This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Generalized Category Discovery

Sagar Vaze* Kai Han† Andrea Vedaldi* Andrew Zisserman*
*Visual Geometry Group, Department of Engineering Science, University of Oxford
†The University of Hong Kong
{sagar,vedaldi,az}@robots.ox.ac.uk kaihanx@hku.hk

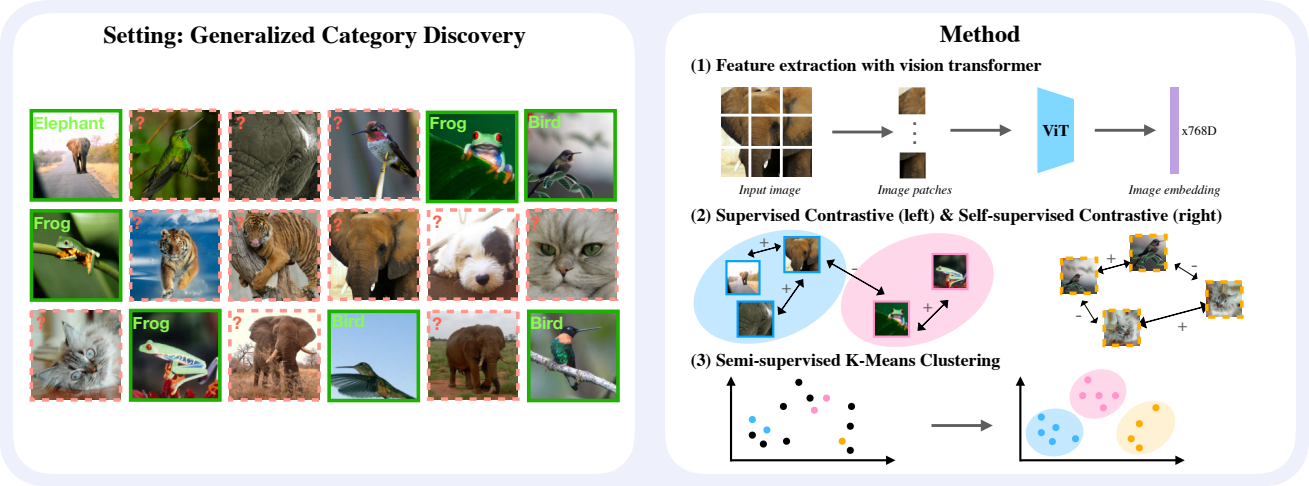


Figure 1. We present a new setting: 'Generalized Category Discovery' and a method to tackle it. Our setting can be succinctly described as: given a dataset, a subset of which has class labels, categorize all unlabelled images in the dataset. The unlabelled images may come from labelled or novel classes. Our method leverages contrastively trained vision transformers to assign labels directly through clustering.

Abstract

In this paper, we consider a highly general image recognition setting wherein, given a labelled and unlabelled set of images, the task is to categorize all images in the unlabelled set. Here, the unlabelled images may come from labelled classes or from novel ones. Existing recognition methods are not able to deal with this setting, because they make several restrictive assumptions, such as the unlabelled instances only coming from known – or unknown – classes, and the number of unknown classes being known a-priori. We address the more unconstrained setting, naming it 'Generalized Category Discovery', and challenge all these assumptions. We first establish strong baselines by taking state-of-the-art algorithms from novel category discovery and adapting them for this task. Next, we propose the use of vision transformers with contrastive representation learning for this open-world setting. We then introduce a simple yet effective semi-supervised k-means method to cluster the unlabelled data into seen and unseen classes automatically, substantially outperforming the baselines. Finally, we also propose a new approach to estimate the number of classes in the unlabelled data. We thoroughly evaluate our approach on public datasets for generic object classification and on fine-grained datasets, leveraging the recent Semantic Shift Benchmark suite. Code: https://www.robots.ox.ac.uk/~vgg/research/gcd

1. Introduction

Consider an infant sitting in a car and observing the world. Object instances will pass the car and, for some of these, the infant may have been told their category ('that is a dog', 'that is a car') and be able to recognize them. There will also be instances that the infant has not seen before (cats and bicycles) and, having seen a number of these instances, we might expect the infant's visual recognition system to cluster these into new categories.

