Consistent Prompt Tuning for Generalized Category Discovery

Muli Yang[1], Jie Yin[2], Yanan Gu[3], Cheng Deng[2*], Hanwang Zhang[4], Hongyuan Zhu[1]

[1]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore.
[2]School of Electronic Engineering, Xidian University, Xi'an, China.
[3]Norinco Group Testing and Research Institute, Xi'an, China.
[4]School of Computer Science and Engineering, Nanyang Technological University, Singapore.

*Corresponding author(s). E-mail(s): chdeng@mail.xidian.edu.cn;
Contributing authors: yangml@i2r.a-star.edu.sg; yinjie@stu.xidian.edu.cn;
yanangu.xd@gmail.com; hanwangzhang@ntu.edu.sg; zhuh@i2r.a-star.edu.sg;

Abstract

Generalized Category Discovery (GCD) aims at discovering both known and unknown classes in unlabeled data, using the knowledge learned from a limited set of labeled data. Despite today's foundation models being trained with Internet-scale multi-modal corpus, we find that they still struggle in GCD due to the ambiguity in class definitions. In this paper, we present Consistent Prompt Tuning (CPT) to disambiguate the classes for large vision-language models (*e.g.*, CLIP). To this end, CPT learns a set of "task + class" prompts for labeled and unlabeled data of both known and unknown classes, with the "task" tokens globally shared across classes, which contain a unified class definition pattern, *e.g.*, "the foreground is an animal named" or "the background scene is". These prompts are optimized with two efficient regularization techniques that encourage consistent global and local relationships between any two matched inputs. CPT is evaluated on various existing GCD benchmarks, as well as in new practical scenarios with fewer annotations and customized class definitions, demonstrating clear superiority and broad versatility over existing state-of-the-art methods. Code will be made available.

Keywords: Category Discovery, Prompt Learning, Multimodal Learning, Transfer Learning

1 Introduction

The deep learning era has witnessed tremendous success in the *closed world* where the unlabeled data is assumed to belong to known classes (Krishna et al., 2017; Kuznetsova et al., 2020; Lin et al., 2014; Russakovsky et al., 2015). However, this assumption limits the applicability of deep models in the *open-world* scenario (Zhu et al., 2024), where the unlabeled data is no longer restricted to known classes. In response to this challenge, *Generalized Category Discovery* (GCD) (Vaze et al., 2022a) has been introduced to efficiently handle unlabeled data in the open world, by clustering unlabeled data (which may contain both known and unknown classes) based on semantic classes. As shown in Fig. 1, distinguishing between a "class" (*e.g.*, `dog`) and the "class-irrelevant" content (*e.g.*, `snow`) is crucial for GCD, which requires fully leveraging class definitions found in labeled data and transferring that

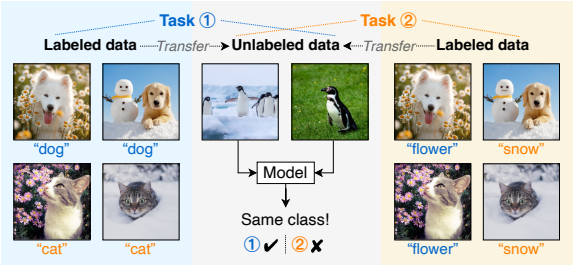

Fig. 1 Generalized Category Discovery (GCD) aims to cluster unlabeled data using the knowledge in labeled data. Although recent VLMs are trained with Internet-scale corpus, they may still struggle in GCD since different class definitions (as depicted in the two different tasks) lead to distinct clustering results.

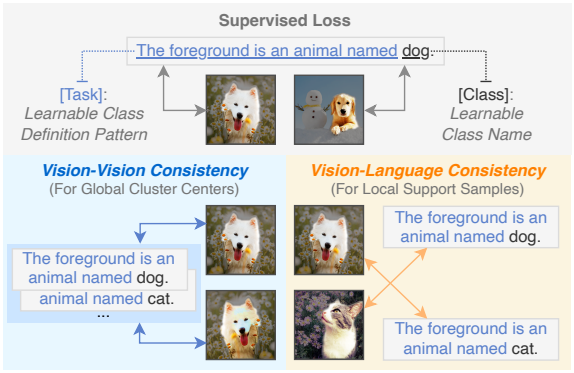

Fig. 2 Conceptual illustration of "task + class" prompts and CPT objective. The supervised loss is applied to labeled data, whereas *vision-vision consistency* (VVC) and *vision-language consistency* (VLC) are applied to both labeled and unlabeled data. See Sec. 3.2 for more details.

knowledge to unknown classes in the unlabeled data.

Undoubtedly, we are now in an era dominated by large foundation models (Alayrac et al., 2022; Brown et al., 2020; Devlin et al., 2019; Radford et al., 2021; Wu et al., 2023a; Yuan et al., 2021). Given the fact that these models are pre-trained on Internet-scale data, has GCD lost its significance? No. Indeed, thanks to the extensive aligned multi-modal knowledge, vision-language models (VLMs) have been born with the surprising zero-shot inference ability. By leveraging predefined text descriptions that specify classes of interest, VLMs can effortlessly recognize unseen images through the comparison of visual and textual features. However, recent findings (Zhou et al., 2022a,b) reveal that the effectiveness of this zero-shot ability highly relies on meticulously crafted text prompts, highlighting the necessity of learning-based prompt tuning. In essence, the ability to "see" does not equate to "knowing"—the diversity in visual content and the ambiguity in class definitions can both impede the superpower of VLMs across various downstream applications (Ren et al., 2023a). Moreover, this challenge becomes more pronounced in the open-world scenario where the classes of interest are even unknown, presenting a greater obstacle in the adaptation of these models for GCD.

In this paper, we seek to disambiguate the class definitions for GCD, leveraging recent advances of VLMs, *e.g.*, CLIP (Radford et al., 2021). For the first time, we propose to learn a set of class descriptions with "task + class" prompts for both known and unknown classes. The "task" tokens are shared for all classes, which can be interpreted as prompting the class definition pattern (*e.g.*, `the foreground is an animal named`) that excludes the class-irrelevant information, hence allowing the "class" tokens to focus on the specific class content (*e.g.*, `dog`). Since the unlabeled data contains novel classes with unknown class names, we initialize the class tokens with random vectors, and tune both task and class tokens to easily adapt CLIP to various GCD tasks with diverse class definitions, instead of fine-tuning the whole model which is computationally prohibitive and prone to distorting the pre-trained knowledge (Kumar et al., 2022).

With the above goal in mind, another question arises: how do we precisely find these prompts for both known and unknown classes without supervision for unlabeled data? We start with the supervision for labeled data of known classes, which helps identify the class definition pattern shared across classes, and capture this knowledge into the "task" tokens. Combined with different learnable "class" tokens that specifically focus on the class names, we can simultaneously recognize known classes and discover unknown ones under the guidance of one unified class definition pattern. As shown in Fig. 2, we propose two consistency regularization techniques to facilitate prompt learning on both labeled and unlabeled data with high efficiency. The "consistency" here refers to the consistent *multi-modal embedding similarity* between any two matched inputs with the same semantic information, in harmony

with CLIP's original training objective that optimizes such similarity for each image-text pair. In particular, *vision-vision consistency* (VVC) is introduced to encourage the consistency of any two matched image inputs *w.r.t. global* "cluster centers", *i.e.*, all the class description prompts. On the other hand, *vision-language consistency* (VLC) is devised to encourage the consistency of any two matched image and text inputs *w.r.t.* a set of *local* support samples (*i.e.*, a training batch; see Sec. 3.2). VVC and VLC constitute strong and complementary global and local regularization for multi-modal relationships, which can be seamlessly integrated into the joint training pipeline for labeled and unlabeled data of both known and unknown classes, enabling high efficacy and broad versatility across various GCD tasks. Our proposed method, dubbed *Consistent Prompt Tuning* (CPT), has been evaluated on a range of GCD benchmarks spanning generic and fine-grained image recognition tasks, as well as in several new practical scenarios with fewer annotations and customized class definitions, achieving up to 10+% accuracy improvement over state-of-the-art methods.

In a nutshell, our main contributions are

1. Successful adaptation of a VLM for Generalized Category Discovery (GCD), by transforming the problem into finding a set of "task + class" prompts to flexibly disambiguate the class definitions for the first time;
2. Two consistency regularization techniques, which align global and local relationships for the two modalities, enabling accurate prompt learning on both labeled and unlabeled data with high efficacy and broad versatility;
3. Extensive experiments on several GCD benchmarks in diverse settings, showing the clear superiority of our proposed method over existing state-of-the-art approaches.

2 Related Work

Generalized Category Discovery. Generalized Category Discovery (GCD) (Vaze et al., 2022a) aims to cluster unlabeled data using the knowledge learned from the labeled data, which assumes that the unlabeled data contains both known and unknown classes (An et al., 2023b,c; Banerjee et al., 2024; Chiaroni et al., 2023; Hao et al., 2023; Otholt et al., 2024; Ouldnoughi et al., 2023; Pu et al., 2023; Rastegar et al., 2023; Vaze et al., 2023; Wang et al., 2024b; Wen et al., 2023; Yang et al., 2023b; Zhang et al., 2023b; Zhao et al., 2023). It was newly introduced as a generalized setting of Novel Class Discovery (NCD) (Han et al., 2019; Hsu et al., 2018, 2019; Troisemaine et al., 2023a), which originally assumes that the unlabeled data only contains unknown classes (Fini et al., 2021; Gu et al., 2023; Han et al., 2020, 2022; Hasan et al., 2023b; Joseph et al., 2022b; Li et al., 2023b,c; Wang et al., 2024d, 2020; Yang et al., 2022; Zhang et al., 2022a,b; Zhao and Han, 2021; Zhong et al., 2021a,b). In addition, category discovery can also be found in various practical scenarios (An et al., 2023a, 2022; Chi et al., 2022; Du et al., 2023; Fan et al., 2024; Fei et al., 2022; Feng et al., 2024, 2023; Gao et al., 2023a; Han et al., 2023; Hasan et al., 2023a; Hayes et al., 2024; Hogan et al., 2023; Jia et al., 2021b; Li et al., 2023f, 2022; Liu et al., 2024; Peng et al., 2023; Riz et al., 2023; Sun and Li, 2023; Sun et al., 2023b; Troisemaine et al., 2022, 2023b; Wang et al., 2024c, 2023b; Weng et al., 2024; Zhao et al., 2022; Zheng et al., 2022), such as semi-supervised learning (Cao et al., 2022; Guo et al., 2022; Liu et al., 2023a; Rizve et al., 2022a,a,b; Sun et al., 2024; Wang et al., 2024g; Ye et al., 2023; Zhang et al., 2023a), multi-domain learning (Yu et al., 2022; Zang et al., 2023; Zhuang et al., 2022), class-incremental learning (Chen et al., 2024; Joseph et al., 2022a; Kim et al., 2023; Liu et al., 2023b; Liu and Tuytelaars, 2022; Marczak et al., 2023; Roy et al., 2022; Wu et al., 2023b; Zhang et al., 2022d; Zhao and Mac Aodha, 2023), long-tailed learning (Bai et al., 2023; Chuyu et al., 2023; Li et al., 2023d,e; Yang et al., 2023a), active learning (Ma et al., 2024), and federated learning (Pu et al., 2024).

Most existing GCD methods resort to fine-tuning a large vision model, *e.g.*, ViT (Dosovitskiy et al., 2021) (pre-trained with DINO (Caron et al., 2021)), usually followed by a second-stage non-parametric clustering process (Chiaroni et al., 2023; Pu et al., 2023; Vaze et al., 2022a; Zhang et al., 2023b), which demand high training costs and often suffer from class ambiguities. In this paper, we introduce a parameter-efficient learning framework for GCD by leveraging the capabilities of large pre-trained vision-language models

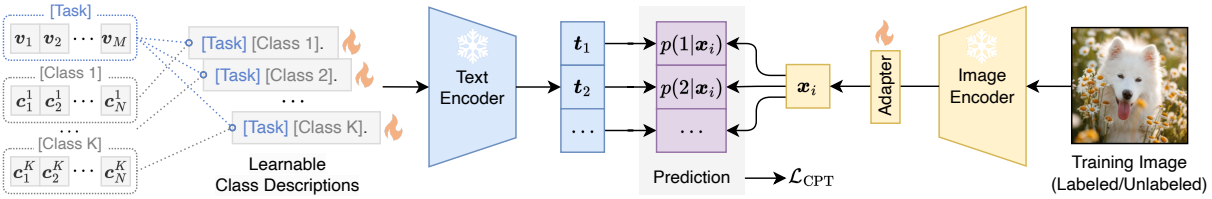

Fig. 3 Overall pipeline of our proposed *Consistent Prompt Tuning* (CPT) with the frozen text and image encoders of CLIP (Radford et al., 2021). CPT tunes K learnable "task + class" prompts and a lightweight adapter for labeled and unlabeled data of both known and unknown classes.

(VLMs), such as CLIP (Radford et al., 2021). Our approach reformulates the problem as identifying a set of "task + class" prompts, enabling flexible and effective disambiguation of class definitions for the first time.

Vision-Language Models. Benefiting from the multi-modal representation ability trained on Internet-scale data, vision-language models (VLMs) like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021a) have demonstrated surprising generalization ability in various downstream applications, such as zero-shot learning (Goel et al., 2022; Nayak et al., 2023; Wang et al., 2023a; Wortsman et al., 2022) and few-shot learning (Khattak et al., 2023; Zang et al., 2022; Zhang et al., 2022c). Such success can be mainly credited to powerful parameter-efficient fine-tuning methods like prompt tuning (Liu et al., 2023c), *i.e.*, automatically finding soft text prompts that help VLMs to better understand the task (Ge et al., 2023; Lester et al., 2021; Li and Liang, 2021; Menon et al., 2023; Qin and Eisner, 2021; Shin et al., 2020; Wang et al., 2024f; Zhou et al., 2022a,b).

These advancements pave the way for further progress in adapting VLMs to a range of open-world challenges. Notable developments include zero-shot adaptation (Li et al., 2023a), which enables the unsupervised adaptation of CLIP to novel tasks, open-vocabulary learning (Ren et al., 2023b), which introduces a task-agnostic soft prompt for a broad set of visual concepts, and image tagging (Zhang et al., 2024b), which facilitates the identification of meaningful objects in any given image. A particularly significant area is vocabulary-free zero-shot learning (Conti et al., 2023; Han et al., 2023; Zhang et al., 2024a), which, while sharing the goal of discovering categories through clustering, specifically targets the identification of class names for unlabeled images with minimal reliance on textual information.

Recent research has also explored the application of VLMs to the GCD task. For instance, Ouldnoughi et al. (2023) propose to enhance visual knowledge by retrieving relevant text descriptions from a vast external text corpus. In contrast, Wang et al. (2024a) introduce a method for synthesizing pseudo text embeddings for unlabeled images. Additionally, Liu et al. (2024) take a further step by integrating multiple VLMs with a Large Language Model (LLM) to reason about unlabeled images. Our work differentiates from them by eliminating the need for external knowledge bases, extra pre-trained models, or multiple training stages, making it both highly effective and computationally efficient, while also demonstrating robustness and generalizability across various backbones.

3 Approach

In this section, we introduce in detail our proposed *Consistent Prompt Tuning* (CPT) on top of a large pre-trained vision-language model, *i.e.*, CLIP (Radford et al., 2021), for the task of *Generalized Category Discovery* (GCD) (Vaze et al., 2022a). We start with the overview of the problem definition and our overall framework, followed by the two proposed consistency regularization techniques for efficient prompt tuning.

3.1 Overview

Problem Definition. In GCD, the training data $\mathcal{D}$ is composed of two subsets, *i.e.*, a labeled subset $\mathcal{D}^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathcal{D}^l|}$ containing labeled images $\boldsymbol{x}$ with one-hot labels $\boldsymbol{y}$ from label space $\mathcal{Y}^l$, and an unlabeled subset $\mathcal{D}^u = \{\boldsymbol{x}_j\}_{j=|\mathcal{D}^l|+1}^{|\mathcal{D}|}$

containing unlabeled images $\boldsymbol{x}$ of both known and unknown classes from label space $\mathcal{Y}$, where $\mathcal{Y}^l \subset \mathcal{Y}$. The goal of GCD is to partition $\mathcal{D}^u$ into K clusters using the knowledge learned from $\mathcal{D}^l$. We follow the common practice (Fini et al., 2021; Han et al., 2022; Wen et al., 2023; Zhong et al., 2021a) to assume the number of total categories known *a priori*, *i.e.*, $K = |\mathcal{Y}|$, which can also be effectively estimated using off-the-shelf methods (Han et al., 2019; Vaze et al., 2022a).

