Abstract

Generalized Category Discovery (GCD) aims to discover novel categories in unlabelled datasets using knowledge learned from labelled samples. Previous studies argued that parametric classifiers are prone to overfitting to seen categories, and endorsed using a non-parametric classifier formed with semi-supervised k-means. However, in this study, we investigate the failure of parametric classifiers, verify the effectiveness of previous design choices when high-quality supervision is available, and identify unreliable pseudo-labels as a key problem. We demonstrate that two prediction biases exist: the classifier tends to predict seen classes more often, and produces an imbalanced distribution across seen and novel categories. Based on these findings, we propose a simple yet effective parametric classification method that benefits from entropy regularisation, achieves state-of-the-art performance on multiple GCD benchmarks and shows strong robustness to unknown class numbers. We hope the investigation and proposed simple framework can serve as a strong baseline to facilitate future studies in this field. Our code is available at: https://github.com/CVMI-Lab/SimGCD.

1. Introduction

With large-scale labelled datasets, deep learning methods can surpass humans in recognising images [23]. However, it is not always possible to collect large-scale human annotations for training deep learning models. Therefore, there is a rich body of recognition models that focus on learning with a large number of unlabelled data. Among them, semi-supervised learning (SSL) [29, 4, 33] is regarded as a promising approach, yet with the assumption that labelled instances are provided for each of the categories the model needs to classify. Generalized category discovery (GCD) [39] is recently formalised to relax this assumption by assuming the unlabelled data can also contain similar yet distinct categories from the labelled data. The goal of GCD is to learn a model that is able to classify the already-seen categories in the labelled data, and more importantly, jointly discover the new categories in the unlabelled data and make correct classifications. Developing a strong method for this problem could help us better utilise the easily available large-scale unlabelled datasets.

Previous works [39, 20, 16, 5] approach this problem from two perspectives: learning generic feature representations to facilitate the discovery of novel categories, and gen-
erating pseudo clusters/labels for unlabelled data to guide the learning of a classifier. The former is often achieved by using self-supervised learning methods [20, 46, 17, 22, 8, 48] to improve the generalization ability of features to novel categories. For constructing the classifier, earlier works [20, 46, 51, 5, 16] adopt a parametric approach that builds a learnable classifier on top of the extracted features. The classifier is jointly optimised with the backbone using labelled data and pseudo-labelled data.

However, recent research shows [39, 15] that parametric classifiers are prone to overfit to seen categories (see Fig. 2) and thus promote using a non-parametric classifier such as $k$-means clustering. Albeit obtaining promising results, the non-parametric classifiers suffer from heavy computation costs on large-scale datasets due to quadratic complexity of the clustering algorithm. Besides, unlike a learnable parametric classifier, the non-parametric method loses the ability to jointly optimise the separating hyperplane of all categories in a learnable manner, potentially being sub-optimal.

This motivates us to revisit the reason that makes previous parametric classifiers fail to recognise novel classes. In a series of investigations (Sec. 3) from the view of supervision quality, we verify the effectiveness of prior design choices in feature representations and training paradigms when strong supervision is available, and conclude that the key to previous parametric classifiers’ degraded performance is unreliable pseudo labels. By diagnosing the statistics of its predictions, we identify severe prediction biases within the model, i.e., the bias towards predicting more ‘Old’ classes than ‘New’ classes (Fig. 5) and the bias of producing imbalanced pseudo-labels across all classes (Fig. 6).

Based on these findings, we thus present a simple parametric classification baseline for generalized category discovery (see Figs. 1 and 7). The representation learning objective follows GCD [39], and the classification objective is simply cross-entropy for labelled samples and self-distillation [8, 2] for unlabelled samples. Besides, an entropy regularisation term is also adopted to overcome biased predictions by enforcing the model to predict more uniformly distributed labels across all possible categories. Empirically, we indeed observe that our method produces more balanced pseudo-labels (Figs. 9 and 10) and achieves a large performance gain on multiple GCD benchmarks (Tabs. 2 to 4), indicating that the two types of biases we identified are the core reason why the parametric-classifier-based approach performs poorly for GCD. Additionally, we observe that the entropy regulariser could also be used to enforce robustness towards an unknown number of categories (Figs. 11 and 12), this could further ease the deployment of parametric classifiers for GCD in real-world scenarios.

Our contributions are summarised as follows: (1) We revisit the design choices of parametric classification and conclude the key factors that make it fail for GCD. (2) Based on the analysis, we propose a simple yet effective parametric classification method. (3) Our method achieves SOTA on multiple popular GCD benchmarks, challenging the recent promotion of non-parametric classification for this task.