This is the problem that we consider in this work: given an image dataset where only some images are labelled with their categories, assign a category label to each of the rest

images, possibly using new categories not observed in the labelled set. We term this problem *Generalized Category Discovery* (GCD), and suggest that this is a realistic use case for many machine vision applications: whether that is recognizing products in a supermarket; pathologies in medical images; or vehicles in autonomous driving. In these and other realistic vision settings, it is often impossible to know if new images come from labelled or novel categories.

In contrast, consider the limitations of existing image recognition settings. In image classification, one of the most widely studied problems, all of the training images come with class labels. Furthermore, all images at test time come from the same classes as the training set. Semi-supervised learning (SSL) [7] introduces the problem of learning from unlabelled data, but still assumes that all unlabelled images come from the same set of classes as the labelled ones. More recently, the tasks of open-set recognition (OSR) [38] and novel-category discovery (NCD) [19] have tackled open-world settings in which the images at test time may belong to new classes. However, OSR aims only to *detect* test-time images which do not belong to one of the classes in the labelled set, but does not require any further classification amongst these detected images. Meanwhile, in NCD, the closest setting to the one tackled in this work, methods learn from labelled and unlabelled images, and aim to discover new classes in the unlabelled set. However, NCD still makes the limiting assumption that *all* of the unlabelled images come from new categories, which is usually unrealistic.

In this paper, we tackle Generalized Category Discovery in a number of ways. Firstly, we establish strong baselines by taking representative methods from NCD and applying them to this task. To do this, we adapt their training and inference mechanisms to account for our more general setting, as well as retrain them with a more robust backbone architecture. We show that existing NCD methods are prone to overfit the labelled classes in this generalized setting.

Next, observing the potential for NCD methods to overfit their classification heads to the labelled classes, we propose a simple but effective method for recognition by clustering. Our key insight is to leverage the strong 'nearest neighbour' classification property of vision transformers along with contrastive learning. We propose the use of contrastive training and a semi-supervised k-means clustering algorithm to recognize images without a parametric classifier. We show that these proposed methods substantially outperform the established baselines, both on generic object recognition datasets and, particularly, on more challenging fine-grained benchmarks. For the latter evaluations, we leverage the recently proposed Semantic Shift Benchmark suite [45], which was designed for the task of identifying semantic novelty.

Finally, we propose a solution to a challenging and under-investigated problem in image recognition: estimating the number of categories in unlabelled data. Almost all methods, including purely unsupervised ones, assume the knowledge of the number of categories, a highly unrealistic assumption in the real world. We propose an algorithm which leverages the labelled set to tackle this problem.

Our contributions can be summarized as follows: (i) the formalization of Generalized Category Discovery (GCD), a new and realistic setting for image recognition; (ii) the establishment of strong baselines by adapting state-of-the-art techniques from standard novel category discovery to this task; (iii) a simple but effective method for GCD, which uses contrastive representation learning and clustering to directly provide class labels, and outperforms the baselines substantially; (iv) a novel method for estimating the number of categories in unlabelled data, a largely understudied problem; and (v) rigorous evaluation on standard image recognition datasets as well as the recent Semantic Shift Benchmark suite [45].

2. Related work

Our work relates to prior work on *semi-supervised learning*, *open-set recognition*, and *novel category discovery*, which we briefly review next.

Semi-supervised learning A number of methods [7, 33, 37, 40, 49] have been proposed to tackle the problem of semi-supervised learning (SSL). SSL assumes that the labelled and unlabelled instances come from the same set of classes. The objective is to learn a robust classification model leveraging both the labelled and unlabelled data during training. Amongst existing methods, consistency based approaches appear to be popular and effective, such as LadderNet [36], PI model [29], Mean-teacher [43]. Recently, with the success of self-supervised learning, methods have also been proposed to improve SSL by augmenting the methods with self-supervised objectives [37, 49].

