# D2G: DEBIASED LEARNING WITH DISTRIBUTION GUIDANCE FOR GENERALIZED CATEGORY DISCOVERY

**Anonymous authors**
Paper under double-blind review

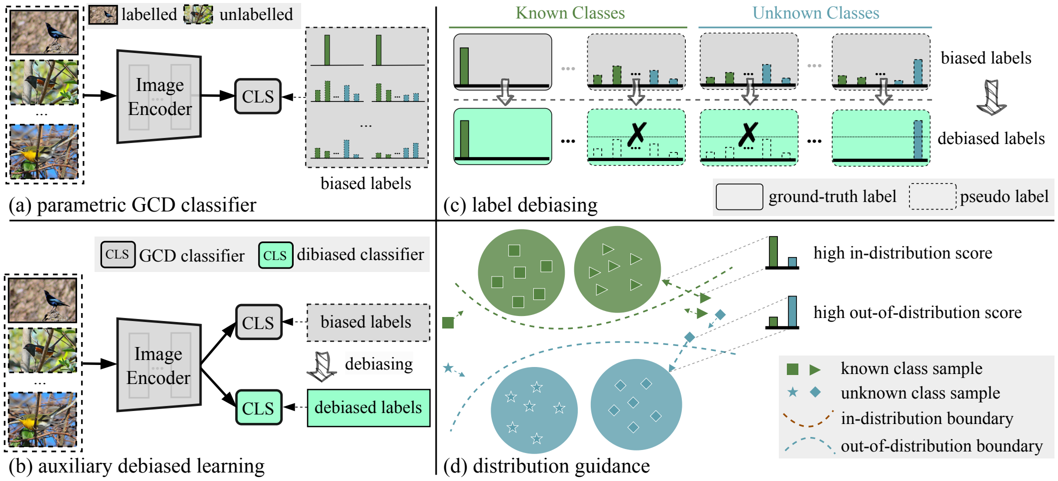
## ABSTRACT

In this paper, we tackle the problem of Generalized Category Discovery (GCD). Given a dataset containing both labelled and unlabelled images, the objective is to categorize all images in the unlabelled subset, irrespective of whether they are from known or unknown classes. In GCD, an inherent label bias exists between known and unknown classes due to the lack of ground-truth labels for the latter. State-of-the-art methods in GCD leverage parametric classifiers trained through self-distillation with soft labels, leaving the bias issue unattended. Besides, they treat all unlabelled samples uniformly, neglecting variations in certainty levels and resulting in suboptimal learning. Moreover, the explicit identification of semantic distribution shifts between known and unknown classes, a vital aspect for effective GCD, has been neglected. To address these challenges, we introduce the **D**ebiased Learning with **D**istribution **G**uidance (**D2G**) framework. Initially, D2G co-trains an auxiliary debiased classifier in the same feature space as the GCD classifier, progressively enhancing the GCD features. Moreover, we introduce a semantic distribution detector in a separate feature space to implicitly boost the learning efficacy of GCD. Additionally, we employ a curriculum learning strategy based on semantic distribution certainty to steer the debiased learning at an optimized pace. Thorough evaluations on GCD benchmarks demonstrate the consistent state-of-the-art performance of our D2G framework, highlighting its superiority.

## 1 INTRODUCTION

Over the years, the field of computer vision has witnessed remarkable progress in diverse tasks such as object detection Girshick (2015); Ren et al. (2015), classification Simonyan & Zisserman (2015); He et al. (2016), and segmentation He et al. (2017); Wang et al. (2020). These advancements have predominantly stemmed from the availability of expansive labelled datasets Deng et al. (2009); Lin et al. (2014). However, the prevalent insufficiency of training data in real-world scenarios is a noteworthy concern. This has engendered a surge in research on semi-supervised learning Chapelle et al. (2009) and self-supervised learning Jing & Tian (2020), yielding promising outcomes in comparison to supervised learning approaches. Recently, the task of category discovery, which was initially studied as novel category discovery (NCD) Han et al. (2019) and subsequently extended to its relaxed variant, generalized category discovery (GCD) Vaze et al. (2022b), has emerged as a research task attracting increasing attention. GCD considers a partially-labelled dataset, where the unlabelled subset may contain instances from both labelled and unseen classes. The objective is to learn to transfer knowledge from labelled data to categorize unlabelled data.

In GCD, there exists an inherent label bias between known and unknown classes due to the absence of ground-truth labels for the latter. This label bias has the potential to cause the model to inadvertently develop a decision rule making confident predictions that inclined to known classes. Similar problem has been identified in the area of long-tailed recognition Tang et al. (2020); Yang et al. (2022). Besides, in other fields such as object classification Choi et al. (2019); Bahng et al. (2020); Geirhos et al. (2020), it is widely known that model performance suffers from task-specific bias. State-of-the-art parametric classifier methods in GCD, such as those proposed by Wen et al. (2023); Zhao et al. (2023); Vaze et al. (2023), leverage the self-distillation Caron et al. (2021) mechanism based on soft labels generated from the model's predictions of another image view. While these methods have shown promising results, they still rely on biased labels for training (as shown



(a) parametric GCD classifier (c) label debiasing (b) auxiliary debiased learning (d) distribution guidance

Figure 1: (a) The commonly used parametric GCD classifier Wen et al. (2023) is trained on labelled and unlabelled images using ground-truth hard labels and soft labels, respectively. (b) The auxiliary debiased learning: training an debiased classifier using debiased labels. (c) The process of label debiasing: keep the ground-truth labels unchanged and transform soft labels to one-hot hard labels using a specified threshold; samples that do not meet the threshold are removed. (d) The illustration of distribution guidance: if a sample receives a high in-distribution/out-of-distribution score, its weight in GCD training will be increased accordingly.

in Fig. 1(a)). The issue of label bias remains an unattended problem in the realm of GCD. Additionally, existing approaches uniformly handle all unlabelled samples without explicitly accounting for their different certainty, which may introduce noise to the model training due to unreliable samples. Moreover, they do not explicitly address semantic shifts, especially in a scenario like GCD involving both known and unknown classes within unlabelled data. Notably, these concerns have been demonstrated to provide significant advantages in related tasks, such as open-world semi-supervised learning Cao et al. (2022). In this area, OpenCon Sun & Li (2022) has attempted to identify novel samples based on their proximity to known prototypes. However, its performance is heavily contingent on predefined distance thresholds, ultimately yielding suboptimal accuracy.

To tackle these challenges, we propose a novel framework, called **D**ebiased Learning with **D**istribution **G**uidance (**D2G**), incorporating several innovative techniques tailored for GCD. Firstly, we introduce a novel auxiliary debiased learning paradigm for GCD (as shown in Fig. 1(b) and (c)). This method entails training an auxiliary debiased classifier in the same feature space as the GCD classifier. Unlike the GCD classifier, both labelled and unlabelled data are trained using one-hot hard labels to prevent label bias between known and unknown classes. Secondly, to discern the semantic distribution of unlabelled samples, we propose to learn a semantic distribution detector in a decoupled normalized feature space, which we empirically find it enhance the learning effect of GCD implicitly. Furthermore, we propose to measure the certainty of a sample based on its semantic distribution detection score. This certainty score then enables the gradual inclusion of unlabelled samples from both known and unknown classes during training, allowing the auxiliary debiased learning to function in a curriculum learning approach (as shown in Fig. 1(d)), thus further enhancing its performance. We develop our D2G framework upon the strong parametric baseline Wen et al. (2023). By effectively incorporating these components into a unified framework, D2G can be trained end-to-end in a single stage while not introducing any additional computational burden during inference. Despite its simplicity, D2G attains unparalleled performance on the public GCD datasets, including the generic classification datasets CIFAR-10 Krizhevsky et al. (2009), CIFAR-100 Krizhevsky et al. (2009), and ImageNet Deng et al. (2009), as well as the fine-grained SSB Vaze et al. (2022a) benchmark.

We make the following key contributions in this work: (1) We propose D2G, a novel framework that addresses the challenging GCD task by considering both label bias and semantic shift, marking the first exploration of these aspects for the challenging GCD task. (2) Within D2G, we propose a novel auxiliary debiased learning paradigm to optimize the clustering feature space, in conjunction with the distribution shift detector in a distinct feature space. They work tightly to enhance the model's discovery capabilities. (3) We introduce a curriculum learning mechanism that steers the debiased learning process using a distribution certainty score, effectively mitigating the negative impact of uncertain samples. (4) Through extensive experimentation on public GCD benchmarks, D2G consistently demonstrates its effectiveness and achieves superior performance.

## 2 RELATED WORK

**Category Discovery.** This task is initially studied as Novel Category Discovery (NCD) Han et al. (2019), aiming to discover categories from unlabelled data consisting of samples from novel categories, by transferring the knowledge from the labelled categories. Many methods have been proposed to tackle NCD, such as Han et al. (2019; 2020; 2021); Fini et al. (2021); Zhao & Han (2021); Joseph et al. (2022). Vaze et al. (2022b) extends NCD Han et al. (2019) to a more relaxed task, Generalized Category Discovery (GCD), wherein unlabelled datasets encompass both known and unknown categories. A baseline method is presented for this task, incorporating self-supervised representation learning and semi-supervised $k$-means clustering, and extending popular NCD methods such as RankStats Han et al. (2020) and UNO Fini et al. (2021) to GCD. CiPR Hao et al. (2024) proposes to bootstrap the representation by leveraging cross-instance positive relations in the partially labelled data for contrastive learning. Cao et al. (2022) addresses a similar problem to GCD from the perspective of semi-supervised learning. SimGCD Wen et al. (2023) introduces a strong parametric baseline achieving promising performance improvements. In Vaze et al. (2023), a new dataset is introduced to illustrate the limitations of unsupervised clustering in GCD. To address these limitations, a method based on the ‘mean-teachers’ approach is proposed. In Rastegar et al. (2023), a category coding approach is introduced, considering category prediction as the outcome of an optimization problem. Recently, SPTNet Wang et al. (2024) is proposed to consider the spatial property of images and presents a spatial prompt tuning method for GCD, enabling the model to better focus on object parts for knowledge transfer.

**Debiased Learning.** The issue of bias in data and the susceptibility of machine learning algorithms to such bias have been widely recognized as crucial challenges across diverse tasks. Numerous methodologies have been developed to address and alleviate biases inherent in training datasets or tasks. The studies by Ponce (2006); Torralba & Efros (2011) elucidate that many training sets impose regularity conditions that are impractical in real-world settings, leading to machine learning models trained on such data failing to generalize in the absence of these conditions. Furthermore, recent research by Hendrycks et al. (2021); Xiao et al. (2021); Li et al. (2021) demonstrate biases in state-of-the-art object recognition models towards specific backgrounds or textures associated with object classes. Additionally, Sagawa et al. (2020) investigate the vulnerability of overparametrized models to spurious correlations, resulting in elevated test errors for minority groups. Notably, large language models also exhibit biased predictions towards certain genders or races, as indicated by Cheng et al. (2021). Furthermore, the severity of biased predictions and fairness concerns related to deployed models are extensively explored across various tasks Zemel et al. (2013); Noble (2018); Bolukbasi et al. (2016). In this paper, we examine the inherent *label bias* in GCD, representing the initial exploration of this issue.

**Out-of-distribution Detection.** In the realm of out-of-distribution (OOD) detection, the objective is to identify samples or data points that originate from a distribution distinct from the one on which the model was trained, encompassing both semantic and domain distributions Yang et al. (2021). The simplest method in this area involves utilizing the predicted softmax class probability to detect OOD samples Hendrycks & Gimpel (2017). ODIN Liang et al. (2018) further enhances this approach by introducing temperature scaling and input pre-processing. Additionally, Bendale & Boult (2016) proposes an alternative approach by calculating the score for an unknown class using a weighted average of all other classes. OOD detection has been applied in various open-set tasks, such as open-set semi-supervised learning Yu et al. (2020) and universal domain adaptation Saito & Saenko (2021), where it is utilized to select in-distribution data during training. In contrast, our focus lies in the exploration of semantic shift detection considering the specific challenges of GCD. OpenCon Sun & Li (2022) has attempted to explore the semantic shift for open-world semi-supervised learning. However, its reliance on a predefined distance threshold to rigidly distinguish inliers and outliers leads to suboptimal accuracy. In contrast, our method takes a distinct approach by avoiding a rigid separation. We subtly utilize the predicted OOD score by our model as a guiding factor for debiased learning, further enabling a curriculum learning scheme.