Overall Framework. Our method is built on top of CLIP (Radford et al., 2021), which is composed of two encoders, one for image input and the other for text input, as shown in Fig. 3. The two encoders are used to transform an input image or a sequence of word tokens into feature vectors. CLIP was originally trained by maximizing (minimizing) the cosine similarity of matched (mismatched) image-text pairs, which can be used for zero-shot recognition by calculating the feature similarity between a test image and text descriptions specifying classes of interest. As discussed in Sec. 1, the inherent zero-shot ability of CLIP is subject to carefully-designed text prompts (Zhou et al., 2022a,b), and hence the crux of the problem in adapting CLIP to GCD is how to find accurate class description prompts that disambiguate the class definitions for both labeled and unlabeled data from known and unknown classes.

In light of this, we propose to learn a set of "task + class" prompts. Since we assume no prior knowledge of the task information, such prompts are firstly initialized as a set of learnable token vectors with the template "`a photo of a [class]`", where "`[class]`" is to be filled with random vectors for K classes, as shown in Fig. 3. Formally, the prompt for the k-th class can be denoted by

$$\boldsymbol{t}_k = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M, \boldsymbol{c}_1^k, \boldsymbol{c}_2^k, \ldots, \boldsymbol{c}_N^k] \,, \quad (1)$$

where $\{\boldsymbol{v}_i\}_{i=1}^M$ are the globally shared token vectors of the task information, initialized by the pre-trained word embeddings of "`a photo of a`", where we use $M = 4$ by default. Similarly, $\{\boldsymbol{c}_n^k\}_{n=1}^N$ are the randomly-initialized token vectors for the k-th class, where $N = 2$. Then we feed each learnable prompt $\boldsymbol{t}_k$ into the frozen text encoder of CLIP, denoted by $g(\cdot)$, to get the ℓ_2-normalized text feature, *i.e.*, $g(\boldsymbol{t}_k)/\|g(\boldsymbol{t}_k)\|$. From now on, we slightly abuse $\boldsymbol{t}_k$ to denote the final text feature of the k-th class for brevity.

Likewise, each input image $\boldsymbol{x}_i$ is passed into the frozen image encoder of CLIP (denoted by $h(\cdot)$) to obtain the image feature as $h(\boldsymbol{x}_i)/\|h(\boldsymbol{x}_i)\|$, and we also slightly abuse $\boldsymbol{x}_i$ to denote the final image feature. Note that here we use a visual adapter (Gao et al., 2023b), *i.e.*, a lightweight MLP together with a residual connection from the original image encoder, to better adapt CLIP to downstream tasks considering the large visual diversity. Accordingly, the prediction probability $\hat{\boldsymbol{y}}_i$ for $\boldsymbol{x}_i$ can be calculated as

$$\hat{y}_{i,k} = p(k|\boldsymbol{x}_i; \tau) = \frac{\exp\left(\boldsymbol{x}_i \cdot \boldsymbol{t}_k / \tau\right)}{\sum_{k'=1}^K \exp\left(\boldsymbol{x}_i \cdot \boldsymbol{t}_{k'} / \tau\right)} \,, \quad (2)$$

where $\hat{y}_{i,k}$ is the k-th element of $\hat{\boldsymbol{y}}_i$, and τ is the temperature parameter (Gou et al., 2021; Hinton et al., 2015) in CLIP. Naturally, the class description prompts $\{\boldsymbol{t}_k\}_{k=1}^K$ can be regarded as *global cluster centers* for each image $\boldsymbol{x}_i$, and we can use a standard cross-entropy loss as the training objective for these prompts (Khattak et al., 2023; Nayak et al., 2023; Zhou et al., 2022a,b):

$$\mathcal{L}_{\text{CE}}(\boldsymbol{x}_i, \boldsymbol{y}_i) = -\sum_{k=1}^K y_{i,k} \log\left(\hat{y}_{i,k}\right) \,. \quad (3)$$

However, due to the insufficient supervision for the unlabeled data $\mathcal{D}^u$, we propose two consistency regularization techniques that help learn accurate "task + class" prompts for both known and unknown classes with high efficiency.

3.2 Consistent Prompt Tuning

We aim to find a set of class description prompts $\{\boldsymbol{t}_k\}_{k=1}^K$ that preserve the consistency between any two matched inputs sharing the same semantic information. In particular, we focus on the *multi-modal embedding similarity*, *i.e.*, to preserve the consistency of image-text feature similarities, which is in line with CLIP's original training objective and is hence naturally suitable for our prompt tuning goal.

Since the inputs can come from either vision or language modality, we can encourage the consistency between any two matched image inputs (*i.e.*, *vision-vision consistency*, VVC), or any two matched image and text inputs (*i.e.*, *vision-language consistency*, VLC), as shown in Fig. 4.

By "matched", we expect two *different* inputs with the *same* semantic information, and thus there is no language-language consistency, as any two different text inputs actually describe distinct classes by our design.

Vision-Vision Consistency. As discussed above, *vision-vision consistency* (VVC) focuses on two different image inputs that share the same semantic information. A simple and effective method is to apply VVC on two different data augmentations of the same input image (Assran et al., 2022, 2021; Caron et al., 2020, 2021; Chen and He, 2021; Sohn et al., 2020), which is easy to implement and applicable to both labeled and unlabeled data.

Concretely, we use two random data augmentations to generate two views of an input image, the features of which are denoted as $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$. As discussed in Sec. 1, VVC encourages the consistency of the two views *w.r.t.* global cluster centers, *i.e.*, $\{\boldsymbol{t}_k\}_{k=1}^K$. We achieve this with an adapted *swapped prediction* (Caron et al., 2020) strategy to encourage consistent multi-modal embedding similarities for $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$:

$$\mathcal{L}_{\text{swap}}(\boldsymbol{x}_i, \boldsymbol{x}'_i) = \frac{1}{2}\big(\mathcal{L}_{\text{CE}}(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}'_i) + \mathcal{L}_{\text{CE}}(\boldsymbol{x}'_i, \tilde{\boldsymbol{y}}_i)\big), \quad (4)$$

where $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{y}}'_i$ are the prediction probability for $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$, respectively, with the k-th element of $\tilde{\boldsymbol{y}}_i$ calculated as $\tilde{y}_{i,k} = p(k|\boldsymbol{x}_i; \tilde{\tau})$ using Eq. (2) ($\tilde{\boldsymbol{y}}'_i$ is computed in the same way), and we stop gradients for both $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{y}}'_i$. Here we follow Assran et al. (2022, 2021); Caron et al. (2021) to set $\tilde{\tau} < \tau$ for sharper target distributions, which helps eliminate the collapsing solutions where all predictions are uniform. On the other hand, maximizing the mean entropy of the prediction probability is also proven necessary (Assran et al., 2022, 2021; Cao et al., 2022; Wen et al., 2023), which prevents the predictions from being dominated by a few classes. Therefore, the final objective for VVC is

$$\mathcal{L}_{\text{VVC}} = \mathcal{L}_{\text{swap}}(\boldsymbol{x}_i, \boldsymbol{x}'_i) - \epsilon H(\bar{\boldsymbol{y}}), \quad (5)$$

in which $H(\bar{\boldsymbol{y}})$ denotes the entropy of the average prediction in a training batch $\mathcal{B}$, *i.e.*, $\bar{\boldsymbol{y}} = \frac{1}{2|\mathcal{B}|}\sum_{\mathcal{B}}(\hat{\boldsymbol{y}}_i + \hat{\boldsymbol{y}}'_i)$, and $\epsilon > 0$ is a trade-off parameter for controlling the strength of entropy regularization.

Vision-Language Consistency. Different from VVC, *vision-language consistency* (VLC) focuses on an image and a text input that share the same semantic information. While VVC optimizes the *global* relationships by encouraging consistent predictions between matched image inputs *w.r.t.* global cluster centers, VLC is in turn designed to regulate *local* relationships by encouraging consistent multi-modal embedding similarities in a local support set, and we directly use the current training batch $\mathcal{B}$ for simplicity.

Formally, we denote all available image views in $\mathcal{B}$ as $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{|\mathcal{B}|} \cup \{\boldsymbol{x}'_i\}_{i=1}^{|\mathcal{B}|}$. As shown in Fig. 4, we collect for each image view in $\mathcal{X}$ a matched text prompt to construct the local support set, *i.e.*, $\mathcal{S} = \{(\boldsymbol{x}_i, \tilde{\boldsymbol{t}}_i)\}_{i=1}^{|\mathcal{B}|} \cup \{(\boldsymbol{x}'_i, \tilde{\boldsymbol{t}}_i)\}_{i=1}^{|\mathcal{B}|}$, and we use $\mathcal{T}$ to denote all the matched text prompts. For the labeled images in $\mathcal{D}^l$, $\tilde{\boldsymbol{t}}_i$ can be selected according to the ground-truth labels. For the unlabeled images in $\mathcal{D}^u$, we instead use the soft *pseudo-labels* initially calculated in Eq. (4). In particular, we employ the average target distribution for the two augmentation views, *i.e.*, $\frac{1}{2}(\tilde{\boldsymbol{y}}_i + \tilde{\boldsymbol{y}}'_i)$, as the pseudo-label for both $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$. Therefore, by gathering all K text prompts into a matrix as $\mathbf{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_K]$, the matched text input $\tilde{\boldsymbol{t}}_i$ for both $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$ can be selected as

$$\tilde{\boldsymbol{t}}_i = \begin{cases} \boldsymbol{t}_k & \text{if } \boldsymbol{x}_i, \boldsymbol{x}'_i \in \mathcal{D}^l \text{ from } k\text{-th class}; \\ \mathbf{T} \cdot \dfrac{\tilde{\boldsymbol{y}}_i + \tilde{\boldsymbol{y}}'_i}{2} & \text{otherwise } (\boldsymbol{x}_i, \boldsymbol{x}'_i \in \mathcal{D}^u). \end{cases} \quad (6)$$

In this way, the generated $\tilde{\boldsymbol{t}}_i$ for each unlabeled image is a linear combination of all K prompts, which respects the uncertainty in the soft pseudo-labels, such that the neighborhood information in the local support set $\mathcal{S}$ can be well preserved, and hence works much better in our experiments than using the hard one-hot pseudo-labels.

We aim to encourage the consistency in multi-modal embedding similarities of matched image-text pairs in the local support set $\mathcal{S}$ for VLC. It is approached by regularizing the similarities between each image $\boldsymbol{x}_i$ and all support texts in $\mathcal{T}$, and vice versa. As shown in Fig. 4, we calculate the probability distribution $\boldsymbol{q}_i^{x\to\mathcal{T}}$ of image-to-text

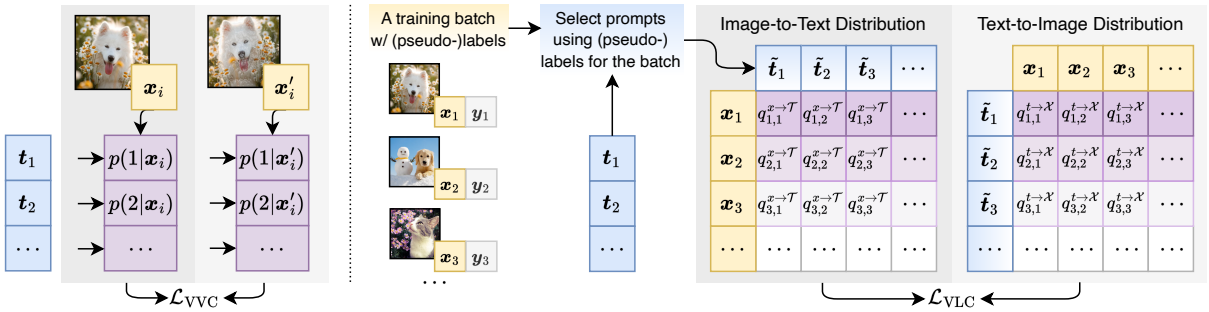

Fig. 4 Illustration of *vision-vision consistency* (VVC) and *vision-language consistency* (VLC).

similarities for image $\boldsymbol{x}_i$:

$$q_{i,j}^{x\to\mathcal{T}} = \frac{\exp\left(\boldsymbol{x}_i \cdot \tilde{\boldsymbol{t}}_j/\tau\right)}{\sum_{\tilde{\boldsymbol{t}}_{j'}\in\mathcal{T}} \exp\left(\boldsymbol{x}_i \cdot \tilde{\boldsymbol{t}}_{j'}/\tau\right)}, \tag{7}$$

and $\boldsymbol{q}_i^{t\to\mathcal{X}}$ of text-to-image similarities for text $\tilde{\boldsymbol{t}}_i$:

$$q_{i,j}^{t\to\mathcal{X}} = \frac{\exp\left(\boldsymbol{x}_j \cdot \tilde{\boldsymbol{t}}_i/\tau\right)}{\sum_{\boldsymbol{x}_{j'}\in\mathcal{X}} \exp\left(\boldsymbol{x}_{j'} \cdot \tilde{\boldsymbol{t}}_i/\tau\right)}, \tag{8}$$

where $q_{i,j}^{x\to\mathcal{T}}$ and $q_{i,j}^{t\to\mathcal{X}}$ are respectively the j-th element of $\boldsymbol{q}_i^{x\to\mathcal{T}}$ and $\boldsymbol{q}_i^{t\to\mathcal{X}}$, and we slightly abuse $\boldsymbol{x}_i$ to denote both $\boldsymbol{x}_i$ and $\boldsymbol{x}_i'$. The two distributions here characterize the local relationships in the support set $\mathcal{S}$ for the two modalities, which are expected to be consistent to align the neighborhood relationships that has been shown critical to the unlabeled data (Assran et al., 2021; Li et al., 2021; Yang et al., 2022; Zhong et al., 2021a). To this end, VLC is realized by minimizing the bidirectional Kullback-Leibler (KL) divergence (Wu et al., 2021) between $\boldsymbol{q}_i^{x\to\mathcal{T}}$ and $\boldsymbol{q}_i^{t\to\mathcal{X}}$ for each pair of matched image $\boldsymbol{x}_i$ and text $\tilde{\boldsymbol{t}}_i$ in a training batch:

$$\mathcal{L}_{\text{VLC}} = \frac{1}{2}\left(D_{\text{KL}}(\boldsymbol{q}_i^{x\to\mathcal{T}} \| \boldsymbol{q}_i^{t\to\mathcal{X}}) + D_{\text{KL}}(\boldsymbol{q}_i^{t\to\mathcal{X}} \| \boldsymbol{q}_i^{x\to\mathcal{T}})\right), \tag{9}$$

where D_{KL} denotes the KL divergence.

3.3 Training and Inference

Training. The overall training objective of our proposed CPT is the combination of supervised cross-entropy loss on labeled data ($\mathcal{D}^l$) and the two consistency regularization losses on both labeled and unlabeled data ($\mathcal{D}$):

$$\mathcal{L}_{\text{CPT}} = \frac{\lambda_1}{|\mathcal{D}^l|}\sum_{\mathcal{D}^l}\mathcal{L}'_{\text{CE}} + \frac{\lambda_2}{|\mathcal{D}|}\sum_{\mathcal{D}}\mathcal{L}_{\text{VVC}} + \frac{\lambda_3}{|\mathcal{D}|}\sum_{\mathcal{D}}\mathcal{L}_{\text{VLC}}, \tag{10}$$

where $\mathcal{L}'_{\text{CE}} = \frac{1}{2}\left(\mathcal{L}_{\text{CE}}(\boldsymbol{x}_i, \boldsymbol{y}_i) + \mathcal{L}_{\text{CE}}(\boldsymbol{x}_i', \boldsymbol{y}_i)\right)$, and λ_1, λ_2, $\lambda_3 > 0$ are trade-off parameters that balance each loss term.

Inference. After training, CPT generates predictions for unlabeled images by simply comparing the cosine similarity between each image and the K learned class description prompts in CLIP's multi-modal embedding space. In particular, for each unlabeled image $\boldsymbol{x}_i$, the prediction can be computed by $\arg\max_{k\in\{1,2,\ldots,K\}} p(k|\boldsymbol{x}_i;\tau)$ using Eq. (2).

4 Experiments

In this section, we present the experimental evaluations of our proposed CPT. We start with the experimental setup, followed by the evaluation results and analysis.