Open-set recognition The problem of open-set recognition (OSR) is formalized in [38], with the objective being to classify unlabelled instances from the same semantic classes as the labelled data, while detecting test instances from unseen classes. OpenMax [3] is the first deep learning method to approach this problem with Extreme Value Theory. GANs are often employed to generate adversarial samples to train an open-set classifier, e.g, [14, 25, 32]. Several methods have been proposed to train models such that images with large reconstruction error are regarded as open-set samples [34, 41, 48]. There are also methods that learns prototypes for the labelled classes, and identify the images from unknown classes by the distances to the prototypes [8,9,39]. More recently, [8,9] proposed to learn reciprocal points which describe 'otherness' with respect to the labelled classes. [50] jointly trains a flow-based density es-

timator and a classification based encoder for OSR. Finally, Vaze et al. [45] study the correlation between the closed-set and open-set performance, showing that state-of-the-art OSR results can be obtained by boosting the closed-set accuracy of the standard cross-entropy baseline.

Novel category discovery The problem of novel category discovery (NCD) is formalized in DTC [19]. Earlier methods that could be applied to this problem include KCL [21] and MCL [22], both of which maintain two models trained with labelled data and unlabelled data respectively, for general task transfer learning. AutoNovel (aka Rankstats) [17, 18] tackles the NCD problem with a three stage method. The model is first trained with self-supervision on all data for low-level representation learning. Then, it is further trained with full supervision on labelled data to capture higher level semantic information. Finally, a joint learning stage is carried out to transfer knowledge from the labelled to unlabelled data with ranking statistics. Zhao and Han [51] propose a model with two branches, one for global feature learning and the other for local feature learning, such that dual ranking statistics and mutual learning are conducted with these two branches for better representation learning and new class discovery. OpenMix [53] mixes the labelled and unlabelled data to avoid the model from over-fitting for NCD. NCL [52] extracts and aggregates the pairwise pseudo-labels for the unlabelled data with contrastive learning and generates hard negatives by mixing the labelled and unlabelled data in the feature space for NCD. Jia et al. [23] propose an end-to-end NCD method for single- and multi-modal data with contrastive learning and winner-takes-all hashing. A unified cross-entropy loss is introduced in UNO [13] to allow the model to be trained on labelled and unlabelled data jointly, by swapping the pseudo-labels from labelled and unlabelled classification heads.

Finally, we highlight the work by Girish et al. [15] that tackles a similar setting to GCD but for the task of *GAN attribution* instead of image recognition, as well as the concurrent work by Cao et al. [4] that tackles a similar setting for image recognition under the name *Open World Semi-Supervised Learning*. Different to our setting, they do not leverage large-scale pretraining or demonstrate performance on the Semantic Shift Benchmark, which better isolates the problem of detecting semantic novelty.

3. Generalized category discovery

We first formalize the task of *Generalized Category Discovery* (GCD). In short, we consider the problem of classifying images in a dataset, a subset of which has known class labels. The task is to assign class labels to all remaining images, using classes that may or may not be observed in the labelled images (see Fig. 1, left).

Formally, we define GCD as follows. We consider a dataset $\mathcal{D}$ comprising two parts $\mathcal{D}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{U}}$, where $\mathcal{Y}_{\mathcal{L}} \subset \mathcal{Y}_{\mathcal{U}}$. During training, the model does not have access to the labels in $\mathcal{D}_{\mathcal{U}}$, and is tasked with predicting them at test time. Furthermore, we assume access to a validation set, $\mathcal{D}_{\mathcal{V}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N'} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{L}}$, which is disjoint from the training set and contains images from the same classes as the labelled set. This formalization allows us to clearly see the distinction with the novel category discovery setting. NCD assumes $\mathcal{Y}_{\mathcal{L}} \cap \mathcal{Y}_{\mathcal{U}} = \emptyset$ and existing methods rely on this prior knowledge during training.

In this section, we describe the methods we propose to tackle GCD. First, we describe our approach to the problem. Leveraging recent progress in self-supervised representation learning, we propose a simple but effective approach based on contrastive learning, with classification performed by a semi-supervised k-means algorithm. Next, we develop a method to estimate the number of categories in the unlabelled data – a challenging task that is understudied in the literature. Finally, we build two strong baselines for GCD by modifying state-of-the-art NCD methods, RankStats [18] and UNO [13], to fit with our setting.