# 3 PRELIMINARIES

## 3.1 PROBLEM STATEMENT

Generalized category discovery (GCD) aims to learn a model that can not only correctly classify the unlabelled samples of known categories but also cluster those of unknown categories. Given an unlabelled dataset $\mathcal{D}_u = \{(\boldsymbol{x}_i^u, y_i^u)\} \in \mathcal{X} \times \mathcal{Y}_u$ and a labelled dataset $\mathcal{D}_l = \{(\boldsymbol{x}_i^l, y_i^l)\} \in \mathcal{X} \times \mathcal{Y}_l$, where $\mathcal{Y}_u$ and $\mathcal{Y}_l$ are their label sets respectively. The unlabelled dataset contains samples from both known and unknown categories, *i.e.*, $\mathcal{Y}_l \subset \mathcal{Y}_u$. The number of labelled categories is $M = |\mathcal{Y}_l|$. We assume the number of categories $K = |\mathcal{Y}_l \cup \mathcal{Y}_u|$ to be known following previous works Han et al. (2021); Wen et al. (2023); Vaze et al. (2023). When it is unknown, methods like Han et al. (2019); Vaze et al. (2022b) can be applied to provide a reliable estimation.

## 3.2 BASELINE

Wen et al. (2023) introduces a robust parametric GCD baseline, which has been widely adopted in the field ever since Vaze et al. (2023); Wang et al. (2024). It employs a parametric classifier, implemented in a self-distillation manner Caron et al. (2021). The classifier is randomly initialized with $K$ normalized category prototypes $\mathcal{C} = \{\boldsymbol{c}_1, ..., \boldsymbol{c}_K\}$. For the randomly augmented view of an image $\boldsymbol{x}_i$ and its corresponding normalized hidden feature vector $\boldsymbol{h}_i = \phi(\boldsymbol{x}_i)/||\phi(\boldsymbol{x}_i)||$, the output probability for the $k$th category is given by:

$$\boldsymbol{p}_i{}^{(k)} = \frac{\exp(\boldsymbol{h}_i \cdot \boldsymbol{c}_k/\tau_s)}{\sum_{j=1}^{K} \exp(\boldsymbol{h}_i \cdot \boldsymbol{c}_j/\tau_s)}, \tag{1}$$

where $\tau_s$ is the scaling temperature for this *'student'* view. The soft label $\boldsymbol{q}_i$ is produced by the *'teacher'* view with a sharper temperature $\tau_t$ using another augmented view in the same fashion. The self-distillation loss of the two views is then simply calculated following the cross-entropy loss $\ell_{ce}(\boldsymbol{q}', \boldsymbol{p}) = -\sum_{j=1}^{K} \boldsymbol{q}'^{(j)} \log \boldsymbol{p}^{(j)}$. Given a mini-batch $\mathcal{B}$ containing both labelled samples $\mathcal{B}_l$ and unlabelled images $\mathcal{B}_u$, the self-distillation loss is calculated using all the samples in the mini-batch:

$$\mathcal{L}_{cls}^{u} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell_{ce}(\boldsymbol{q}_i', \boldsymbol{p}_i) - \xi H(\overline{\boldsymbol{p}}), \tag{2}$$

where $\overline{\boldsymbol{p}} = \frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} (\boldsymbol{p}_i + \boldsymbol{p}_i')$ denotes the mean prediction within a batch and its entropy $H(\overline{\boldsymbol{p}}) = -\sum_{j=1}^{K} \overline{\boldsymbol{p}}^{(j)} \log \overline{\boldsymbol{p}}^{(j)}$ weighted by $\xi$. For the labelled samples, the supervised classification loss is written as $\mathcal{L}_{cls}^{s} = \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} \ell_{ce}(\boldsymbol{p}_i, \boldsymbol{y}_i)$, where $\boldsymbol{y}_i$ represents the one-hot vector corresponding to the ground-truth label $y_i$. The whole classification objective is $\mathcal{L}_{cls} = (1 - \lambda_b^{gcd})\mathcal{L}_{cls}^{u} + \lambda_b^{gcd} \mathcal{L}_{cls}^{s}$. Combining with the representation learning loss $\mathcal{L}_{rep}$ adopted from Vaze et al. (2022b), the overall training objective becomes:

$$\mathcal{L}_{gcd} = \mathcal{L}_{cls} + \mathcal{L}_{rep}. \tag{3}$$

Through training with $\mathcal{L}_{gcd}$ on both $\mathcal{D}_l$ and $\mathcal{D}_u$, the classifier can directly predict the labels for unlabelled samples after training.

# 4 DEBIASED LEARNING WITH DISTRIBUTION-GUIDANCE FOR GCD

In this section, we present our Debiased Learning with Distribution-Guidance (D2G) framework for GCD (see Fig. 2). First, in Sec. 4.1, we present the semantic distribution learning on the GCD task. Next, in Sec. 4.2, we demonstrate the training paradigm of the debiased classifier. Finally, we describe the joint training and inference process of our full framework in Sec. 4.3.

## 4.1 LEARNING SEMANTIC DISTRIBUTION

OOD detection methods have been employed in tasks like universal domain adaptation Saito & Saenko (2021) and open-set semi-supervised learning Yu et al. (2020), obtaining improved performance. In these tasks, the identified OOD samples are treated as a single *background* class to avoid affecting the recognition of unlabelled samples from the labelled classes, and the distribution shifts can be of any type. In GCD, we are particularly interested in identifying the semantic shifts. The instances from the labelled classes are considered in-distribution (ID) samples, while the instances

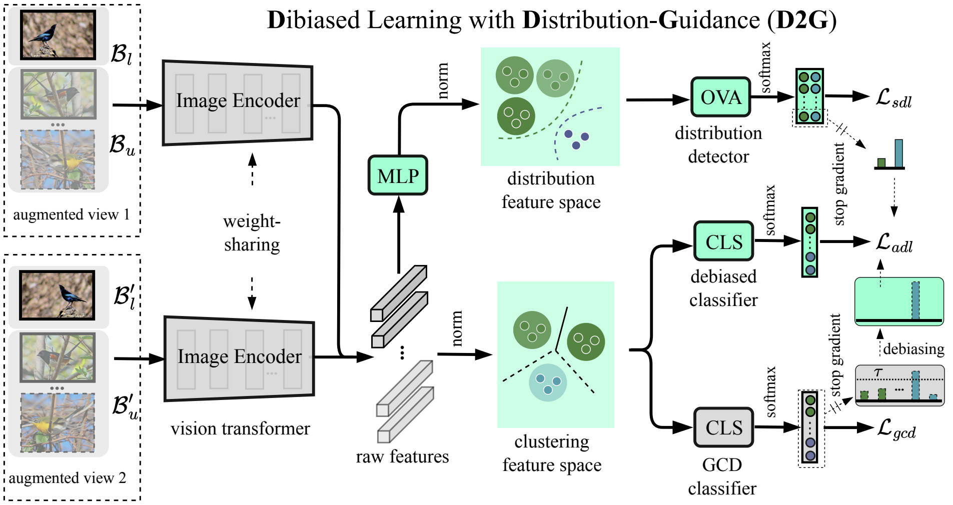


Figure 2: Debiased Learning with Distribution-Guidance (D2G) framework. Our model employs a Siamese architecture to handle samples through two augmented image views. Raw features in the upper branch undergo nonlinear transformation using a MLP, followed by normalization into a feature space for semantic distribution learning employing a one-vs-all (OVA) classifier. In the lower branch, the GCD classifier is trained on normalized features. Predictions from both branches collectively contribute to the training of the debiased classifier. As D2G aligns with prior work in the representation learning branch, it's not explicitly depicted here.

from the novel classes are considered OOD samples. However, the potential of effectively introducing OOD techniques for GCD remains under-explored. An intuitive approach for OOD detection is to use the self-entropy of the GCD classifier. In common practice, the maximum score or logit from a closed-set classifier can serve as a good indicator of OOD Vaze et al. (2022a). However, this is not suitable for the GCD classifier, which contains an important mean entropy regularization term in the loss function to prevent biased predictions towards known classes. Nevertheless, we find that it also results in the classifier's predictions on known categories being less confident, thereby degrading the OOD detection performance. Moreover, self-entropy-based OOD methods need to manually establish a threshold Geng et al. (2020) for rejecting "unknown" samples, which relies on validation or a pre-defined ratio of "unknown" samples, making them impractical for the GCD task where we do not have such validation samples. One-vs-all (OVA) classifier Saito & Saenko (2021), which has consistently shown promise in the literature Saito et al. (2021); Fan et al. (2023); Li et al. (2023), can be a more suitable option. Moreover, in the context of OOD, the objective is not to differentiate between multiple distinct unknown categories, as we do in GCD; rather, we aim to distinguish all unknown samples from the known classes, effectively framing this as a binary classification problem. This need prompted us to introduce a different feature space that is better suited for this task. Therefore, as depicted in Fig. 2, we introduce an additional multi-layer perceptron (MLP) projection network $\rho_s$, to project raw features into another embedding space, followed by $\ell_2$-normalization to attain the embedding space for distribution discrimination. Different from the prior works applying OOD in the magnitude-aware feature space for other tasks Yu et al. (2020); Saito et al. (2021); Li et al. (2023), we empirically found that the $\ell_2$-normalized feature space aligns more seamlessly with the DINO pre-trained weights in GCD. Subsequently, we devise $M$ $\ell_2$-normalized binary classifiers, denoted as $\chi = \{\chi_1, \chi_2, ..., \chi_M\}$, for semantic OOD detection in GCD.

Given the augmented view $\boldsymbol{x}_i$ of an image, its corresponding $\ell_2$-normalized feature in the semantic distribution feature space is denoted as $\boldsymbol{f}_i = \rho_s(\phi(\boldsymbol{x}_i))/||\rho_s(\phi(\boldsymbol{x}_i))||$. Subsequently, the output of the $k$-th binary classifier is $\boldsymbol{o}_{i,k} = \text{softmax}(\chi_k(\boldsymbol{f}_i))$, where $\boldsymbol{o}_{i,k} = (o^+_{i,k}, o^-_{i,k})$ and $o^+_{i,k} + o^-_{i,k} = 1$. For labelled samples, a multi-binary cross-entropy loss with a hard-negative sampling strategy Saito et al. (2021) is employed:

$$\mathcal{L}^s_{sdl} = \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (-\log(o^+_{i,y_i}) - \min_{k \neq y_i} \log(o^-_{i,k})), \tag{4}$$

where $y_i$ represents the ground-truth category label of the sample $\boldsymbol{x}_i$. For unlabelled samples, an entropy minimization technique Saito & Saenko (2021) is applied to improve low-density separation:

$$\mathcal{L}^u_{sdl} = -\frac{1}{\mathcal{B}_u}\sum_{i\in\mathcal{B}_u}\sum_{j=1}^{M}(o^+_{i,j}\log(o^+_{i,j}) + o^-_{i,k}\log(o^-_{i,k})), \tag{5}$$

where $\mathcal{B}_u$ denotes the unlabelled subset in current mini-batch. The loss function for the semantic distribution learning is defined as:

$$\mathcal{L}_{sdl} = \mathcal{L}^s_{sdl} + \mathcal{L}^u_{sdl}. \tag{6}$$

By optimizing $\mathcal{L}_{sdl}$, our detector distinctly segregates the feature distributions between known and unknown categories. Additionally, it generates a predicted score based on the maximum output from all $M$ binary classifiers, denoted as:

$$s_i = o^-_{i,y_p}, y_p = \arg\max_j o^+_{i,j}. \tag{7}$$

This score will serve as a crucial cue for the debiased learning to be introduced next.

