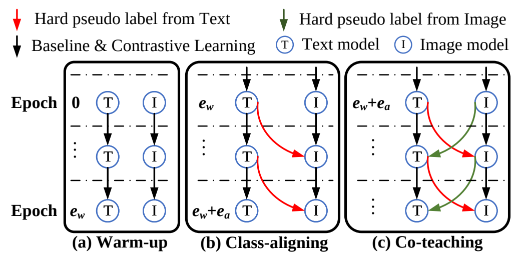

Fig. 4: Schema of cross-modality co-teaching.

$$\mathcal{L}_{\mathrm{p}}^{\mathbf{I}} = \frac{1}{|B^s|} \sum_{i \in B^s} \ell(\hat{y}_i^{\mathbf{T}}, \boldsymbol{p}_i^{\mathbf{I}}), \tag{6}$$

where $|B^s|$ indicates the number of selected samples in the mini-batch.

Stage III: Co-teaching. Following the warm-up and class-aligning stages, we establish cross-modality co-teaching to facilitate the mutual benefit between the image and text models, as shown in Fig 4(c). Similar to the class-aligning stage, we also generate high-confidence samples from the image model, which is defined as $\mathrm{Top}_k^{\mathbf{I}}(s)$. For each sample i in $\mathrm{Top}_k^{\mathbf{I}}(s)$, the hard pseudo label is $\hat{y}_i^{\mathbf{I}} = k$, and thus the pseudo-labeling loss for the text classifier is formulated as:

$$\mathcal{L}_{\mathrm{p}}^{\mathbf{T}} = \frac{1}{|B^s|} \sum_{i \in B^s} \ell(\hat{y}_i^{\mathbf{I}}, \boldsymbol{p}_i^{\mathbf{T}}). \tag{7}$$

The total loss of Stage III is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \mathcal{L}_{\mathrm{con}} + \mathcal{L}_{\mathrm{p}}^{\mathbf{I}} + \mathcal{L}_{\mathrm{p}}^{\mathbf{T}}. \tag{8}$$

Inference. When determining the category of object i, we employ soft voting to merge the category perceptions from both modalities:

$$\boldsymbol{P}_i = \boldsymbol{p}_i^{\mathbf{I}} + \boldsymbol{p}_i^{\mathbf{T}}. \tag{9}$$

The predicted category is the one that achieves the largest value in $\boldsymbol{P}_i$, *i.e.*, $\arg\max \boldsymbol{P}_i$. This integrated approach leverages the strengths of both modalities, leading to more accurate classification.

4 Experiments

Datasets. We conduct comprehensive experiments on four generic image classification datasets, *i.e.*, CIFAR-10, CIFAR-100 [18], ImageNet-100, and ImageNet-1K [8], as well as on four fine-grained image classification datasets, *i.e.*, CUB [36], Stanford Cars [15], Oxford Pets [28], and Flowers102 [25]. Following [35], for each dataset, we select half of the classes as the known classes and the remaining as the unseen classes. We select 50% of samples from the known classes as the labeled dataset $\mathcal{D}_L$ and regard the remaining samples as unlabeled dataset $\mathcal{D}_U$ that are from both known and unseen classes.

Evaluation Protocol. Following [35], we use the clustering accuracy (ACC) to evaluate the performance of each algorithm. Specifically, the ACC is calculated on the unlabeled data, formulated as $ACC = \frac{1}{|\mathcal{D}_U|} \sum_{i=1}^{|\mathcal{D}_U|} \mathbb{I}(y_i = C(\overline{y}_i))$, where $\overline{y} = \arg\max \boldsymbol{P}$ represents the predicted labels and y denotes the ground truth. The C denotes the optimal permutation.

Implementation Details. We construct the tag lexicon by incorporating tags from semantic segmentation and object detection datasets [11, 19, 22], image classification datasets [4, 7, 16, 20, 28, 33, 39] and visual genome dataset [17]. The CLIP [30] model with a ViT-B-16 architecture serves as the backbone, ensuring a fair comparison to methods utilizing the DINO [6] model with the same architecture. A linear classifier layer is added after the backbone for each modality. During training, only the last layers of both the text and image encoders in the backbone are fine-tuned, using a learning rate of 0.0005. The learning rates for classifiers start at 0.1 and decrease following a cosine annealing schedule. We train the model for 200 epochs on all datasets, using a batch size of 128 and processing dual views of randomly augmented images and texts. The temperature parameters τ_s and τ_u are set at 0.1 and 0.05, respectively. τ_t starts at 0.035 and decreases to 0.02, except for CIFAR-10 where it varies from 0.07 to 0.04 due to fewer classes. We use the logit scale value to set τ_c as in the CLIP model. The balancing factor λ is set to 0.2. Warm-up and class-aligning stages last for 10 and 5 epochs, respectively, denoted as $e_w = 10$ and $e_a = 5$. For selecting high-confidence samples, we choose $s = r * \frac{|\mathcal{D}_U|}{K}$ samples per category, with r set at 0.6. The average result over the three runs is reported.