3.1. Our approach

The key insight of our approach for image recognition in an open-world setting is to remove the need for parametric classification heads. Instead, we perform clustering directly in the feature space of a deep network (see Fig. 1, right). Classification heads (typically, linear classifiers on top of a learned embedding) are best trained with the cross-entropy loss, which has been shown to be susceptible to noisy labels [12]. Furthermore, when training a linear classifier for unlabelled classes, a typical method is to generate (noisy) pseudo-labels for the unlabelled instances. This would suggest that parametric heads are susceptible to performance deterioration on the unlabelled classes. Finally, we note that, by necessity, classification heads must be trained from scratch, which further makes them vulnerable to overfitting on the labelled classes.

Meanwhile, self-supervised contrastive learning has been widely used as pre-training to achieve robust representations in NCD [23, 52]. Furthermore, when combined with vision transformers, it generates models which are good nearest neighbour classifiers [6]. Inspired by this, we find that constrastively training a ViT model allows us to directly cluster in the model's feature space, thereby removing the need for a linear head which could lead to overfitting. Specifically, we train the representation with a noise contrastive loss [16] on all images *without using any labels*. This is important because it avoids overfitting the features to the subset of classes that are (partially) labelled. We add a further supervised contrastive component [24] for the labelled instances to make use of the labelled data (see Fig. 1, middle row on the right).

3.1.1 Representation learning

We use, for all methods, a vision transformer (ViT-B-16) [11] pretrained with DINO [6] self-supervision on (unlabelled) ImageNet [10] as our backbone. This is motivated firstly because the DINO model is a strong nearest neighbour classifier, which suggests non-parametric clustering in its feature space would work well. Secondly, self-supervised vision transformers have demonstrated the attractive quality of learning to attend to salient parts of an object without human annotation. We find this feature to be useful for this task, because *which* object parts are important for classification is likely to transfer well from the labelled to the unlabelled categories (see Sec. 4.5).

Finally, we wish to reflect a realistic and practical setting. In the NCD literature, it is standard to train a ResNet-18 [20] backbone from scratch for the target task. However, in a real-world setting, a model is often initialized with large-scale pretrained weights to optimize performance (often ImageNet supervised pretraining). In order to avoid conflicts with our experimental setting (which assumes a finite labelled set), we use self-supervised ImageNet weights. To enhance the representation such that it is more tailored for the labelled and unlabelled data we have, we further fine-tune the representation on our target data jointly with supervised contrastive learning on the labelled data, and unsupervised contrastive learning on *all* the data.

Formally, let $\mathbf{x}_i$ and $\mathbf{x}_i'$ be two views (random augmentations) of the same image in a mini-batch B. The unsupervised contrastive loss is written as:

$$\mathcal{L}_i^u = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_i'/\tau\right)}{\sum_n \mathbb{1}_{[n\neq i]} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_n/\tau\right)}, \tag{1}$$

where $\mathbf{z}_i = \phi(f(\mathbf{x}_i))$ and $\mathbb{1}_{[n\neq i]}$ is an indicator function evaluating to 1 *iff* $n \neq i$, and τ is a temperature value. f is the feature backbone, and ϕ is a multi-layer perceptron (MLP) projection head.

The supervised contrastive loss is written as:

$$\mathcal{L}_i^s = -\frac{1}{|\mathcal{N}(i)|} \sum_{q\in\mathcal{N}(i)} \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_q/\tau\right)}{\sum_n \mathbb{1}_{[n\neq i]} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_n/\tau\right)}, \tag{2}$$

where $\mathcal{N}(i)$ denotes the indices of other images having the same label as $\mathbf{x}_i$ in the mini-batch B. Finally, we construct the total loss over the batch as:

$$\mathcal{L}^t = (1-\lambda)\sum_{i\in B} \mathcal{L}_i^u + \lambda \sum_{i\in B_\mathcal{L}} \mathcal{L}_i^s \tag{3}$$

where $B_\mathcal{L}$ corresponds to the labelled subset of B and λ is a weight coefficient. Using the labels *only* in a contrastive framework, rather than in a cross-entropy loss, means that unlabelled and labelled data are treated similarly. The supervised contrastive component is only used to nudge the network towards a semantically meaningful representation, thereby minimizing overfitting on the labelled classes.