2. Related Works

Semi-Supervised Learning (SSL) has been an important research topic where a number of methods have been proposed [4, 33, 36]. SSL assumes that the labelled instances are available for all possible categories in the unlabelled dataset; the objective is to learn a model to perform classification using both the labelled samples as well as the large-scale available unlabelled data. One of the most effective methods for SSL is the consistency-based method, where the model is forced to learn consistent representations of two different augmentations of the same image [33, 4, 36]. Furthermore, it is also shown that self-supervised representation learning is helpful for the task of SSL [45, 30] as it can provide a strong representation for the task.

Open-Set Semi-Supervised Learning considers the case where the unlabelled data may contain outlier data points that do not belong to any of the categories in the labelled training set. The goal is to learn a classifier for the labelled categories from a noisy unlabelled dataset [44, 32, 10, 19]. As this problem only focuses on the performance of the labelled categories, the outlier from novel categories are simply rejected and no further classification is needed.

Generalized Category Discovery (GCD) is a relatively new problem recently formalised in Vaze et al. [39], and is also studied in a parallel line of work termed open-world semi-supervised learning [5, 34]. Different from the common assumption of SSL [29], GCD does not assume the unlabelled dataset comes from the same class set as the labelled dataset, posing a greater challenge for designing an effective model. GCD can be seen as a natural extension of the novel category discovery (NCD) problem [21]
where it is assumed that the unlabelled dataset and the labelled dataset do not have any class overlap, thus baselines for NCD [20, 46, 51, 50, 16] can be adopted for the GCD problem by extending the classification head to have more outputs [39]. The incremental setting of GCD is also explored [47, 31]. It is pointed out in [39] that a non-parametric classifier formed using semi-supervised $k$-means can outperform strong parametric classification baselines from NCD [20, 16] because it can alleviate the overfitting to seen categories in the labelled set. In this paper, we revisit this claim and show that parametric classifiers can reach stronger performance than non-parametric classifiers.

**Deep Clustering** aims at learning a set of semantic prototypes from unlabelled images with deep neural networks. Considering that no label information is available, the focus is on how to obtain reliable pseudo-labels. While early attempts rely on hard labels produced by $k$-means [6], there has been a shift towards soft labels produced by optimal transport [1, 7], and more recently sharpened predictions from an exponential moving average-updated teacher model [8, 2]. Deep clustering has shown strong potential for unsupervised representation learning [6, 1, 7, 8, 2], unsupervised semantic segmentation [11, 23], semi-supervised learning [3], and novel category discovery [16]. In this work, we study the techniques that make strong parametric classifiers for GCD with inspirations from deep clustering.

### 3. On the Failure of Parametric Classification

In order to explore the reason that makes previous parametric classifiers fail to recognize ‘New’ classes for generalized category discovery, this section presents preliminary studies to reveal the role of two major components: representation learning (Sec. 3.2) and pseudo-label quality on unseen classes (Sec. 3.3). These have led to conflicting choices of previous works, but why? We show a unified viewpoint (Figs. 3 and 4), and emphasis that taking pseudo-label quality into account is important for selecting the suitable design choice. This then led to our diagnosis of what makes the degenerated pseudo-labels (Sec. 3.4), and motivated our de-biased pseudo-labeling strategy.

#### 3.1. Investigation Setting

**Generalized category discovery.** Given an unlabelled dataset $D^u = \{(x^u_i, y^u_i)\} \in \mathcal{X} \times \mathcal{Y}_u$ where $\mathcal{Y}_u$ is the label space of the unlabelled samples, the goal of GCD is to learn a model to categorize the samples in $D^u$ using the knowledge from a labelled dataset $D^l = \{(x^l_i, y^l_i)\} \in \mathcal{X} \times \mathcal{Y}_l$ where $\mathcal{Y}_l$ is the label space of labelled samples and $\mathcal{Y}_l \subset \mathcal{Y}_u$. We denote the number of categories in $\mathcal{Y}_u$ as $K_u = |\mathcal{Y}_u|$, it is common to assume the number of categories is known a-priori [20, 46, 51, 16], or can be estimated using off-the-shelf methods [21, 39].

**Representation learning.** For representation learning, we follow GCD [39], which applies supervised contrastive learning [24] on labelled samples, and self-supervised contrastive learning [9] on all samples (detailed in Sec. 4.1).

