Parametric Information Maximization for Generalized Category Discovery

Florent Chiaroni*
ÉTS Montreal & Thales Digital Solutions
Montreal, Canada

Jose Dolz
ÉTS Montreal
Montreal, Canada

Ziko Imtiaz Masud
Thales Digital Solutions
Montreal, Canada

Amar Mitiche
INRS
Montreal, Canada

Ismail Ben Ayed
ÉTS Montreal
Montreal, Canada

Abstract

We introduce a Parametric Information Maximization (PIM) model for the Generalized Category Discovery (GCD) problem. Specifically, we propose a bi-level optimization formulation, which explores a parameterized family of objective functions, each evaluating a weighted mutual information between the features and the latent labels, subject to supervision constraints from the labeled samples. Our formulation mitigates the class-balance bias encoded in standard information maximization approaches, thereby handling effectively both short-tailed and long-tailed data sets. We report extensive experiments and comparisons demonstrating that our PIM model consistently sets new state-of-the-art performances in GCD across six different datasets, more so when dealing with challenging fine-grained problems. Our code: `https://github.com/ThalesGroup/pim-generalized-category-discovery`.

1. Introduction

Deep learning methods are driving progress in a wide span of computer vision tasks, particularly when large labeled datasets are easily accessible for training. Obtaining such large datasets is a cumbersome process, which is often a limiting factor impeding the scalability of these models. To alleviate this limitation, semi-supervised learning (SSL) has emerged as an appealing alternative, which leverages both labeled and unlabeled data to boost the performance of deep models. Despite recent success, SSL approaches work under the *closed-set* assumption, in which the categories in the labeled and unlabeled subsets share the same underlying class label space. Nevertheless, this assumption rarely holds in real scenarios, where novel categories may emerge in conjunction with known classes, which typically results in significant drops in the performances of standard supervised deep learning models. Thus, the ability to detect whether the input of a deep learning model belongs or not to a set of *known* classes seen during training is essential for robust deployment in a breadth of critical application areas, such as medicine, security, finance, agriculture, marketing and engineering [4, 34]. Thus, devising novel learning models that can address the realistic *open-set* scenario is of paramount importance.

Novel category discovery (NCD) [18, 1] tackles this problem by exploiting the knowledge learned from a set of relevant known classes to improve clustering into unknown categories. Nevertheless, NCD assumes the two sets of classes to be disjoint, which means that the unlabeled dataset contains only instances belonging to the set of novel categories. *Generalized category discovery (GCD)* [42] considers a more general scenario, where unlabeled data contain instances from both seen and novel classes. This scenario is particularly challenging, as learning is performed under class distribution mismatch, and the unlabeled data may contain categories never encountered in the available labeled set.

Contributions: In this work, we address the generalized category discovery task from an information-theoretic perspective. Our contributions are summarized as follows:

- We introduce a *Parametric Information Maximization* (PIM) model for GCD. Specifically, we propose a bi-level optimization formulation, which explores a parameterized family of objective functions, each evalu-

*Corresponding author: `florent.chiaroni.ai@gmail.com` (permanent address) ating a weighted mutual information between the features and the latent labels, subject to supervision constraints from the labeled samples. Our formulation mitigates the class-balance bias encoded in standard information maximization, deals effectively with both short-tailed and long-tailed data sets, and is model-agnostic (i.e., could be used in conjunction with any feature extractor).

- We report extensive experiments and comparisons demonstrating that PIM consistently sets new state-of-the-art performances across six different datasets, with larger gaps on the more challenging fine-grained benchmarks. It outperforms both specialized GCD methods and standard information-maximization approaches.

2. Related work

Semi-supervised learning (SSL) has been widely explored in the machine learning and computer vision community. This learning paradigm aims at leveraging large unlabeled datasets that contain the same set of classes as the labeled samples. Due to their satisfactory performance, consistency-based approaches have gained popularity recently, such as Mean-Teacher [39], MixMatch [5], UDA [45] or FixMatch [37]. An interesting alternative is self-training, which relies on the generation of pseudo-labels from a small amount of labeled data [36, 49], or in solving surrogate classification tasks [15, 46]. Nevertheless, the main limitation is that most existing SSL models rely on the *closed-set* assumption, as they do not consider unlabeled data points sampled from novel semantic categories.

Novel Class Discovery (NCD), which was formalized in [18], relaxes the closed-set assumption, as it focuses on discovering new categories in the unlabeled set by leveraging the knowledge learned from the labeled set. AutoNovel [17] (also referred to as RankStats) resorts to ranking statistics as an efficient approach for NCD. First, a good embedding is learned in a self-supervised manner for learning the early feature representation layers, which is followed by a supervised fine-tuning step with labeled samples for learning high level feature representations. Finally, to determine whether two instances from the unlabeled set are from the same category, a robust ranking statistics approach is introduced. The authors of [48] proposed a method based on dual ranking statistics, coupled with mutual knowledge distillation. OpenMix [51] showed that mixing up both labeled and unlabeled data prevents the representation learning model from overfitting the labeled categories. Other methods [23, 50] adopted contrastive-learning strategies for the task of novel category discovery. UNO [14] uses a cross-entropy loss to train the model with both the labeled and unlabeled data. Despite the good performance in discovering new categories, these methods assume that the test dataset contains instances drawn solely from the novel classes. A recent work by [47] presented a method based on a mutual-information measure, which is different from the discriminative and constrained mutual information we introduce in this work. The mutual information in [47] evaluates the relation between the old and novel categories in the label space, arguing that maximizing such a measure promotes transferring semantic knowledge. In our case, we introduce a parametric, bi-level optimization of the mutual information between the feature and label spaces, on both labeled and unlabeled samples.