3.1.2 Label assignment with semi-supervised k-means

Given the learned representation for the data, we can now assign class or cluster labels for each unlabelled data point, either from the labelled classes or unseen new classes. Instead of performing this parametrically as is common in NCD (and risk overfitting to the labelled data) we propose to use a non-parametric method. Namely, we propose to modify the classic k-means into a constraint algorithm by forcing the assignment of the instances in $\mathcal{D}_\mathcal{L}$ to the correct cluster based on their ground-truth labels. Note, here we assume knowledge of the number of clusters, k. We tackle the problem of estimating this parameter in Sec. 3.2. The initial $|\mathcal{Y}_\mathcal{L}|$ centroids for $\mathcal{D}_\mathcal{L}$ are obtained based on the ground-truth class labels, and an additional $|\mathcal{Y}_\mathcal{U} \setminus \mathcal{Y}_\mathcal{L}|$ (number of new classes) initial centroids are obtained from $\mathcal{D}_\mathcal{U}$ with k-means++ [1], constrained on the centroids of $\mathcal{D}_\mathcal{L}$. During each centroid update and cluster assignment cycle, instances from the same class in $\mathcal{D}_\mathcal{L}$ are always forced to have the same cluster assignment, while each instance in $\mathcal{D}_\mathcal{U}$ can be assigned to any cluster based on the distance to different centroids. After the semi-supervised k-means converges, each instance in $\mathcal{D}_\mathcal{U}$ can be assigned a cluster label. We provide a clear diagram of this in Appendix B.

3.2. Estimating the class number in unlabelled data

Here, we tackle the problem of finding the number of classes in the unlabelled data. In the NCD and unsupervised clustering settings, prior knowledge of the number of categories in the dataset is often assumed, but this is unrealistic in the real world given that the labels themselves are unknown. To estimate the number of categories in $\mathcal{D}_\mathcal{U}$, we leverage the information available in $\mathcal{D}_\mathcal{L}$. Specifically, we perform k-means clustering on the entire dataset, $\mathcal{D}$, before evaluating clustering accuracy on *only* the labelled subset (see Sec. 4.1 for the metric's definition).

Clustering accuracy is evaluated by running the Hungarian algorithm [28] to find the optimal assignment between the set of cluster indices and ground truth labels. If the number of clusters is higher than the total number of classes, the extra clusters are assigned to the null set, and all instances assigned to those clusters are said to have been predicted incorrectly. Conversely, if the number of clusters is lower than the number of classes, extra classes are assigned to the null set, and all instances with those ground truth labels are said to have been to be predicted incorrectly. Thus, we assume that, if the clustering (across $\mathcal{D}$) is performed with k too high or too low, then this will be reflected in a sub-optimal clustering accuracy on $\mathcal{D}_\mathcal{L}$. In other words, we assume clustering accuracy on the labelled set will be maximized when $k = |\mathcal{Y}_\mathcal{L} \cup \mathcal{Y}_\mathcal{U}|$. This intuition leads us to use the clustering accuracy as a 'black box' scoring function, $ACC = f(k; \mathcal{D})$, which we optimize with Brent's algorithm to find the optimal k. Different to the method in [18],

which exhaustively iterates through all possible values of k, we find that the black-box optimization allows our method to scale to datasets with many categories.

Finally, we highlight that labelled sets with different granularities would induce different estimates of the number of classes. However, we suggest that the labelled set defines the system of categorization — that the granularity of a real-world dataset is not an intrinsic property of the images, but rather a framework imposed by the labels. For instance, in Stanford Cars, the dataset could equally be labelled at the 'Manufacturer', 'Model' or 'Variant' level, with the categorization system defined by the labels assigned.

3.3. Two strong baselines

We adapt two methods from the nearest image recognition sub-field, novel category discovery (NCD), for our generalized category discovery (GCD) task. *RankStats* [18] is widely used as a competitive baseline for novel category discovery, while *UNO* [13] is to the best of our knowledge the state-of-the-art method for NCD.

Baseline: RankStats+ RankStats trains two classifiers on top of a shared feature representation: the first head is fed instances from the labelled set and is trained with the cross-entropy loss, while the second head sees only instances from unlabelled classes (again, in the NCD setting, the labelled and unlabelled classes are disjoint). In order to adapt RankStats to GCD, we train it with a single classification head for the total number of classes in the dataset. We then train the first $|\mathcal{Y}_{\mathcal{L}}|$ elements of the head with the cross-entropy loss, and train the *entire* head with the binary cross-entropy loss with pseudo-labels.