4.1 Experimental Setup

Datasets. We use six GCD benchmark datasets for evaluation, including three datasets for *generic* image recognition, *i.e.*, CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), and ImageNet-100 (Deng et al., 2009), and another three for *fine-grained* image recognition, *i.e.*, CUB-200 (Wah et al., 2011), Stanford-Cars (Krause et al., 2013), and Oxford-Pets (Parkhi et al., 2012). Following standard GCD protocol (Vaze et al., 2022a; Wen et al., 2023), a portion of all available classes in each

dataset is firstly selected as the labeled known classes (*old*), half images from which are further used as the labeled subset $\mathcal{D}^l$. Another half, as well as the images from the unknown classes (*new*), are used as the unlabeled subset $\mathcal{D}^u$. The statistics of these datasets are detailed in Sec. B.1.

Evaluation Metric. The *clustering accuracy* metric (Vaze et al., 2022a) is adopted for GCD evaluation, which is calculated as

$$\text{ClusterAcc} = \max_{r \in \mathcal{P}(\mathcal{Y})} \frac{1}{|\mathcal{D}^u|} \sum\nolimits_{\mathcal{D}^u} \mathbb{1}\big(y_i = r(\hat{y}_i)\big)\,, \tag{11}$$

where y_i and $\hat{y}_i$ are respectively the ground-truth label and the model's cluster assignment prediction for each sample $\boldsymbol{x}_i$ in $\mathcal{D}^u$; $\mathcal{P}(\mathcal{Y})$ denotes the set of all possible permutations of the class labels in $\mathcal{Y}$, for which the optimal permutation can be obtained using the Hungarian algorithm (Kuhn, 1955). It is worth noting that the Hungarian assignment is computed only once, after which we can measure the accuracy in *all classes*, *old classes*, and *new classes*, respectively.

Implementation Details. Our proposed CPT is implemented on top of a pre-trained CLIP (Radford et al., 2021), in which a ViT-B/16 (Dosovitskiy et al., 2021) is used as the image encoder, and a Transformer (Vaswani et al., 2017) the text encoder. Unless otherwise specified, we default to using OpenAI's CLIP implementation and model weights (Radford et al., 2021). During training, we only update the K "task + class" prompts as well as the visual adapter (*i.e.*, a two-layer MLP with ReLU activation), and keep the whole CLIP model frozen.

For all datasets, we use an SGD optimizer with the batch size of 256 and the learning rate of 0.05, and train for 100 epochs with a cosine annealing schedule. Task and class token lengths are set to $M = 4$ and $N = 2$ in Eq. (1), where the task tokens are initialized with the pre-trained word embeddings of "`a photo of a`", and the class tokens are randomly initialized. The default temperature parameter is inherited from CLIP, *i.e.*, $\tau = 0.01$ in Eqs. (2), (3), (7) and (8). As to $\tilde{\tau}$ in Eq. (4), we follow (Caron et al., 2021; Wen et al., 2023) to use a warm-up schedule for it, *i.e.*, decreasing from 0.007 to 0.004 for the first 15 epochs. We also follow (Wen et al., 2023) to set the value of ϵ in Eq. (5) for fair comparisons, *i.e.*, $\epsilon = 4$ for CIFAR100, $\epsilon = 2$ for CUB-200, and $\epsilon = 1$ for all the rest. As for the trade-off parameters in Eq. (10), we follow the recent works (Pu et al., 2023; Vaze et al., 2022a; Wen et al., 2023; Zhang et al., 2023b) to balance the supervised and self-supervised objectives for fair comparisons, *i.e.*, setting $\lambda_1 = 0.35, \lambda_2 = 0.65$, while λ_3 is chosen based on the performance in the validation set of labeled classes. We use PyTorch (Paszke et al., 2019) framework for implementation, and all runs are performed on one NVIDIA GeForce RTX 2080 Ti GPU. Sec. B also provides more implementation details.

Table 1 Ablation study†. GCD accuracy (%) is reported.

Dataset →					CIFAR100			CUB-200		
#	$\boldsymbol{v}$	Adapter	$\mathcal{L}_{\text{VVC}}$	$\mathcal{L}_{\text{VLC}}$	All	Old	New	All	Old	New
(1)					57.5	72.5	27.6	36.5	48.8	30.3
(2)			✓		75.3	79.7	66.6	55.7	68.0	49.6
(3)	✓		✓		77.6	79.6	73.7	60.5	67.9	56.8
(4)	✓	✓	✓		78.5	79.6	76.5	61.0	66.9	58.1
(5)	✓		✓	✓	80.5	80.6	80.3	68.5	72.4	66.5
(6)	✓	✓	✓	✓	**81.3**	**81.3**	**81.2**	**70.1**	**73.5**	**68.4**

†By default, class tokens $\boldsymbol{c}$ are learnable, and the supervised cross-entropy loss $\mathcal{L}'_{\text{CE}}$ in Eq. (10) is used.

4.2 Ablation Study

We conduct ablation studies to evaluate the effectiveness of each module in our proposed CPT. Specifically, we evaluate the effects of learnable task tokens $\boldsymbol{v}$ in Eq. (1), the visual adapter on top of the CLIP's image features, and the two consistency regularization techniques. Besides, we aim to provide more insight into our design of the "task + class" prompt with more detailed ablation studies results, including prompt styles and token initialization. Tabs. 1 and 2 summarize the ablation study results on two representative datasets for both generic and fine-grained image recognition tasks.

Effect of Task Tokens. We can see from the results in Rows (2) and (3) of Tab. 1 that the learnable task tokens $\boldsymbol{v}$ do bring consistent performance boost on top of the class tokens $\boldsymbol{c}$, especially in the fine-grained CUB-200 dataset specialized in bird species. This coincides with our hypothesis that accurate and unified task information helps disambiguate the classes in GCD. To better understand the role of task tokens, we present detailed

Table 2 Ablation study of task tokens. GCD accuracy (%) with different task token settings is reported.

Dataset →	CIFAR100			CUB-200		
Task Token Setting ↓	All	Old	New	All	Old	New
† *Task/class token number:*						
0 task + 2 class tokens	78.8	80.1	76.2	63.9	70.7	60.4
0 task + 6 class tokens	79.7	81.1	77.1	65.3	70.8	62.5
4 task + 2 class tokens	**81.3**	**81.3**	**81.2**	**70.1**	**73.5**	**68.4**
‡ *Task token initialization:*						
Random (*fixed*)	79.4	80.9	76.3	63.4	71.3	59.5
Random (*learnable*)	81.1	81.3	80.9	69.2	73.5	67.0
"a photo of a" (*fixed*)	80.4	81.1	79.2	66.5	71.7	64.0
"a photo of a" (*learnable*)	**81.3**	**81.3**	**81.2**	**70.1**	**73.5**	**68.4**

†Class tokens are learnable by default.
‡4 task tokens and 2 learnable class tokens are used by default.

Table 3 GCD accuracy (%) using semi-supervised k-means (Vaze et al., 2022a) on CLIP's image encoder.

Dataset →	CIFAR100			CUB-200		
Method ↓	All	Old	New	All	Old	New
GCD† (w/ SS-k-means)	71.9	77.8	60.2	52.5	58.6	49.5
SimGCD† + SS-k-means	78.8	**82.6**	71.5	64.6	**68.8**	62.4
CPT + SS-k-means	**80.0**	81.6	**76.7**	**65.7**	66.0	**65.6**

†Reimplemented using CLIP's image encoder.

ablation studies in Tab. 2 without altering other settings, as discussed below.

On one hand, we explore different prompt style choices, *i.e.*, task/class token numbers, to validate the effectiveness of our "task + class" prompt design for GCD. We can see from the results that task tokens are indeed requisite to unlock the potential of VLMs, especially for the fine-grained CUB-200 dataset that is focused on a more specialized task. In this regard, we observe that the benefit of task tokens is most obvious in the accuracy of "New" classes, indicating that these task tokens are essential for transferring the class definition pattern to better discover unknown classes.

On the other hand, we further explore how different initialization methods and the extra learnability work on the task tokens. Note that for the class tokens, we use random initialization for both known and unknown classes for simplicity and practicability, without the need for prior knowledge of the task at hand. As shown in Tab. 2, while an accurate initialization of task tokens helps disambiguate GCD, we found that their learnability plays a more vital role in achieving better performance, even under random initialization. In a nutshell, learning accurate "task + class" prompt guidance has proved necessary and highly effective in unlocking the potential of VLMs for GCD. Sec. A.1 also provides more ablation studies on the effect of token length.

Effect of Visual Adapter. We use a lightweight visual adapter on top of CLIP's original image features to cope with the large visual diversity in downstream tasks. Trained with our CPT objective, the adapter helps establish the consistency for both image and text modalities. By comparing the results in Rows (3) and (4), we can see the positive effect of adding such a lightweight adapter. We can also see from Rows (5) and (6) that the adapter benefits more from the full set of consistency regularization, which promotes the multi-modal synergy in both global and local relationships.

Effect of Vision-Vision Consistency. The VVC regularization is designed to encourage consistent global predictions *w.r.t.* all class description prompts for two augmentation views of each input image. As these prompts can be regarded as a series of cluster centers, VVC naturally serves as an effective clustering objective for unlabeled data of both known and unknown classes. The results in Rows (1) and (2) verify its validity towards such an objective, which can be further amplified with other proposed techniques.

Effect of Vision-Language Consistency. As an important complement to VVC, the VLC regularization aims to regulate neighborhood relationships within a local support set, which is essential for finding accurate class description prompts. It is achieved by maintaining consistent local similarities of matched image-text pairs in a training batch, from both image-to-text and text-to-image directions. We can clearly see the effectiveness of VLC from Rows (5) and (6), which brings a significant improvement on top of VVC, and also works in harmony with other techniques such as the visual adapter. Overall, with the contribution of each proposed module, CPT learns accurate "task + class" tokens, which effectively disambiguates the class definitions and hence achieves superior GCD performance. Sec. A.3 also provides more ablation studies on the effect of VVC and VLC.

Table 4 GCD accuracy (%) on generic image recognition datasets. Methods are based on either DINO (Caron et al., 2021) or CLIP (Radford et al., 2021) pre-trained ViT-B/16 (Dosovitskiy et al., 2021) backbone.

Dataset →		CIFAR10			CIFAR100			ImageNet-100		
Method ↓	Backbone ↓	All	Old	New	All	Old	New	All	Old	New
k-means (1965)	ViT-B/16 (DINO)	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ (2022)	ViT-B/16 (DINO)	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ (2021)	ViT-B/16 (DINO)	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	95.0	57.9
ORCA (2022)	ViT-B/16 (DINO)	81.8	86.2	79.6	69.0	77.4	52.0	73.5	92.6	63.9
GCD (2022a)	ViT-B/16 (DINO)	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
XCon (2022)	ViT-B/16 (DINO)	96.0	97.3	95.4	74.2	81.2	60.3	77.6	93.5	69.7
DCCL (2023)	ViT-B/16 (DINO)	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
PromptCAL (2023b)	ViT-B/16 (DINO)	**97.9**	96.6	**98.5**	81.2	84.2	75.3	83.1	92.7	78.3
SimGCD (2023)	ViT-B/16 (DINO)	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9
GPC (2023)	ViT-B/16 (DINO)	92.2	98.2	89.1	77.9	85.0	63.0	76.9	94.3	71.0
PIM (2023)	ViT-B/16 (DINO)	94.7	97.4	93.3	78.3	84.2	66.5	83.1	95.3	77.0
k-means†	ViT-B/16 (CLIP)	74.4	72.1	75.5	46.0	46.1	45.8	62.7	57.6	65.3
GCD†	ViT-B/16 (CLIP)	94.6	97.5	93.1	71.9	77.8	60.2	75.2	88.6	68.5
PromptCAL†	ViT-B/16 (CLIP)	94.9	96.4	94.1	69.4	77.3	53.5	75.2	87.0	69.3
SimGCD†	ViT-B/16 (CLIP)	96.9	95.7	97.5	80.9	**85.2**	72.5	88.6	94.6	85.6
GPC†	ViT-B/16 (CLIP)	93.9	97.2	90.6	65.0	72.5	35.3	76.1	92.4	59.9
PIM†	ViT-B/16 (CLIP)	94.4	97.1	93.1	73.7	82.9	55.3	80.7	**95.9**	73.0
CLIP‡	ViT-B/16 (CLIP)	89.4	87.8	90.2	67.1	67.1	67.3	85.5	89.2	83.6
CLIP-GCD (2023)	ViT-B/16 (CLIP)	96.6	97.2	96.4	**85.2**	85.0	**85.6**	84.0	95.5	78.2
CPT (Ours)	ViT-B/16 (CLIP)	96.3	96.2	96.4	81.3	81.3	81.2	**89.2**	95.2	**86.1**

†Reimplemented using CLIP's image encoder. ‡Zero-shot CLIP with oracle class descriptions.

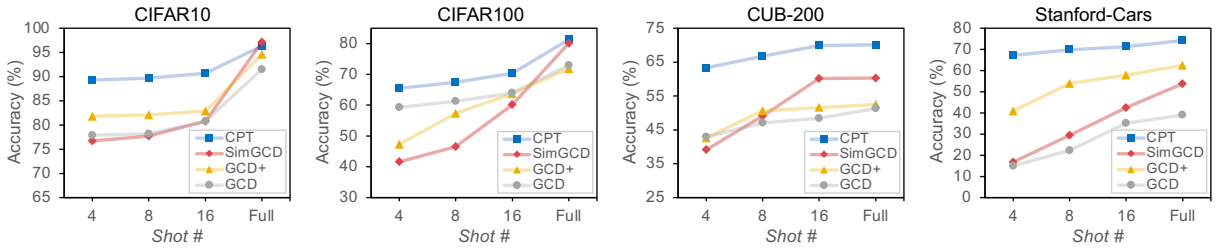

Fig. 5 GCD accuracy (reported in *all classes*) in the *few-shot* setting. "GCD+": GCD (Vaze et al., 2022a) reimplemented using CLIP's image encoder.

Effect of Single Image Encoder. To assess the effectiveness of our proposed CPT in enhancing the visual representation capabilities of the image encoder, we conduct semi-supervised k-means (SS-k-means) inference on the CLIP ViT model whose last block has been fine-tuned using CPT, following the approach outlined in GCD (Vaze et al., 2022a). As a comparison, we also conduct SS-k-means on the CLIP ViT model fine-tuned using GCD (Vaze et al., 2022a) and SimGCD (Wen et al., 2023), respectively.

As shown in Tab. 3, CPT + SS-k-means consistently achieves competitive results, suggesting that the training scheme of CPT not only aids in finding accurate "task + class" prompts, but also improves visual representation capabilities of the image encoder.

4.3 Main Results

Results on Existing GCD Benchmarks. We show in Tabs. 4 and 5 the results of the current state-of-the-art methods (SoTAs) and our proposed CPT on several GCD benchmarks including generic and fine-grained image recognition tasks. We also implement another three baselines using CLIP backbone to better assess the effectiveness of our proposed CPT, *i.e.*, (1) k-means†: applying k-means (MacQueen, 1965) on top of

Table 5 GCD accuracy (%) on fine-grained image recognition datasets. Methods are based on either DINO (Caron et al., 2021) or CLIP (Radford et al., 2021) pre-trained ViT-B/16 (Dosovitskiy et al., 2021) backbone.

Dataset →		CUB-200			Stanford-Cars			Oxford-Pets		
Method ↓	Backbone ↓	All	Old	New	All	Old	New	All	Old	New
k-means (1965)	ViT-B/16 (DINO)	34.3	38.9	32.1	12.8	10.6	13.8	77.1	70.1	80.7
RankStats+ (2022)	ViT-B/16 (DINO)	33.3	51.6	24.2	28.3	61.8	12.1	–	–	–
UNO+ (2021)	ViT-B/16 (DINO)	35.1	49.0	28.1	35.5	70.5	18.6	–	–	–
ORCA (2022)	ViT-B/16 (DINO)	35.3	45.6	30.2	23.5	50.1	10.7	–	–	–
GCD (2022a)	ViT-B/16 (DINO)	51.3	56.6	48.7	39.0	57.6	29.9	80.2	85.1	77.6
XCon (2022)	ViT-B/16 (DINO)	52.1	54.3	51.0	40.5	58.8	31.7	86.7	91.5	84.1
DCCL (2023)	ViT-B/16 (DINO)	63.5	60.8	64.9	43.1	55.7	36.2	88.1	88.2	88.0
PromptCAL (2023b)	ViT-B/16 (DINO)	62.9	64.4	62.1	50.2	70.1	40.6	78.6	90.2	72.5
SimGCD (2023)	ViT-B/16 (DINO)	60.3	65.6	57.7	53.8	71.9	45.0	84.8	86.2	84.1
GPC (2023)	ViT-B/16 (DINO)	55.4	58.2	53.1	42.8	59.2	32.8	80.5	90.0	70.4
PIM (2023)	ViT-B/16 (DINO)	62.7	75.7	56.2	43.1	66.9	31.6	83.9	92.3	79.5
k-means†	ViT-B/16 (CLIP)	46.2	39.5	49.5	42.5	41.9	42.8	57.7	49.5	62.0
GCD†	ViT-B/16 (CLIP)	52.5	58.6	49.5	62.5	73.5	57.1	75.8	83.8	71.5
PromptCAL†	ViT-B/16 (CLIP)	53.7	61.4	49.9	60.1	77.9	51.5	88.8	93.9	86.1
SimGCD†	ViT-B/16 (CLIP)	68.6	71.4	67.2	70.2	84.8	63.1	87.3	88.9	86.4
GPC†	ViT-B/16 (CLIP)	52.0	61.4	42.6	55.7	72.7	39.2	76.3	88.4	63.3
PIM†	ViT-B/16 (CLIP)	58.7	70.8	52.6	62.3	79.0	54.2	83.5	**96.3**	76.8
CLIP‡	ViT-B/16 (CLIP)	55.7	51.8	57.6	62.7	67.7	60.2	85.8	91.7	82.7
CLIP-GCD (2023)	ViT-B/16 (CLIP)	62.8	**77.1**	55.7	70.6	**88.2**	62.2	79.9	83.7	77.8
CPT (Ours)	ViT-B/16 (CLIP)	**70.1**	73.5	**68.4**	**74.2**	84.3	**69.3**	**92.5**	93.9	**91.7**

†Reimplemented using CLIP's image encoder. ‡Zero-shot CLIP with oracle class descriptions.