4.1 Comparison with the State of the Art

We conduct a comparative analysis between our proposed TextGCD and leading methods in GCD, as delineated in Tab. 1 and Tab. 2. These methods encompass GCD [35], SimGCD [38], DCCL [29], GPC [45], PromptCAL [44], CLIP-GCD [27] and SPTNet [37]. Except for CLIP-GCD [27], which uses CLIP as the backbone, the others utilize the DINO as the backbone. To better ensure the fairness of our comparison, we also reproduce SimGCD by using CLIP as the backbone. Clearly, our TextGCD outperforms all compared methods in terms of All accuracy on all datasets. Importantly, compared to the best competitor, SimGCD, our method achieves a significant improvement on the ImageNet-1K and all the fine-grained datasets. Specifically, our method outperforms SimGCD by 7.7% in All accuracy and 11.4% in New accuracy on the ImageNet-1K dataset, and, by 11.1% in All accuracy and 12.7% in New accuracy averaged on fine-grained datasets. This demonstrates the superiority of our method over existing visual-based methods. In comparison to CLIP-GCD, which merely concatenates features from both modalities, TextGCD achieves higher results. For instance, TextGCD surpasses CLIP-GCD by 13.8%, 16.3%, and 10.9% in terms of All accuracy on the CUB, Stanford Cars, and Flowers102 datasets, respectively.

Table 1: Results on generic datasets. The best results are highlighted in **bold**.

Methods	Backbone	CIFAR-10			CIFAR-100			ImageNet-100			ImageNet-1K		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD [35]	DINO	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3	52.5	72.5	42.2
SimGCD [38]	DINO	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9	57.1	77.3	46.9
DCCL [29]	DINO	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2	-	-	-
GPC [45]	DINO	92.2	**98.2**	89.1	77.9	85.0	63.0	76.9	94.3	71.0	-	-	-
PromptCAL [44]	DINO	97.9	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3	-	-	-
SPTNet [37]	DINO	97.3	95.0	**98.6**	81.3	84.3	75.6	85.4	93.2	81.4	-	-	-
CLIP-GCD [27]	CLIP	96.6	97.2	96.4	85.2	85.0	**85.6**	84.0	**95.5**	78.2	-	-	-
SimGCD [38]	CLIP	96.6	94.7	97.5	81.6	82.6	79.5	86.1	94.5	81.9	48.2	72.7	36.0
TextGCD	CLIP	**98.2**	98.0	**98.6**	**85.7**	**86.3**	84.6	**88.0**	92.4	**85.2**	**64.8**	**77.8**	**58.3**

Table 2: Results on fine-grained datasets. The best results are highlighted in **bold**.

Methods	Backbone	CUB			Stanford Cars			Oxford Pets			Flowers102		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD [35]	DINO	51.3	56.6	48.7	39.0	57.6	29.9	80.2	85.1	77.6	74.4	74.9	74.1
SimGCD [38]	DINO	60.3	65.6	57.7	53.8	71.9	45.0	87.7	85.9	88.6	71.3	80.9	66.5
DCCL [29]	DINO	63.5	60.8	64.9	43.1	55.7	36.2	88.1	88.2	88.0	-	-	-
SPTNet [37]	DINO	65.8	68.8	65.1	59.0	79.2	49.3	-	-	-	-	-	-
CLIP-GCD [27]	CLIP	62.8	77.1	55.7	70.6	**88.2**	62.2	-	-	-	76.3	88.6	70.2
SimGCD [38]	CLIP	62.0	76.8	54.6	75.9	81.4	73.1	88.6	75.2	95.7	75.3	87.8	69.0
TextGCD	CLIP	**76.6**	**80.6**	**74.7**	**86.9**	87.4	**86.7**	**95.5**	**93.9**	**96.4**	**87.2**	**90.7**	**85.4**

4.2 Ablation Study

Components Ablation. In Tab. 3, we present an ablation study of the key components in TextGCD, specifically cross-modality co-teaching, contrastive learning, and soft voting. Beginning with the Baseline method, we incrementally incorporate these three components into the framework to evaluate their impact on the performance of both image and text classifiers. We also report the results of 1) employing the first tag derived from the visual lexicon as the GCD labels and 2) performing k-means [23] using the image or text features without training. It becomes evident that the "First-Tag" and "k-means" methods yield inferior

Table 3: Ablation study on TextGCD components. "T" and "I" denote the output results of the text and image classifiers, respectively. "First-Tag" refers to the accuracy achieved using the most similar tag from the visual lexicon as the label. "k-means" indicates applying k-means on the initial backbone features.