Baseline: UNO+ Similarly to RankStats, UNO is trained with classification heads for labelled and unlabelled data. The model is then trained in a SwAV-like manner [5]. First, multiple views (random augmentations) of a batch are generated and fed to the same model. For the labelled images in the batch, the labelled head is trained with the cross-entropy loss using the ground truth labels. For the unlabelled images, predictions (logits from the unlabelled head) are gathered for a given view and used as pseudo-labels with which to optimize the loss from *other* views. To adapt this mechanism, we simply concatenate both the labelled and unlabelled heads, thus allowing generated pseudo-labels for the unlabelled samples to belong to any class in the dataset.

4. Experiments

4.1. Experimental setup

Data We demonstrate results on six datasets in our proposed setting. For each dataset, we take the training set and sample a set of classes for which we have labels during training. We further sub-sample 50% of the images from these classes to constitute the labelled set $\mathcal{D}_{\mathcal{L}}$. The remaining instances from these classes, along with all instances from the other classes, constitute $\mathcal{D}_{\mathcal{U}}$. We further construct the validation set for the labelled classes from the test or validation split of each dataset.

Table 1. Datasets used in our experiments. We show the number of classes in the labelled and unlabelled sets ($|\mathcal{Y}_{\mathcal{L}}|, |\mathcal{Y}_{\mathcal{U}}|$), as well as the number of images ($|\mathcal{D}_{\mathcal{L}}|, |\mathcal{D}_{\mathcal{U}}|$).

	CIFAR10	CIFAR100	ImageNet-100	CUB	SCars	Herb19
$\lvert\mathcal{Y}_{\mathcal{L}}\rvert$	5	80	50	100	98	341
$\lvert\mathcal{Y}_{\mathcal{U}}\rvert$	10	100	100	200	196	683
$\lvert\mathcal{D}_{\mathcal{L}}\rvert$	12.5k	20k	31.9k	1.5k	2.0k	8.9k
$\lvert\mathcal{D}_{\mathcal{U}}\rvert$	37.5k	30k	95.3k	4.5k	6.1k	25.4k

We first demonstrate results on three generic object recognition datasets: CIFAR10 [27], CIFAR100 [27] and ImageNet-100 [10]. ImageNet-100 refers to the ImageNet dataset with 100 classes randomly subsampled. These datasets establish the methods' performance on well-known datasets in the standard image recognition literature.

We further evaluate on the recently proposed Semantic Shift Benchmark [45] (SSB, including CUB [46] and Stanford Cars [26]), as well as on Herbarium19 [42]. SSB provides fine-grained evaluation datasets with clear 'axes of semantic variation' and further provides categories for $\mathcal{D}_{\mathcal{U}}$ which are delineated from $\mathcal{D}_{\mathcal{L}}$ in a semantically coherent fashion. Thus, the user can be confident that the recognition system is identifying new classes based on a true semantic signal, rather than simply responding to low-level distributional shifts in the data, as may be the case for generic object recognition datasets. The long-tailed nature of Herbarium19 adds an additional challenge to the evaluation.

The fine-grained datasets further reflect many real-world use cases for image recognition systems, which are deployed in constrained environments with many similar objects (*e.g.* products in a supermarket, traffic monitoring, or animal tracking in the wild). In fact, the Herbarium19 dataset itself represents a real-world use case for GCD: while we are aware of roughly 400k species of plants, and estimate that there are around 80k yet to be discovered, it currently takes roughly 35 years from plant collection to plant species description if performed manually [42]. We summarize the dataset splits used in our evaluations in Tab. 1, and provide more details in Appendix A.

Evaluation protocol For each dataset, we train the models on $\mathcal{D}$ (without access to the ground truth labels in $\mathcal{D}_{\mathcal{U}}$). At test-time, we measure the clustering accuracy between the ground truth labels y_i and the model's predictions $\hat{y}_i$ as:

$$ACC = \max_{p \in \mathcal{P}(\mathcal{Y}_{\mathcal{U}})} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{y_i = p(\hat{y}_i)\} \tag{4}$$

Here, $M = |\mathcal{D}_{\mathcal{U}}|$ and $\mathcal{P}(\mathcal{Y}_{\mathcal{U}})$ is the set of all permutations of the class labels in the unlabelled set. Our main metric is ACC on 'All' instances, indicating image recognition accu-

racy across the entire unlabelled set $\mathcal{D}_{\mathcal{U}}$. We further report values for both the 'Old' classes subset (instances in $\mathcal{D}_{\mathcal{U}}$ belonging to classes in $\mathcal{Y}_{\mathcal{L}}$) and 'New' classes subset (instances in $\mathcal{D}_{\mathcal{U}}$ belonging to classes in $\mathcal{Y}_{\mathcal{U}} \setminus \mathcal{Y}_{\mathcal{L}}$).