Generalized Category Discovery (GCD) extends NCD by allowing both the old and new classes to coexist in the unlabeled dataset, which we tackle in this work. This pragmatic yet challenging scenario was recently introduced in [42] and triggered several other recent studies of the GCD problem. In [42], the authors proposed to fine-tune a pre-trained DINO ViT [11] with one supervised and one self-supervised contrastive term. Then, they used a semi-supervised clustering for label assignment. Note that, while UNO [14] and RankStats [17] are originally investigated for the NCD task, they are adapted for GCD in the recent study in [42], yielding UNO+ and RankStats+, respectively. Another recent approach, referred to as ORCA [10], addressed a similar problem, naming it *open world semi-supervised learning*. ORCA consists of controlling the intra-class variance of the seen classes to align and reduce the learning gap w.r.t. novel categories.

Maximizing the mutual information. Our discriminative partitioning approach (PIM) is built on the general and well-known *InfoMax* principle [32], which prescribes maximizing the mutual information (MI) between the inputs and outputs of a system. Several variants of this general principle have been recently used in machine learning and computer vision tasks, including deep clustering [22, 28, 21], few-shot learning [8, 43, 6], representation learning [40, 20, 2, 25], deep metric learning [7] and domain adaptation [35, 31, 3]. To the best of our knowledge, addressing the GCD problem from an information-theoretic perspective remains unexplored.

The pioneering discriminative clustering model in [28] and the recent transductive few-shot method in [8] are closely related to our work, as they both maximize the mutual information between the features and the latent labels. However, as we shall see in our experiments, the direct application of information maximization [28, 8] to GCD may not be highly competitive. First, the standard mutual-information objective has a strong encoded bias towards balanced partitions, via its marginal-entropy term, which might be detrimental to performances. In this work, we introduce a parametric family of mutual-information objectives, which we tackle with a bi-level optimization formulation, thereby estimating automatically the weight of the marginal-entropy term. Our parametric information maximization effectively deals with both short-tailed and long-tailed data sets, mitigating the class-balance bias. Secondly, the InfoMax models in [28, 8] were designed in the scenario where the unlabeled set contains examples from the classes seen in the available labeled set. Finally, in [28, 8], the mutual-information objective is defined over the set of unlabeled samples. In contrast, we propose a constrained mutual-information formulation defined over both labeled and unlabeled samples, thereby capturing the distribution of the whole data set in the context of GCD. As we will see in our experiments, our parametric, bi-level information maximization substantially outperforms [28, 8] in the GCD scenario.

3. Generalized Category Discovery problem

Problem definition. Assume we are given a dataset $\mathcal{D}$ composed of two subsets so that $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$. First, $\mathcal{D}_L = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ refers to a labeled subset containing N images from a set of known classes in $\mathcal{Y}_L$. For each image $\boldsymbol{x}_i$ in $\mathcal{D}_L$, we have access to its corresponding one-hot vector label $\boldsymbol{y}_i = (y_{i,k})_{1 \leq k \leq K^{\text{old}}}$, where $K^{\text{old}} = |\mathcal{Y}_L|$ is the number of classes in $\mathcal{Y}_L$. $y_{i,k} = 1$ if $\boldsymbol{x}_i$ belongs to class k, and 0 otherwise. Now, let $\mathcal{D}_U = \{\boldsymbol{x}_i\}_{i=1}^M$ denote the unlabeled subset, which contains M images from a set of classes $\mathcal{Y}_U$ composed of *known* classes, as well as *novel* classes, i.e., $\mathcal{Y}_L \subset \mathcal{Y}_U$. Note that, during inference, $K = |\mathcal{Y}_U|$ is the total number of classes, which contains both known and novel categories. Given this setting, the Generalized Category Discovery (GCD) task introduced in [42] consists in partitioning the images in the unlabeled set into separate clusters at test time. Each obtained cluster is supposed to represent a separate known or novel category. In other words, the GCD problem amounts to jointly solving (i) a semi-supervised classification task for the known classes; and (ii) a clustering task for the novel classes.

Notation. Let us denote $g_{\boldsymbol{\theta}} : \mathcal{D} \to \mathcal{Z} \subset \mathbb{R}^D$ as the trained encoder responsible for mapping an input image $\boldsymbol{x}_i$ into a feature vector $\boldsymbol{z}_i$ of dimension D, with θ the set of trainable parameters and $\mathcal{Z}$ the set of all embedded features, for both the labeled and unlabeled samples. We now define a soft partitioning model $f_{\boldsymbol{W}} : \mathcal{Z} \to [0,1]^K$, which is parameterized by weight matrix $\boldsymbol{W} = (\boldsymbol{w}_k)_{1 \leq k \leq K}$, where $\boldsymbol{w}_k = (w_{k,n})_{1 \leq n \leq D}$ denote its trainable parameters. For each input feature map $\boldsymbol{z}_i$, $f_{\boldsymbol{W}}$ outputs a softmax prediction vector $\boldsymbol{p}_i = (p_{i,k})_{1 \leq k \leq K}$ of dimension K, defined on the standard $(K-1)$-probability simplex domain $\Delta^{K-1} = \{\boldsymbol{p}_i \in [0,1]^K \mid \boldsymbol{p}_i^T \mathbf{1} = 1\}$. Note that, similarly to the prior work in [42], we assume the number of clusters during the partitioning task to be known.

Let $Z \in \mathbb{R}^D$ denote a random variable representing the feature map. Z follows $\mathbb{P}(Z)$, which denotes the distribution of the set of embedded features $\mathcal{Z}$. Hence, each feature map data point $\boldsymbol{z}_i$ is a realization of Z. Furthermore, let $Y \in \mathcal{Y} = \{1, \ldots, K\}$ be the random variable following the dataset label distribution $\mathbb{P}(Y)$.