Methods	Classifier	CIFAR-100			CUB		
		All	Old	New	All	Old	New
First-Tag	-	12.0	12.0	11.9	47.8	43.4	52.1
k-means	T	70.2	66.6	77.5	67.6	64.5	69.2
	I	46.5	45.6	48.5	46.7	50.6	44.7
Baseline	T	81.8	83.6	78.3	67.0	75.9	62.6
	I	76.3	80.9	67.0	49.0	64.2	41.3
+Co-Teaching	T	84.4	86.2	81.1	74.4	79.3	72.0
	I	82.4	85.0	77.3	71.7	79.4	67.8
+Contrastive	T	84.6	84.8	84.1	74.8	78.6	73.0
	I	83.1	84.0	81.1	73.9	80.4	70.7
+Soft Voting	-	**85.7**	**86.3**	**84.6**	**76.6**	**80.6**	**74.7**

results compared to the baseline. On the other hand, the "k-means" method and the baseline results demonstrate that using the textual information achieves clearly higher results than using the image cues, thereby highlighting the superior quality of the generated textual descriptions. Compared to the baseline, integrating our proposed co-teaching approach markedly elevates the performance across both modalities, particularly notable within the context of the fine-grained CUB dataset. This substantiates the efficacy of our co-teaching strategy in fostering synergistic learning and collaborative advancement between text and image models. Furthermore, image-text contrastive learning consistently increases the All and New accuracies, underlining the necessity of aligning the two modalities. Soft voting can further improve performance, illustrating that the integration of information from both modalities leads to more precise category discrimination.

Table 4: Ablation study on TextGCD training stages. "Warm-up" refers to the warm-up stage and "Cls. Align" denotes the class-aligning stage. "T→I" signifies aligning the image model using the text model, whereas "I→T" indicates aligning the text model using the image model.

Warm-up	Cls. Align		Co-Teaching	CIFAR-100			ImageNet-100			CUB			Stanford Cars		
	I→T	T→I		All	Old	New	All	Old	New	All	Old	New	All	Old	New
			✓	83.3	86.8	76.4	80.6	92.8	74.5	74.8	77.4	73.5	83.9	88.1	80.6
✓			✓	82.8	**87.2**	74.1	81.6	92.6	76.0	76.5	**81.2**	74.1	85.3	**87.8**	84.0
✓	✓		✓	84.6	86.6	80.6	87.5	**93.5**	84.4	76.1	80.9	73.6	**87.5**	86.4	**87.9**
✓		✓	✓	**85.7**	86.3	**84.6**	**88.0**	92.4	**85.2**	**76.6**	80.6	**74.7**	86.9	87.4	86.7

Training Stages Ablation. In Tab. 4, we assess the impact of each stage within the proposed CCT phase, *i.e.,* warm-up, class-aligning, and co-teaching. Results show that using co-teaching solely achieves lower results than our full method. Specifically, the warm-up process is important for the fine-grained dataset, *i.e.*, CUB and SCars. On the other hand, the class-aligning consistently improves the performance on all four datasets, which obtains 10.5%, 9.2%, 0.6%, and 2.7% in New accuracy on CIFAR-100, ImageNet-100, CUB and SCars, respectively. This observation highlights the vital role of class-aligning stage in resolving conflicts between the two classifiers. To further investigate the proposed class-aligning strategy, we evaluate the variant of using the image model to guide the text model, denoted as "I→T". In terms of All and New accuracies, we observe that 1) both strategies can achieve consistent improvements (except for "I→T" on CUB) and that 2) the proposed text-guide-image strategy produces higher results than the image-guide-text variant in most cases. These advantages can be attributed to the more representative initial features generated by the textual descriptions. Therefore, the discriminative textual features can generate high-quality pseudo-labels for effectively guiding visual model (text-guide-image) or can be robust to noisy labels when guided by the visual model (image-guide-text). This further demonstrates the importance of introducing textual modality into our co-teaching framework.