**Classifier.** We follow UNO [16] to adopt a prototypical classifier. Take $f(x)$ as the feature vector of an image $x$ extracted using from the backbone $f$, the procedure for producing logits is $l = \frac{1}{2} (w||w||^T (f(x)||f(x)||)$. Here $\tau$ is the temperature value that scales up the norm of $l$ and facilitates optimisation of the cross-entropy loss [42].

**Training settings.** We train with varying supervision qualities. The minimal supervision setting utilises only the labels in $D^l$, while the oracle supervision setting assumes all samples are labelled (both $D^l$ and $D^u$). Besides, we study two practical settings that adopt pseudo labels for unlabelled samples in $D^u$: self-label that predicts pseudo-labels with the Sinkhorn Knopp algorithm following [16], and self-distil, which depicts another pseudo-labeling strategy as in Fig. 7 and will be introduced in detail in Sec. 4.2. For all settings, we only employ a cross-entropy loss on the (pseudo-)labelled samples on hand for classification. Note that unless otherwise stated, this is done on decoupled features, thus representation learning is unaffected.

#### 3.2. Which Representation to Build Your Classifier?

**Motivation.** Following the trend of deep clustering that focuses on self-supervised representation learning [7], previous parametric classification work UNO [16] fed the classifier with representations taken from the projector. While in GCD [39], significantly stronger performance is achieved with a non-parametric classifier built upon representations taken from the backbone. We revisit this choice as follows.

**Setting.** Consider $f$ as the feature backbone, and $g$ is a multi-layer perceptron (MLP) projection head. Given an input image $x_i$, the representation from the backbone can be written as $f(x_i)$, and that from the projector is $g(f(x_i))$.

![Figure 3. Results with different representations](attachment:image.png)

Table 1. CIFAR10 Results with different representations. We build the classifier on post-backbone or post-projector representations, and train with varying supervision quality. Results on 'Old' class consistently benefit from the post-backbone representations regardless of the supervision quality, while unleashing its potential on 'New' class requires stronger pseudo labels.
Result & discussion. As in Fig. 3, the post-backbone feature space has a clearly higher upper bound for learning prototypical classifiers than the post-projector feature space. Using a projector in self-supervised learning lets the projector focus on solving pretext tasks and allows the backbone to keep as much information as possible (which facilitates downstream tasks) [12]. But when good classification performance is all you need, our results suggest that the classification objective should build on post-backbone representations directly. The features post the projector might focus more on solving the pretext task and not be necessarily useful for the classification objective. Note that high-quality pseudo labels are necessary to unleash the post-backbone representations’ potential to recognise novel categories.

3.3. Decoupled or Joint Representation Learning?

Motivation. Previous parametric classification methods, e.g., UNO [16], commonly tune the representations jointly with the classification objective. On the contrary, in the two-stage non-parametric method GCD [39] where the performance in ‘New’ classes is notably higher, classification/clustering is fully decoupled from representation learning, and the representations can be viewed as unaltered by classification. In this part, we study whether the joint learning strategy contributes to previous parametric classifiers’ degraded performance in recognising ‘New’ classes.

Setting. Consider $f(x)$ as the representation fed to the classifier, decoupled training, as the previous settings adopted, indicates $f(x)$ is decoupled when computing the logits $l$, thus the classification objective won’t supervise representation learning. While for joint training, the representations are jointly optimised by classification.

Result & discussion. The results are illustrated in Fig. 4. When adopting the self-labeling strategy, there is a sharp drop in ‘Old’ class performance on both datasets, while for the ‘New’ classes, it can improve by 13 points on CIFAR100, and drop by a small margin on CUB. In contrast, when a stronger pseudo-labelling strategy (self-distillation) or even oracle labels are utilised, we observe consistent gains from joint training. This means that the joint training strategy does not necessarily result in UNO [16]’s low performance in ‘New’ classes; on the contrary, it can even boost ‘New’ class performance by a notable margin. Our overall explanation is that UNO’s framework could not make reliable pseudo-labels, thus restricting its ability to benefit from joint training. The joint training strategy is not to blame and is, in fact, helpful. When switching to a more advanced pseudo-labelling paradigm that produces high-quality pseudo-labels, the help from joint training can be even more significant.

3.4. The Devil Is in the Biased Predictions

Motivation. In Secs. 3.2 and 3.3, we verified the effectiveness of two design choices when high-quality pseudo labels are available, and concluded the key to previous work’s degraded performance is unreliable pseudo labels. We then further diagnose the statistics of its predictions as follows.