4. Background on information maximization

Marginal distributions. Let $\boldsymbol{\pi} = (\pi_k)_{1 \leq k \leq K}$, where $\pi_k = \mathbb{P}(Y = k; \boldsymbol{W})$ denote the marginal distributions that one can approximate by the soft[1] proportion of points within each cluster, via Monte-Carlo estimation, as follows:

$$\begin{aligned} \pi_k &= \int_{\boldsymbol{z}} \mathbb{P}(Z = \boldsymbol{z}) \mathbb{P}(Y = k | Z = \boldsymbol{z}; \boldsymbol{W}) \mathrm{d}\boldsymbol{z} \\ &\approx \frac{1}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} \mathbb{P}(Y = k | Z = \boldsymbol{z}_i; \boldsymbol{W}) = \frac{1}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} p_{i,k} \end{aligned} \tag{1}$$

Mutual Information. The mutual information between the labels and the features maps can be written as follows:

$$I(Y, Z) = \mathcal{H}(Y) - \mathcal{H}(Y|Z), \tag{2}$$

with $\mathcal{H}(Y)$ referring to the entropy of the marginal distributions $\mathbb{P}(Y = k; \boldsymbol{W})$, and $\mathcal{H}(Y|Z)$ referring to the entropy of the conditional probability distribution $\mathbb{P}(Y|Z; \boldsymbol{W})$.

Marginal entropy. The marginal entropy term, $\mathcal{H}(Y)$, could be estimated using the soft marginal distribution approximation in (1), as follows:

$$\begin{aligned} \mathcal{H}(Y) &= -\sum_{k=1}^{K} \mathbb{P}(Y = k; \boldsymbol{W}) \log \mathbb{P}(Y = k; \boldsymbol{W}) \\ &= -\sum_{k=1}^{K} \pi_k \log \pi_k \end{aligned} \tag{3}$$

The class-balance bias in InfoMax approaches [8, 28]. It is common in the literature to maximize the unsupervised mutual information in Eq. (2), which is often defined over unlabeled samples. This is the case, for instance, of the discriminative clustering in [28] (RIM) or the transductive few-shot inference in [8] (TIM). A closer look at the marginal-entropy term in (3) enables to write it, up to a constant, as a Kullback-Leibler (KL) divergence between the marginal probabilities of predictions and the uniform distribution:

$$\mathcal{H}(Y) \stackrel{c}{=} -\mathcal{D}_{KL}(Y || \mathcal{U}_K), \tag{4}$$

[1] We use the term *soft* because the proportions are directly estimated on the softmax predictions, instead of using hard labels.

where $\stackrel{\mathrm{c}}{=}$ stands for equality up to additive and/or non-negative multiplicative constant, and $\mathcal{U}_K$ is the uniform distribution over K classes. Thus, maximizing $\mathcal{H}(Y)$ pushes the marginal distribution towards the uniform distribution, as made explicit by the previous equation, thereby encoding a strong bias towards balanced partitions. Note that this standard mutual information objective lacks a mechanism to explicitly control the weight of the marginal entropy. Therefore, this term has the potential to harm the performance in the case of imbalanced scenarios, where the underlying class distribution is no longer uniform. Based on the above-identified limitation of the mutual information, we introduce a parametric family of mutual-information objectives, which we tackle with a bi-level optimization formulation, thereby estimating the relative weight of the marginal-entropy term.

Beyond discriminative InfoMax clustering approaches [28, 22], it is worth noting that standard generative clustering objectives, such the ubiquitous K-means and it probabilistic generalizations [24], also have a well-known bias towards balanced partitions [24, 9]. We note that, in the context of GCD, the study in [42], which introduced the task, used K-means clustering.

5. Proposed bi-level and constrained InfoMax

Constrained mutual information We propose to maximize a constrained version of the mutual information presented in (2), integrating supervision constraints on the conditional probabilities $\boldsymbol{p}_i$ of the samples within the labeled set. Our constrained information maximization reads:

$$\max_{\boldsymbol{W}} \mathcal{H}(Y) - \mathcal{H}(Y|Z) \quad \text{s.t.} \quad \boldsymbol{y}_i = \boldsymbol{p}_i \quad \forall \boldsymbol{z}_i \in \mathcal{Z}_L \tag{5}$$

where $\mathcal{Z}_L$ denotes the set of embedded features for the labeled samples. It is straightforward to notice that by plugging the equality constraints in (5) into the mutual information, one could write the objective as follows:

$$\min_{\boldsymbol{W}} \sum_{k=1}^{K} \pi_k \log \pi_k - \frac{1}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} \sum_{k=1}^{K} h_{i,k} \log p_{i,k}, \tag{6}$$

where $h_{i,k} = y_{i,k}$ if $\boldsymbol{z}_i \in \mathcal{Z}_L$ and $h_{i,k} = p_{i,k}$ otherwise. That is, for $\boldsymbol{y}_i = \boldsymbol{p}_i \quad \forall \boldsymbol{z}_i \in \mathcal{Z}_L$, the objectives in (5) and (6) are equal to each other. Interestingly, the terms corresponding to $h_{i,k} = y_{i,k}$ in (6) yield the standard cross-entropy (CE) loss for the labeled samples. This CE loss could be viewed as a *penalty* function for imposing constraints $\boldsymbol{y}_i = \boldsymbol{p}_i \quad \forall \boldsymbol{z}_i \in \mathcal{Z}_L$, as it reaches its minimum when these constraints are satisfied. Therefore, we do not need to impose explicitly the equality constraints in (5). Notice that, for the labeled samples, CE in (6) replaced the conditional entropy term in the mutual information. This is reasonable as CE enables to jointly impose the supervision constraints while encouraging implicitly confident predictions, as it pushes them toward one vertex of the simplex. Both CE and conditional entropy reach their minima at the vertices of the simplex.