The maximum over the set of permutations is computed via the Hungarian optimal assignment algorithm [28]. Importantly, we compute the Hungarian assignment only once, across all categories $\mathcal{Y}_{\mathcal{U}}$, and measure classification accuracy on 'Old' and 'New' subsets only afterwards. The interaction between when the Hungarian assignment is performed, and the resultant ACC on the subsets, can be unintuitive and is elaborated upon in Appendix E.

Implementation details All methods are trained with a ViT-B-16 backbone with DINO pre-trained weights, and use the output `[CLS]` token as the feature representation. All methods were trained for 200 epochs, with the best model selected using accuracy on the validation set. We fine-tune the final transformer block for all methods.

For our method, we fine-tune the final block of the vision transformer with an initial learning rate of 0.1 which we decay with a cosine annealed schedule. We use a batch size of 128 and $\lambda = 0.35$ in the loss (see Eq. (3)). Furthermore, following standard practise in self-supervised learning, we project the model's output through a non-linear projection head before applying the contrastive loss. We use the same projection head as in [6] and discard it at test-time. For the baselines from NCD, we follow the original implementations and learning schedules as far as possible, referring to the original papers for details [13, 18].

Finally, in order to estimate k, we run our k-estimation method on DINO features extracted from each of the considered benchmarks. We run Brent's algorithm on a constrained domain for k, with the minimum set at $|\mathcal{Y}_{\mathcal{L}}|$ and the maximum set at 1000 classes for all datasets.

4.2. Comparison with the baselines

We report results for all compared methods in Tab. 2 and Tab. 3. As an additional baseline, we also report results when running k-means directly on top of raw DINO features (reported as k-means). Tab. 2 presents results on the generic object recognition datasets, while Tab. 3 shows results on SSB and Herbarium19. We further show results on the FGVC-Aircraft [31] evaluation from SSB in Appendix D.

Overall (across 'All' instances in $\mathcal{D}_{\mathcal{U}}$), our method outperforms RankStats+ and UNO+ baselines on the standard image recognition datasets by 9.3% in absolute terms, and 11.5% in proportional terms. Meanwhile, on the more challenging fine-grained evaluations, our method outperforms the baselines by 8.9% in absolute terms and 27.0% in proportional terms.

We find that on categories with labelled examples ('Old' classes), the baselines which use parametric classifiers can outperform our method, but this is at the expense of ACC on the 'New' categories. We also found that, if the baselines were trained for longer, they would begin to sacrifice ACC on 'Old' categories for ACC on 'New' ones, but that best overall performance was achieved using early stopping by monitoring the performance on the validation set.

Table 2. Results on generic image recognition datasets.

	CIFAR10			CIFAR100			ImageNet-100		
Classes	All	Old	New	All	Old	New	All	Old	New
k-means [30]	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	**71.3**
RankStats+	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+	68.6	**98.3**	53.8	69.5	**80.6**	47.2	70.3	**95.0**	57.9
Ours	**91.5**	97.9	**88.2**	**70.8**	77.6	**57.0**	**74.1**	89.8	66.3

Table 3. Results on SSB [45] and Herbarium19 [42].

	CUB			Stanford Cars			Herbarium19		
Classes	All	Old	New	All	Old	New	All	Old	New
k-means [30]	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+	33.3	51.6	24.2	28.3	61.8	12.1	27.9	**55.8**	12.8
UNO+	35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	53.7	14.7
Ours	**51.3**	**56.6**	**48.7**	**39.0**	57.6	**29.9**	**35.4**	51.0	**27.0**

4.3. Estimating the number of classes

We report results on estimating the number of classes in Tab. 4. We find that on the generic object recognition datasets, we can come very close to the ground truth number of categories in the unlabelled set, with a maximum error of 10%. On the fine-grained datasets, we report an average discrepancy of 18.9%. We note the highly challenging nature of these datasets, with many constituent classes which are visually similar.