### 4.2 AUXILIARY DEBIASED LEARNING

As depicted in Fig. 2, the raw features are normalized to the clustering feature space in the lower branch, wherein novel categories are discovered. In order to minimize the unintended negative impact of biased labels while maintaining the basic probability constraints Assran et al. (2022) and consistency regularization Caron et al. (2021) in the GCD classifier, we propose an auxiliary debiased learning mechanism. Specifically, a parallel debiased classifier $\psi_s$ initialized with $K$ normalized prototypes $\mathcal{C}^a = \{\boldsymbol{c}^a_1, ..., \boldsymbol{c}^a_K\}$, is trained in the same embedding space using debiased labels. Note that in our experiment, we only finetune the last two transformer blocks of the DINO Caron et al. (2021) pre-trained ViT backbone. The $k$-th softmax score of sample $\boldsymbol{x}_i$ is given by:

$$\boldsymbol{p}_i^{a(k)} = \frac{\exp(\boldsymbol{h}_i\cdot\boldsymbol{c}^a_k/\tau_a)}{\sum_{j=1}^K \exp(\boldsymbol{h}_i\cdot\boldsymbol{c}^a_j/\tau_a)}, \tag{8}$$

where $\tau_a$ is the scaling temperature. The maximum classification score has demonstrated promising performance in several semi-supervised learning methods and we find it also a good indicator of sample quality in the context of GCD task. For an augmented view $\boldsymbol{x}_i$ and its GCD classifier prediction $\boldsymbol{p}_i$, a debiasing threshold $\tau$ is set on the $\max(\boldsymbol{p}_i)$, with only samples surpassing $\tau$ being utilized to train the debiased classifier, expressed as $\mathbb{1}(\max(\boldsymbol{p}_i) > \tau)$. Additionally, given that the semantic distribution detector and the GCD classifier are learned in different feature spaces and paradigms, it is essential to ensure the alignment of their predictions. Consequently, we introduce a function to indicate the task consistency of these two tasks, defined as:

$$\mathcal{F}(\hat{y}_i, s_i) = \mathbb{1}(\hat{y}_i \in \mathcal{Y}_u \wedge s_i > 0.5) \vee \mathbb{1}(\hat{y}_i \in \mathcal{Y}_l \wedge s_i < 0.5) \tag{9}$$

where $\hat{y}_i = \arg\max(\boldsymbol{p}_i)$ represents the predicted category index by the GCD classifier, and $\hat{\boldsymbol{y}}_i$ denotes its corresponding one-hot vector. This function aims to selectively filter out samples with identical distribution predictions across the two tasks.

Furthermore, as previously stated, given the inclusion of both known (in-distribution) and unknown (out-of-distribution) samples in the unlabelled data, it is imperative to devise a learning strategy based on semantic distribution information. With the training progresses, the semantic OOD scores gradually approach the two extremes (*i.e.*, 0 and 1). The score of the unknown class sample steadily increases to 1, while the score of the known class gradually decreases to 0. Prior techniques Saito et al. (2021); Li et al. (2023) simply employ a threshold to determine whether the sample belongs to the known or unknown. Such a naïve method is unreliable and may introduce many noises to the model training for GCD. In our approach, we prioritize samples with distinct distributions for self-training, aligning with the principles of curriculum learning. To establish a consistent metric for assessing sample discriminability, we introduce a normalized distribution certainty score:

$$d_i = |2 \times s_i - 1|, \tag{10}$$

which approaches the value 0 for ambiguous samples and the value 1 for certain samples. This score, to a certain extent, indicates the learning status of samples and can serve as a crucial cue for our debiased classifier. Therefore, the auxiliary debiased learning loss for unlabelled samples is written as:

$$\mathcal{L}^u_{adl} = \frac{1}{\mathcal{B}_u}\sum_{i\in\mathcal{B}_u} \mathbb{1}(\max(\boldsymbol{p_i}) > \tau) \times \mathcal{F}(\hat{y}_i, s_i) \times d_i \times \ell_{ce}(\boldsymbol{p}^a_i, \hat{\boldsymbol{y}}_i). \tag{11}$$

**Algorithm 1** End-to-end Training Algorithm for D2G.

```
Input: Set of labelled data $\mathcal{D}_l = \{(\boldsymbol{x}_i^l, y_i^l)\}$, set of unlabelled data $\mathcal{D}_u = \{(\boldsymbol{x}_i^u, y_i^u)\}$. Data augmentation function $\mathcal{A}$. Model parameters $w$, learning rate $\eta$, epoch $E_{max}$, iteration $I_{max}$, trade-off parameters, $\lambda_{sdl}$, $\lambda_{adl}$;
for $Epoch = 1\ \ to\ \ E_{max}$ do
    for $Iteration = 1\ \ to\ \ I_{max}$ do
        Sample labelled data $\mathcal{B}_l$, unlabelled data $\mathcal{B}_u$; $i \in \mathcal{B}_u$
        Compute model prediction $\boldsymbol{p}_i$, $\boldsymbol{p}_i^a$, $s_i$; loss function $\mathcal{L}_{gcd}$, $\mathcal{L}_{sdl}$      // Eq.3,6,8
        Compute debiased label $\hat{y}_i$; task consistency $\mathcal{F}(\hat{y}_i, s_i)$      // Eq.9
        Compute loss function $\mathcal{L}_{adl}^s$, $\mathcal{L}_{adl}^u$, $\mathcal{L}_{adl}$      // Eq.11,12,13
        Compute loss function $\mathcal{L}_{all} = \mathcal{L}_{gcd} + \lambda_{sdl}\mathcal{L}_{sdl} + \lambda_{adl}\mathcal{L}_{adl}$
        Update model parameters $w = w - \eta \bigtriangledown_w \mathcal{L}_{all}$
    end
end
Output: Model parameter $w$.
```

In this manner, the training of the debiased classifier transforms into a curriculum learning process, where easily identifiable samples that are clearly semantic in-distribution or out-of-distribution are given higher priority for learning. Moreover, our debiased classifier also retains the prior knowledge from the labelled data. For the labelled samples, it's is simply trained with the cross-entropy loss:

$$\mathcal{L}_{adl}^s = \frac{1}{\mathcal{B}_l} \sum_{i \in \mathcal{B}_l} \ell_{ce}(\boldsymbol{p}_i^a, \boldsymbol{y}_i). \tag{12}$$

Finally, the overall training loss for the debiased classifier is:

$$\mathcal{L}_{adl} = \mathcal{L}_{adl}^s + \mathcal{L}_{adl}^u. \tag{13}$$

In this manner, all the samples are trained using one-hot hard labels, irrespective of their belongings to known or unknown categories. Operating within the same feature space, our debiased classifier collaborates closely with the GCD classifier, thereby facilitating the joint optimization of the clustering feature space.

### 4.3 LEARNING AND INFERENCE FRAMEWORK

Based on the baseline GCD classifier, our framework is designed to be trained in a multi-task manner. Different from previous approaches in the open-set literature Yu et al. (2020), our D2G framework employs a *one-stage* training process, eliminating the necessity for task-specific warm-up phases. Consequently, the three tasks can be jointly trained end-to-end with the overall loss:

$$\mathcal{L}_{all} = \mathcal{L}_{gcd} + \lambda_{sdl}\mathcal{L}_{sdl} + \lambda_{adl}\mathcal{L}_{adl}, \tag{14}$$

where $\lambda_{sdl}$ and $\lambda_{adl}$ denote the loss weights for the semantic distribution detector and debiased classifier, respectively. The complete training pipeline of the framework is illustrated in Algorithm 1.

Throughout the joint training process, the three branches are collectively optimized in an end-to-end manner. During inference, only the GCD classifier is retained. This indicates that our method does not impose any additional computational overhead compared to the baseline approach during inference, further emphasizing its simplicity and efficiency.

## 5 EXPERIMENTS

In this section, we present a comprehensive evaluation of the proposed D2G framework and further perform meticulous ablation studies to showcase the effectiveness of its individual components. More results and analysis can be found in the Appendix.

### 5.1 EXPERIMENTAL SETUP

**Datasets.** We conduct a comprehensive evaluation of our method across diverse benchmarks, encompassing the generic image recognition benchmark (CIFAR-10/100 Krizhevsky et al. (2009), ImageNet-100 Deng et al. (2009)), the Semantic Shift Benchmark (SSB) Vaze et al. (2022c) comprising fine-grained datasets CUB Wah et al. (2011), Stanford Cars Krause et al. (2013), and FGVC-Aircraft Maji et al. (2013), along with the challenging ImageNet-1K Deng et al. (2009). For each

Table 2: Comparison of state-of-the-art GCD methods on SSB Vaze et al. (2022c) benchmark. Results are reported in ACC across the 'All', 'Old' and 'New' categories.

| | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | All | Old | New | All | Old | New | All | Old | New | All |
| $k$-means MacQueen et al. (1967) | 34.3 | 38.9 | 32.1 | 12.8 | 10.6 | 13.8 | 16.0 | 14.4 | 16.8 | 21.1 |
| RankStats+ Han et al. (2021) | 33.3 | 51.6 | 24.2 | 28.3 | 61.8 | 12.1 | 26.9 | 36.4 | 22.2 | 29.5 |
| UNO+ Fini et al. (2021) | 35.1 | 49.0 | 28.1 | 35.5 | 70.5 | 18.6 | 40.3 | 56.4 | 32.2 | 37.0 |
| ORCA Cao et al. (2022) | 35.3 | 45.6 | 30.2 | 23.5 | 50.1 | 10.7 | 22.0 | 31.8 | 17.1 | 26.9 |
| GCD Vaze et al. (2022b) | 51.3 | 56.6 | 48.7 | 39.0 | 57.6 | 29.9 | 45.0 | 41.1 | 46.9 | 45.1 |
| XCon Fei et al. (2022) | 52.1 | 54.3 | 51.0 | 40.5 | 58.8 | 31.7 | 47.7 | 44.4 | 49.4 | 46.8 |
| OpenCon Sun & Li (2022) | 54.7 | 63.8 | 54.7 | 49.1 | 78.6 | 32.7 | - | - | - | - |
| PromptCAL Zhang et al. (2023) | 62.9 | 64.4 | 62.1 | 50.2 | 70.1 | 40.6 | 52.2 | 52.2 | 52.3 | 55.1 |
| DCCL Pu et al. (2023) | 63.5 | 60.8 | 64.9 | 43.1 | 55.7 | 36.2 | - | - | - | - |
| GPC Zhao et al. (2023) | 52.0 | 55.5 | 47.5 | 38.2 | 58.9 | 27.4 | 43.3 | 40.7 | 44.8 | 44.5 |
| SimGCD Wen et al. (2023) | 60.3 | 65.6 | 57.7 | 53.8 | 71.9 | 45.0 | 54.2 | 59.1 | 51.8 | 56.1 |
| $\mu$GCD Vaze et al. (2023) | 65.7 | 68.0 | 64.6 | 56.5 | 68.1 | <u>50.9</u> | 53.8 | 55.4 | 53.0 | 58.7 |
| InfoSieve Rastegar et al. (2023) | **69.4** | **77.9** | **65.2** | 55.7 | 74.8 | 46.4 | 56.3 | <u>63.7</u> | 52.5 | 60.5 |
| CiPR Hao et al. (2024) | 57.1 | 58.7 | 55.6 | 47.0 | 61.5 | 40.1 | - | - | - | - |
| SPTNet Wang et al. (2024) | 65.8 | 68.8 | <u>65.1</u> | <u>59.0</u> | <u>79.2</u> | 49.3 | <u>59.3</u> | 61.8 | <u>58.1</u> | <u>61.4</u> |
| **D2G(ours)** | <u>66.3</u> | <u>71.8</u> | 63.5 | **65.3** | **81.6** | **57.4** | **61.7** | **63.9** | **60.6** | **64.4** |

dataset, we adhere to the data split scheme detailed in Vaze et al. (2022b). The method involves sampling a subset of all classes as the known ('Old') classes $\mathcal{Y}_l$. Subsequently, 50% of the images from these known classes are utilized to construct $\mathcal{D}_l$, while the remaining images are designated as the unlabelled data $\mathcal{D}_u$. The statistics can be seen in Tab. 1.