Bi-level optimization To mitigate the bias of the mutual information towards balanced partitions, we propose to explore a family of weighted versions of the objective in (6), which we parameterize with a variable parameter λ and tackle as a bi-level optimization problem:

$$\begin{aligned} F(\boldsymbol{W}, \lambda) = & \underbrace{\sum_{k=1}^{K} \pi_k \log \pi_k}_{\mathcal{H}(Y)} \underbrace{- \frac{1}{|\mathcal{Z}_L|} \sum_{i \in \mathcal{Z}_L} \sum_{k=1}^{K} y_{i,k} \log p_{i,k}}_{\text{CE}} \\ & \underbrace{- \frac{\lambda}{|\mathcal{Z}_U|} \sum_{i \in \mathcal{Z}_U} \sum_{k=1}^{K} p_{i,k} \log p_{i,k}}_{\propto \mathcal{H}(Y|Z)} \end{aligned} \tag{7}$$

where $\mathcal{Z}_U$ denotes the set of embedded features for the unlabeled samples (i.e., $\mathcal{Z} = \mathcal{Z}_U \cup \mathcal{Z}_L$). Variable $\lambda \in (0, 1]$ controls the effect of the unsupervised loss terms in Eq. (7), i.e., confidence vs. class balance. Therefore, as we will see in our experiments, learning λ from the labeled set, via a bi-level optimization, yields highly competitive performances in the GCD setting, more so when dealing with long-tailed (imbalanced) data sets. Our bi-level formulation reads:

$$\min_{\boldsymbol{W}} F(\boldsymbol{W}, \lambda) \quad \text{s.t} \quad \lambda \in \arg\max_{\lambda \in (0,1]} A_L(\lambda), \tag{8}$$

where F is the upper-level objective and A_L is the lower-level objective defined by the clustering accuracy[2] on the set of labeled samples:

$$A_L(\lambda) = \frac{1}{|\mathcal{Z}_L|} \sum_{i=1}^{|\mathcal{Z}_L|} \mathbb{1}_{\{\hat{\boldsymbol{y}}_i(\lambda) = \boldsymbol{y}_i\}}, \tag{9}$$

and $\hat{\boldsymbol{y}}_i(\lambda)$ are the one-hot vector predictions on labeled samples maximizing parametric mutual information:

$$G(\boldsymbol{W}, \lambda) = \sum_{k=1}^{K} \pi_k \log \pi_k - \frac{\lambda}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} \sum_{k=1}^{K} p_{i,k} \log p_{i,k} \tag{10}$$

To tackle our problem, we explore a finite set[3] $\boldsymbol{\lambda}$ of uniformly-spaced values of variable λ in $(0, 1]$. For each of these values of λ, we optimize $G(\boldsymbol{W}, \lambda)$ in Eq. (10) w.r.t to linear-classifier parameters $\boldsymbol{W}$ via standard gradient steps, thereby obtaining predictions $\hat{\boldsymbol{y}}_i(\lambda)$. Note that, although

[2] We used the Hungarian algorithm to align labels of the most consistent K^{old} clusters (among the total K clusters) with the K^{old} class labels.

[3] In our experiments, we used 19 values of λ in $[5e^{-2}, 1]$, i.e. the cardinalty of set $\boldsymbol{\lambda}$ is 19.

we explore several values of λ, this remains computationally efficient as the feature encoder parameters are fixed and only classifier parameters $\boldsymbol{W}$ are updated. For initializing $\boldsymbol{W}$, which could be viewed as class prototypes, we use the K-means++ algorithm. This process yields a prediction of the optimal λ as: $\lambda_{\text{opt}} = \arg\max_{\lambda \in \boldsymbol{\lambda}} A_L(\lambda)$. Finally, the partitioning solution of our GCD problem in Eq. (8) is obtained by optimizing the upper-level objective $F(\boldsymbol{W}, \lambda_{\text{opt}})$ via gradient steps.

6. Experiments

6.1. Experiments setting

Datasets. We evaluate and compare our approach to GCD state-of-the-art approaches across six different natural image datasets. More concretely, this includes three well-known generic object recognition datasets, CIFAR10 [30], CIFAR100 [30] and ImageNet-100 [12], as well as the recent semantic shift benchmark suite (SSB) [41]. The latter is composed of three fine-grained datasets, CUB [44], Stanford Cars [29] and Herbarium19 [38], which bring an additional challenge to the performance of the baselines. CUB and Stanford Cars contain fine-grained categories, which are arguably harder to distinguish than generic object classes. Herbarium19 is a long-tailed dataset, which reflects a real-world use case with severe class imbalance along with large intra-class and low inter-class variations.

We follow the original GCD setting [42] by splitting the original training set of each dataset into labeled and unlabeled subsets. More precisely, half of the image samples belonging to the K^{old} known classes is assigned to the labeled subset, whereas the remaining half is assigned to the unlabeled subset. The latter also contains all the image samples from the remaining classes present in the original dataset, which we consider as the novel classes. In this way, the unlabeled subset is composed of instances from K different classes. Table 2 provides, for each dataset, the number of classes as well as the number of samples, within the labeled and unlabeled subsets.

Evaluation protocol. We follow the evaluation protocol presented in GCD [42]. In particular, for the partitioning task, we first employ the Hungarian algorithm to find cluster-to-class assignment jointly on all samples, both within the known and novel classes. Then, we use this optimal label-assignment solution to estimate the overall partitioning accuracy (ACC) for all classes (ALL), for known classes (OLD), and for novel classes (NEW)[4]. In the evaluation, we also report the estimated number of classes $\hat{K}$ in the unlabeled set and the corresponding error $Err = \frac{|\hat{K}-K|}{K}$, with K the real number of classes for each dataset.

Implementation details:

- **Encoder $g_{\boldsymbol{\theta}}$:** As in [42], we employ the vision transformer ViT-B-16 [13] as our backbone encoder $g_{\boldsymbol{\theta}}$ (i.e. the feature extractor). It is first pre-trained on the unlabeled dataset ImageNet [12] with DINO [11] self-supervision. Then, it is fine-tuned on each GCD dataset of interest with a semi-supervised contrastive loss composed of an unsupervised noise contrastive term [16] and a supervised contrastive term [26]. It is empirically demonstrated in [42] that this pre-training procedure achieves robust feature representations. The ensuing feature dimension is 768 per input image.