CLIP's image encoder; (2) GCD†: reimplementing GCD (Vaze et al., 2022a) using CLIP's image encoder; and (3) CLIP‡: zero-shot CLIP with oracle class descriptions, where we follow (Zhou et al., 2022b) to conduct zero-shot inference using the class names of both known and unknown classes (which are originally unavailable in GCD) in a specially-designed template prompt for each dataset. For fair comparisons, we directly use a CLIP pre-trained ViT-B/16 in lieu of the DINO pre-trained ViT-B/16, and fine-tune the last ViT block as originally required by the baseline (Vaze et al., 2022a), with all other hyperparameters unchanged.

Compared with existing SoTAs, CPT shows considerable performance improvement in most cases. CPT is on par with current SoTAs on generic datasets like CIFAR10/100, mainly due to the very low resolution of the images and the lack of scale-based data augmentation in CLIP (Radford et al., 2021), yet it still performs well on other generic dataset like ImageNet. Additionally, we also observe relatively poor performance on some fine-grained datasets like FGVC-Aircraft and Herbarium-19 due to potential task misalignment in CLIP's pre-trained knowledge (see Secs. A.7 and C for details). However, we believe CPT can also be applied to stronger VLMs that better handle these cases. Moreover, it is clear that simply switching to the CLIP backbone does not always show a positive effect on baselines in all cases, indicating that our improvement is not trivially from a stronger backbone (Mayilvahanan et al., 2024), but because of the efficacy of our proposed prompt learning strategy and consistent tuning objective.

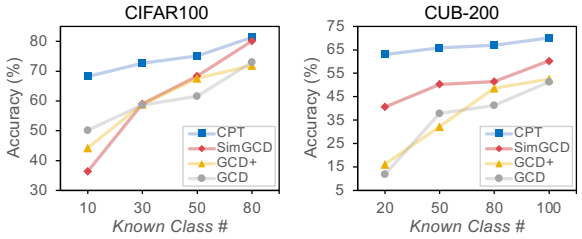

Fig. 6 GCD accuracy (reported in *all classes*) in the *few-known-class* setting on CIFAR100 and CUB-200 datasets. "GCD+": GCD (Vaze et al., 2022a) reimplemented using CLIP's image encoder.

Table 6 GCD accuracy (%) in the *customized* setting on NICO dataset.

Dataset →	NICO-10			NICO-33			NICO-100		
Method ↓	All	Old	New	All	Old	New	All	Old	New
GCD (2022a)	95.1	97.3	94.0	38.6	58.4	25.1	48.0	67.0	39.1
PromptCAL (2023b)	89.6	80.9	93.6	30.5	42.4	22.5	45.3	53.4	41.5
SimGCD (2023)	97.3	97.0	97.4	45.9	67.7	31.0	46.3	65.1	37.4
PIM (2023)	96.4	97.2	96.0	42.2	67.6	24.9	49.5	67.5	41.0
GCD†	98.4	98.0	98.8	39.3	65.0	21.8	51.6	54.5	49.2
PromptCAL†	97.5	98.5	97.1	29.2	41.8	20.6	48.9	64.4	41.7
SimGCD†	**98.5**	97.5	**98.9**	46.6	72.2	29.1	52.5	69.4	44.6
PIM†	97.5	98.5	97.0	47.6	71.8	31.0	50.0	73.0	39.2
CPT (Ours)	**98.5**	**98.6**	98.5	**53.5**	**75.7**	**38.4**	**61.2**	**73.5**	**55.4**

†Reimplemented using CLIP's image encoder.

Table 7 GCD accuracy (%) on Rare-Species dataset. Methods are based on either DINO (Caron et al., 2021) or OpenCLIP (Cherti et al., 2023) pre-trained ViT-B/16 (Dosovitskiy et al., 2021) backbone.

Dataset →		Rare-Species		
Method ↓	Backbone ↓	All	Old	New
GCD (2022a)	DINO	42.0	50.2	37.9
PromptCAL (2023b)	DINO	47.8	54.2	44.6
SimGCD (2023)	DINO	45.1	59.5	37.9
PIM (2023)	DINO	44.9	52.2	41.3
GCD†	OpenCLIP	42.5	53.5	37.1
PromptCAL†	OpenCLIP	50.6	57.1	47.4
SimGCD†	OpenCLIP	47.5	61.0	40.8
PIM†	OpenCLIP	43.6	58.0	36.4
CPT (Ours)	OpenCLIP	**54.2**	**65.0**	**48.8**

†Reimplemented using OpenCLIP's image encoder.

Results in Few-Annotation Settings. We further evaluate CPT's generalization ability in two more challenging few-annotation settings, *i.e.*, *few-shot* and *few-known-class* GCD. In few-shot GCD, the quantity of labeled data in each known class is restricted to a small number (*shots*). We show in Fig. 5 the results with 4, 8, 16 shots, as well as the results with full shots (original GCD setting) for reference. For few-known-class GCD, we instead restrict the number of known classes (by converting a certain number of known classes into unknown ones) without altering the shots in each class, as shown in Fig. 6. Compared with current SoTAs (including the reimplemented one with CLIP backbone), CPT shows surprisingly strong generalization capability in both settings, demonstrating its versatility in practical scenarios with very limited annotations.

Results on Customized GCD. As discussed earlier in Sec. 1, the key challenge of GCD lies in the ambiguity of class definitions. To further verify whether existing methods and our proposed CPT are able to handle highly customized class definitions, here we explore another new scenario, termed *customized* GCD. In particular, we construct different class definitions for the same image data using extra annotations available in the animal subset of NICO (He et al., 2021) dataset. In this dataset, each image is annotated with an animal label (*e.g.*, "dog") and a context label (*e.g.*, "on grass"). We use the two sets of annotations to construct three types of class definitions, *i.e.*, (1) NICO-10: using 10 animal labels (5/5 known/unknown classes); (2) NICO-33: using 33 context labels (17/16 known/unknown classes); (3) NICO-100: using 100 animal-context labels (50/50 known/unknown classes). Details are presented in Sec. B.2.

As shown in Tab. 6, all involved methods show excellent performance in NICO-10, where the class definition is animal-centric and hence less ambiguous. However, the performance drops considerably in NICO-33 where we use context labels to define the classes, which are animal-agnostic and hence more ambiguous for such an animal dataset. A similar phenomenon can also be observed in NICO-100. Compared with current SoTAs that are prone to be biased to the foreground-object-centric class definitions, CPT shows clear superiority in handling customized class definitions thanks to the flexible "task + class" prompt learning paradigm and its consistent tuning objective.

Results on "Truly Novel" Data. Given that the training data for CLIP (Radford et al., 2021) is not publicly available, it might be challenging to confirm whether the discovered categories are truly novel. To evaluate the ability of our proposed method to discover "truly novel" categories, we curate a new dataset with categories never found in the LAION-400M dataset (Schuhmann et al., 2021), and reimplement our proposed CPT using OpenCLIP (Cherti et al., 2023), which is pre-trained on LAION-400M.

To achieve this, we utilize the newly introduced "Rare-Species" dataset (Stevens et al., 2024),

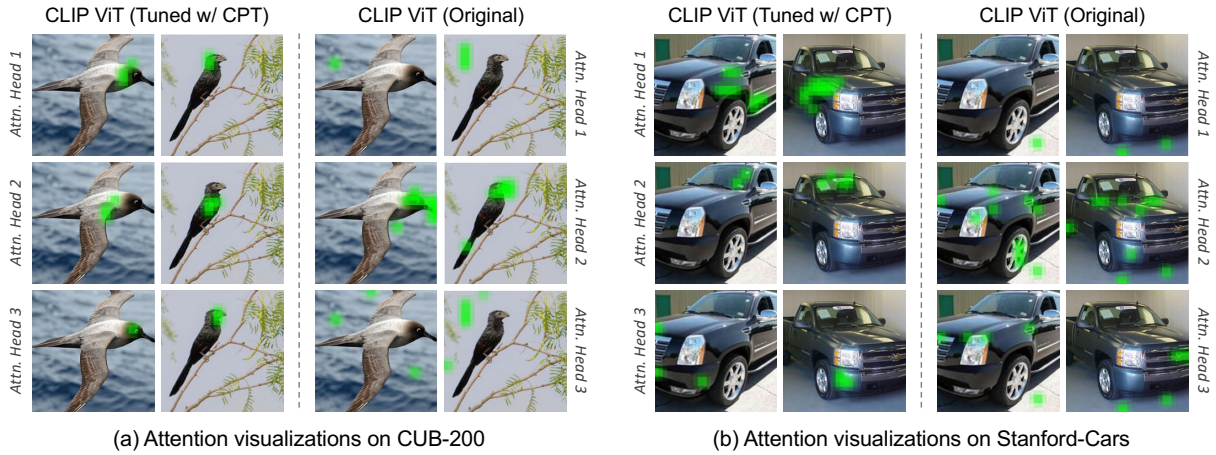

Fig. 7 Attention map visualizations on CLIP ViT-B/16 before or after fine-tuning with our proposed CPT.

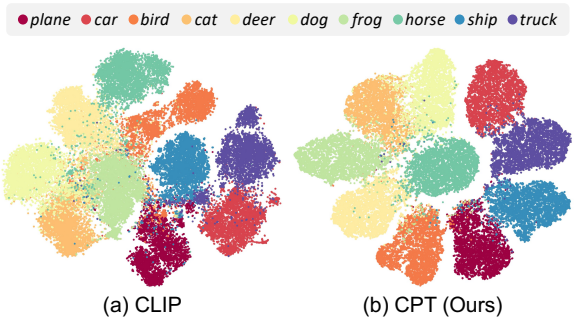

Fig. 8 t-SNE (Van Der Maaten, 2014) visualizations. The first 5 classes are known.

which consists of images of 400 endangered biological species. This dataset represents a collection of species categories that are arguably as "truly novel" as possible, making it an ideal and practically significant test case for our method. After traversing every text caption in LAION-400M, we identified 180 truly novel categories within the Rare-Species dataset. We designate these 180 categories as the unknown (new) categories and randomly select additional 180 categories in Rare-Species as the known (old) ones, creating a 360-category dataset for subsequent GCD evaluation (see details in Sec. B.3).

As shown in Tab. 7, we evaluate the GCD performance of our proposed CPT against several state-of-the-art baselines on the Rare-Species dataset using the OpenCLIP backbone. CPT consistently surpasses the baselines across all cases, demonstrating a significant advantage in discovering "truly novel" categories.

Visualizations. Fig. 8 compares the visualizations between CLIP's original visual space and CPT's output space on CIFAR10. While in Fig. 8(a) we can roughly observe different clusters in CLIP's original visual space, CPT further improves them with much clearer decision boundaries among classes (Fig. 8(b)). This is credited to our consistent prompt tuning objective, which explicitly aligns global and local relationships for both modalities.

Fig. 7 shows the attention map visualizations on CLIP ViT before and after fine-tuning with our proposed CPT (where we fine-tune only the last ViT block following Vaze et al. (2022a)). Initially, the original CLIP ViT exhibits very few correctly activated attention heads—typically just one or two—which leads to irregular and ambiguous attention patterns across most heads. However, after applying our proposed CPT, the attention heads are better able to focus on consistent and semantically meaningful visual patterns, such as heads, bodies, and eyes of the birds. This is particularly beneficial in GCD tasks, as it aids in pinpointing the critical differences that separate various categories, thereby improving the generalization ability in category discovery.

5 Conclusions

We present *Consistent Prompt Tuning* (CPT) in this paper, leveraging the power of large pre-trained vision-language models for Generalized Category Discovery (GCD). We highlight that the key challenge of GCD lies in the ambiguity in

class definitions, and hence propose to learn a set of "task + class" prompts to better disambiguate the classes. To ensure efficient prompt learning on both labeled and unlabeled data, CPT utilizes two regularization techniques that encourage the consistency of multi-modal embedding similarities to optimize global and local relationships. CPT has been evaluated on various existing GCD benchmarks as well as in several new practical scenarios, showing clear superiority over current state-of-the-art methods.

Limitations and Future Work. The main limitation of our proposed method comes from its large reliance on CLIP (Radford et al., 2021). Despite being trained on web-scale data, CLIP may still fall short in specialized scenarios. This can be seen from the relatively poor performance on FGVC-Aircraft and Herbarium-19, as shown in Tab. A6. The results suggest a misalignment between CLIP's pre-trained knowledge and the demands of these downstream tasks. This misalignment may stem from the multi-modal contrastive training objective of the CLIP model, which relies on precise alignment between visual and textual representations. If the training texts lack the fine-grained details needed to capture specific distinctions, the corresponding visual representations are likely to be affected as well. Therefore, investigating improved training techniques to enhance the visual representation capabilities of CLIP on various downstream tasks is a promising future direction.

Another limitation is that the proposed method lacks the ability to estimate the number of unknown classes, which is a key challenge in GCD. Even though this can be done with off-the-shelf tools (Han et al., 2019; Vaze et al., 2022a), we still deem that an inherent estimation mechanism facilitates VLMs' fast adaption in practical scenarios. In this regard, a potential extension could involve using "dynamic prompts". These prompts can be designed to expand or contract according to class hierarchy, allowing the model to automatically adjust and discover the appropriate number of classes in real-time during the training process. This extension constitutes a promising future direction.

Data Availability Statements

In this paper, we use data from nine existing benchmarks, which are available at

1. CIFAR10 (Krizhevsky et al., 2009): link
2. CIFAR100 (Krizhevsky et al., 2009): link
3. ImageNet (Deng et al., 2009): link
4. CUB-200 (Wah et al., 2011): link
5. Stanford-Cars (Krause et al., 2013): link
6. Oxford-Pets (Parkhi et al., 2012): link
7. FGVC-Aircraft (Maji et al., 2013): link
8. Herbarium-19 (Tan et al., 2019): link
9. NICO (He et al., 2021): link
10. Rare-Species (Stevens et al., 2024): link

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 23716–23736.

An, W., Shi, W., Tian, F., Lin, H., Wang, Q., Wu, Y., Cai, M., Wang, L., Chen, Y., Zhu, H., et al. (2023a). Generalized category discovery with large language models in the loop. *arXiv preprint arXiv:2312.10897*.

An, W., Tian, F., Chen, P., Tang, S., Zheng, Q., and Wang, Q. (2022). Fine-grained category discovery under coarse-grained supervision with hierarchical weighted self-contrastive learning. *arXiv preprint arXiv:2210.07733*.

An, W., Tian, F., Shi, W., Chen, Y., Wu, Y., Wang, Q., and Chen, P. (2023b). Transfer and alignment network for generalized category discovery. *arXiv preprint arXiv:2312.16467*.

An, W., Tian, F., Zheng, Q., Ding, W., Wang, Q., and Chen, P. (2023c). Generalized category discovery with decoupled prototypical network. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 12527–12535.

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. (2022). Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision (ECCV)*, pages 456–473.

Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., and Rabbat, M. (2021). Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8443–8452.

Bai, J., Liu, Z., Wang, H., Chen, R., Mu, L., Li, X., Zhou, J. T., Feng, Y., Wu, J., and Hu, H. (2023). Towards distribution-agnostic generalized category discovery. *arXiv preprint arXiv:2310.01376*.