Setting. We categorise the model’s errors into four types: “True Old”, “False New”, “False Old”, and “True New” according to the relationship between predicted and ground-truth class. E.g., “True New” refers to predicting a ‘New’ class sample to another ‘New’ class, while “False Old” indicates predicting a ‘New’ class sample as some ‘Old’ class.

Figure 5. Prediction bias between ‘Old’/‘New’ classes. We simplify the results to binary classification and categorise errors in ‘All’ ACC into four types. Both works, especially UNO+, are prone to make “False Old” predictions, and many samples corresponding to ‘New’ classes are misclassified as an ‘Old’ class.

Figure 6. Prediction bias across ‘Old’/‘New’ classes. We show the per-class prediction distributions. Both works, especially UNO+, are prone to make biased predictions. Across all classes, the predictions are unexpectedly biased towards the head classes.
Result & discussion. We observe two types of prediction bias. In Fig. 5, both works, especially UNO+ [16], are prone to make “False Old” predictions. In other words, their predictions are biased towards ‘Old’ classes. Besides, the “True New” errors are also notable, indicating that misclassification within ‘New’ classes is also common. We then depict the predictions’ overall distribution across ‘Old’/’New’ classes in Fig. 6, and both works show highly biased predictions. This double-bias phenomenon then motivated the prediction entropy regularisation design in our method.

4. Method

In this section, we present the whole picture of this simple yet effective method (see Fig. 7), a one-stage framework that builds on GCD [39], and jointly trains a parametric classifier with self-distillation and entropy regularisation. And in Sec. 5.3, we discuss the step-by-step changes that lead a simple baseline to our solution.

4.1. Representation Learning

Our representation learning objective follows GCD [39], which is supervised contrastive learning [24] on labelled samples, and self-supervised contrastive learning [9] on all samples. Formally, given two views (random augmentations) \( x_i \) and \( x'_i \) of the same image in a mini-batch \( B \), the self-supervised contrastive loss is written as:

\[
L^u_{\text{rep}} = \frac{1}{|B|} \sum_{i \in B} - \log \frac{\exp(z^\top_i z'_i / \tau_u)}{\sum_i \exp(z^\top_i z'_i / \tau_u)},
\]

where the feature \( z_i \) is normalised, \( g, f \) denote the backbone and the projection head, and \( \tau_u \) is a temperature value. The supervised contrastive loss is similar, and the major difference is that positive samples are matched by their labels, formally written as:

\[
L^\text{cls} = \frac{1}{|B|} \sum_{i \in B} \frac{1}{|\mathcal{N}_i|} \sum_{q \in \mathcal{N}_i} - \log \frac{\exp(z^\top_i z'_i / \tau_c)}{\sum_i \exp(z^\top_i z'_i / \tau_c)},
\]

where \( \mathcal{N}_i \) indexes all other images in the same batch that hold the same label as \( x_i \). The overall representation learning loss is balanced with \( \lambda \): \( L_{\text{rep}} = (1 - \lambda)L^u_{\text{rep}} + \lambda L^\text{cls} \), where \( B^l \) corresponds to the labelled subset of \( B \).

4.2. Parametric Classification

Our parametric classification paradigm follows the self-distillation [8, 2] fashion. Formally, with \( K = |\mathcal{Y}_l| \cup \mathcal{Y}_u| \) denoting the total number of categories, we randomly initialise a set of prototypes \( \mathcal{C} = \{c_1, \ldots, c_k\} \), each standing for one category. During training, we calculate the soft label for each augmented view \( x_i \), by softmax on cosine similarity between the hidden feature \( h_i = f(x_i) \) and the prototypes \( c_k \).

\[
L = \frac{1}{|B|} \sum_{i \in B} \sum_{k \neq n} \log \frac{\exp(z^\top_i p_k / \tau)}{\sum_i \exp(z^\top_i p_k / \tau)} - \varepsilon H(p),
\]

where \( \varepsilon H(p) = -\sum_k p_k \log p_k \) denotes the entropy of the predictions; \( H(p) = -\sum_k p_k \log p_k \) denotes the mean prediction of a batch, and the entropy \( H(p) \) is equal to the overall objective is simply \( L_{\text{rep}} + L_{\text{cls}} \).

Discussions. Please note that this work doesn’t aim to promote new methods but to examine existing solutions, provide insights into their failures and build a simple yet strong baseline solution. The paradigm of producing pseudo-labels from sharpened predictions of another augmented view appears to resemble consistency-based methods [33, 4, 36] in the SSL community. However, despite differences in aug-

![Figure 7. The overall framework of our method.](image-url)
mentation strategies and soft/hard pseudo-labels, our approach jointly performs category discovery and self-training style learning, while the SSL methods purely focus on bootstrapping itself with unlabelled data, and does not discover novel categories. Besides, entropy regularisation is also explored in deep clustering to avoid trivial solution [2]. In contrast, our method shows its help in overcoming the prediction bias between and within seen/novel classes (Figs. 9 and 10), and enforcing robustness to unknown numbers of categories (Fig. 11).