- **Partitioning model $f_{\boldsymbol{W}}$:** Our partitioning model follows the architecture of a standard linear classifier. We first initialize prototypes $\boldsymbol{W}$ with the centroids produced by the semi-supervised K-means (ssKM) clustering modelon the entire feature map set $\mathcal{Z}$. The maximum number of clustering iterations for ssKM is set to 100. Then, we train $f_{\boldsymbol{W}}$ with the standard Adam optimizer [27], with a learning rate of 0.001 and a weight decay of 0.01, for 1000 epochs during the partitioning task, but only for 500 epochs during the search for $\hat{K}$, in order to reduce the computational cost. We set the training batch size equal to the size of the dataset, which is quite feasible in terms of memory and computation, since our approach requires only the pre-computed feature maps.

- **Conditional entropy weight λ:** During the search of the number of classes, we simply set $\lambda = 1$ for the unsupervised discriminative clustering step. However, during the partitioning task where the number of classes is fixed, i.e. with K assumed to be known or equal to $\hat{K}$, we automatically select the optimal value for λ in the interval $(0, 1]$, as previously detailed in Sec. 5.

6.2. Main results

In this section, we perform a comprehensive empirical evaluation of our method and compare it to GCD [42], as well as to several adapted state-of-the-art approaches. In particular, RankStats+ and UNO+ are the adapted versions from RankStats [17] and UNO [14], which were originally developed for the NCD task. Furthermore, we report the results obtained when applying *K-means* [33] on the raw extracted features from DINO. The scores for K-means, RankStats+, UNO+ and GCD [42] are reported from [42].

[4] Note that our partially supervised strategy already enables to correctly align beforehand the clusters corresponding to the known classes with real-class labels (See Eq. 7). W.r.t this interesting property, the standard classification accuracy metric could also be directly employed to estimate the OLD Acc of our method.

		CUB			Stanford Cars			Herbarium19		
Approach	InfoMax	All	Old	New	All	Old	New	All	Old	New
k-means		34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+ [17] (TPAMI-21)		33.3	51.6	24.2	28.3	61.8	12.1	27.9	55.8	12.8
UNO+ [14] (ICCV-21)		35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	53.7	14.7
ORCA [10] (ICLR-22)		27.5	20.1	31.1	15.9	17.1	15.3	22.9	25.9	21.3
ORCA [10] - ViTB16		38.0	45.6	31.8	33.8	52.5	25.1	25.0	30.6	19.8
GCD [42] (CVPR-22)		51.3	56.6	48.7	39.0	57.6	29.9	35.4	51.0	27.0
RIM [28] (NeurIPS-10) (semi-sup.)	✓	52.3	51.8	52.5	38.9	57.3	30.1	40.1	**57.6**	30.7
TIM [8] (NeurIPS-20)	✓	53.4	51.8	54.2	39.3	56.8	30.8	40.1	57.4	30.7
PIM (proposed)	✓	**62.7**	**75.7**	**56.2**	**43.1**	66.9	**31.6**	**42.3**	56.1	**34.8**

		CIFAR10			CIFAR100			ImageNet-100		
Approach	InfoMax	All	Old	New	All	Old	New	All	Old	New
k-means		83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ [17] (TPAMI-21)		46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [14] (ICCV-21)		68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	95.0	57.9
ORCA [10] (ICLR-22)		88.9	88.2	89.2	55.1	65.5	34.4	67.6	90.9	56.0
ORCA [10] - ViTB16		**97.1**	96.2	**97.6**	69.6	76.4	56.1	76.5	92.2	68.9
GCD [42] (CVPR-22)		91.5	97.9	88.2	70.8	77.6	57.0	74.1	89.8	66.3
RIM [28] (NeurIPS-10) (semi-sup.)	✓	92.4	98.1	89.5	73.8	78.9	63.4	74.4	91.2	66.0
TIM [8] (NeurIPS-20)	✓	93.1	98.0	90.6	73.4	78.3	63.4	76.7	93.1	68.4
PIM (proposed)	✓	94.7	97.4	93.3	**78.3**	**84.2**	**66.5**	**83.1**	**95.3**	**77.0**

Table 1. **Generalized Category Discovery partitioning.** Partitioning ACC scores across fine-grained and generic datasets.

	CIFAR10	CIFAR100	ImageNet-100	CUB	SCars	Herbarium19
$\lvert\mathcal{Y}_L\rvert$	5	80	50	100	98	341
$\lvert\mathcal{Y}_U\rvert$	10	100	100	200	196	683
$\lvert\mathcal{D}_L\rvert$	12.5k	20k	31.9k	1.5k	2.0k	8.9k
$\lvert\mathcal{D}_U\rvert$	37.5k	30k	95.3k	4.5k	6.1k	25.4k

Table 2. Composition of the datasets used.

We have also evaluated ORCA with its original ResNet architecture [19] by using the code provided by the authors[5], as well as with the more competitive ViT-B-16 architecture [13], as both GCD [42] and our method use this architecture.

In order to highlight the superior performance of the proposed approach in comparison to previous mutual-information strategies, we have also adapted the InfoMax approaches RIM [28] and TIM [8], previously discussed in Sec. 2, to this novel setting. Specifically, TIM and the semi-supervised version of RIM were originally designed to deal with semi-labeled datasets, where both the labeled and unlabeled sets contain examples from the same classes. Thus, we have expanded the number of prototypes in RIM and TIM, and their resulting prediction output vector from K^{old} to K. For the sake of fairness, we use the same training hyper-parameters and initialization prototypes as in our approach, as detailed in Sec. 6.1. Furthermore, we also applied RIM and TIM on top of the same fixed feature extractor, similarly to ssKM in GCD [42] and our approach.