**Evaluation metrics.** We assess the GCD performance using the clustering accuracy (ACC) in accordance with established conventions Vaze et al. (2022b). For evaluation, the ACC on $\mathcal{D}_l$ is computed as follows, given the ground truth $y_i$ and the predicted labels $\hat{y}_i$:

$$\text{ACC} = \frac{1}{|\mathcal{D}_u|} \sum_{i=1}^{|\mathcal{D}_u|} \mathbb{1}(y_i = h(\hat{y}_i)), \quad (15)$$

where $h$ represents the optimal permutation that aligns the predicted cluster assignments with the ground-truth class labels. ACC for 'All' classes, 'Old' classes and 'New' classes are reported for comprehensive assessment.

Table 1: Overview of dataset, including the classes in the labelled and unlabelled sets ($|\mathcal{Y}_l|$, $|\mathcal{Y}_u|$) and counts of images ($|\mathcal{D}_l|$, $|\mathcal{D}_u|$). 'FG' denotes fine-grained.

| Dataset | FG | $\lvert\mathcal{D}_l\rvert$ | $\lvert\mathcal{Y}_l\rvert$ | $\lvert\mathcal{D}_u\rvert$ | $\lvert\mathcal{Y}_u\rvert$ |
|---|---|---|---|---|---|
| CIFAR-10 Krizhevsky et al. (2009) | ✗ | 12.5K | 5 | 37.5K | 10 |
| CIFAR-100 Krizhevsky et al. (2009) | ✗ | 20.0K | 80 | 30.0K | 100 |
| ImageNet-100 Deng et al. (2009) | ✗ | 31.9K | 50 | 95.3K | 100 |
| CUB Wah et al. (2011) | ✓ | 1.5K | 100 | 4.5K | 200 |
| Stanford Cars Krause et al. (2013) | ✓ | 2.0K | 98 | 6.1K | 196 |
| FGVC-Aircraft Maji et al. (2013) | ✓ | 1.7K | 50 | 5.0K | 100 |
| ImageNet-1K Deng et al. (2009) | ✗ | 321K | 500 | 960K | 1000 |

**Implementation details.** Following previous attempts in GCD Vaze et al. (2022b); Wen et al. (2023), our model is structured with a ViT-B/16 Dosovitskiy et al. (2021) backbone pre-trained using DINO Caron et al. (2021), and the feature space centers around the 768-dimensional classification token. The projection networks for representation learning and semantic distribution detection comprise three-layer and five-layer MLPs, respectively. The model is trained with a batch size of 128, initiating with an initial learning rate of $10^{-1}$ which decays to $10^{-4}$ using a cosine schedule over 200 epochs. Notably, the loss weights $\lambda_{sdl}$ and $\lambda_{adl}$ are set to 0.01 and 1.0, while the loss balancing weight $\lambda_b^{gcd}$ is assigned to 0.35 following Wen et al. (2023). Regarding the temperature parameters, the initial temperature $\tau_t$ is established at 0.07, subsequently warmed up to 0.04 employing a cosine schedule during the first 30 epochs, whereas the other temperatures are set to 0.1.

### 5.2 BENCHMARK RESULTS

We present benchmark results of our method and compare it with state-of-the-art techniques in generalized category discovery (including ORCA Cao et al. (2022), GCD Vaze et al. (2022b), XCon Fei et al. (2022), OpenCon Sun & Li (2022), PromptCAL Zhang et al. (2023), DCCL Pu et al. (2023), GPC Zhao et al. (2023), CiPR Hao et al. (2024), SimGCD Wen et al. (2023), $\mu$GCD Vaze et al. (2023), InfoSieve Rastegar et al. (2023), and SPTNet Wang et al. (2024)), as well as robust baselines derived from novel category discovery (RankStats+Han et al. (2021), UNO+Fini et al. (2021), and $k$-means MacQueen et al. (1967)). All methods are based on the DINO Caron et al. (2021) pre-trained backbone. This comparative evaluation encompasses performance on the fine-grained SSB

Table 3: Comparison of state-of-the-art GCD methods on generic datasets. It includes CIFAR-10 Krizhevsky et al. (2009), CIFAR-100 Krizhevsky et al. (2009), ImageNet-100 Deng et al. (2009), and ImageNet-1K Deng et al. (2009) dataset.

| | CIFAR-10 | | | CIFAR-100 | | | ImageNet-100 | | | ImageNet-1K | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| $k$-means MacQueen et al. (1967) | 83.6 | 85.7 | 82.5 | 52.0 | 52.2 | 50.8 | 72.7 | 75.5 | 71.3 | - | - | - |
| RankStats+ Han et al. (2021) | 46.8 | 19.2 | 60.5 | 58.2 | 77.6 | 19.3 | 37.1 | 61.6 | 24.8 | - | - | - |
| UNO+ Fini et al. (2021) | 68.6 | **98.3** | 53.8 | 69.5 | 80.6 | 47.2 | 70.3 | **95.0** | 57.9 | - | - | - |
| ORCA Cao et al. (2022) | 69.0 | 77.4 | 52.0 | 73.5 | **92.6** | 63.9 | 81.8 | 86.2 | 79.6 | - | - | - |
| GCD Vaze et al. (2022b) | 91.5 | 97.9 | 88.2 | 73.0 | 76.2 | 66.5 | 74.1 | 89.8 | 66.3 | 52.5 | 72.5 | 42.2 |
| XCon Fei et al. (2022) | 96.0 | 97.3 | 95.4 | 74.2 | 81.2 | 60.3 | 77.6 | 93.5 | 69.7 | - | - | - |
| OpenCon Sun & Li (2022) | - | - | - | - | - | - | 84.0 | 93.8 | 81.2 | - | - | - |
| PromptCAL Zhang et al. (2023) | **97.9** | 96.6 | 98.5 | 81.2 | 84.2 | 75.3 | 83.1 | 92.7 | 78.3 | - | - | - |
| DCCL Pu et al. (2023) | 96.3 | 96.5 | 96.9 | 75.3 | 76.8 | 70.2 | 80.5 | 90.5 | 76.2 | - | - | - |
| GPC Zhao et al. (2023) | 90.6 | 97.6 | 87.0 | 75.4 | 84.6 | 60.1 | 75.3 | 93.4 | 66.7 | - | - | - |
| SimGCD Wen et al. (2023) | 97.1 | 95.1 | 98.1 | 80.1 | 81.2 | 77.8 | 83.0 | 93.1 | 77.9 | 57.1 | 77.3 | 46.9 |
| InfoSieve Rastegar et al. (2023) | 94.8 | 97.7 | 93.4 | 78.3 | 82.2 | 70.5 | 80.5 | 93.8 | 73.8 | - | - | - |
| CiPR Hao et al. (2024) | 97.7 | 97.5 | 97.7 | 81.5 | 82.4 | 79.7 | 80.5 | 84.9 | 78.3 | - | - | - |
| SPTNet Wang et al. (2024) | 97.3 | 95.0 | **98.6** | 81.3 | 84.3 | 75.6 | 85.4 | 93.2 | 81.4 | - | - | - |
| **D2G(ours)** | 97.2 | 94.8 | 98.4 | **83.0** | 84.6 | **79.9** | **85.9** | 94.3 | **81.6** | **65.0** | **82.0** | **56.5** |

Table 4: Ablations. The results regarding the different components in our framework on Stanford Cars Krause et al. (2013). ACC of ‘All’, ‘Old’ and ‘New’ categories are listed.

| | Debiased Learning | Auxiliary Classifier | Semantic Dist. Learning | Dist. Guidance | Stanford Cars | | |
|---|---|---|---|---|---|---|---|
| | | | | | All | Old | New |
| (1) | ✗ | ✗ | ✗ | ✗ | 53.8 | 71.9 | 45.0 |
| (2) | ✓ | ✗ | ✗ | ✗ | 51.3 | 72.8 | 40.9 |
| (3) | ✓ | ✓ | ✗ | ✗ | 58.5 | 78.7 | 48.8 |
| (4) | ✗ | ✗ | ✓ | ✗ | 56.5 | 73.3 | 48.3 |
| (5) | ✓ | ✓ | ✓ | ✗ | 60.7 | 78.1 | 52.3 |
| (6) | ✓ | ✓ | ✓ | ✓ | **65.3** | **81.6** | **57.4** |

benchmark Vaze et al. (2022c) and generic image recognition datasets Krizhevsky et al. (2009); Deng et al. (2009), as shown in Tab. 2 and Tab. 9.

**Results on SSB.** As shown in Tab. 2, D2G demonstrates superior performance across all datasets, achieving an average ACC of 64.4 on ‘All’ categories, surpassing the second-best by 3%. It maintains the best on both Stanford Cars and FGVC-Aircraft, while ranking second on CUB, where it is outperformed only by InfoSieve Rastegar et al. (2023), a hierarchical encoding method specifically designed for fine-grained GCD. In contrast, D2G aims for broader improvements across both generic and fine-grained datasets. These results reveal D2G’s exceptional ability to uncover new categories, while also showcasing remarkable performance in recognizing known categories.

**Results on generic datasets.** In Tab. 9, we report results on three widely used generic datasets (CIFAR-10, CIFAR-100 and ImageNet-100) in GCD, as well as the challenging ImageNet-1K. Our method attains superior performance in terms of ACC across ‘All’ categories, establishing the new state-of-the-art, except CIFAR-10, on which the performance is nearly saturated (over 97% ACC) for our method and other most competitive methods. On the challenging ImageNet-1K, containing $1,000$ classes with diverse images, D2G also establishes the new state-of-the-art, surpassing the previous best-performing method by $7.9\%$. These results validate the effectiveness and robustness of our method for generalized category discovery on generic datasets.

### 5.3 ANALYSIS

In this section, we provide ablations regarding the key components within our framework. Besides, we study the impact of the debiasing threshold $\tau$ and labelled data.

**Framework components.** Starting with the baseline method trained using $\mathcal{L}_{gcd}$ (Row (1)), we gradually incorporate our proposed techniques on the Stanford Cars dataset, as depicted in Tab. 4. An intuitive approach is to apply debiased learning to the original classifier as in Row (2). However, this still produces a biased supervision signal because it relies on the original GCD loss for that classifier. It turns out that such a naïve approach may even hurt the performance. Rows (1) and (2) indicate that directly applying debiased learning to the GCD classifier can lead to a decrease in performance, particularly affecting novel categories. The introduction of an auxiliary classifier in

Row (3) demonstrates significant performance enhancements. Similarly, our semantic distribution learning alone results in a 2.7% improvement across all categories in Row (4). Row (5) highlights that co-training the debiased classifier and semantic distribution detector further boosts performance. Notably, guiding the debiased learning with semantic distribution certainty and task consistency function yields a notable 4.6% performance increase in Row (6).

**Loss function.** In addition, we explore the impact of the data and the respective loss functions employed during the training of debiased classifier, denoted as $\mathcal{L}^s_{adl}$ and $\mathcal{L}^u_{adl}$, targeting the labelled and unlabelled datasets, respectively. These experiments are undertaken on the FGVC-Aircraft Maji et al. (2013) using various subset combinations. Solely training with $\mathcal{L}^s_{adl}$ introduces bias towards known categories, leading to a notable performance decline. Conversely, exclusive training with $\mathcal{L}^u_{adl}$ fails to reach optimal performance levels, underscoring the essential role of knowledge derived from labelled data. These outcomes demonstrate the vital significance of both $\mathcal{L}^s_{adl}$ and $\mathcal{L}^u_{adl}$ in optimizing the debiased classifier.