5. Experiments

5.1. Experimental Setup

Datasets. We validate the effectiveness of our method on the generic image recognition benchmark (including CIFAR10/100 [26] and ImageNet-100 [37]), the recently proposed Semantic Shift Benchmark [40] (SSB, including CIFAR10/100 [26], ImageNet-100 [37], and the harder Herbarium 19 [35] and ImageNet-1K [13]. For each dataset, we follow [39] to sample a subset of all classes as the labelled ('Old') classes \( \mathcal{Y}_l \); 50% of the images from these labelled classes are used to construct \( \mathcal{D}^u \), and the remaining images are regarded as the unlabelled data \( \mathcal{D}^u \).

See Tab. 1 for statistics of the datasets we evaluate on.

Evaluation protocol. We evaluate the model performance with clustering accuracy (ACC) following standard practice [39]. During evaluation, given the ground truth \( y^* \) and the predicted labels \( \hat{y} \), the ACC is calculated as \( \text{ACC} = \frac{1}{M} \sum_{i=1}^{M} 1(\hat{y}_i = p(y^*_i)) \) where \( M = |\mathcal{D}^u| \), and \( p \) is the optimal permutation that matches the predicted cluster assignments to the ground truth class labels.

Implementation details. Following GCD [39], we train all methods with a ViT-B/16 backbone [14] pre-trained with DINO [8]. We use the output of \([CLS]\) token with a dimension of 768 as the feature for an image, and only fine-tune the last block of the backbone. We train with a batch size of 128 for 200 epochs with an initial learning rate of 0.1 decayed with a cosine schedule on each dataset. Aligning with [39], the balancing factor \( \lambda \) is set to 0.35, and the temperature values \( \tau_u, \tau_l \) as 0.07, 1.0, respectively. For the classification objective, we set \( \tau_s \) to 0.1, and \( \tau_l \) is initialised to 0.07, then warmed up to 0.04 with a cosine schedule in the starting 30 epochs. All experiments are done with an NVIDIA GeForce RTX 3090 GPU.

5.2. Comparison With the State of the Arts

We compare with state-of-the-art methods in generalized category discovery (ORCA [5] and GCD [39]), strong baselines derived from novel category discovery (RS+ [20] and UNO+ [16]), and \( k \)-means [27] on DINO [8] features. On both the fine-grained SSB benchmark (Tab. 2) and generic image recognition datasets (Tab. 3), our method achieves notable improvements in recognising ‘New’ classes (across the instances in \( \mathcal{D}^u \) that belong to classes in \( \mathcal{Y}_u \)) outperforming the SOTAs by around 10%. The results in old classes are also competing against the best-performing baselines. Given that the ability to discover ‘New’ classes is a more desirable ability, the results are quite encouraging.

In Tab. 4, we also report the results on Herbarium 19 [35], a naturally long-tailed fine-grained dataset that is closer to the real-world application of generalized category discovery, and ImageNet-1K [13], a large-scale generic classification dataset. Still, our method shows consistent improvements in all metrics.
As shown in column (+SD), we achieve consistent improvements across all datasets by a large margin (e.g., 26% in CIFAR100, 13% in CUB) in ‘New’ classes. We then further adopt a teacher temperature warmup strategy (+TW) to lower the confidence of the pseudo-labels at an earlier stage. The intuition is that at the beginning, both the classifier and the representation are not well fitted to the target data, thus the pseudo-labels are not quite reliable. This is shown to be helpful for fine-grained classification datasets, while for generic classification datasets, which are similar to the pre-training data (ImageNet), the unreliable pseudo label is not a problem, thus lowering the confidence does not show help. For simplicity, we keep the training strategy consistent.

### Jointly training the representation

Previous settings adopt a decoupled training strategy for consistent representations with GCD [39] and fair comparison. Finally, as confirmed in Sec. 3.3, we jointly supervise the representation with the classification objective (+JT). This results in a consistent improvement in ‘New’ classes for all datasets. Changes in ‘Old’ classes are mostly neutral or positive, with a notable drop in CIFAR100. Our intuition is that the original representations are already good enough for ‘Old’ classes in this dataset, and some incorrect pseudo labels lead to sight degradation in this case.