We first focus on the partitioning task (Tab. 1), which evaluates the label assignment performance, ACC, on the unlabeled samples. Following the original GCD setting [42], we assume the real number of classes, K, to be known.

Comparisons with the state-of-the-art in GCD. Overall, one could observe that PIM significantly outperforms state-of-the-art methods UNO+, RankStat+ and GCD [42], with a consistent performance improvement ranging from 3% on CIFAR10 to up to 11% on CUB, when considering all categories. Considering ORCA, the improvement gain brought by our method is particularly significant on the fine-grained datasets. It ranges from 9.3% on Stanford Cars to up to 24.7% on CUB. In contrast, the performance differences are less remarkable across the general datasets. In particular, PIM does not outperform ORCA-ViTB16 in the simpler CIFAR-10 dataset, but brings 8-9% in performance gains in CIFAR-100 and ImageNet-100, which are arguably more complex datasets. These results suggest that our approach is more suitable in scenarios where the total number of classes is relatively large, potentially presenting a severe degree of class imbalance.

Comparisons with adapted RIM [28] and TIM [8]. The results obtained by the adapted semi-supervised RIM, TIM, and our approach PIM show the overall superiority of mutual information based methods compared to GCD [42]. In

[5] https://github.com/snap-stanford/orca

addition, one may observe that RIM and TIM yielded very similar results. This behaviour could be explained by the fact that these two adaptations to the GCD problem amount to use the same fixed loss function, and the slight performance differences might be due to the classifier choice for the conditional model. Indeed, RIM uses a multiclass logistic regression, whereas the soft-classifier of TIM measures the l2 norm between prototypes (i.e. classifier weights) and uses L2-normalized embedded features. Last, and more importantly, we can observe that methods based on the standard mutual information yielded performances lower than the proposed approach. We hypothesize that the reason for these differences is two-fold: (i) RIM and TIM compute the marginal entropy term exclusively over the unlabeled data, while we compute it over the entire distribution. In other words, PIM maximizes the mutual information over the whole dataset, which enables to better capture the entire data distribution; (ii) thanks to the proposed bi-level optimization process, the optimal lambda parameter for each dataset can be estimated automatically (as detailed in Section 5). We stress that our hypothesis is supported by empirical evidence in Section 6.3.

Interval error estimates to assess the test uncertainty. Figure 1 plots additional results with **error bars** (conventional 95% confidence intervals) for `ALL` ACC on fine-grained datasets for GCD, adapted RIM and TIM, as well as our approach PIM which are all applied on top of the same fixed feature extractor. We used the arbitrary five different seed values $\{1, 2, 3, 4, 5\}$ to initialize models weights. One may observe that there are no error bar overlaps between PIM and the other methods, which suggests statistically significant differences with the state-of-the-art.

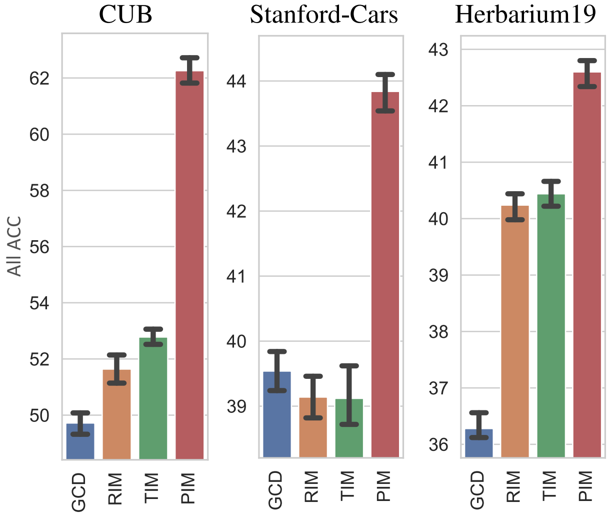


Figure 1. **Error bars** (conventional 95% confidence intervals) for All ACC on fine-grained datasets using the arbitrary five different seed values $\{1, 2, 3, 4, 5\}$.

6.3. Ablation studies

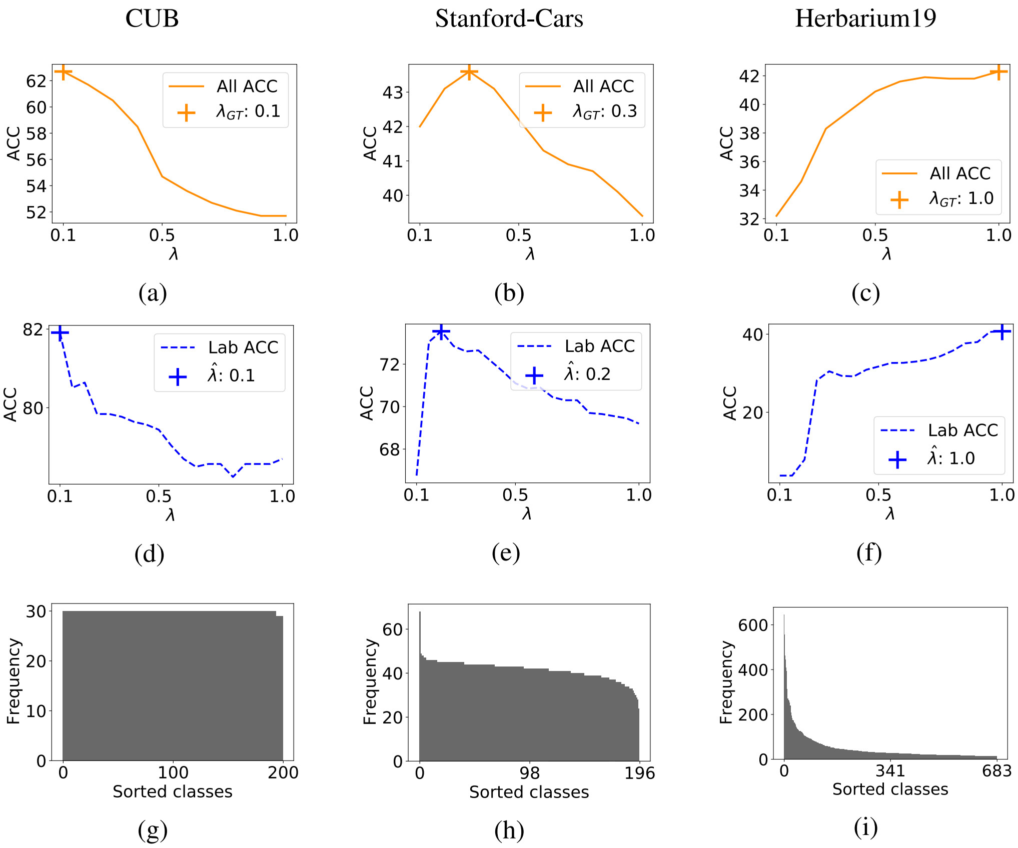
Along the following ablation studies, we focus our attention on the challenging fine-grained datasets, as these provide good illustrations of the interest of each of our technical choices.