4.3 Evaluation

Backbone Evaluation. In Tab. 5, we evaluate the impact of the image encoder in TextGCD by substituting the image encoder of CLIP with that of DINO, while maintaining the text encoder from CLIP. In TextGCD(DINO), we omit the cross-modal contrastive loss $\mathcal{L}_{\text{con}}$. Despite a decreased accuracy on CUB, TextGCD(DINO) still significantly outperforms SimGCD, which utilizes the same image encoder. Moreover, we compare SimGCD and TextGCD using

Table 5: Evaluation on different pretrained ViT-B backbones.

Methods	Backbone	CIFAR-100			CUB		
		All	Old	New	All	Old	New
SimGCD	DINO	80.1	81.2	77.8	60.3	65.6	57.7
TextGCD	DINO	**86.1**	**88.7**	81.0	73.7	80.3	70.4
TextGCD	CLIP	85.7	86.3	**84.6**	**76.6**	**80.6**	**74.7**

the ViT-H-based CLIP for a more fair comparison, as we use the ViT-H for text generation. Tab. 6 shows that our TextGCD outperforms SimGCD by a large margin on both datasets. Interestingly, both SimGCD and our method show a decrease in performance on the generic dataset CIFAR-100, in terms of New accuracy, compared to the ones using ViT-B-16 as the backbone (see Tab. 1). This suggests that fine-tuning on the generic dataset may adversely affect the ability of the large model. In addition, we can observe a small margin between our methods with ViT-B-16 and ViT-H-14. This indicates that our advantage mainly benefited from the high-quality text description and the co-teaching strategy, enabling our approach to adapt to both smaller and larger models.

Table 6: Evaluation on ViT-H-14 backbone.

Methods	CLIP Backbone	CIFAR-100			CUB		
		All	Old	New	All	Old	New
SimGCD	ViT-H-14	78.1	80.0	74.4	69.1	76.3	65.4
TextGCD	ViT-H-14	**86.4**	**89.3**	**80.7**	**78.6**	**81.5**	**77.1**

Text Generation Evaluation. In Tab. 7, we analyze the impact of the number of tags and attributes for constructing category descriptions in the RTG phase. Results show that richer textual content (more tags and attributes) leads to higher category recognition accuracy. Due to the token limitation of CLIP, we can use up to only three tags and two attributes in the input of the text model. We will handle this drawback and investigate the impact of including more tags and attributes in future work.

Table 7: Evaluation on the number of tags and attributes.

# Tag	# Attribute	CIFAR-100			CUB		
		All	Old	New	All	Old	New
1	0	81.6	81.6	81.8	67.9	69.6	67.0
2	0	83.5	83.4	83.7	68.6	75.3	65.3
3	0	83.7	84.7	81.6	69.9	74.9	67.3
3	1	84.7	85.4	83.4	74.7	80.5	71.7
3	2	**85.7**	**86.3**	**84.6**	**76.6**	**80.6**	**74.7**

Auxiliary Model Evaluation. In Tab. 8, we explore the effect of substituting the auxiliary model with FLIP [21] and CoCa [42]. Results show that our method consistently produces high performance on both datasets and that using a more powerful model (FLIP) leads to higher results.

Table 8: Evaluation on the auxiliary model.

Auxiliary Model	Backbone	CIFAR-100			CUB		
		All	Old	New	All	Old	New
CLIP [30]	ViT-H-14	85.7	86.3	84.6	76.6	80.6	74.7
CoCa [42]	ViT-L-14	85.2	85.8	83.8	73.6	81.7	69.5
FLIP [21]	ViT-G-14	**87.6**	**87.7**	**87.5**	**79.5**	**83.5**	**77.4**

Co-Teaching Strategy Evaluation. To demonstrate the effectiveness of the cross-modal co-teaching strategy, we conduct a comparative analysis with a co-teaching approach utilizing a single modality. Results in Tab. 9 show that single-modality co-teaching variants obtain lower results than the proposed cross-modality co-teaching. This finding highlights the critical role of ensuring the model diversity in co-teaching.

Table 9: Evaluation on co-teaching strategies. "I↔T" is our cross-modality co-teaching. "T↔T" and "I↔I" indicate the single-modality variant using text modality and image modality, respectively.

Methods	CIFAR-100			CUB		
	All	Old	New	All	Old	New
T↔T	79.8	83.6	72.1	71.7	73.0	71.1
I↔I	75.5	80.5	65.7	60.3	79.5	50.7
I↔T (Ours)	**85.7**	**86.3**	**84.6**	**76.6**	**80.6**	**74.7**

4.4 Hyper-parameters Analysis

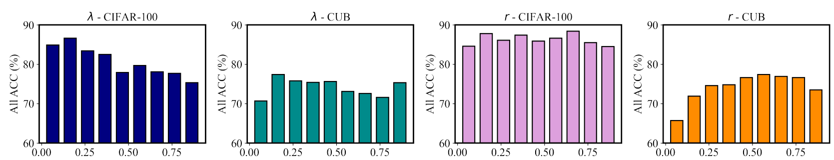

Fig. 5: Impact of hyper-parameters.