Table 5: Experimental results on distillation data by using different loss functions.

| $\mathcal{L}^s_{adl}$ | $\mathcal{L}^u_{adl}$ | FGVC-Aircraft | | |
|---|---|---|---|---|
| | | All | Old | New |
| | | 54.2 | 59.1 | 51.8 |
| ✓ | | 53.1 | 60.5 | 49.4 |
| | ✓ | 57.9 | 60.1 | 56.9 |
| ✓ | ✓ | **61.7** | **63.9** | **60.6** |

**Debiasing threshold $\tau$.** Similar to self-training approaches Sohn et al. (2020); Zhang et al. (2021), the selection of the threshold for generating pseudo-labels also plays a crucial role in our approach. Consistent with the methods outlined in Wen et al. (2023) and Vaze et al. (2022b), we calibrate the threshold based on its performance on a separate validation set of the labelled data. Detailed results regarding different thresholds on the FGVC-Aircraft Wah et al. (2011) dataset, covering performance on both the unlabelled training dataset and the validation set, are presented. As shown in Tab. 6, the threshold is incrementally adjusted in intervals of 0.05. Notably, the performance trends for both datasets align, with optimal performance achieved when the threshold is set to 0.85.

Table 6: Experimental results regarding threshold $\tau$ on the unlabelled set and validation set of FGVC-Aircraft Maji et al. (2013) dataset.

| | Unlabelled Set | | | Validation Set | | |
|---|---|---|---|---|---|---|
| $\tau$ | All | Old | New | All | Old | New |
| 0.9 | 59.4 | **64.7** | 56.7 | 58.9 | 61.1 | 56.8 |
| 0.85 | **61.7** | 63.9 | **60.6** | **61.1** | **62.0** | **60.3** |
| 0.8 | 60.7 | 61.5 | 60.3 | 60.6 | 61.6 | 59.6 |

### 5.4 VISUALIZATION RESULTS

Additionally, we explore the visual representation of the baseline and our method using t-SNE Van der Maaten & Hinton (2008). Specifically, we randomly select a set of 20 classes, including 10 from the 'Old' categories and 10 from the 'New' categories. The clearly distinguishable clusters depicted in Fig. 3 indicate that the features obtained within our framework form notably cohesive groupings compared to those of the baseline. This effectively demonstrates the optimization impacts induced by our method on the clustering feature space.

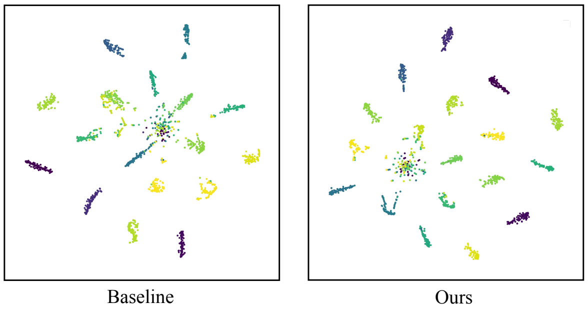


Figure 3: T-SNE Van der Maaten & Hinton (2008) visualization on the category discovery features of 20 classes randomly sampled from the CIFAR-100 Krizhevsky et al. (2009) dataset.

## 6 CONCLUSION

This paper presents D2G, a distribution-guided debiased learning framework for GCD, comprising three primary components. Firstly, we introduce an auxiliary debiased learning mechanism by concurrently training a parallel classifier with the GCD classifier, thereby facilitating optimization in the GCD feature space. Secondly, a semantic distribution detector is introduced to explicitly identify semantic shifts and implicitly enhance performance. Lastly, we propose a semantic distribution certainty score that enables a curriculum-based learning approach, promoting effective learning for both seen and unseen classes. Despite its simplicity, D2G showcases superior performance, as evidenced by comprehensive evaluation on seven public benchmarks.

# REFERENCES

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 2009.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *ICLR*, 2021.

Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Ssb: Simple but strong baseline for boosting performance of open-set semi-supervised learning. In *ICCV*, 2023.

Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *BMVC*, 2022.

Enrico Fini, Enver Sangineto, Stéphane Lathuiliere, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2020.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE TPAMI*, 2020.

Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE TPAMI*, 2021.

Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *TMLR*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE TPAMI*, 2020.

KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *ECCV*, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *ICLR*, 2021.

Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *ICCV*, 2023.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

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

Safiya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

Jean Ponce. *Toward category-level object recognition*, volume 4170. Springer Science & Business Media, 2006.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, 2023.

Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *NeurIPS*, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, 2020.

Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *ICCV*, 2021.

Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. In *NeurIPS*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

Yiyou Sun and Yixuan Li. Opencon: Open-world contrastive learning. *TMLR*, 2022.

Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *ICLR*, 2022a.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022b.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. The semantic shift benchmark. In *ICML workshop*, 2022c.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *NeurIPS*, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, 2024.

Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *IJCV*, 2022.

Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, 2020.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 2021.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, 2023.

Xinwei Zhang, Jianwen Jiang, Yutong Feng, Zhi-fan Wu, Xibin Zhao, Hai Wan, Mingqian Tang, Rong Jin, and Yue Gao. Grow and merge: a unified framework for continuous categories discovery. In *NeurIPS*, 2022.

Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021.

Bingchen Zhao and Oisin Mac Aodha. Incremental generalized category discovery. In *ICCV*, 2023.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *ICCV*, 2023.

# APPENDIX

**A Implementation Details** **16**

**B Results on Additional Datasets** **17**

**C Experiments with the Stronger DINOv2 Representations** **18**

**D Category Discovery with Estimated Category Numbers** **19**

**E Extension to Incremental Generalized Category Discovery** **20**

**F Utilization Ratio of Unlabelled Data** **21**

**G GCD Classifier *vs.* Debiased Classifer** **22**

**H Performance of the Semantic Distribution Detector** **23**

**I Analysis of Attention Maps** **24**

**J Ablation Studies on More Datasets** **26**

**K Impact of Hyperparameters** **27**

**L Stability Analysis** **29**

**M Prediction Error Analysis** **30**

# A IMPLEMENTATION DETAILS

We adopt the class splits of labelled ('Old') and unlabelled ('New') categories in Vaze et al. (2022b) for generic object recognition datasets (including CIFAR-10 Krizhevsky et al. (2009) and CIFAR-100 Krizhevsky et al. (2009)) and the fine-grained Semantic Shift Benchmark Vaze et al. (2022c) (comprising CUB Wah et al. (2011), Stanford Cars Krause et al. (2013), and FGVC-Aircraft Maji et al. (2013)). Specifically, for all these datasets except CIFAR-100, $50\%$ of all classes are selected as 'Old' classes ($\mathcal{Y}_l$), while the remaining classes are treated as 'New' classes ($\mathcal{Y}_u \backslash \mathcal{Y}_l$). For CIFAR-100, $80\%$ of the classes are designated as 'Old' classes, while the remaining $20\%$ as 'New' classes. Furthermore, for ImageNet-1K Deng et al. (2009), which is not covered in Vaze et al. (2022b), we follow Wen et al. (2023) to select the first $500$ classes sorted by class ID as the labelled classes. For all the datasets, $50\%$ of the images from the labelled classes are randomly sampled to form the labelled dataset $\mathcal{D}_l$, and all remaining images are regarded as the unlabelled dataset $\mathcal{D}_u$. Moreover, following Vaze et al. (2022b) and Wen et al. (2023), the model's hyperparameters are chosen based on its performance on a hold-out validation set, formed by the original test splits of labelled classes in each dataset. All experiments utilize the PyTorch framework on a workstation with an Intel i7 CPU and eight Nvidia Tesla V100 GPUs. The models are trained with a batch size of 128 on a single GPU, except for the the model on ImageNet-1K dataset, for which the training is performed with eight GPUs.

## B RESULTS ON ADDITIONAL DATASETS

To assess the performance of the proposed method comprehensively, we conducted evaluations on two more fine-grained datasets: Oxford-Pet Parkhi et al. (2012) and Herbarium 19 Tan et al. (2019). Oxford-Pet is a challenging dataset featuring various species of cats and dogs with limited data. Herbarium19, on the other hand, is a botanical research dataset encompassing diverse plant types, known for its long-tailed distribution and fine-grained categorization. The outcomes of our experiments on these datasets are detailed in Tab. 7. The results of SimGCD Wen et al. (2023) on Oxford-Pet are obtained through the execution of the officially released code. Our D2G model consistently demonstrates superior performance on both datasets.

Table 7: Comparison with state-of-the-art GCD methods on Herbarium19 Tan et al. (2019) and Oxford-Pet Parkhi et al. (2012).

| | Oxford-Pet | | | Herbarium19 | | |
|---|---|---|---|---|---|---|
| Method | All | Old | New | All | Old | New |
| $k$-means MacQueen et al. (1967) | 77.1 | 70.1 | 80.7 | 13.0 | 12.2 | 13.4 |
| RankStats+ Han et al. (2021) | - | - | - | 27.9 | 55.8 | 12.8 |
| UNO+ Fini et al. (2021) | - | - | - | 28.3 | 53.7 | 14.7 |
| ORCA Cao et al. (2022) | - | - | - | 24.6 | 26.5 | 23.7 |
| GCD Vaze et al. (2022b) | 80.2 | 85.1 | 77.6 | 35.4 | 51.0 | 27.0 |
| XCon Fei et al. (2022) | 86.7 | 91.5 | 84.1 | - | - | - |
| OpenCon Sun & Li (2022) | - | - | - | 39.3 | 58.9 | 28.6 |
| DCCL Pu et al. (2023) | 88.1 | 88.2 | 88.0 | - | - | - |
| SimGCD Wen et al. (2023) | 91.7 | 83.6 | 96.0 | 44.0 | 58.0 | 36.4 |
| $\mu$GCD Vaze et al. (2023) | - | - | - | **45.8** | **61.9** | **37.2** |
| InfoSieve Rastegar et al. (2023) | 90.7 | **95.2** | 88.4 | 40.3 | 59.0 | 30.2 |
| **D2G(ours)** | **93.0** | 86.4 | **96.5** | 44.7 | 59.4 | 36.8 |

## C EXPERIMENTS WITH THE STRONGER DINOV2 REPRESENTATIONS

To further evaluate the robustness of the proposed method, we also evaluate the performance of D2G utilizing the stronger DINOv2 Oquab et al. (2023) pre-trained weights. Like in Vaze et al. (2023), in Tab. 8, we also compare our method with the k-means MacQueen et al. (1967) baseline, and SimGCD Wen et al. (2023), $\mu$GCD Vaze et al. (2023). Our method outperforms other methods on CUB Wah et al. (2011) and FGVC-Aircraft Maji et al. (2013) on 'All', 'Old' and 'New' classes consistently. On Stanford Cars Krause et al. (2013), our method outperforms other methods on 'New' classes, while performing the second-best on 'All' and 'Old' classes. Moreover, for the average performance of 'All' classes across the three datasets, D2G outperforms the SimGCD baseline by about 6% and $\mu$GCD by about 3%. Additionally, we also evaluate our model on generic datasets and compare it with the SimGCD baseline in Tab. 9, demonstrating consistent improvement. The results on both fine-grained and generic datasets validate the robustness of our proposed method on the stronger DINOv2 representations, further showcasing its effectiveness.