6.3.1 Automatic finding of optimal λ: Handling both short-tailed and long-tailed datasets

We now motivate the interest of estimating the most appropriate λ value for each dataset by using the proposed automatic finding strategy presented in Sec. 5. λ_{GT} in Figures (a), (b), (c) is the λ value obtained when using the ground-truth labels, i.e. the value that provides the maximum performance on the unlabeled set, and $\hat{\lambda}$ in Figures (d), (e) and (f) is the estimated optimal value. Interestingly, Figures 2 (a), (b) and (c) show how the performance `ALL` ACC could vary significantly depending on the selected λ, hence motivating the interest of finding an optimal λ value for each dataset. Figures 2 (d), (e) and (f) validate our hypothesis: Selecting the λ value that maximizes the Lab ACC on the labeled data in our unconstrained conditional version also maximizes the ACC on all the unlabeled data (`ALL` ACC) when using the constrained version instead, showing the correlation between both. In regard to the classes frequencies observed on each dataset (See Figures 2 (g), (h) and (i)), it is also interesting to note that a small λ value provides better results on a dataset with uniform class distributions such as CUB, whereas a higher λ value is more appropriate on the imbalanced dataset Herbarium19. Indeed, λ controls the relative effect of the marginal entropy term in (7). Thus, these results show that automatically selecting λ can mitigate the class-balance bias encoded in the standard mutual information by giving more importance to the conditional entropy term in long-tailed (imbalanced) scenarios.

6.3.2 Effect of each loss term

We now evaluate the contribution of each term in our learning objective (7). In particular, Tab. 3 highlights the effect of the conditional entropy, the marginal entropy and the constraint penalty (i.e. replacing conditional entropy with CE on the labeled points). From these results, we can draw three different observations: 1) Minimizing only the conditional entropy term yields degenerated solutions, as expected. 2) Maximizing the marginal entropy term and enforcing the proposed constraint prevent these undesired degenerated solutions. 3) Maximizing the marginal entropy term on the entire dataset, and hence maximizing as well the mutual information on the entire dataset, as we propose, further enhances the performance.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2. λ **effect analysis on fine-grained datasets.** The first row represents the ACC on all the unlabeled points depending on λ value. The second row represents the ACC on the labeled points depending on λ value. The third row represents the frequency of examples per class, in a sorted order.

			CUB			Stanford Cars			Herbarium19		
Loss terms used	$\mathcal{H}(Y)$ on	s.t. $\boldsymbol{y}_i = \boldsymbol{p}_i \; \forall \boldsymbol{z}_i \in \mathcal{Z}_L$	All	Old	New	All	Old	New	All	Old	New
$-\mathcal{H}(Y\|Z)$	not used	✗	6.1	0.0	9.1	2.5	0.0	3.6	4.0	4.0	4.1
$-\mathcal{H}(Y\|Z)$	not used	✓	38.6	46.2	34.8	29.8	51.3	19.5	34.6	45.4	28.9
$\mathcal{H}(Y) - \mathcal{H}(Y\|Z)$	$\mathcal{Z}_U$	✗	53.7	55.5	52.9	37.4	49.5	31.6	35.2	39.0	33.2
$\mathcal{H}(Y) - \mathcal{H}(Y\|Z)$	$\mathcal{Z} = \mathcal{Z}_L \cup \mathcal{Z}_U$	✗	56.6	66.4	51.7	40.8	60.2	31.5	36.1	41.0	33.4
$\mathcal{H}(Y) - \mathcal{H}(Y\|Z)$	$\mathcal{Z}_U$	✓	58.3	72.8	51.1	41.4	66.6	29.2	40.1	**57.3**	30.8
$\mathcal{H}(Y) - \mathcal{H}(Y\|Z)$	$\mathcal{Z} = \mathcal{Z}_L \cup \mathcal{Z}_U$	✓	**62.7**	**75.7**	**56.2**	**43.1**	**66.9**	**31.6**	**42.3**	56.1	**34.8**

Table 3. **Effect of the loss terms and the constraint** on the predictive performances (ACC) of PIM on fine-grained datasets.

6.4. Towards a practical setting

6.4.1 Estimating the number of classes

In order to find the number of classes, we follow the strategy proposed in GCD [42], which we refer to as Max-ACC (GCD), but we replace the K-means clustering stage with the unconstrained (i.e. unsupervised) version of our method PIM, referred to as Max-ACC (PIM). The results from these methods are reported in Table 4. Overall, the proposed combination Max-ACC (PIM) is more appropriate than Max-ACC (GCD) [42] on both generic and fine-grained datasets, except on CIFAR-100 where Max-ACC (GCD) [42] finds the real number of classes.