Balancing Factor λ and Proportion Coefficient r. The impact of the balancing factor λ and the proportion coefficient r is illustrated in Fig. 5. We use All accuracy as the evaluation metric. An increase in λ suggests that the classifiers of both models increasingly prioritize labeled data from known categories. Results show that $\lambda = 0.2$ achieves the best accuracy on both datasets. The parameter r determines the fraction of high-confidence samples chosen from each category for the text and image models. A higher value of r is typically advantageous, and the best result is achieved when $r = 0.6$. Note that we use the same hyper-parameters (λ and r) for all datasets to avoid over-tuning.

5 Conclusion

In this paper, we introduce a novel approach called TextGCD for Generalized Category Discovery. Compared to previous methods that only rely on visual patterns, we additionally integrate textual information into the framework. Specifically, we first construct a visual lexicon based on off-the-shelf category tags and enrich it with attributes by Large Language Models (LLMs). Following this, a Retrieval-based Text Generation (RTG) approach is proposed to assign each sample with likely tags and corresponding attributes. Building upon this, we design a Cross-modal Co-Teaching (CCT) strategy to fully take the mutual benefit between visual and textual information, enabling robust and complementary model training. Experiments on eight benchmarks show that our TextGCD produces new state-of-the-art performance. We hope this study could bring a new perspective, which assists the framework with foundational model and textual information, for the GCD community.

Acknowledgement

We acknowledge the support of the MUR PNRR project iNEST-Interconnected Nord-Est Innovation Ecosystem (ECS00000043) funded by the NextGenerationEU. Also, this work was partially supported by the EU Horizon projects ELIAS (No. 101120237) and AI4Trust (No. 101070190).

References

1. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning. In: European Conference on Computer Vision (2022)
2. Berrios, W., Mittal, G., Thrush, T., Kiela, D., Singh, A.: Towards language models that can see: Computer vision through the lens of natural language. arXiv preprint arXiv:2306.16410 (2023)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100 (1998)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision (2014)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
7. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2014)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
11. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
12. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in Neural Information Processing Systems (2018)
13. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Autonovel: Automatically discovering and learning novel visual categories. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

14. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2013)
16. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2013)
17. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (2017)
18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical Report (2009)
19. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision (2020)
20. Li, F.F., Andreeto, M., Ranzato, M., Perona, P.: Caltech 101 (2022)
21. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014)
23. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967)
24. Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022)
25. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (2008)
26. Novack, Z., McAuley, J., Lipton, Z.C., Garg, S.: Chils: Zero-shot image classification with hierarchical label sets. In: International Conference on Machine Learning. PMLR (2023)
27. Ouldnoughi, R., Kuo, C.W., Kira, Z.: Clip-gcd: Simple language guided generalized category discovery. arXiv preprint arXiv:2305.10420 (2023)
28. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2012)
29. Pu, N., Zhong, Z., Sebe, N.: Dynamic conceptional contrastive learning for generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)

31. Roy, S., Krivosheev, E., Zhong, Z., Sebe, N., Ricci, E.: Curriculum graph co-teaching for multi-target domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
32. Roy, S., Liu, M., Zhong, Z., Sebe, N., Ricci, E.: Class-incremental novel class discovery. In: European Conference on Computer Vision (2022)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (2015)
34. Sloutsky, V.M.: From perceptual categories to concepts: What develops? Cognitive science (2010)
35. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
36. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Computation & Neural Systems Technical Report (2011)
37. Wang, H., Vaze, S., Han, K.: Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In: The Twelfth International Conference on Learning Representations (2023)
38. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
39. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2010)
40. Yang, F., Li, K., Zhong, Z., Luo, Z., Sun, X., Cheng, H., Guo, X., Huang, F., Ji, R., Li, S.: Asymmetric co-teaching for unsupervised cross-domain person re-identification. In: Proceedings of the AAAI conference on artificial intelligence (2020)
41. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
42. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
43. Yuan, Y., Chen, C.S., Liu, Z., Neiswanger, W., Liu, X.S.: Importance-aware co-teaching for offline model-based optimization. Advances in Neural Information Processing Systems (2024)
44. Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.S.: 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 (2023)
45. Zhao, B., Wen, X., Han, K.: Learning semi-supervised gaussian mixture models for generalized category discovery. arXiv preprint arXiv:2305.06144 (2023)
46. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
47. Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N.: Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)