Banerjee, A., Kallooriyakath, L. S., and Biswas, S. (2024). Amend: Adaptive margin and expanded neighborhood for efficient generalized category discovery. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2101–2110.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 1877–1901.

Cao, K., Brbic, M., and Leskovec, J. (2022). Open-world semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. (2021). Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568.

Chen, G., Peng, P., Huang, Y., Geng, M., and Tian, Y. (2024). Adaptive discovering and merging for incremental novel class discovery. *arXiv preprint arXiv:2403.03382*.

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829.

Chi, H., Liu, F., Yang, W., Lan, L., Liu, T., Han, B., Niu, G., Zhou, M., and Sugiyama, M. (2022). Meta discovery: Learning to discover novel classes given very limited data. In *International Conference on Learning Representations (ICLR)*.

Chiaroni, F., Dolz, J., Masud, Z. I., Mitiche, A., and Ben Ayed, I. (2023). Parametric information maximization for generalized category discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1729–1739.

Chuyu, Z., Ruijie, X., and Xuming, H. (2023). Novel class discovery for long-tailed recognition. *Transactions on Machine Learning Research*.

Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y., and Ricci, E. (2023). Vocabulary-free image classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 30662–30680.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Du, R., Chang, D., Liang, K., Hospedales, T., Song, Y.-Z., and Ma, Z. (2023). On-the-fly category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11691–11700.

Fan, J., Liu, D., Chang, H., Huang, H., Chen, M., and Cai, W. (2024). Seeing unseen: Discover novel biomedical concepts via geometry-constrained probabilistic modeling. *arXiv preprint arXiv:2403.01053*.

Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. (2023). Improving clip training with language rewrites. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 35544–35575.

Fei, Y., Zhao, Z., Yang, S., and Zhao, B. (2022). Xcon: Learning with experts for fine-grained category discovery. In *The British Machine Vision Conference (BMVC)*, page 96.

Feng, J., Yang, Y., Xie, Y., Li, Y., Guo, Y., Guo, Y., He, Y., Xiang, L., and Ding, G. (2024). Debiased novel category discovering and localization. *arXiv preprint arXiv:2402.18821*.

Feng, W., Ju, L., Wang, L., Song, K., and Ge, Z. (2023). Towards novel class discovery: A study in novel skin lesions clustering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–33. Springer.

Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., and Ricci, E. (2021). A unified objective for novel class discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9284–9292.

Gao, F., Zhong, W., Cao, Z., Peng, X., and Li, Z. (2023a). Opengcd: Assisting open world recognition with generalized category discovery. *arXiv preprint arXiv:2308.06926*.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. (2023b). CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15.

Ge, C., Huang, R., Xie, M., Lai, Z., Song, S., Li, S., and Huang, G. (2023). Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., and Grover, A. (2022). Cyclip: Cyclic contrastive language-image pretraining. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6704–6719.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Gu, P., Zhang, C., Xu, R., and He, X. (2023). Class-relation knowledge distillation for novel class discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16474–16483.

Guo, L.-Z., Zhang, Y.-G., Wu, Z.-F., Shao, J.-J., and Li, Y.-F. (2022). Robust semi-supervised learning when not all classes have labels. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3305–3317.

Han, K., Li, Y., Vaze, S., Li, J., and Jia, X. (2023). What's in a name? beyond class indices for image recognition. *arXiv preprint*

arXiv:2304.02364.

Han, K., Rebuffi, S.-A., Ehrhardt, S., Vedaldi, A., and Zisserman, A. (2020). Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations (ICLR)*.

Han, K., Rebuffi, S.-A., Ehrhardt, S., Vedaldi, A., and Zisserman, A. (2022). Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781.

Han, K., Vedaldi, A., and Zisserman, A. (2019). Learning to discover novel visual categories via deep transfer clustering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8401–8409.

Hao, S., Han, K., and Wong, K.-Y. K. (2023). Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*.

Hasan, Z., Ahmed, M., Faridee, A. Z. M., Purushotham, S., Kwon, H., Lee, H., and Roy, N. (2023a). Nev-ncd: Negative learning, entropy, and variance regularization based novel action categories discovery. In *IEEE International Conference on Image Processing (ICIP)*, pages 2720–2724.

Hasan, Z., Faridee, A. Z. M., Ahmed, M., Purushotham, S., Kwon, H., Lee, H., and Roy, N. (2023b). Novel categories discovery from probability matrix perspective. *arXiv preprint arXiv:2307.03856*.

Hayes, T. L., de Souza, C. R., Kim, N., Kim, J., Volpi, R., and Larlus, D. (2024). Pandas: Prototype-based novel class discovery and detection. *arXiv preprint arXiv:2402.17420*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

He, Y., Shen, Z., and Cui, P. (2021). Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hogan, W., Li, J., and Shang, J. (2023). Open-world semi-supervised generalized relation discovery aligned in a real-world setting. *arXiv preprint arXiv:2305.13533*.

Hsu, Y.-C., Lv, Z., and Kira, Z. (2018). Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations (ICLR)*.

Hsu, Y.-C., Lv, Z., Schlosser, J., Odom, P., and Kira, Z. (2019). Multi-class classification without multi-class labels. In *International Conference on Learning Representations (ICLR)*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021a). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916.

Jia, X., Han, K., Zhu, Y., and Green, B. (2021b). Joint representation learning and novel category discovery on single-and multi-modal data. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 610–619.

Joseph, K., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., and Balasubramanian, V. N. (2022a). Novel class discovery without forgetting. In *European Conference on Computer Vision (ECCV)*, pages 570–586.

Joseph, K., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., and Balasubramanian, V. N. (2022b). Spacing loss for discovering novel categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3761–3766.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. (2023). Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

Kim, H., Suh, S., Kim, D., Jeong, D., Cho, H., and Kim, J. (2023). Proxy anchor-based unsupervised learning for continuous generalized category discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16688–16697.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 554–561.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using

crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Li, J., Savarese, S., and Hoi, S. (2023a). Masked unsupervised self-training for label-free image classification. In *International Conference on Learning Representations (ICLR)*.

Li, J., Xiong, C., and Hoi, S. C. (2021). Comatch: Semi-supervised learning with contrastive graph regularization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9475–9484.

Li, Q., Ma, Q., Nie, W., and Liu, A. (2023b). Reinforcement learning based multi-modal feature fusion network for novel class discovery. *arXiv preprint arXiv:2308.13801*.

Li, W., Fan, Z., Huo, J., and Gao, Y. (2023c). Modeling inter-class and intra-class constraints in novel class discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3449–3458.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Li, Z., Dai, B., Simsek, F., Meinel, C., and Yang, H. (2023d). Imbagcd: Imbalanced generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Li, Z., Meinel, C., and Yang, H. (2023e). Generalized categories discovery for long-tailed recognition. *arXiv preprint arXiv:2401.05352*.

Li, Z., Otholt, J., Dai, B., Hu, D., Meinel, C., and Yang, H. (2023f). Supervised knowledge may hurt novel class discovery performance. *Transactions on Machine Learning Research*.

Li, Z., Otholt, J., Dai, B., Meinel, C., Yang, H., et al. (2022). A closer look at novel class discovery from the labeled set. *arXiv preprint arXiv:2209.09120*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755.

Liu, J., Wang, Y., Zhang, T., Fan, Y., Yang, Q., and Shao, J. (2023a). Open-world semi-supervised novel class discovery. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4002–4010.

Liu, M., Roy, S., Li, W., Zhong, Z., Sebe, N., and Ricci, E. (2024). Democratizing fine-grained visual recognition with large language models. In *International Conference on Learning Representations (ICLR)*.

Liu, M., Roy, S., Zhong, Z., Sebe, N., and Ricci, E. (2023b). Large-scale pre-trained models are surprisingly strong in incremental novel class discovery. *arXiv preprint arXiv:2303.15975*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023c). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Y. and Tuytelaars, T. (2022). Residual tuning: Toward novel category discovery without labels. *IEEE Transactions on Neural Networks and Learning Systems*.

Ma, S., Zhu, F., Zhong, Z., Zhang, X.-Y., and Liu, C.-L. (2024). Active generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16890–16900.

MacQueen, J. (1965). Some methods for classification and analysis of multivariate observations.

In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, page 281.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. (2013). Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Marczak, D., Rypeść, G., Cygert, S., Trzciński, T., and Twardowski, B. (2023). Generalized continual category discovery. *arXiv preprint arXiv:2308.12112*.

Mayilvahanan, P., Wiedemer, T., Rusak, E., Bethge, M., and Brendel, W. (2024). Does CLIP's generalization performance mainly stem from high train-test similarity? In *International Conference on Learning Representations (ICLR)*.

Menon, S., Chandratreya, I. P., and Vondrick, C. (2023). Task bias in contrastive vision-language models. *International Journal of Computer Vision*, pages 1–15.

Nayak, N. V., Yu, P., and Bach, S. H. (2023). Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations (ICLR)*.

Otholt, J., Meinel, C., and Yang, H. (2024). Guided cluster aggregation: A hierarchical approach to generalized category discovery. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2618–2627.

Ouldnoughi, R., Kuo, C.-W., and Kira, Z. (2023). CLIP-GCD: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Peng, Z., Tian, Q., Xu, J., Jin, Y., Lu, X., Tan, X., Xie, Y., and Ma, L. (2023). Generalized category discovery in semantic segmentation. *arXiv preprint arXiv:2311.11525*.

Pu, N., Li, W., Ji, X., Qin, Y., Sebe, N., and Zhong, Z. (2024). Federated generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28741–28750.

Pu, N., Zhong, Z., and Sebe, N. (2023). Dynamic conceptional contrastive learning for generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3488.

Qin, G. and Eisner, J. (2021). Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

Rastegar, S., Doughty, H., and Snoek, C. G. (2023). Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Ren, S., Li, L., Ren, X., Zhao, G., and Sun, X. (2023a). Delving into the openness of CLIP. In *Findings of the Association for Computational Linguistics (ACL)*, pages 9587–9606.

Ren, S., Zhang, A., Zhu, Y., Zhang, S., Zheng, S., Li, M., Smola, A. J., and Sun, X. (2023b). Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 12569–12588.

Riz, L., Saltori, C., Ricci, E., and Poiesi, F. (2023). Novel class discovery for 3d point cloud semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9393–9402.

Rizve, M. N., Kardan, N., Khan, S., Khan, F. S., and Shah, M. (2022a). Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 382–401.

Rizve, M. N., Kardan, N., and Shah, M. (2022b). Towards realistic semi-supervised learning. In

European Conference on Computer Vision (ECCV), pages 437–455.

Roy, S., Liu, M., Zhong, Z., Sebe, N., and Ricci, E. (2022). Class-incremental novel class discovery. In *European Conference on Computer Vision (ECCV)*, pages 317–333.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 596–608.

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., et al. (2024). BioCLIP: A vision foundation model for the tree of life. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19412–19424.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. (2023a). Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Sun, Y. and Li, Y. (2023). Opencon: Open-world contrastive learning. *Transactions on Machine Learning Research*.

Sun, Y., Shi, Z., and Li, Y. (2024). A graph-theoretic framework for understanding open-world semi-supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Sun, Y., Shi, Z., Liang, Y., and Li, Y. (2023b). When and how does known class help discover unknown ones? provable understanding through spectral analysis. In *International Conference on Machine Learning (ICML)*, pages 33014–33043.

Tan, K. C., Liu, Y., Ambrose, B., Tulig, M., and Belongie, S. (2019). The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*.

Troisemaine, C., Flocon-Cholet, J., Gosselin, S., Vaton, S., Reiffers-Masson, A., and Lemaire, V. (2022). A method for discovering novel classes in tabular data. In *IEEE International Conference on Knowledge Graph (ICKG)*, pages 265–274.

Troisemaine, C., Lemaire, V., Gosselin, S., Reiffers-Masson, A., Flocon-Cholet, J., and Vaton, S. (2023a). Novel class discovery: an introduction and key concepts. *arXiv preprint arXiv:2302.12028*.

Troisemaine, C., Reiffers-Masson, A., Gosselin, S., Lemaire, V., and Vaton, S. (2023b). A practical approach to novel class discovery in tabular data. *arXiv preprint arXiv:2311.05440*.

Van Der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. (2022a). Generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7492–7501.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. (2022b). Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations (ICLR)*.

Vaze, S., Vedaldi, A., and Zisserman, A. (2023). No representation rules them all in category discovery. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*.

Wang, E., Peng, Z., Xie, Z., Liu, X., and Cheng, M.-M. (2024a). Get: Unlocking the multi-modal potential of clip for generalized category discovery. *arXiv preprint arXiv:2403.09974*.

Wang, H., Vaze, S., and Han, K. (2024b). Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *International Conference on Learning Representations (ICLR)*.

Wang, H., Yang, M., Wei, K., and Deng, C. (2023a). Hierarchical prompt learning for compositional zero-shot recognition. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1470–1478.

Wang, J., Zhang, L., Liu, J., Guo, T., and Wu, W. (2024c). Learning from semi-factuals: A debiased and semantic-aware framework for generalized relation discovery. *arXiv preprint arXiv:2401.06327*.

Wang, L., Liu, C., Guo, J., Dong, J., Wang, X., Huang, H., and Zhu, Q. (2023b). Federated continual novel class learning. *arXiv preprint arXiv:2312.13500*.

Wang, W., Lei, T., Chen, Q., and Liu, Y. (2024d). Semantic-guided novel category discovery. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Wang, W., Sun, Q., Zhang, F., Tang, Y., Liu, J., and Wang, X. (2024e). Diffusion feedback helps clip see better. *arXiv preprint arXiv:2407.20171*.

Wang, Y., Yu, Z., Wang, J., Heng, Q., Chen, H., Ye, W., Xie, R., Xie, X., and Zhang, S. (2024f). Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237.

Wang, Y., Zhong, Z., Qiao, P., Cheng, X., Zheng, X., Liu, C., Sebe, N., Ji, R., and Chen, J. (2024g). Discover and align taxonomic context priors for open-world semi-supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, Z., Salehi, B., Gritsenko, A., Chowdhury, K., Ioannidis, S., and Dy, J. (2020). Open-world class discovery with kernel networks. In *IEEE International Conference on Data Mining (ICDM)*, pages 631–640.

Wen, X., Zhao, B., and Qi, X. (2023). Parametric classification for generalized category discovery: A baseline study. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16590–16600.

Weng, T., Xiao, J., and Jiang, H. (2024). Decompose novel into known: Part concept learning for 3d novel class discovery. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. (2022). Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. (2023a). Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., Liu, T.-Y., et al. (2021). R-drop: Regularized dropout for neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 10890–10905.

Wu, Y., Chi, Z., Wang, Y., and Feng, S. (2023b). Metagcd: Learning to continually learn in generalized category discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1655–1665.

Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. (2024). Demystifying clip data. In *International Conference on Learning Representations (ICLR)*.

Yang, M., Wang, L., Deng, C., and Zhang, H. (2023a). Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3459–3468.

Yang, M., Zhu, Y., Yu, J., Wu, A., and Deng, C. (2022). Divide and conquer: Compositional experts for generalized novel class discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14268–14277.

Yang, X., Pan, X., King, I., and Xu, Z. (2023b). Generalized category discovery with clustering assignment consistency. *arXiv preprint arXiv:2310.19210*.

Ye, B., Gan, K., Wei, T., and Zhang, M.-L. (2023). Bridging the gap: Learning pace synchronization for open-world semi-supervised learning.

arXiv preprint arXiv:2309.11930.

Yu, Q., Ikami, D., Irie, G., and Aizawa, K. (2022). Self-labeling framework for novel category discovery over domains. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3161–3169.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. (2021). Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zang, Y., Li, W., Zhou, K., Huang, C., and Loy, C. C. (2022). Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*.

Zang, Z., Shang, L., Yang, S., Wang, F., Sun, B., Xie, X., and Li, S. Z. (2023). Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11858–11867.

Zhang, C., Hu, C., Xu, R., Gao, Z., He, Q., and He, X. (2022a). Mutual information-guided knowledge transfer for novel class discovery. *arXiv preprint arXiv:2206.12063*.

Zhang, J., Ma, X., Guo, S., and Xu, W. (2023a). Towards unbiased training in federated open-world semi-supervised learning. In *International Conference on Machine Learning (ICML)*, pages 41498–41509.

Zhang, L., Qi, L., Yang, X., Qiao, H., Yang, M.-H., and Liu, Z. (2022b). Automatically discovering novel visual categories with self-supervised prototype learning. *arXiv preprint arXiv:2208.00979*.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. (2022c). Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision (ECCV)*, pages 493–510.

Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., and Khan, F. (2023b). Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2775.

Zhang, S., Naseer, M., Chen, G., Shen, Z., Khan, S., Zhang, K., and Khan, F. S. (2024a). S3a: Towards realistic zero-shot classification via self structural semantic alignment. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 7278–7286.

Zhang, X., Jiang, J., Feng, Y., Wu, Z.-F., Zhao, X., Wan, H., Tang, M., Jin, R., and Gao, Y. (2022d). Grow and merge: A unified framework for continuous categories discovery. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 27455–27468.

Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al. (2024b). Recognize anything: A strong image tagging model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1724–1732.

Zhao, B. and Han, K. (2021). Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 22982–22994.

Zhao, B. and Mac Aodha, O. (2023). Incremental generalized category discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19137–19147.

Zhao, B., Wen, X., and Han, K. (2023). Learning semi-supervised gaussian mixture models for generalized category discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16623–16633.

Zhao, Y., Zhong, Z., Sebe, N., and Lee, G. H. (2022). Novel class discovery in semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349.

Zheng, J., Li, W., Hong, J., Petersson, L., and Barnes, N. (2022). Towards open-set object detection and discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3961–3970.

Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., and Sebe, N. (2021a). Neighborhood contrastive learning for novel class discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10867–10875.

Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., and Sebe, N. (2021b). Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9462–9470.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Zhu, F., Ma, S., Cheng, Z., Zhang, X.-Y., Zhang, Z., and Liu, C.-L. (2024). Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*.

Zhuang, J., Chen, Z., Wei, P., Li, G., and Lin, L. (2022). Open set domain adaptation by novel class discovery. In *IEEE International Conference on Multimedia and Expo (ICME)*.

Appendix A Additional Results

We aim to provide more insight into our proposed CPT with additional experimental results, including more ablation studies and results on more datasets and settings.

A.1 Effect of Token Length

We use fixed token lengths in Eq. (1) for all of our experiments, *i.e.*, $M = 4$ for task tokens, and $N = 2$ for class tokens. Here we experiment with different token lengths to better assess their effectiveness.

Task Tokens. We test four different token lengths for task tokens, *i.e.*, $M = 4, 8, 16, 32$, respectively, and conduct the experiments with all other parameters set to their default. As shown in Fig. A1, we can observe that a larger task token length M can generally lead to better performance. This is explainable since a larger task token length allows a stronger representation of task information (class definition pattern), which coincides with our hypothesis that better task information helps disambiguate the classes in GCD, and hence leads to better performance. In our experiments, we still use a moderate context token length, *i.e.*, $M = 4$, both for simplicity and to be consistent with existing prompt tuning methods in other fields (Nayak et al., 2023; Zhou et al., 2022a,b).

Class Tokens. Similarly, we also respectively test four different token lengths for class tokens, *i.e.*, $N = 1, 2, 4, 8$, and conduct the experiments with all other parameters set to their default. As shown in Fig. A1, $N = 2$ can generally lead to a good performance, and larger class token length does not necessarily result in better performance, which is different from the task tokens. This is in line with our design for class tokens by focusing on the specific class information for GCD. As a result, benefiting from the class definition pattern provided by task tokens, a moderate class token length is good enough for GCD performance.

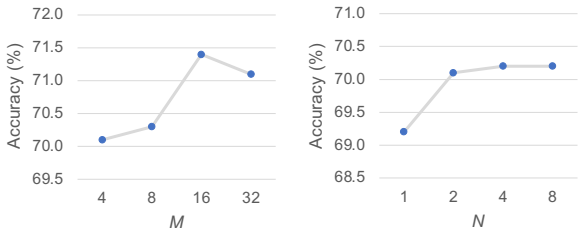

Fig. A1 Effect of token length with varying M (length of task tokens) or N (length of class tokens) in Eq. (1). GCD accuracy of *all class* on the CUB-200 dataset is reported.

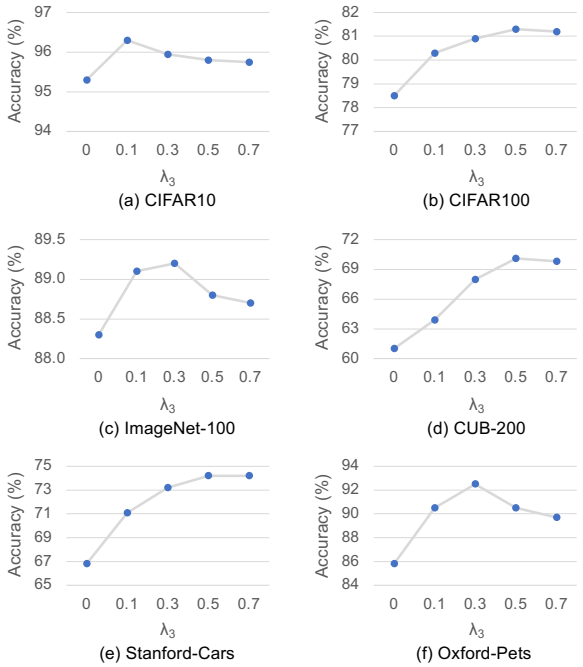

Fig. A2 Effect of *vision-language consistency* (VLC) with varying trade-off parameter λ_3 in Eq. (10). GCD accuracy of *all class* is reported.

A.2 Effect of Class Token Initialization

By default, both the task and class tokens are randomly initialized at the start of training. However, since some classes are labeled, we can initialize the class tokens directly using their ground truth (GT) class names. Tab. A1 presents a comparison of the experimental results with and without initialization using the GT class names.

We observe that performance remains relatively stable whether or not GT class name initialization is used, demonstrating the effectiveness of our proposed CPT in searching for accurate "task + class" prompts. However, in some cases, a more precise initialization with GT class names can slightly improve final performance, particularly for fine-grained datasets.

Table A1 Ablation study of class token initialization. GCD accuracy (%) with different class token initialization methods is reported.

Dataset →	CIFAR100			CUB-200		
Class Token Initialization ↓	All	Old	New	All	Old	New
Labeled + Unlabeled Classes:						
GT Init. + Random Init.	81.2	81.2	81.2	71.5	72.9	70.8
Random Init. + Random Init.	81.3	81.3	81.2	70.1	73.5	68.4

Table A2 Ablation study of *vision-vision consistency* (VVC) and *vision-language consistency* (VLC) regularization. GCD accuracy (%) on CIFAR100 and CUB-200 datasets is reported.

Dataset →	CIFAR100			CUB-200		
Method ↓	All	Old	New	All	Old	New
$\boldsymbol{v}$ + Adapter	57.8	73.9	25.4	36.3	47.3	30.8
$\boldsymbol{v}$ + Adapter + $\mathcal{L}_{\text{VVC}}$	78.5	79.6	76.5	61.0	66.9	58.1
$\boldsymbol{v}$ + Adapter + $\mathcal{L}_{\text{VVC}}$ + $\mathcal{L}_{\text{VLC}}$	**81.3**	**81.3**	**81.2**	**70.1**	**73.5**	**68.4**

Table A3 Ablation study of batch sizes. GCD accuracy (%) on CIFAR100 and CUB-200 datasets are reported.

Dataset →		CIFAR100			CUB-200		
Method ↓	Batch Size ↓	All	Old	New	All	Old	New
GCD (2022a)	128	73.0	76.2	66.5	51.3	56.6	48.7
GCD (2022a)	256	72.8	77.2	64.1	50.8	52.2	50.2
SimGCD (2023)	128	80.1	81.2	77.8	60.3	65.6	57.7
SimGCD (2023)	256	80.1	78.0	84.3	59.3	53.9	62.0
CPT (Ours)	128	80.7	81.1	80.1	70.9	73.6	69.6
CPT (Ours)	256	81.3	81.3	81.2	70.1	73.5	68.4

A.3 Effect of VVC and VLC

As supplementary to the ablation study results in Tab. 1, we clearly show the effectiveness of *vision-vision consistency* (VVC) and *vision-language consistency* (VLC) regularization in Tab. A2. It is also worth noting that, VVC serves as the essential training objective for unlabeled data by providing pseudo-labels for them, and hence VVC is a prerequisite for VLC as pseudo-labels are necessary for constructing local support set for a training batch, as discussed in Sec. 3.2.

In our proposed CPT, VVC regularization corresponds to a self-supervised loss term similar to the unsupervised contrastive loss used in GCD (Vaze et al., 2022a), and thus we follow

Table A4 GCD accuracy (%) on generic and fine-grained recognition datasets with different CLIP backbones.

Dataset →		CIFAR10			CIFAR100			ImageNet-100		
Method ↓	Backbone ↓	All	Old	New	All	Old	New	All	Old	New
CPT (Ours)	ResNet-50 (CLIP)	63.4	74.2	58.0	36.3	41.2	26.4	77.9	87.0	73.3
CPT (Ours)	ResNet-101 (CLIP)	61.2	73.6	55.0	34.7	39.3	25.4	80.6	88.7	76.5
CPT (Ours)	ViT-B/32 (CLIP)	94.3	95.1	93.9	76.3	78.4	71.9	86.5	93.1	83.1
CPT (Ours)	ViT-B/16 (CLIP)	96.3	96.2	96.4	81.3	81.3	81.2	89.2	95.2	86.1
CPT (Ours)	ViT-L/14 (CLIP)	**98.2**	**97.8**	**98.4**	**84.9**	**86.1**	**82.5**	**93.7**	**96.8**	**92.2**
Dataset →		CUB-200			Stanford-Cars			Oxford-Pets		
Method ↓	Backbone ↓	All	Old	New	All	Old	New	All	Old	New
CPT (Ours)	ResNet-50 (CLIP)	28.2	35.9	24.4	19.8	31.6	14.1	59.0	67.6	54.5
CPT (Ours)	ResNet-101 (CLIP)	31.1	39.4	26.9	20.1	32.3	14.2	64.0	72.2	59.6
CPT (Ours)	ViT-B/32 (CLIP)	59.2	65.4	56.1	62.0	74.9	54.6	86.1	91.0	83.5
CPT (Ours)	ViT-B/16 (CLIP)	70.1	73.5	68.4	74.2	84.3	69.3	92.5	93.9	91.7
CPT (Ours)	ViT-L/14 (CLIP)	**75.5**	**79.2**	**73.7**	**82.0**	**88.5**	**78.9**	**95.4**	**96.4**	**94.9**

recent works (Pu et al., 2023; Vaze et al., 2022a; Wen et al., 2023; Zhang et al., 2023b) to fix the ratio of supervised and unsupervised loss terms for fair comparisons, *i.e.*, setting $\lambda_1 = 0.35$ and $\lambda_2 = 0.65$ in Eq. (10).

For VLC, we show its effectiveness in Fig. A2 by varying the trade-off parameter λ_3 in Eq. (10). We can observe that VLC is clearly helpful for GCD performance, and a comparable trade-off parameter λ_3 to λ_1 and λ_2, *e.g.*, $\lambda_3 = 0.3$ or 0.5, is good enough for the performance.

A.4 Effect of Training Batch Size

Unlike most existing baselines that typically use a batch size of 128 and 200 training epochs, we employ a larger batch size of 256 and reduce the number of training epochs to 100 to accelerate the training process for CPT. To provide a thorough comparison, we experiment with different batch sizes to evaluate their impact on final performance. As shown in Tab. A3, both the baseline methods and our proposed CPT exhibit relatively stable performance across different batch sizes. This indicates that batch size might not be a highly sensitive hyperparameter for the existing methods.

A.5 Effect of CLIP ViT Variants

We follow recent works (Zhou et al., 2022a,b) to use a CLIP (Radford et al., 2021) pre-trained ViT-B/16 (Dosovitskiy et al., 2021) to encode images, which is consistent with existing GCD methods (Pu et al., 2023; Vaze et al., 2022a; Wen et al., 2023; Zhang et al., 2023b) that use a DINO (Caron et al., 2021) pre-trained ViT-B/16. In Tab. A4, we explore the effect of different visual backbones for our proposed CPT, *i.e.*, ResNet-50 (He et al., 2016), ResNet-101 (He et al., 2016), ViT-B/32 (Dosovitskiy et al., 2021), ViT-B/16 (Dosovitskiy et al., 2021), and ViT-L/14 (Dosovitskiy et al., 2021). Not surprisingly, CPT with the largest visual backbone, *i.e.*, ViT-L/14, consistently achieves the best GCD performance benefiting from our prompt learning strategy and consistent tuning objective.

A.6 Results w/ Different CLIP Implementations

Although the training data for CLIP (Radford et al., 2021) is not publicly available, subsequent works such as OpenCLIP (Cherti et al., 2023) and MetaCLIP (Xu et al., 2024) have curated their own training datasets and managed to reproduce CLIP's performance, with some even surpassing the original results.

For a thorough assessment, we examine the performance of our proposed CPT alongside several state-of-the-art baselines across various CLIP implementations. The experimental results, utilizing different backbones such as CLIP, OpenCLIP, and MetaCLIP, are presented in Tab. A5. We also reproduce the recent CLIP-based method, CLIP-GCD (Ouldnoughi et al., 2023), which enhances the original GCD method (Vaze et al., 2022a) by retrieving the nearest text descriptions from large

Table A5 GCD accuracy (%) with different CLIP implementations. Methods are based on either OpenAI's CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), or MetaCLIP (Xu et al., 2024), all with ViT-B/16 (Dosovitskiy et al., 2021) image encoder.

Dataset →		CIFAR10			CIFAR100		
Method ↓	Backbone ↓	All	Old	New	All	Old	New
GCD†	CLIP	94.6	97.5	93.1	71.9	77.8	60.2
SimGCD†	CLIP	96.9	95.7	97.5	80.9	85.2	72.5
CLIP-GCD	CLIP	96.6	97.2	96.4	85.2	85.0	85.6
CPT (Ours)	CLIP	96.3	96.2	96.4	81.3	81.3	81.2
GCD†	OpenCLIP	94.7	97.8	93.1	72.6	77.0	63.8
SimGCD†	OpenCLIP	96.6	95.1	97.4	80.5	83.4	74.6
CLIP-GCD†	OpenCLIP	94.5	97.8	92.8	73.0	79.1	60.9
CPT (Ours)	OpenCLIP	96.4	95.1	97.1	81.9	81.4	82.9
GCD†	MetaCLIP	96.9	**98.7**	95.9	77.6	82.4	68.1
SimGCD†	MetaCLIP	**98.3**	97.1	**98.8**	84.6	**87.7**	78.2
CLIP-GCD†	MetaCLIP	96.6	**98.7**	95.6	77.1	80.6	70.1
CPT (Ours)	MetaCLIP	98.0	97.9	98.0	**87.1**	86.8	**87.6**
Dataset →		CUB-200			Stanford-Cars		
Method ↓	Backbone ↓	All	Old	New	All	Old	New
GCD†	CLIP	52.5	58.6	49.5	62.5	73.5	57.1
SimGCD†	CLIP	68.6	71.4	67.2	70.2	84.8	63.1
CLIP-GCD	CLIP	62.8	**77.1**	55.7	70.6	88.2	62.2
CPT (Ours)	CLIP	70.1	73.5	68.4	74.2	84.3	69.3
GCD†	OpenCLIP	54.9	60.1	52.4	67.0	74.9	63.2
SimGCD†	OpenCLIP	63.2	69.0	60.3	70.6	78.3	66.8
CLIP-GCD†	OpenCLIP	56.5	66.6	51.4	65.6	77.9	59.7
CPT (Ours)	OpenCLIP	65.6	69.5	63.7	79.3	88.3	75.0
GCD†	MetaCLIP	55.6	62.0	52.3	65.9	74.0	62.1
SimGCD†	MetaCLIP	70.3	76.0	67.4	73.1	84.0	67.9
CLIP-GCD†	MetaCLIP	57.6	67.1	52.9	64.4	77.5	58.1
CPT (Ours)	MetaCLIP	**72.9**	76.5	**71.1**	**81.9**	**89.0**	**78.4**

†Reimplemented using (Open/Meta)CLIP's image encoder.

text caption corpora (in our experiments, we use captions from CC-12M (Changpinyo et al., 2021), identified as the best-performing dataset in CLIP-GCD's ablation study). However, we encountered challenges in replicating its strong performance when changing the CLIP backbone for CLIP-GCD. In contrast, we observed that CPT generally delivers better and more stable performance across different CLIP backbones, demonstrating its robustness and superior generalization capabilities.