#### Improving the representations

As suggested in Sec. 3.2, we build the classifier on the backbone (+BR). This further makes notable improvements in ‘Old’ classes, while changes in ‘New’ classes vary across datasets. This indicates that the pseudo labels’ quality is insufficient to benefit from the post-backbone representations (Fig. 3).

#### Improving the pseudo labels

We start by replacing the self-labelling strategy with our self-distillation paradigm. As shown in column (+SD), we achieve consistent improvements across all datasets by a large margin (e.g., 26% in CIFAR100, 13% in CUB) in ‘New’ classes. We then further adopt a teacher temperature warmup strategy (+TW) to lower the confidence of the pseudo-labels at an earlier stage. The intuition is that at the beginning, both the classifier and the representation are not well fitted to the target data, thus the pseudo-labels are not quite reliable. This is shown to be helpful for fine-grained classification datasets, while for generic classification datasets, which are similar to the pre-training data (ImageNet), the unreliable pseudo label is not a problem, thus lowering the confidence does not show help. For simplicity, we keep the training strategy consistent.

### Jointly training the representation

Previous settings adopt a decoupled training strategy for consistent representations with GCD [39] and fair comparison. Finally, as confirmed in Sec. 3.3, we jointly supervise the representation with the classification objective (+JT). This results in a consistent improvement in ‘New’ classes for all datasets. Changes in ‘Old’ classes are mostly neutral or positive, with a notable drop in CIFAR100. Our intuition is that the original representations are already good enough for ‘Old’ classes in this dataset, and some incorrect pseudo labels lead to sight degradation in this case.

#### Improving the representations

As suggested in Sec. 3.2, we build the classifier on the backbone (+BR). This further makes notable improvements in ‘Old’ classes, while changes in ‘New’ classes vary across datasets. This indicates that the pseudo labels’ quality is insufficient to benefit from the post-backbone representations (Fig. 3).

#### Improving the pseudo labels

We start by replacing the self-labelling strategy with our self-distillation paradigm. As shown in column (+SD), we achieve consistent improvements across all datasets by a large margin (e.g., 26% in CIFAR100, 13% in CUB) in ‘New’ classes. We then further adopt a teacher temperature warmup strategy (+TW) to lower the confidence of the pseudo-labels at an earlier stage. The intuition is that at the beginning, both the classifier and the representation are not well fitted to the target data, thus the pseudo-labels are not quite reliable. This is shown to be helpful for fine-grained classification datasets, while for generic classification datasets, which are similar to the pre-training data (ImageNet), the unreliable pseudo label is not a problem, thus lowering the confidence does not show help. For simplicity, we keep the training strategy consistent.

### Jointly training the representation

Previous settings adopt a decoupled training strategy for consistent representations with GCD [39] and fair comparison. Finally, as confirmed in Sec. 3.3, we jointly supervise the representation with the classification objective (+JT). This results in a consistent improvement in ‘New’ classes for all datasets. Changes in ‘Old’ classes are mostly neutral or positive, with a notable drop in CIFAR100. Our intuition is that the original representations are already good enough for ‘Old’ classes in this dataset, and some incorrect pseudo labels lead to sight degradation in this case.

#### Improving the representations

As suggested in Sec. 3.2, we build the classifier on the backbone (+BR). This further makes notable improvements in ‘Old’ classes, while changes in ‘New’ classes vary across datasets. This indicates that the pseudo labels’ quality is insufficient to benefit from the post-backbone representations (Fig. 3).

#### Improving the pseudo labels

We start by replacing the self-labelling strategy with our self-distillation paradigm. As shown in column (+SD), we achieve consistent improvements across all datasets by a large margin (e.g., 26% in CIFAR100, 13% in CUB) in ‘New’ classes. We then further adopt a teacher temperature warmup strategy (+TW) to lower the confidence of the pseudo-labels at an earlier stage. The intuition is that at the beginning, both the classifier and the representation are not well fitted to the target data, thus the pseudo-labels are not quite reliable. This is shown to be helpful for fine-grained classification datasets, while for generic classification datasets, which are similar to the pre-training data (ImageNet), the unreliable pseudo label is not a problem, thus lowering the confidence does not show help. For simplicity, we keep the training strategy consistent.

### Jointly training the representation

Previous settings adopt a decoupled training strategy for consistent representations with GCD [39] and fair comparison. Finally, as confirmed in Sec. 3.3, we jointly supervise the representation with the classification objective (+JT). This results in a consistent improvement in ‘New’ classes for all datasets. Changes in ‘Old’ classes are mostly neutral or positive, with a notable drop in CIFAR100. Our intuition is that the original representations are already good enough for ‘Old’ classes in this dataset, and some incorrect pseudo labels lead to sight degradation in this case.