Table 8: Comparison with state-of-the-art GCD methods on SSB leveraging DINOv2 Oquab et al. (2023) pre-trained weights.

| | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | All | Old | New | All | Old | New | All | Old | New | All |
| $k$-means MacQueen et al. (1967) | 67.6 | 60.6 | 71.1 | 29.4 | 24.5 | 31.8 | 18.9 | 16.9 | 19.9 | 38.6 |
| GCD Vaze et al. (2022b) | 71.9 | 71.2 | 72.3 | 65.7 | 67.8 | 64.7 | 55.4 | 47.9 | 59.2 | 64.3 |
| CiPR Hao et al. (2024) | **78.3** | 73.4 | **80.8** | 66.7 | 77.0 | 61.8 | 59.2 | 65.0 | 56.3 | 68.1 |
| SimGCD Wen et al. (2023) | 71.5 | 78.1 | 68.3 | 71.5 | 81.9 | 66.6 | 63.9 | 69.9 | 60.9 | 69.0 |
| $\mu$GCD Vaze et al. (2023) | 74.0 | 75.9 | 73.1 | **76.1** | **91.0** | 68.9 | 66.3 | 68.7 | 65.1 | 72.1 |
| SPTNet Wang et al. (2024) | 76.3 | 79.5 | 74.6 | - | - | - | - | - | - | - |
| **D2G(ours)** | 77.5 | **80.8** | 75.8 | 75.4 | 87.7 | **69.5** | **71.9** | **76.0** | **69.8** | **74.9** |

Table 9: Comparison with state-of-the-art GCD methods on generic datasets leveraging DINOv2 Oquab et al. (2023) pre-trained weights.

| | CIFAR-10 | | | CIFAR-100 | | | ImageNet-100 | | | ImageNet-1K | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| GCD Vaze et al. (2022b) | 97.8 | **99.0** | 97.1 | 79.6 | 84.5 | 69.9 | 78.5 | 89.5 | 73.0 | - | - | - |
| CiPR Hao et al. (2024) | **99.0** | 98.7 | 99.2 | **90.3** | 89.0 | **93.1** | 88.2 | 87.6 | 88.5 | - | - | - |
| SimGCD Wen et al. (2023) | 98.7 | 96.7 | **99.7** | 88.5 | 89.2 | 87.2 | 89.9 | 95.5 | 87.1 | 58.0 | 66.9 | 53.2 |
| SPTNet Wang et al. (2024) | - | - | - | - | - | - | 90.1 | 96.1 | 87.1 | - | - | - |
| **D2G(ours)** | 98.9 | 97.5 | 99.6 | 90.1 | **90.9** | 88.6 | **93.2** | **97.0** | **91.2** | **71.7** | **86.2** | **64.5** |

## D CATEGORY DISCOVERY WITH ESTIMATED CATEGORY NUMBERS

Following the majority of the literature, we experiment mainly using the ground-truth category numbers. In this section, we report the results of D2G using the number of categories estimated utilizing an off-the-shelf method Vaze et al. (2022b), to showcase the performance with the ground-truth category numbers are not available. Tab. 10 reports the estimated numbers. We compare D2G with SimGCD Wen et al. (2023), $\mu$GCD Vaze et al. (2023), and GCD Vaze et al. (2022b) in Tab. 11. For both CUB Wah et al. (2011) and Stanford Cars Krause et al. (2013), despite a discrepancy of approximately $15\%$ between the ground-truth and estimated category numbers, our method exhibits a smaller decline in performance compared to GCD and SimGCD. The same trend is also observed on Imagenet-100 Deng et al. (2009). D2G remains the most competitive method on 'All' classes using the same estimated category numbers on all four datasets, which clearly demonstrates the robustness and effectiveness of our proposed method.

Table 10: Estimated class numbers in the unlabelled data using method proposed in Vaze et al. (2022b).

| | CUB | Stanford Cars | CIFAR-100 | ImageNet-100 |
|---|---|---|---|---|
| Ground-truth $K$ | 200 | 196 | 100 | 100 |
| Estimated $K$ | 231 | 230 | 100 | 109 |

Table 11: Results with the estimated number of categories. The estimated class numbers in Tab. 10 are adopted for all methods.

| | CUB | | | Stanford Cars | | | CIFAR-100 | | | ImageNet-100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| GCD Vaze et al. (2022b) | 47.1 | 55.1 | 44.8 | 35.0 | 56.0 | 24.8 | 73.0 | 76.2 | 66.5 | 72.7 | 91.8 | 63.8 |
| SimGCD Wen et al. (2023) | 61.5 | 66.4 | 59.1 | 49.1 | 65.1 | 41.3 | 80.1 | 81.2 | 77.8 | 81.7 | 91.2 | 76.8 |
| $\mu$GCD Vaze et al. (2023) | 62.0 | 60.3 | **62.8** | 56.3 | 66.8 | 51.1 | - | - | - | - | - | - |
| **D2G (ours)** | **64.5** | **68.5** | 62.5 | **63.3** | **78.6** | **55.8** | **83.0** | **84.6** | **79.9** | **84.9** | **93.3** | **80.7** |

# E EXTENSION TO INCREMENTAL GENERALIZED CATEGORY DISCOVERY

To further assess the effectiveness of D2G, we extend it to the more challenging task of Incremental Generalized Category Discovery (IGCD) Zhao & Mac Aodha (2023). This presents a challenging category-incremental learning scenario, wherein the objective is to construct models capable of accurately classifying images from previously encountered categories while also identifying new ones. Learning takes place over a sequence of time steps during which the model acquires new labelled and unlabelled data, and discards old data at each iteration. Both D2G and the baseline SimGCD Wen et al. (2023) can be expanded to this incremental learning setup by integrating them with iCaRL Rebuffi et al. (2017). We compare these two extended methods with approaches specifically created for IGCD, including GM Zhang et al. (2022) and the method proposed in Zhao & Mac Aodha (2023), on the fine-grained dataset CUBWah et al. (2011) and the generic dataset CIFAR-100 Krizhevsky et al. (2009). It can be observed from Tab. 12 that our method yields the best performance on $M_d$ while maintaining comparable $M_f$ with the state-of-the-art methods. Across both generic and fine-grained datasets, D2G achieves an improvement of $2.2\%$ to $4.8\%$ in terms of $M_d$ and $M_f$. The results demonstrate the adaptability of D2G in more challenging settings such as IGCD, thereby further underscoring its advantages.

Table 12: Results on mixed incremental setting of IGCD Zhao & Mac Aodha (2023).

| | CUB | | CIFAR-100 | |
|---|---|---|---|---|
| Method | $M_f \downarrow$ | $M_d \uparrow$ | $M_f \downarrow$ | $M_d \uparrow$ |
| GM Zhang et al. (2022) | **3.6** | 30.6 | 6.8 | 26.7 |
| IGCD Zhao & Mac Aodha (2023) | 4.0 | 31.2 | **6.7** | 29.4 |
| SimGCD Wen et al. (2023)+iCaRL | 9.4 | 29.4 | 10.7 | 28.3 |
| D2G+iCaRL | 6.0$^{-3.4}$ | **34.2**$^{+4.8}$ | 8.1$^{-2.6}$ | **30.5**$^{+2.2}$ |

# F UTILIZATION RATIO OF UNLABELLED DATA

The data utilization ratio is a notable index for pseudo-labeling methods, offering clear insights into the data efficiency. Our examination encompasses the utilization ratio of unlabelled data from both the 'Old' and 'New' classes during the training of the debiased classifier on FGVC-Aircraft Maji et al. (2013) and Stanford Cars Krause et al. (2013), as depicted in Fig. 4. Initially, the majority of data from the unknown categories remains untapped. Subsequently, after approximately 20 epochs, samples from unknown categories start to be incorporated. The utilization ratio keeps growing, reaching a ratio of around $40\%$ at the 100th epoch. Ultimately, more than $60\%$ of the known categories' samples and nearly half of the unknown categories' samples are utilized.

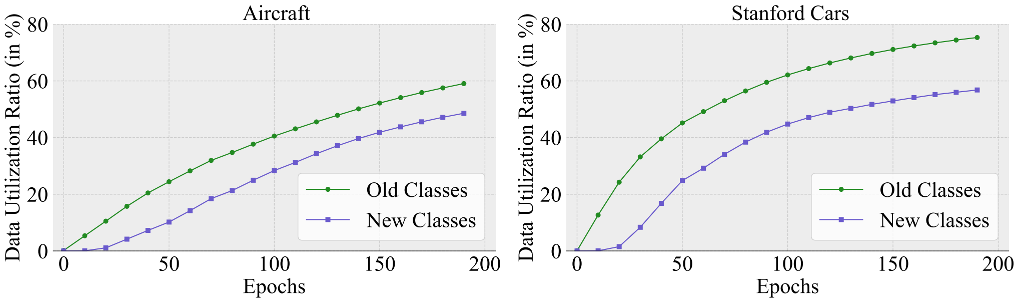


Figure 4: Unlabelled data utilization ratios for 'Old' and 'New' classes during training on FGVC-Aircraft Maji et al. (2013) (left) and Stanford Cars Krause et al. (2013) (right) datasets.

## G GCD CLASSIFIER *vs.* DEBIASED CLASSIFER

We compare the performance between the two classifiers, the GCD Classifier and the debiased classifier, in our framework. We report the ACC results across different epochs in Fig. 5 when training on Stanford Cars Krause et al. (2013), including unlabelled data from both training and the validation splits of the original dataset. Initially, the debiased classifier exhibits bias towards the 'Old' classes, given that the training data primarily comprises labelled data from known categories. However, as predicted scores of the unlabelled samples, particularly those from the unknown categories, progressively surpass the debiasing threshold, the performance on the unknown categories gradually improves and eventually matches with the labelled categories. Ultimately, upon convergence of the model, the performance on both known and unknown categories converges to that of the GCD classifier.

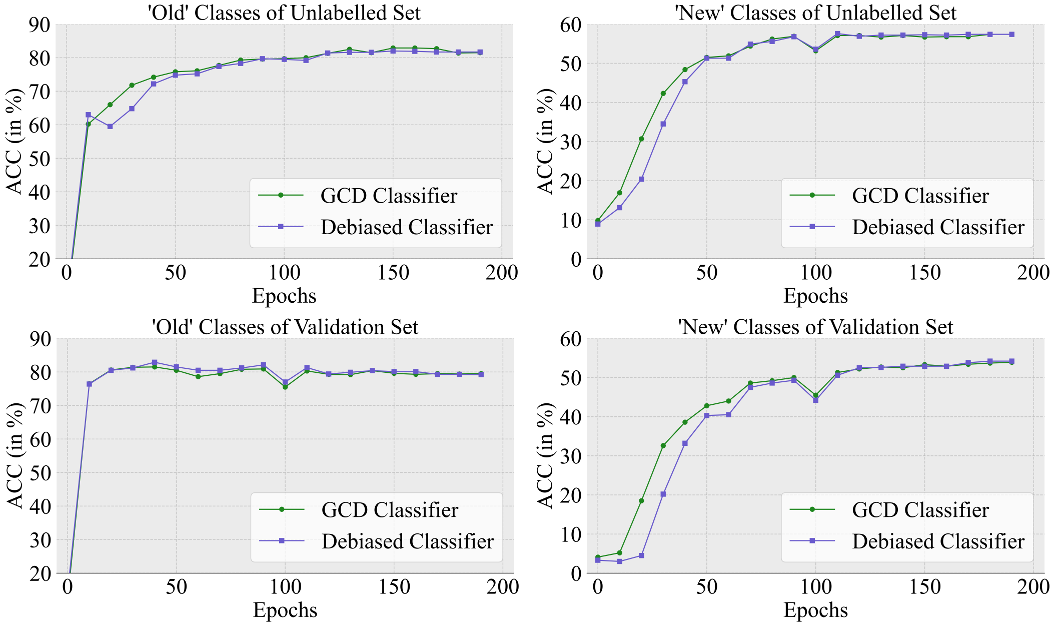


Figure 5: ACC evolution on both the 'Old' and 'New' classes of GCD Classifier and debiased classifier during training on Stanford Cars dataset Krause et al. (2013). The top two figures depict ACC on the unlabelled training set, while the bottom two illustrate ACC on the validation set.

## H PERFORMANCE OF THE SEMANTIC DISTRIBUTION DETECTOR

We evaluate the OOD detection performance of our semantic distribution detector in D2G, using the threshold-free Area Under the Receiver-Operator curve (AUROC) as the evaluation metric, which is widely used in the OOD detection literature. A comparison of the OOD performance between training the entire framework and training solely the distribution detector is presented in Tab. 13. A significant improvement in OOD performance is obtained by training jointly the GCD classifier and debiased classifier. This aligns with the results presented in Tab. 4 of the main paper, which demonstrate the mutual benefits among the three branches (tasks) in our framework. Additionally, we visualize the distribution of the score $s_i$ on the challenging SSB datasets in Fig. 6 which shows that our method can successfully distinguish samples from ‘Old’ and ‘New’ classes in the unlabelled data of both the training and validation splits of the original dataset.