6.4.2 Performance when the number of classes is unknown

While we followed the standard practices for the partitioning task in the experiments of Section 6.2, we argue that having access to the number of expected classes is an unre-

	CUB	Stanford Cars	Herbarium19	mean
	$\hat{K}(Err)$	$\hat{K}(Err)$	$\hat{K}(Err)$	(Err)
Ground truth	200(-)	196(-)	683(-)	(-)
Max-ACC (GCD) [42]	231 (16%)	230 (15%)	520 (24%)	(18%)
Max-ACC (PIM)	**227 (14%)**	**169 (13%)**	**563 (18%)**	**(15%)**

	CIFAR10	CIFAR100	ImageNet-100	mean
	$\hat{K}(Err)$	$\hat{K}(Err)$	$\hat{K}(Err)$	(Err)
Ground truth	10 (-)	100 (-)	100 (-)	(-)
Max-ACC (GCD) [42]	9 (10%)	**100 (0%)**	109 (9%)	(6%)
Max-ACC (PIM)	**10 (0%)**	95 (5%)	**102 (2%)**	**(2%)**

Table 4. **Estimation of the number of classes** in the unlabeled set using Brent's algorithm as in [42]. Max-ACC (GCD) [42] results are reported from [42].

alistic assumption. Thus, we now relax this assumption by repeating the partitioning experiments with the estimated value of $\hat{K}$ (See Tab. 4) instead of the real value K. For the GCD method [42], we used the code provided by the authors[6], except on CIFAR-100, for which we directly report the scores from Tab. 1 because $\hat{K} = K$. These results, which are reported on Tab. 5, demonstrate the superiority of our method even in this more challenging scenario. This suggests that our formulation serves as a more robust solution in the absence of prior knowledge about K for the GCD task.

	CUB			Stanford Cars			Herbarium19		
	All	Old	New	All	Old	New	All	Old	New
GCD [42]	51.1	56.4	48.4	39.1	58.6	29.7	37.2	51.7	29.4
PIM	**62.0**	**75.7**	**55.1**	**42.4**	**65.3**	**31.3**	**42.0**	**55.5**	**34.7**

	CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
GCD [42]	80.5	**97.9**	71.8	70.8	77.6	57.0	77.9	91.1	71.3
PIM	**94.7**	97.4	**93.3**	**75.6**	**81.6**	**63.6**	**83.0**	**95.3**	**76.9**

Table 5. **Realistic GCD partitioning.** ACC scores are obtained by assuming $\hat{K}$ (See Tab. 4) is the number of expected classes.

7. Conclusion

In this work, we proposed a simple yet effective alternative for Generalized Category Discovery. In particular, we introduce a parametric family of mutual information objectives, which we tackle with a bi-level optimization formulation. Our solution allows to estimate the relative weight of the marginal-entropy term automatically, which mitigates the class-balance bias inherent in standard information maximization. Our empirical validation demonstrates that by learning the optimal weight that controls the relative effect of the marginal-entropy, our model deals effectively with both short-tailed and long-tailed datasets. Indeed, our formulation achieves new state-of-the-art results in GCD tasks, outperforming existing solutions across the different benchmarks by a significant margin. It is worth noting that our formulation is flexible as it could be coupled with any trained feature extractor. Thus, we hope that the proposed framework will be useful for future research and development to solve the GCD problem for real-world applications.

Limitations. A common limitation in the current GCD learning paradigm stems from the fact that the models require access to the entire target unlabeled dataset at test time. Needless to say, this strong assumption might hinder the use of these approaches when the target set is composed of a small number of images, or when samples appear sequentially. As in [42], we assume that the optimal values for λ and K are obtained when the ACC is maximized on the available labeled points. Despite that our extensive experiments empirically confirm that this assumption is promising in practice, it could be interesting to find a metric that could simultaneously consider the novel classes.

References

[1] Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Neural Information Processing Systems (NeurIPS)*, 2019.

[3] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82:102617, 2022.

[4] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Neural Information Processing Systems (NeurIPS)*, 2019.

[6] Malik Boudiaf, Hoel Kervadec, Imtiaz Masud Ziko, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European Conference on Computer Vision (ECCV)*, 2020.

[8] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Transductive informa-

[6] `https://github.com/sgvaze/generalized-category-discovery`

tion maximization for few-shot learning. *Neural Information Processing Systems (NeurIPS)*, 2020.

[9] Yuri Boykov, Hossam N. Isack, Carl Olsson, and Ismail Ben Ayed. Volumetric bias in segmentation and reconstruction: Secrets and solutions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[10] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

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

[14] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[17] Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2022.

[18] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference for Learning Representations (ICLR)*, 2019.

[21] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International Conference on Machine Learning (ICML)*, 2017.

[22] Mohammed Jabi, Marco Pedersoli, Amar Mitiche, and Ismail Ben Ayed. Deep clustering: On the link between discriminative models and k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1887–1896, 2021.

[23] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[24] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 1997.

[25] Mete Kemertas, Leila Pishdad, Konstantinos G Derpanis, and Afsaneh Fazly. Rankmi: A mutual information maximizing ranking loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[28] Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. In *Neural Information Processing Systems (NeurIPS)*, 2010.

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2013.

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

[31] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, 2020.

[32] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[33] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

[34] David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano L. I. Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection. In *International Joint Conference on Neural Networks (IJCNN)*, 2021.

[35] Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. Exploring category-agnostic clusters for open-set domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[37] 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 *Neural Information Processing Systems (NeurIPS)*, 2020.

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

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[40] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[41] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations (ICLR)*, 2021.

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.

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

[45] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[46] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[47] Chuyu Zhang, Chuanyang Hu, Ruijie Xu, Zhitong Gao, Qian He, and Xuming He. Mutual information-guided knowledge transfer for novel class discovery. *arXiv preprint arXiv:2206.12063*, 2022.

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

[49] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[51] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.