#### Improving the representations

As suggested in Sec. 3.2, we build the classifier on the backbone (+BR). This further makes notable improvements in ‘Old’ classes, while changes in ‘New’ classes vary across datasets. This indicates that the pseudo labels’ quality is insufficient to benefit from the post-backbone representations (Fig. 3).

#### Improving the pseudo labels

We start by replacing the self-labelling strategy with our self-distillation paradigm. As shown in column (+SD), we achieve consistent improvements across all datasets by a large margin (e.g., 26% in CIFAR100, 13% in CUB) in ‘New’ classes. We then further adopt a teacher temperature warmup strategy (+TW) to lower the confidence of the pseudo-labels at an earlier stage. The intuition is that at the beginning, both the classifier and the representation are not well fitted to the target data, thus the pseudo-labels are not quite reliable. This is shown to be helpful for fine-grained classification datasets, while for generic classification datasets, which are similar to the pre-training data (ImageNet), the unreliable pseudo label is not a problem, thus lowering the confidence does not show help. For simplicity, we keep the training strategy consistent.

### Jointly training the representation

Previous settings adopt a decoupled training strategy for consistent representations with GCD [39] and fair comparison. Finally, as confirmed in Sec. 3.3, we jointly supervise the representation with the classification objective (+JT). This results in a consistent improvement in ‘New’ classes for all datasets. Changes in ‘Old’ classes are mostly neutral or positive, with a notable drop in CIFAR100. Our intuition is that the original representations are already good enough for ‘Old’ classes in this dataset, and some incorrect pseudo labels lead to sight degradation in this case.

#### Improving the representations

As suggested in Sec. 3.2, we build the classifier on the backbone (+BR). This further makes notable improvements in ‘Old’ classes, while changes in ‘New’ classes vary across datasets. This indicates that the pseudo labels’ quality is insufficient to benefit from the post-backbone representations (Fig. 3).

#### Improving the pseudo labels

We start by replacing the self-labelling strategy with our self-distillation paradigm. As shown in column (+SD), we achieve consistent improvements across all datasets by a large margin (e.g., 26% in CIFAR100, 13% in CUB) in ‘New’ classes. We then further adopt a teacher temperature warmup strategy (+TW) to lower the confidence of the pseudo-labels at an earlier stage. The intuition is that at the beginning, both the classifier and the representation are not well fitted to the target data, thus the pseudo-labels are not quite reliable. This is shown to be helpful for fine-grained classification datasets, while for generic classification datasets, which are similar to the pre-training data (ImageNet), the unreliable pseudo label is not a problem, thus lowering the confidence does not show help. For simplicity, we keep the training strategy consistent.

### Jointly training the representation

Previous settings adopt a decoupled training strategy for consistent representations with GCD [39] and fair comparison. Finally, as confirmed in Sec. 3.3, we jointly supervise the representation with the classification objective (+JT). This results in a consistent improvement in ‘New’ classes for all datasets. Changes in ‘Old’ classes are mostly neutral or positive, with a notable drop in CIFAR100. Our intuition is that the original representations are already good enough for ‘Old’ classes in this dataset, and some incorrect pseudo labels lead to sight degradation in this case.
5.4. Analyses And Discussions

Entropy regularisation helps overcome prediction bias. We verify the effectiveness of entropy regularisation in overcoming prediction bias by diagnosing the model’s classification errors and class-wise prediction distributions. Fig. 9 shows that this term consistently helps reduce “False New” and “False Old” errors, which refer to predicting an ‘Old’ class sample to a ‘New’ class, and vice-versa. Besides, Fig. 10 shows proper entropy regularisation helps overcome the imbalanced pseudo labels across all classes, and approach the ground truth (GT) class distribution.

Entropy regularisation enforces robustness to unknown numbers of categories. The main text assumed the category number is known a-priori following prior works [20, 46, 51, 16], which is impractical [49]. In Fig. 11, we present the results with different numbers of categories on five representative datasets. A category number lower than the ground truth significantly limits the ability to discover ‘New’ categories, and the model tends to focus more on the ‘Old’ classes. On the other hand, increasing the category number results in less harm to the generic image recognition datasets and can even be helpful for some datasets. When a stronger entropy penalty is imposed, the model shows strong robustness to the category number. Interestingly, further analysis in Fig. 12 shows the network prefers to keep the number of active prototypes low and close to the real category number. This finding is inspiring and could ease the deployment of GCD in real-world scenarios.