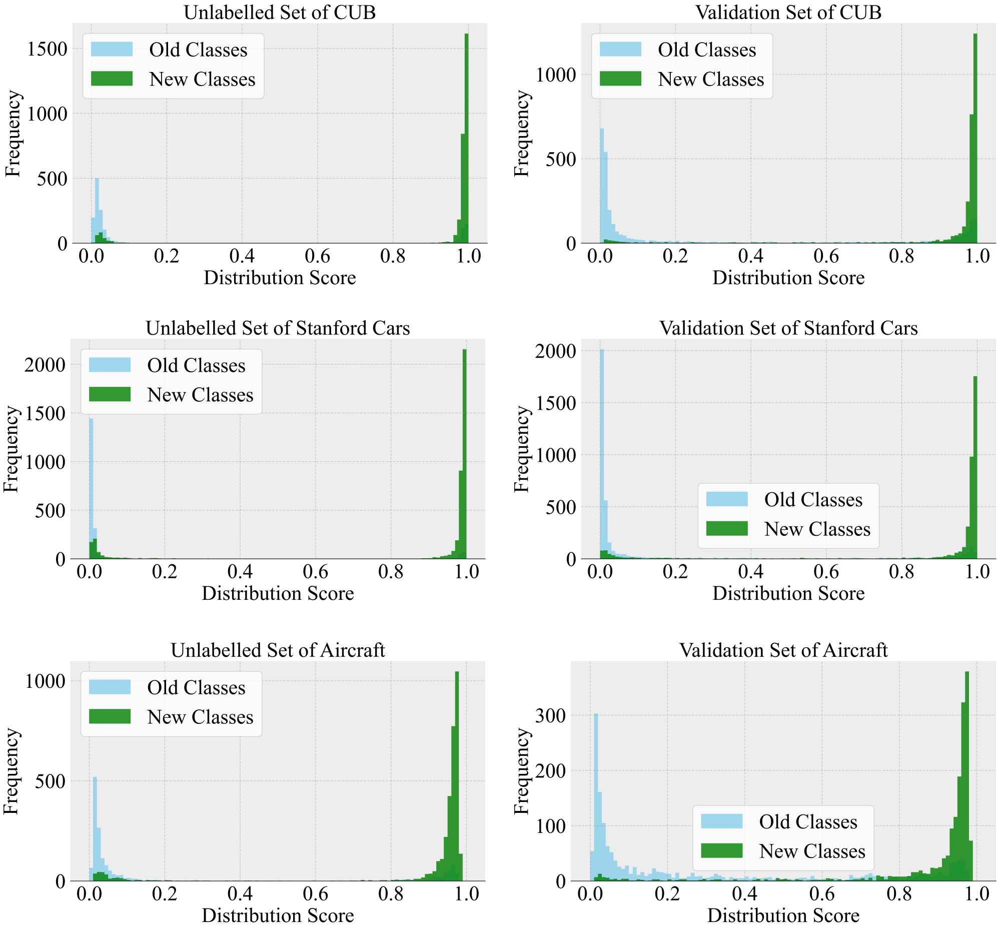


Figure 6: Histograms of the distribution scores $s_i$ for datasets in SSB Vaze et al. (2022c).

Table 13: OOD performance in terms of AUROC on unlabelled data, including CIFAR-10 Krizhevsky et al. (2009), CIFAR-100 Krizhevsky et al. (2009), ImageNet-100 Deng et al. (2009), CUB Wah et al. (2011), Stanford Cars Krause et al. (2013), and FGVC-Aircraft Maji et al. (2013).

| | CIFAR-10 | CIFAR-100 | ImageNet-100 | CUB | Stanford Cars | FGVC-Aircraft |
|---|---|---|---|---|---|---|
| $\mathcal{L}_{sdl}$ | 66.1 | 90.8 | 96.5 | 77.5 | 78.6 | 76.2 |
| $\mathcal{L}_{sdl}$+$\mathcal{L}_{gcd}$+$\mathcal{L}_{adl}$ | **97.5** | **94.8** | **99.5** | **86.8** | **89.6** | **86.3** |

# I ANALYSIS OF ATTENTION MAPS

In our D2G framework, both the backbone embedding space and the GCD classifier are optimized. Thus, the `CLS` token is indirectly optimized. We can glean insights from its attention with the patch embeddings. In Fig. 7, we visualize the attention maps from the final transformer block in the DINO backbone Caron et al. (2021) on the three fine-grained datasets in SSB benchmark Vaze et al. (2022c). Within this final block, a multi-head self-attention layer with 12 attention heads attends to the input features, producing 12 attention maps between the `CLS` token and patch embeddings at a resolution of $14 \times 14$. Following Caron et al. (2021), we compute the mean value of these attention maps and upsample them to the image size to visualize the most prominent regions. The visualization demonstrates that the attention maps generated by our model predominantly focus on the object of interest, effectively ignoring spurious factors and background clutter, while those of the DINO baseline are more scattered over the entire image.

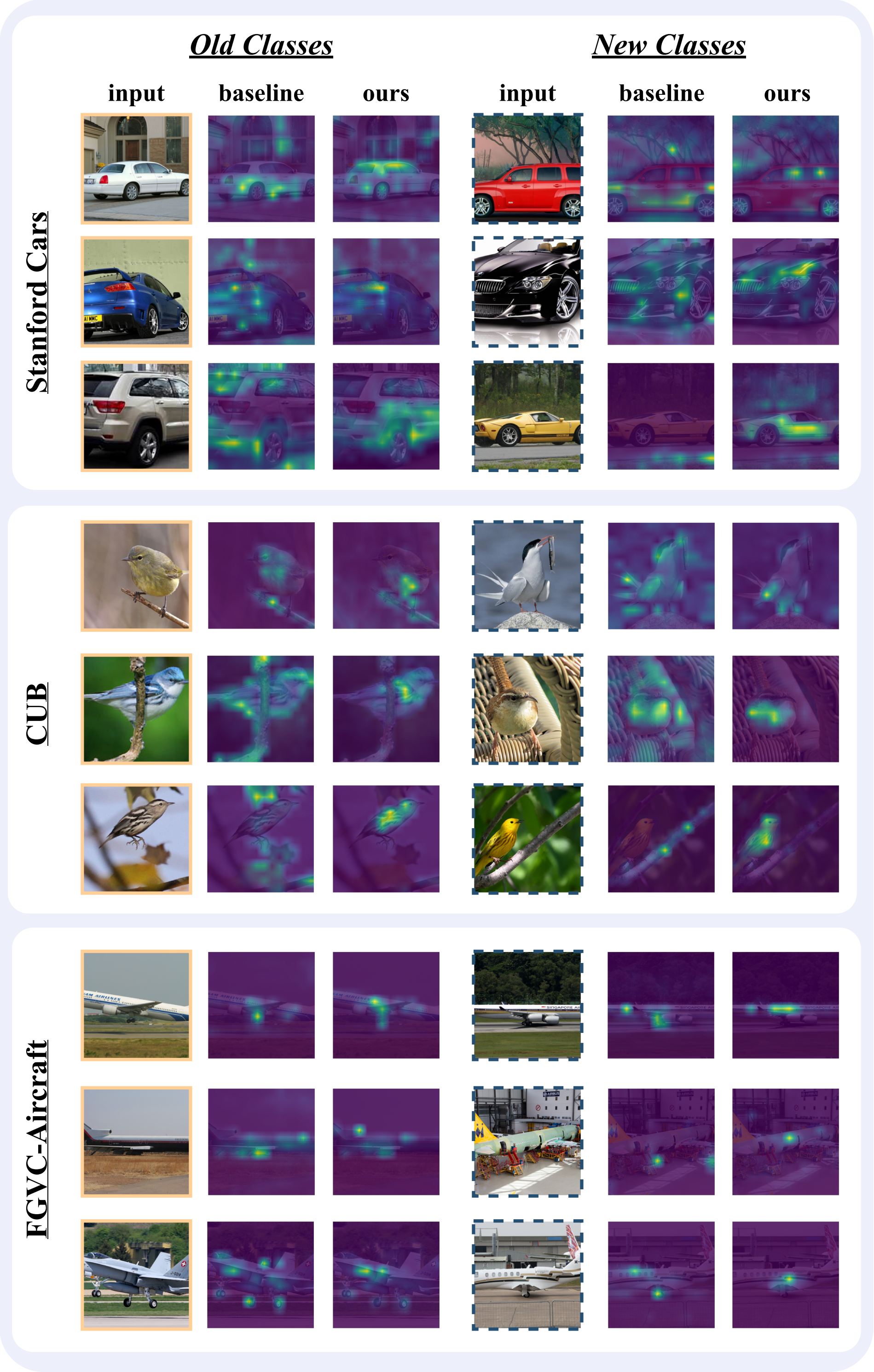


Figure 7: Visualization of attention maps. Our method successfully directs its attention towards foreground objects, irrespective of whether they belong to the 'Old' or 'New' classes. The baseline denotes the pre-trained DINO.

## J ABLATION STUDIES ON MORE DATASETS

In addition to the Stanford Cars dataset, we present ablation results on additional datasets to validate the effectiveness of the proposed components. These include the other two datasets from the SSB benchmark: CUB Wah et al. (2011) and FGVC-Aircraft Maji et al. (2013), as well as the generic dataset ImageNet-100 Deng et al. (2009), detailed in Tab. 14. The results indicate that directly applying debiased learning to the original GCD classifier results in a performance decline across all three datasets (Row (1) *vs.* Row (2)). In contrast, utilizing an auxiliary classifier leads to performance improvements of 3.3%, 3.5%, and 1.7% on the three datasets, respectively, as observed in Row (1) *vs.* Row (3). This underscores the importance of the auxiliary classifier in achieving effective debiased learning. Moreover, the joint training of the debiased classifier and the OOD detector provides further enhancements (Row (3) *vs.* Row (5)). Lastly, the incorporation of distribution guidance results in additional performance improvements. These findings align with those observed on the Stanford Cars dataset, as demonstrated in Tab. 4.

Table 14: Ablations on more datasets, including CUB Wah et al. (2011), FGVC-Aircraft Maji et al. (2013) and ImageNet-100 Deng et al. (2009). ACC of ‘All’, ‘Old’ and ‘New’ categories are listed.

| | Debiased Learning | Auxiliary Classifier | Semantic Dist. Learning | Dist. Guidance | CUB | | | FGVC-Aircraft | | | ImageNet-100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | All | Old | New | All | Old | New | All | Old | New |
| (1) | ✗ | ✗ | ✗ | ✗ | 60.3 | 65.6 | 57.7 | 54.2 | 59.1 | 51.8 | 83.0 | 93.1 | 77.9 |
| (2) | ✓ | ✗ | ✗ | ✗ | 58.6 | **72.3** | 51.7 | 53.7 | 62.9 | 49.1 | 82.8 | 94.1 | 77.2 |
| (3) | ✓ | ✓ | ✗ | ✗ | 63.8 | 69.3 | 61.1 | 57.7 | 59.8 | 56.5 | 84.7 | 94.0 | 80.0 |
| (4) | ✗ | ✗ | ✓ | ✗ | 61.3 | 69.4 | 57.3 | 56.6 | **64.8** | 52.5 | 83.5 | 92.4 | 78.9 |
| (5) | ✓ | ✓ | ✓ | ✗ | 64.9 | 70.9 | 61.9 | 59.4 | 64.4 | 56.9 | 85.0 | 93.8 | 80.3 |
| (6) | ✓ | ✓ | ✓ | ✓ | **66.3** | 71.8 | **63.5** | **61.7** | 63.9 | **60.6** | **85.9** | **94.3** | **81.6** |

## K IMPACT OF HYPERPARAMETERS

In this section, we analyze the impact of hyperparameters in our D2G framework, including the depth of the projection network $\rho_s$, loss weights, and the number of tuned blocks.