A.7 Results on More Datasets

Here we evaluate our proposed CPT on another two GCD benchmark datasets that focus on fine-grained visual recognition in specialized domains, *i.e.*, FGVC-Aircraft (Maji et al., 2013) and Herbarium-19 (Tan et al., 2019). These two

Table A6 GCD accuracy (%) on FGVC-Aircraft and Herbarium-19 datasets. Methods are based on either DINO (Caron et al., 2021) or CLIP (Radford et al., 2021) pre-trained ViT-B/16 (Dosovitskiy et al., 2021) backbone.

Dataset →		FGVC-Aircraft			Herbarium-19		
Method ↓	Backbone ↓	All	Old	New	All	Old	New
k-means (1965)	DINO	16.0	14.4	16.8	13.0	12.2	13.4
RankStats+ (2022)	DINO	26.9	36.4	22.2	27.9	55.8	12.8
UNO+ (2021)	DINO	40.3	56.4	32.2	28.3	53.7	14.7
ORCA (2022)	DINO	22.0	31.8	17.1	20.9	30.9	15.5
GCD (2022a)	DINO	45.0	41.1	46.9	35.4	51.0	27.0
XCon (2022)	DINO	47.7	44.4	49.4	–	–	–
DCCL (2023)	DINO	–	–	–	–	–	–
PromptCAL (2023b)	DINO	52.2	52.2	52.3	37.0	52.0	28.9
SimGCD (2023)	DINO	54.2	59.1	51.8	44.0	58.0	36.4
GPC (2023)	DINO	46.3	42.5	47.9	36.5	51.7	27.9
PIM (2023)	DINO	48.7	61.9	42.1	42.3	56.1	34.8
k-means†	CLIP	27.3	22.8	29.6	9.3	9.7	9.0
GCD†	CLIP	41.2	39.3	42.2	35.5	52.4	26.5
PromptCAL†	CLIP	42.2	48.4	39.0	37.4	50.6	30.3
SimGCD†	CLIP	55.1	59.4	53.0	47.4	**64.2**	38.4
PIM†	CLIP	44.2	53.2	39.7	41.5	58.2	32.5
CLIP‡	CLIP	30.6	23.9	33.9	5.6	5.2	5.8
CLIP-GCD (2023)	CLIP	50.0	56.6	46.5	–	–	–
CPT* (Ours)	CLIP	53.3	56.1	51.9	45.3	61.2	36.7
GCD†	MetaCLIP	47.9	47.6	48.0	39.1	55.1	30.4
SimGCD†	MetaCLIP	60.2	65.9	57.4	50.2	63.9	42.8
CPT* (Ours)	MetaCLIP	**61.1**	**68.1**	**57.6**	**51.3**	63.4	**44.8**

†Reimplemented using (Meta)CLIP's image encoder.
‡Zero-shot CLIP with oracle class descriptions.
*Tuning the last ViT block (as done by the baselines).

datasets are found to be challenging, and hence the performance is either unpromising or even absent in some most recent GCD works, as shown in Tab. A6. We also found that baselines using the CLIP backbone do not achieve promising results compared with the DINO-based ones. Similarly, our proposed CPT does not outperform the existing state-of-the-art method, SimGCD (Wen et al., 2023), yet can achieve competitive results compared with both CLIP- and DINO-based baselines. Nevertheless, we show that tuning more parameters for CPT (*i.e.*, tuning the last ViT block, as also done by the baselines) brings a considerable performance boost, demonstrating its potential in highly specialized domains.

Given the fact that lower performance on FGVC-Aircraft has also been reported in other CLIP-based literature (Khattak et al., 2023; Zang et al., 2022; Zhou et al., 2022a,b), one possible explanation can be that the granularity of text descriptions used in CLIP's pre-training may not be fine enough to precisely characterize the

Table A7 Estimated category number K.

Dataset →	CF10	CF100	ImgNet	CUB	S-Cars	O-Pets
Truth	10	100	100	200	196	37
GCD (2022a)	9	100	109	231	230	40
CLIP† (2021)	9	102	104	193	212	39

†Category number estimation by applying GCD (Vaze et al., 2022a) on CLIP's image encoder.

Table A8 GCD accuracy (%) with ground-truth /estimated category numbers.

Dataset →	ImageNet-100			CUB-200			Stanford-Cars		
Method ↓	All	Old	New	All	Old	New	All	Old	New
GCD† (2022a)	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9
SimGCD† (2023)	83.0	93.1	77.9	60.3	65.6	57.7	53.8	71.9	45.0
GPC† (2023)	76.9	94.3	71.0	55.4	58.2	53.1	42.8	59.2	32.8
GCD*†	75.2	88.6	68.5	52.5	58.6	49.5	62.5	73.5	57.1
SimGCD*†	88.6	94.6	85.6	68.6	71.4	67.2	70.2	**84.8**	63.1
GPC*†	76.1	92.4	59.9	52.0	61.4	42.6	55.7	72.7	39.2
CPT† (Ours)	**89.2**	**95.2**	**86.1**	**70.1**	**73.5**	**68.4**	**74.2**	84.3	**69.3**
GCD‡ (2022a)	72.7	91.8	63.8	47.1	55.1	44.8	35.0	56.0	24.8
SimGCD‡ (2023)	81.7	91.2	76.8	61.5	66.4	59.1	49.1	65.1	41.3
GPC‡ (2023)	75.3	93.4	66.7	52.0	55.5	47.5	38.2	58.9	27.4
GCD*‡	74.4	92.4	65.4	54.3	61.4	50.7	61.4	76.3	54.2
SimGCD*‡	85.9	94.9	81.2	68.3	67.9	**68.5**	69.9	83.7	63.3
GPC*‡	74.7	89.4	60.0	48.9	59.5	38.4	53.8	68.8	39.2
CPT‡ (Ours)	**86.5**	**95.0**	**82.3**	**68.8**	**69.5**	**68.5**	**73.1**	**85.5**	**67.1**

†Training with ground-truth category number.
‡Training with estimated category number.
*Reimplemented using CLIP's image encoder.

subtle visual differences among classes in FGVC-Aircraft and Herbarium-19. In this regard, even though utilizing more sophisticated CLIP model variants (Fan et al., 2023; Sun et al., 2023a; Wang et al., 2024e; Xu et al., 2024) and/or tuning more parameters can already yield competitive results, adapting VLMs in a more efficient way remains an open challenge for future work, as discussed in Sec. C.

A.8 Results w/ Unknown Category Number

In our experiments, we follow common practice (Fini et al., 2021; Han et al., 2022; Wen et al., 2023; Zhong et al., 2021a) to assume the number of total categories known *a priori*. Here we explore the scenario where such knowledge is unknown. In particular, we conduct experiments by firstly estimating the category numbers using the off-the-shelf method (Vaze et al., 2022a), where a semi-supervised k-means is introduced to sweep over the candidate category numbers to find the optimal one that maximizes the clustering accuracy on the labeled subset.

Tab. A7 compares the estimation results with original image features (Vaze et al., 2022a) and with CLIP's image encoder. We can observe that the estimated category numbers with CLIP's image encoder are slightly more accurate, which can be credited to CLIP's stronger feature representation ability.

Tab. A8 shows the results of existing methods and our proposed CPT trained with the estimated category numbers in Tab. A7, where we adopt the category numbers estimated with original image features (Vaze et al., 2022a) for all involved methods for fair comparisons. We can observe that the performance with estimated category numbers slightly degenerates compared with the one under ground-truth category numbers in most cases. Still, our proposed CPT consistently achieves the best performance across all evaluation metrics, demonstrating its superiority in practical scenarios where the ground-truth category number is unknown.

Appendix B Implementation Details

Here we provide more implementation details in our experiments, including dataset details and training details.

B.1 GCD Benchmark Details

Here we elaborate on the dataset construction of GCD benchmarks, whose statistics are shown in Tab. A9. For all datasets, half of all classes are sampled as "Old" classes ($\mathcal{Y}^l$), and the rest of them are regarded as "New" classes ($\mathcal{Y}\backslash\mathcal{Y}^l$), except the CIFAR100 dataset, for which we use 80/20 classes as Old/New classes following the existing works. For each dataset, we directly use the first $|\mathcal{Y}^l|$ classes as Old classes according to the class index, except for CUB-200 and Stanford-Cars datasets, where we use the data splits provided in (Vaze et al., 2022a,b). For GCD training, we randomly sample half of the images from Old classes in the training split of each dataset as labeled subset $\mathcal{D}^l$, and regard all other training

Table A9 Dataset statistics of GCD benchmarks used in our experiments.

Data Type →	Generic			Fine-Grained		
Dataset →	CIFAR10	CIFAR100	ImageNet-100	CUB-200	Stanford-Cars	Oxford-Pets
Labeled Classes	5	80	50	100	98	19
Labeled Images	12,500	20,000	31,860	1,498	2,000	942
Unlabeled Classes	10	100	100	200	196	37
Unlabeled Images	37,500	30,000	95,255	4,496	6,144	2,738

Table A10 GCD accuracy (%) in the *customized* setting on NICO dataset.

Dataset →	NICO-10 (Animal)			NICO-10 (Context)			NICO-33			NICO-100		
Method ↓	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD (2022a)	95.1	97.3	94.0	45.4	66.4	29.4	38.6	58.4	25.1	48.0	67.0	39.1
SimGCD (2023)	97.3	97.0	97.4	55.4	66.8	46.8	45.9	67.7	31.0	46.3	65.1	37.4
GCD†	98.4	98.0	98.8	48.3	73.7	28.9	39.3	65.0	21.8	51.6	54.5	49.2
SimGCD†	**98.5**	97.5	**98.9**	57.6	73.9	45.1	46.6	72.2	29.1	52.5	69.4	44.6
CPT (Ours)	**98.5**	**98.6**	98.5	**63.1**	**75.4**	**53.8**	**53.5**	**75.7**	**38.4**	**61.2**	**73.5**	**55.4**

†Reimplemented using CLIP's image encoder.

Table B11 Dataset details of NICO-10 (Animal).

Old Classes	sheep, bird, rat, cat, horse		
New Classes	elephant, dog, bear, cow, monkey		
Old Classes #	5	Labeled Images #	2,395
Total Classes #	10	Unlabeled Images #	7,501

images as unlabeled subset $\mathcal{D}^u$. In order to search for hyperparameters, we also use a validation set, which comes from the Old classes in the test split of each dataset, disjoint from the labeled and unlabeled subsets in training.

B.2 NICO Dataset Details

As we introduce a practical setting of GCD in Sec. 4.3 (*i.e.*, *customized* GCD) to better assess existing methods under customized class definitions, the NICO (He et al., 2021) dataset is adopted to satisfy such a purpose. In this dataset, each of the 10 animal classes has 9 or 10 contexts, with 33 contexts in total. There are common contexts that are shared by different animal classes, *e.g.*, "on snow", "on grass", "in river", and also unique contexts that are owned by only one animal class, *e.g.*, "in circus", "at sunset", "flying". 17 out of 33 contexts are common contexts, and thus the correlation between classes and contexts should be less obvious.

We make use of the animal class (*e.g.*, "dog") and the background context (*e.g.*, "on grass") annotations for each image in the animal subset of NICO, and construct three benchmarks, *i.e.*, NICO-10, NICO-33, and NICO-100, all with the same training images yet different class definitions. The benchmark details are shown in Tabs. B11, B13 and B14. For comprehensive evaluations, we also construct another benchmark by merging some visually similar contexts into one class to reduce the context class number (to 10, the same as the number of animal classes in NICO-10, to isolate the effect of different class numbers), as shown in Tab. B12. We denote this new benchmark as "NICO-10 (Context)" to distinguish it from the original "NICO-10 (Animal)", and the GCD performance is shown in Tab. A10, where the baselines tend to be biased to the foreground-object-centric class definitions, whereas our proposed CPT demonstrates consistent superiority in handling customized class definitions.

B.3 Rare-Species Dataset Details

As detailed in Sec. 4.3, we adapted the Rare-Species dataset (Stevens et al., 2024) by identifying "truly novel" categories that are not found in OpenCLIP's training set, specifically LAION-400M (Schuhmann et al., 2021). LAION-400M is a publicly available dataset comprising approximately 400 million English (image URL, text caption) pairs. The Rare-Species dataset contains

Table B12 Dataset details of NICO-10 (Context).

Old Classes	{on_grass, eating_grass, spotted, at_sunset}, {at_home, in_zoo, in_cage, in_hole}, {in_forest, on_tree, climbing, on_branch}, {in_water, on_beach, in_river}, eating		
New Classes	{walking, running, on_road, in_street, on_ground}, {lying, sitting, standing}, {aside_people, on_shoulder, in_hand, in_circus}, {on_snow, white, flying}, {black, brown}		
Old Classes #	5	Labeled Images #	2,989
Total Classes #	10	Unlabeled Images #	6,907

Table B13 Dataset details of NICO-33.

Old Classes	in_water, lying, in_cage, on_ground, on_branch, on_shoulder, standing, flying, in_hand, at_home, in_hole, running, eating, in_street, on_tree, in_river, on_beach		
New Classes	in_circus, in_zoo, black, brown, white, eating_grass, on_snow, spotted, sitting, climbing, on_road, in_forest, on_grass, at_sunset, walking, aside_people		
Old Classes #	17	Labeled Images #	2,851
Total Classes #	33	Unlabeled Images #	7,045

around 12,000 images representing 400 endangered biological species. To identify the Rare-Species categories that do not appear in LAION-400M, we performed a regex string match between the 400 category names in Rare-Species and all the text captions in LAION-400M.

Specifically, each species category in Rare-Species is linked to both a scientific name (*e.g.*, Myliobatis goodei) and a common name (*e.g.*, southern eagle ray). It is important to note that a single common name can occasionally refer to multiple closely related species with distinct scientific names, as highlighted by Stevens et al. (2024). To maintain a thorough and rigorous screening process, any species found by either their scientific or common name in LAION-400M's text captions are excluded from our final list of "truly novel" categories.

After that, we identified 188 categories never found in LAION-400M. From these, we selected 180 categories with sufficient images to serve as the final unknown (new) categories. Another 180 categories were chosen from the remaining 212 categories to constitute the known (old) categories. This resulted in a dataset consisting of 360 categories, evenly divided between known and unknown, for the subsequent GCD evaluation. The final statistics of our adapted Rare-Species dataset can be found in Tab. B15.

B.4 Training Details

For fair comparisons, we follow existing GCD works (Vaze et al., 2022a; Wen et al., 2023) to sample the same amount of labeled and unlabeled data for each training batch. We also follow them to use the fixed ratio of supervised and self-supervised objectives, *i.e.*, setting $\lambda_1 = 0.35$ and $\lambda_2 = 0.65$ for Eq. (10), which can trade off the effect of balanced sampling and emphasize the training on unlabeled data. We further select optimal λ_3 in Eq. (10) according to the performance on the validation set mentioned in Sec. 4.1. Specifically, we use $\lambda_3 = 0.1, 0.5, 0.3, 0.5, 0.5, 0.3$ respectively for CIFAR10, CIFAR100, ImageNet-100, CUB-200, Stanford-Cars, Oxford-Pets, and the effect is discussed in Sec. A.3. The reimplementation of the state-of-the-art methods (Vaze et al., 2022a; Wen et al., 2023) all follows their default parameter settings. Other training details are also discussed in Sec. 4.1.

Appendix C More Discussions

Broader Impacts. VLMs like CLIP (Radford et al., 2021) are trained on massive data collected from the Internet, which contains inherent biases present in society, and can thus lead to biased or discriminatory outputs that may reinforce social inequalities. Therefore, the use of VLMs should be under the circumstance of comprehensively evaluated, especially in critical scenarios where people are directly affected.

Table B14 Dataset details of NICO-100.

Old Classes	sheep in_water, sheep lying, sheep eating, sheep on_snow, sheep on_road, sheep in_forest, sheep on_grass, sheep at_sunset, sheep walking, sheep aside_people, bird in_water, bird in_cage, bird on_ground, bird on_branch, bird on_shoulder, bird standing, bird flying, bird in_hand, bird eating, bird on_grass, rat in_water, rat lying, rat on_beach, rat on_tree, rat in_river, rat on_beach, rat eating, rat on_snow, rat in_forest, rat on_grass, cat in_water, cat on_ground, cat at_home, cat eating, cat in_street, cat on_tree, cat in_river, cat on_snow, cat on_grass, cat running, horse lying, horse in_home, horse on_beach, horse in_circus, horse in_river, horse on_beach, horse on_snow, horse in_forest, horse on_grass, horse running		
New Classes	elephant lying, elephant sitting, elephant eating, elephant in_street, elephant in_river, elephant on_beach, elephant in_zoo, elephant on_snow, elephant in_forest, elephant on_grass, dog in_water, dog lying, dog in_cage, dog at_home, dog on_beach, dog eating, dog on_tree, dog on_beach, dog on_snow, dog on_grass, bear in_water, bear lying, bear in_stree, bear on_tree, bear in_river, bear on_beach, bear in_circus, bear in_zoo, bear on_snow, bear in_forest, cow lying, cow eating_grass, cow at_home, cow eating, cow in_river, cow on_snow, cow in_hole, cow in_forest, cow on_grass, cow running, monkey in_water, monkey in_cage, monkey eating, monkey on_beach, monkey on_snow, monkey in_street, monkey on_tree, monkey in_forest, monkey on_grass, monkey running		
Old Classes #	50	Labeled Images #	2,395
Total Classes #	100	Unlabeled Images #	7,501

Table B15 Dataset details of Rare-Species. Class names are in the format of "scientific name (common name)".