What makes for the significant improvements over GCD given identical representations? One interesting message from Fig. 8 is that, even with the same representations (col. +TW), we can already improve GCD by a large margin. We thus study the classification predictions and the major components that lead to the performance gap. As shown in Fig. 13, the non-parametric classifier (semi-supervised $k$-means) adopted by GCD produces highly imbalanced predictions, while our method better fits the true distribution; Right: our method significantly improves GCD’s tail classes.

How does the classification objective change the representations? In Fig. 8, we have shown that jointly training the representations with the classification objective can lead to $\sim 15\%$ boost in ‘New’ classes on CIFAR100. We study this difference by visualising the representations before and after tuning with t-SNE [38]. As in Fig. 14, jointly tuning the feature leads to less ambiguity, larger margins, and compact clusters. Concerning why this is not as helpful for CUB: we hypothesise that one important factor lies in how transferable the features learned in ‘Old’ classes are to ‘New’ classes. While it may be easier for a cat classifier to be adapted to dogs, things can be different for fine-grained bird recognition. Besides, the small scale of CUB, which contains only 6k images while holding a large class split (200), might also make it hard to learn transferable features.
Trade-off between ‘Old’ and ‘New’ categories. We plot the performance evolution throughout the model learning process in Fig. 15. It can be observed that the performance on the ‘Old’ categories first climbs to the highest point at the early stage of training and then slowly degrades as the performance on the ‘New’ categories improves. We believe this demonstrates an important aspect of the design of models for the GCD problem: the performance on the ‘Old’ categories may be in odd with the performance on the ‘New’ categories, how to achieve a better trade-off between these two could be an interesting investigation for future works.

6. Limitations and Potential Future Works

Representation learning. This paper mainly targets improving the classification ability for generalized category discovery. The representation learning, however, follows the prior work GCD [39]. It is expectable that the quality of representation learning can be improved. For instance, generally, by using more advanced geometric and photometric data augmentations [18], and even multiple local crops [7]. Further, can the design of data augmentations be better aligned with the classification criteria of the target data? For another example, using a large batch size has been shown to be critical to the performance of contrastive learning-based frameworks [9]. However, the batch size adopted by GCD [39] is only 128, which might limit the quality of learned representations. Moreover, is the supervised contrastive learning plus self-supervised contrastive learning paradigm the ultimate answer to form the feature manifold? We believe that advances in representation learning can lead to further gains.

Alignment to human-defined categories. This paper follows the common practice of previous works where human labels in seen categories implicitly define the metric for unseen ones, which can be viewed as an effort to align algorithm-discovered categories with human-defined ones. However, labels in seen categories may not be good guidance when there is a gap between seen ones and the novel categories we want to discover, e.g., how to use the labelled images in ImageNet to discover novel categories in CUB? For another example, when we use a very big class vocabulary (e.g., the full ImageNet-22K [13]), categories could overlap with each other, and be in different granularities. Further, assigning text names to the discovered categories still requires a matching process, what if further utilising the relationship between class names, and directly predicting the novel categories in the text space? We believe the alignment between algorithm-discovered categories and human-defined categories is of high research value for future works.

Ethical considerations. Current methods commonly suffer from low-data or long-tailed scenarios. Depending on the data and classification criteria of specific tasks, discrimination against minority categories or instances is possible.

7. Conclusion

This study investigates the reasons behind the failure of previous parametric classifiers in recognizing novel classes in GCD and uncovers that unreliable pseudo-labels, which exhibit significant biases, are the crucial factor. We propose a simple yet effective parametric classification method that addresses these issues and achieves state-of-the-art performance on multiple GCD benchmarks. Our findings provide insights into the design of robust classifiers for discovering novel categories and we hope our proposed framework will serve as a strong baseline to facilitate future studies in this field and contribute to the development of more accurate and reliable methods for category discovery.

Acknowledgements

This work has been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 17202422), and RGC Matching Fund Scheme (RMGS). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust. The authors acknowledge SmartMore and MEGVII for partial computing support, and Zhisheng Zhong for professional suggestions.
References


[16] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In ICCV, 2021. 1, 2, 3, 4, 5, 6, 7, 8


[34] Yiyou Sun and Yixuan Li. Opencon: Open-world contrastive learning. TMLR, 2023. 2


[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. JMLR, 2008. 8, 9

[39] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In CVPR, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9

[40] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In ICLR, 2022. 6