**Depth of projection network $\rho_s$.** As discussed in the paper, it is essential to disentangle the OOD and GCD feature spaces due to the differing learning objectives of these two tasks. To assess the impact of the depth of the projection network $\rho_s$, we conduct an experiment on the SSB benchmark, focusing on the number of layers in this MLP network. Here, a depth of 0 denotes the absence of a projection network, meaning that the two tasks are optimized within the same feature space. As shown in Tab. 19, incorporating a one-layer $\rho_s$ results in performance improvements by 1.3%, 1.6% and 1.1% on CUB, Stanford Cars, and FGVC-Aircraft, respectively. The average GCD performance across all categories of D2G gradually improves as the number of MLP layers increases from 0 to 5. However, extending the MLP to 7 layers yields little to no further improvement in performance. In our implementation, we therefore adopt a 5-layer MLP for $\rho_s$ in our framework.

**Loss weights $\lambda_{sdl}$ and $\lambda_{adl}$.** For these two loss weights, we first intuitively set the default value based on existing literature and our hypothesis. Our rationale for selecting values for the loss weights is as follows: For $\lambda_{sdl}$, we take inspiration from the previous literature using OVA classifier Saito & Saenko (2021). In the paper, the model is fine-tuned with a learning rate of $10^{-3}$ , while the learning rate in the SimGCD baseline is 0.1 (which is 100 times larger than $10^{-3}$). To achieve a similar learning effect, as validated in Saito & Saenko (2021), we scale our $\lambda_{sdl}$ value from 1.0 down to 1/100. Therefore, we set $\lambda_{sdl} = 0.01$ by default. For $\lambda_{adl}$, the weight of the debiased classifier, we expect it to play an important role similar to that of the original GCD classifier (where the loss weight is set to 1.0). Thus, we have defaulted this value to 1.0. After determining the default values, we conducted experiments on the SSB benchmark regarding the two loss weights by exploring values around the defaults. For $\lambda_{sdl}$, the range was (0.005, 0.01, 0.02). As for $\lambda_{adl}$, the range was (0.5, 1.0, 2.0). The impact of $\lambda_{sdl}$ is detailed below in Tab. 16, with $\lambda_{adl}$ set to 1.0. The impact of $\lambda_{adl}$ is illustrated below in Tab. 17, with $\lambda_{sdl}$ set to 0.01. The results are in line with our hypothesis, indicating that our selected hyperparameters are indeed reasonable.

Table 15: GCD performance on SSB Vaze et al. (2022c) using different number of layers in $\rho_s$.

| | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| MLP layer | All | Old | New | All | Old | New | All | Old | New | All |
| 0 | 63.6 | **75.2** | 57.8 | 62.3 | 76.2 | 54.1 | 59.6 | 62.2 | 58.3 | 61.8 |
| 1 | 64.9 | 71.6 | 61.6 | 63.9 | 80.2 | 56.0 | 60.7 | 63.7 | 59.2 | 63.1 |
| 3 | 66.0 | 73.5 | 62.3 | 64.7 | **82.2** | 56.2 | 61.1 | 64.2 | 59.5 | 63.9 |
| 5 | **66.3** | 71.8 | **63.5** | **65.3** | 81.6 | **57.4** | 61.7 | 63.9 | **60.6** | **64.4** |
| 7 | 65.8 | 72.0 | 62.7 | 64.8 | 80.5 | 57.3 | **61.9** | **65.2** | 60.3 | 64.1 |

Table 16: GCD performance on SSB Vaze et al. (2022c) using different values of $\lambda_{sdl}$.

| | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_{sdl}$ | All | Old | New | All | Old | New | All | Old | New | All |
| 0.02 | 65.5 | **73.2** | 61.6 | 64.3 | 79.2 | 57.1 | 60.6 | 63.5 | 59.1 | 63.5 |
| 0.01 | **66.3** | 71.8 | **63.5** | **65.3** | **81.6** | **57.4** | 61.7 | 63.9 | **60.6** | **64.4** |
| 0.005 | 65.8 | 72.4 | 62.5 | 64.9 | 81.2 | 57.0 | **62.1** | **65.4** | 60.3 | 64.3 |

Table 17: GCD performance on SSB Vaze et al. (2022c) using different values of $\lambda_{adl}$.

| | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_{adl}$ | All | Old | New | All | Old | New | All | Old | New | All |
| 0.5 | 64.3 | **72.2** | 60.3 | 63.6 | 79.3 | 56.1 | 60.2 | 63.5 | 58.6 | 62.7 |
| 1.0 | **66.3** | 71.8 | **63.5** | **65.3** | 81.6 | **57.4** | **61.7** | **63.9** | **60.6** | **64.4** |
| 2.0 | 65.5 | 70.8 | 62.8 | 64.1 | **83.0** | 55.0 | 60.4 | 63.5 | 58.8 | 63.3 |

**Number of tuned blocks.** In the baseline configuration Wen et al. (2023), only the last transformer block of the ViT-B/16 backbone is fine-tuned during training. In contrast, our framework incorporates additional tasks, including OOD detection and debiased learning, which would require different embedding spaces, thus calling for the need of more trainable parameters. In our experiments on both fine-grained and generic datasets, we explore tuning the last two blocks, and we note that tuning more than two blocks may lead to instability during training. Furthermore, we observe that increasing the number of tuned blocks can improve performance on specific datasets, particularly those that are fine-grained. As shown in Table 18, tuning one additional transformer block leads to a performance improvement of over $1\%$ on the fine-grained datasets. In contrast, the performance enhancement on the generic datasets is more modest, at no more than $0.6\%$. Similar strategies have also been employed in previous methods, such as Infosieve Rastegar et al. (2023).

Table 18: GCD performance of SimGCD and D2G by tuning different numbers of transformer blocks.

| | | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | ImageNet-100 | | | CIFAR100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | tuned blocks | All | Old | New | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| SimGCD | 1 | 60.3 | 65.6 | 57.7 | 53.8 | 71.9 | 45.0 | 54.2 | 59.1 | 51.8 | 83.0 | 93.1 | 77.9 | 80.1 | 81.2 | 77.8 |
| SimGCD | 2 | 60.8 | 65.8 | 58.4 | 53.6 | 67.6 | 49.8 | 52.8 | 56.8 | 50.8 | 83.2 | 92.9 | 78.3 | 79.4 | 80.1 | 77.3 |
| D2G | 1 | 65.1 | 70.9 | 62.2 | 63.0 | 80.2 | 54.7 | 60.4 | **65.0** | 58.1 | 85.7 | 94.0 | 81.5 | 82.4 | 83.6 | 79.5 |
| D2G | 2 | **66.3** | **71.8** | **63.5** | **65.3** | **81.6** | **57.4** | **61.7** | 63.9 | **60.6** | **85.9** | **94.3** | **81.6** | **83.0** | **84.6** | **79.9** |

## L STABILITY ANALYSIS

Following the baseline established in Wen et al. (2023), we also assess the stability of the proposed method across all datasets utilized in our experiments. Tab. **??** reports the average results over three independent runs together with the standard deviations. Compared to the baseline results reported in Wen et al. (2023), we observe that the variance is even smaller, despite achieving significantly higher performance.

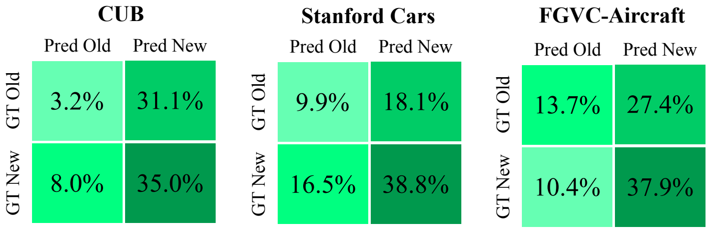
Table 19: Complete results of D2G and SimGCD over three independent runs.

|  | SimGCD |  |  | D2G |  |  |
|---|---|---|---|---|---|---|
| Dataset | All | Old | New | All | Old | New |
| CUB | 60.3±0.1 | 65.6±0.9 | 57.7±0.4 | 66.4±0.4 | 72.9±0.6 | 63.2±0.4 |
| Stanford Cars | 53.8±2.2 | 71.9±1.7 | 45.0±2.4 | 65.2±0.7 | 81.7±1.2 | 57.3±0.6 |
| FGVC-Aircraft | 54.2±1.9 | 59.1±1.2 | 51.8±2.3 | 61.7±0.5 | 65.9±1.2 | 59.5±1.1 |
| CIFAR-10 | 97.1±0.0 | 95.1±0.1 | 98.1±0.1 | 97.3±0.1 | 95.0±0.2 | 98.4±0.1 |
| CIFAR-100 | 80.1±0.9 | 81.2±0.4 | 77.8±2.0 | 83.1±0.7 | 84.7±0.7 | 80.0±0.9 |
| ImageNet-100 | 83.0±1.2 | 93.1±0.2 | 77.9±1.9 | 86.1±0.6 | 94.5±0.5 | 81.8±0.6 |
| ImageNet-1K | 57.1±0.1 | 77.3±0.1 | 46.9±0.2 | 64.9±0.3 | 82.1±0.2 | 56.4±0.4 |
| Oxford-Pet | - | - | - | 93.2±0.2 | 86.3±0.1 | 96.8±0.3 |
| Herbarium19 | 44.0±0.4 | 58.0±0.4 | 36.4±0.8 | 44.9±0.3 | 59.3±0.3 | 37.1±0.5 |

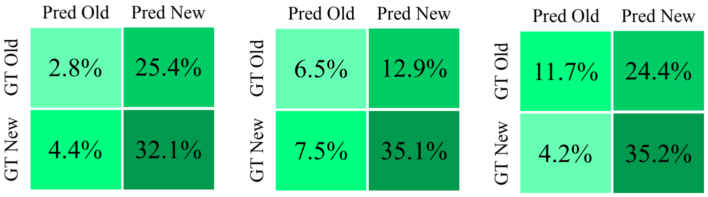
## M PREDICTION ERROR ANALYSIS

In this section, we provide quantitative analysis on the improvements brought by our method from the perspective of prediction errors. Particularly, we examine the baseline model's prediction by categorizing the errors into four types based on the relationship between the predicted and ground-truth classes: 'True Old', 'False New', 'False Old', and 'True New'. For example, 'True New' refers to incorrectly predicting a 'New' class sample to another 'New' class, while 'False Old' indicates incorrectly predicting a 'New' class sample as some 'Old' class. From this perspective, our debiased learning method primarily aims to mitigate the label bias between 'Old' and 'New' classes, thereby reducing the likelihood of 'New' class samples being predicted as 'Old'. Consequently, this reduction in bias leads to a decrease in 'False Old' predictions while reducing the errors of all the other three types.

In Fig. 8, we present the ratios of the four types of *prediction errors* as a proportion of the total number of samples in the new or old categories across three datasets in the SSB benchmark. As shown in Fig. 8(a), the error distributions vary significantly across datasets. Notably, the Stanford Cars dataset exhibits the highest number (16.5%) of 'False Old' samples, explaining why our method demonstrates the most substantial performance improvement on this dataset. In contrast, the CUB dataset shows the fewest (8.0%) 'False Old' samples, indicating relatively limited potential for performance enhancement. Comparing Fig. 8(a) and Fig. 8(b), we can see a significant reduction on the ratio of 'False Old' as well as other three types of errors on all the three datasets.



(a) Ratios of the four error types in SimGCD



(b) Ratios of the four error types in D2G

Figure 8: Ratios of the four types of prediction errors in GCD on SSB benchmark using SimGCD and D2G with DINO Caron et al. (2021) pretrained backbone. 'Pred' and 'GT' refer to the predicted and ground-truth results, respectively.