Old Classes: acropora acuminata (staghorn coral), acropora cervicornis (staghorn coral), acropora echinata (ice fire), acropora florida (branch coral), acropora millepora (staghorn coral), addax nasomaculatus (addax), adetomyrma venatrix (dracula ant), agalychnis lemur (lemur leaf frog), ailuropoda melanoleuca (bamboo bear), albula vulpes (bonefish), alopias pelagicus (pelagic thresher), alouatta caraya (black howler), ambystoma mexicanum (axolotl), anaxyrus californicus (arroyo toad), andigena laminirostris (plate-billed mountain toucan), andrias davidianus (chinese giant salamander), andrias japonicus (japanese giant salamander), anser canagicus (emperor goose), anthus nilghiriensis (nilgiri pipit), ateles hybridus (variegated spider monkey), ateles paniscus (black spider monkey), balaeniceps rex (shoebill), balaenoptera bonaerensis (antarctic minke whale), balistes capriscus (gray triggerfish), balistes vetula (queen triggerfish), bathytoshia centroura (roughtail stingray), bettongia penicillata (woylie), bombus dahlbomii (giant bumble bee), bombycilla japonica (japanese waxwing), brachypelma auratum (mexican flameknee tarantula), brachypteracias leptosomus (short-legged ground roller), brachyteles hypoxanthus (northern muriqui), branta ruficollis (red-breasted goose), buceros bicornis (great hornbill), callorhinchus callorynchus (elephantfish), callorhinus ursinus (northern fur seal), capricornis sumatraensis (serow), carcharhinus acronotus (blacknose shark), carcharhinus albimarginatus (silvertip shark), carcharhinus falciformis (silky shark), carcharhinus melanopterus (blacktip reef shark), carcharhinus obscurus (dusky shark), cebuella pygmaea (pygmy marmoset), centrochelys sulcata (african spurred tortoise), cercopithecus mona (mona monkey), cetorhinus maximus (basking shark), chaetodon rainfordi (rainford's butterflyfish), chlidonias albostriatus (black-fronted tern), choerodon schoenleinii (blackspot tuskfish), choeronycteris mexicana (mexican long-tongued bat), crocodylus palustris (marsh crocodile), crocodylus rhombifer (cuban crocodile), cyclura lewisi (grand cayman blue iguana), cynarina lacrymalis (button coral), cynomys parvidens (utah prairie dog), dalatias licha (kitefin shark), dasyurus maculatus (tiger quoll), daubentonia madagascariensis (aye-aye), diploastrea heliopora (honeycomb coral), dolomedes plantarius (fen raft spider), drepanis coccinea (iiwi), eliomys quercinus (garden dormouse), epinephelus morio (red grouper), equus zebra (cape mountain zebra), esacus magnirostris (beach stone-curlew), esacus recurvirostris (great stone-curlew), eudyptes chrysolophus (macaroni penguin), eudyptes moseleyi (northern rockhopper penguin), eurycea rathbuni (texas blind salamander), euryceros prevostii (helmet vanga), falco deiroleucus (orange-breasted falcon), fimbriaphyllia ancora (hammer coral), formica lugubris (hairy wood ant), fulica alai (hawaiian coot), galaxea astreata (octopus coral), galeocerdo cuvier (tiger shark), gallinago media (great snipe), gavialis gangeticus (gavial), glyptemys muhlenbergii (bog turtle), gubernatrix cristata (yellow cardinal), gyps coprotheres (cape vulture), halcyon pileata (black-capped kingfisher), harpagoxenus canadensis (slave-making ant), harpia harpyja (harpy eagle), heliopora coerulea (blue coral), hemitragus jemlahicus (himalayan tahr), himantopus novaezelandiae (black stilt), hippocampus reidi (longsnout seahorse), hippocampus spinosissimus (spiny seahorse), hyaena hyaena (striped hyena), hylobates agilis (agile gibbon), hymenolaimus malacorhynchos (blue duck), iguana delicatissima (lesser antillean iguana), lamna nasus (porbeagle), larosterna inca (inca tern), lasiorhinus latifrons (southern hairy-nosed wombat), lathamus discolor (swift parrot), latimeria chalumnae (west indian ocean coelacanth), leipoa ocellata (malleefowl), leptopelis vermiculatus (peacock tree frog), leptoptilos javanicus (lesser adjutant), lutjanus cyanopterus (cubera snapper), macrochelys temminckii (alligator snapping turtle), manis javanica (sunda pangolin), merluccius bilinearis (silver hake), mesocricetus auratus (golden hamster), microcebus rufus (brown mouse lemur), mimus macdonaldi (hood mockingbird), montipora capitata (rice coral), myadestes obscurus ('oma'o), mycteria cinerea (milky stork), mycteroperca bonaci (black grouper), myliobatis aquila (spotted eagle ray), myotis grisescens (gray bat), myotis leibii (eastern small-footed bat), negaprion acutidens (sicklefin lemon shark), negaprion brevirostris (lemon shark), nothomyrmecia macrops (dinosaur ant), numenius tahitiensis (bristle-thighed curlew), ognorhynchus icterotis (yellow-eared parrot), okapia johnstoni (okapi), orcaella brevirostris (irrawaddy dolphin), oryx dammah (scimitar-horned oryx), oryx leucoryx (arabian oryx), pagophila eburnea (ivory gull), palinurus elephas (common spiny lobster), pangasianodon hypophthalmus (iridescent shark), pangshura sylhetensis (assam roofed turtle), pangshura tecta (indian roofed turtle), paradisaea rubra (red bird-of-paradise), pardalotus quadragintus (forty-spotted pardalote), phoebastria albatrus (short-tailed albatross), phoebastria irrorata (waved albatross), phoebetria palpebrata (light-mantled albatross), phyllobates vittatus (golfo dulce poison-dart frog), piliocolobus kirkii (zanzibar red colobus), podocnemis unifilis (yellow-spotted river turtle), polyergus breviceps (broad-headed slave-making ant), pongo abelii (sumatran orangutan), pontoporia blainvillei (franciscana), priodontes maximus (giant armadillo), pristis pectinata (smalltooth sawfish), procapra picticaudata (tibetan gazelle), procellaria aequinoctialis (white-chinned petrel), procellaria parkinsoni (black petrel), pseudodiploria strigosa (symmetrical brain coral), pteronura brasiliensis (giant otter), pyrrhula murina (azores bullfinch), python regius (ball python), raja asterias (starry skate), raja undulata (painted ray), rhyticeros undulatus (wreathed hornbill), saguinus oedipus (cotton-top tamarin), saiga tatarica (saiga), salvelinus confluentus (bull trout), scaphirhynchus albus (pallid sturgeon), setophaga kirtlandii (kirtland's warbler), siderastrea siderea (massive starlet coral), somateria fischeri (spectacled eider), somniosus microcephalus (gray shark), spheniscus humboldti (humboldt penguin), spheniscus mendiculus (galapagos penguin), sphyrna mokarran (great hammerhead), sphyrna tiburo (bonnethead), squalus acanthias (dogfish), squatina squatina (angel shark), starnoenas cyanocephala (blue-headed quail-dove), stenotomus chrysops (porgy), stephanoaetus coronatus (crowned eagle), strigops habroptila (kakapo), sylvilagus brasiliensis (tapeti), sylvilagus transitionalis (coney), taricha rivularis (red-bellied newt), thalassarche chrysostoma (grey-headed albatross), thamnophis gigas (giant garter snake), thelenota ananas (pineapple sea cucumber), thinornis cucullatus (hooded dotterel), trachelophorus giraffa (giraffe weevil), tremarctos ornatus (andean bear), tridacna squamosa (fluted giant clam)

New Classes: abronia deppii (deppe's arboreal alligator lizard), abronia graminea (sierra de tehuacan arboreal alligator lizard), acipenser sturio (baltic sturgeon), acropora digitifera (acropora digitifera), agaricia humilis (low relief lettuce coral), agaricia lamarcki (lamarck's sheet coral), allochrocebus lhoesti (l'hoest's monkey), alouatta belzebul (red-handed howler), amazona collaria (yellow-billed amazon), amazona vinacea (vinaceous-breasted amazon), ambystoma barbouri (streamside salamander), aneuretus simoni (sri lankan relict ant), anolis baracoae (baracoa anole), anolis barbouri (hispaniolan hopping anole), anolis bitectus (roof anole), anolis bonairensis (ruthven's anole), anolis cooki (cook's anole), anolis koopmani (koopmans anole), anolis luciae (saint lucian anole), anolis macrolepis (big-scaled anole), anolis marcanoi (red-fanned stout anole), anolis monteverde (puntarenas anole), anorrhinus austeni (austen's brown hornbill), anthracoceros coronatus (malabar pied hornbill), antrostomus vociferus (eastern whip-poor-will), aotus griseimembra (gray-legged night monkey), aotus lemurinus (gray-bellied night monkey), atelopus certus (darien stubfoot toad), atelopus flavescens (cayenne stubfoot toad), atelopus glyphus (pirri harlequin frog), athene blewitti (forest owlet), batrachoseps wrighti (oregon slender salamander), berenicornis comatus (white-crowned hornbill), bitis nasicornis (rhinoceros viper), bolitoglossa hartwegi (hartweg's mushroomtongue salamander), bolitoglossa helmrichi (coban climbing salamander), bombus crotchii (crotch bumble bee), bombus mexicanus (mexican bumble bee), bombus morrisoni (morrison bumblebee), bombus variabilis (variable cuckoo bumble bee), botaurus poiciloptilus (australasian bittern), callithrix aurita (buffy-tufted marmoset), campylopterus ensipennis (white-tailed sabrewing), campylopterus villaviscensio (napo sabrewing), capra sibirica (siberian ibex), carettochelys insculpta (fly river turtle), catharus bicknelli (bicknell's thrush), cebus unicolor (spix's white-fronted capuchin), ceratogomphus triceraticus (cape thorntail), ceratophora tennentii (rhinoceros agama), certhidea olivacea (green warbler-finch), ceruchus chrysomelinus (ceruchus chrysomelinus), chaetodon trifascialis (rightangled butterflyfish), chiloscyllium plagiosum (whitespotted bamboo shark), chioglossa lusitanica (gold-striped salamander), chrysoblephus laticeps (roman seabream), clonophis kirtlandii (kirtland's snake), conolophus pallidus (barrington island iguana), conus mercator (merchant cone), copsychus sechellarum (seychelles magpie-robin), ctenomys pearsoni (pearson's tuco-tuco), ctenosaura bakeri (baker's spinytail iguana), cyclonaias tuberculata (purple wartyback), desmognathus aeneus (seepage salamander), desmognathus imitator (imitator salamander), dichocoenia stokesii (domed star coral), dorymyrmex insanus (crazy pyramid ant), egretta vinaceigula (slaty egret), elaphrus viridis (delta green ground beetle), elaphurus davidianus (père david's deer), epinephelus polyphekadion (comouflage grouper), etheostoma lepidum (greenthroat darter), eupodotis caerulescens (blue korhaan), fungia fungites (common mushroom coral), galaxea fascicularis (galaxy coral), glareola nordmanni (black-winged pratincole), graptemys oculifera (ringed map turtle), gymnura marmorata (california butterfly ray), haliotis roei (roe's abalone), helioseris cucullata (sunray lettuce coral), hemigomphus theischingeri (rainforest vicetail), hippoglossoides platessoides (american dab), hydnophora exesa (spine coral), hydromantes brunus (limestone salamander), hydromedusa maximiliani (brazilian snake-necked turtle), hylobates funereus (east bornean gray gibbon), hynobius quelpaertensis (jeju salamander), iberolacerta bonnali (pyrenean rock lizard), kinyongia multituberculata (west usambara two-horned chamaeleon), lacerta schreiberi (iberian emerald lizard), lepidochelys kempii (atlantic ridley), lepidurus packardi (vernal pool tadpole shrimp), leptobasis melinogaster (cream-tipped swampdamsel), lophura swinhoii (swinhoe's pheasant), luehdorfia japonica (japanese luehdorfia), macaca thibetana (milne-edwards's macaque), macrodontia cervicornis (giant jawed sawyer), madracis formosa (eight-ray finger coral), manis crassicaudata (indian pangolin), martes gwatkinsii (nilgiri marten), mauremys reevesii (reeves' turtle), microlophus bivittatus (san cristobal lava lizard), miopithecus ogouensis (northern talapoin monkey), mokopirirakau granulatus (gray's sticky-toed gecko), monias benschi (subdesert mesite), myliobatis freminvillei (bullnose eagle ray), myliobatis goodei (southern eagle ray), naja siamensis (indo-chinese spitting cobra), naultinus stellatus (starry tree gecko), neochen jubata (orinoco goose), neophema chrysogaster (orange-bellied parakeet), nesospingus speculiferus (puerto rican tanager), neurergus kaiseri (lorestan newt), notophthalmus perstriatus (striped newt), notorynchus cepedianus (bluntnose sevengill shark), odontophorus melanonotus (black-backed wood quail), oligosoma macgregori (macgregor's skink), onychoprion aleuticus (aleutian tern), orbicella annularis (lobed star coral), paramesotriton hongkongensis (hong kong warty newt), parantica nilgiriensis (nilgiri tiger), pateobatis fai (pink whipray), pateobatis jenkinsii (jenkins' whipray), pavona decussata (cactus coral), penelope ochrogaster (chestnut-bellied guan), peruphasma schultei (black beauty stick insect), phaenicophaeus sumatranus (chestnut-bellied malkoha), pheidole neokohli (pheidole neokohli), phelsuma standingi (banded day gecko), piliocolobus tephrosceles (ashy red colobus), pipile pipile (common piping guan), pitta megarhyncha (mangrove pitta), plethodon shenandoah (shenandoah salamander), plethodon stormi (siskiyou mountains salamander), pleurodeles waltl (spanish newt), polyergus nigerrimus (black amazon ant), potamilus amphichaenus (texas heelsplitter), presbytis melalophos (black-crested sumatran langur), presbytis thomasi (north sumatran leaf monkey), propithecus deckenii (van der decken's sifaka), protosticta sanguinostigma (red spot reedtail), pseudodiploria clivosa (knobby brain coral), pseudophilautus wynaadensis (waynad bush frog), ptyas korros (chinese ratsnake), puffinus huttoni (hutton's shearwater), python sebae (african rock python), rana iberica (iberian frog), rana sierrae (sierra nevada yellow-legged frog), rhacophorus lateralis (winged gliding frog), rhinoderma darwinii (darwin's frog), rhyacotriton cascadae (cascade torrent salamander), rhyacotriton olympicus (olympic salamander), rhyticeros narcondami (narcondam hornbill), rostroraja alba (bottlenosed skate), saara hardwickii (hardwick's spiny-tailed lizard), saga pedo (predatory bush-cricket), saguinus leucopus (silvery-brown bare-face tamarin), sapajus cay (hooded capuchin), sapajus nigritus (black capuchin), scomberomorus commerson (narrow-barred spanish mackerel), sericossypha albocristata (white-capped tanager), sorex alpinus (alpine shrew), sotalia guianensis (guiana dolphin), sousa plumbea (indian humpback dolphin), stichopus herrmanni (herrmann's sea cucumber), taeniurops meyeni (round ribbontail ray), thalassarche carteri (indian yellow-nosed albatross), thalassarche chlororhynchos (atlantic yellow-nosed albatross), thinornis novaeseelandiae (shore dotterel), trachypithecus auratus (east javan langur), trionyx triunguis (african softshell turtle), triturus pygmaeus (pygmy marbled newt), trogon bairdii (baird's trogon), turbinaria reniformis (yellow cabbage coral), tursiops aduncus (indian ocean bottlenose dolphin), tylototriton shanjing (yunnan newt), tylototriton verrucosus (yunnan newt), urolophus kapalensis (kapala stingaree), urosaurus auriculatus (socorro island tree lizard), zonites algirus (algiers snail)

Old Classes #	180	Labeled Images #	2,700
Total Classes #	360	Unlabeled Images #	8,100