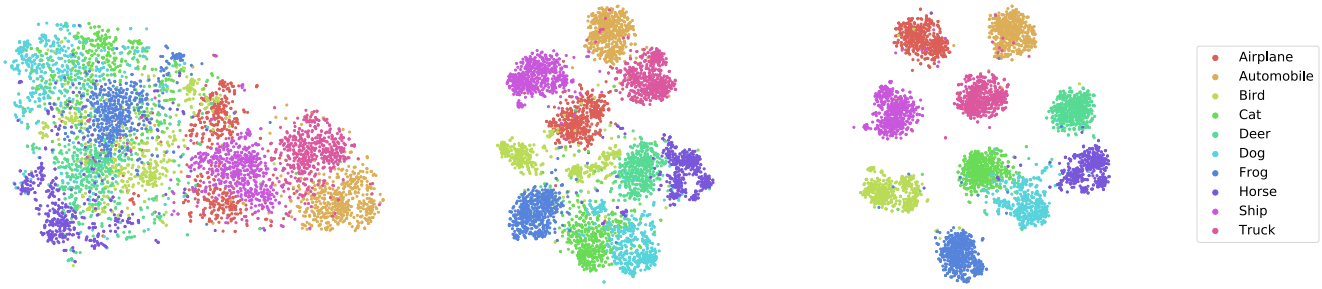
4.4. Ablation study

In Tab. 5, we inspect the contributions of the various elements of our proposed approach. Specifically, we identify the importance of the following components of the method: ViT backbone; contrastive fine-tuning (regular and supervised); and semi-supervised k-means clustering.

ViT Backbone Rows (1) and (2) show the effect of the ViT model for the clustering task, as (1) and (2) represent a ResNet-50 model and ViT-B-16 trained with DINO respectively. The ResNet model performs nearly 20% worse aggregated over 'Old' and 'New' classes. To disambiguate this from the general capacity of the architecture, note that the ImageNet linear probe discrepancy (the standard evaluation protocol for self-supervised models) is roughly 3% [6]. Meanwhile, the discrepancy in their k-NN accuracies on

Table 4. Estimation of the number of classes in unlabelled data.

	CIFAR10	CIFAR100	ImageNet-100	CUB	SCars	Herb19
Ground truth	10	100	100	200	196	683
Ours	9	100	109	231	230	520
Error	10%	0%	9%	16%	15%	28%



ResNet50 (DINO) **ViT (DINO)** **ViT (Ours)**

Figure 2. TSNE visualization of instances in CIFAR10 for features generated by a ResNet-50 and ViT model trained with DINO self-supervision on ImageNet, and a ViT model after fine-tuning with our approach.

Table 5. Ablation study on the different components of our approach.

	ViT Backbone	Contrastive Loss	Sup. Contrastive Loss	Semi-Sup k-means	CIFAR100			Herbarium19		
					All	Old	New	All	Old	New
(1)	✗	✗	✗	✗	34.0	34.8	32.4	12.1	12.5	11.9
(2)	✓	✗	✗	✗	52.0	52.2	50.8	12.9	12.9	12.8
(3)	✓	✓	✗	✗	54.6	54.1	53.7	14.3	15.1	13.9
(4)	✓	✗	✓	✗	60.5	72.2	35.0	17.8	22.7	15.4
(5)	✓	✓	✓	✗	71.1	**78.3**	56.6	28.7	32.1	26.9
(6)	✓	✓	✓	✓	**73.0**	76.2	**66.5**	**35.4**	**51.0**	**27.0**

ImageNet is roughly 9% [6], suggesting why the ViT model performs so much better for the clustering task.

Contrastive fine-tuning Rows (2)-(5) show the effects of introducing different combinations of contrastive fine-tuning on the target dataset. We find that including any of the contrastive methods alone gives relatively marginal improvements over using the raw DINO features. We find that the full benefit is only realized when combining the self-supervised and supervised contrastive losses on the target dataset. Specifically, the combination of the contrastive losses allows us to boost aggregated clustering accuracy by a further 19% on CIFAR100 and by 16% on Herbarium19 (more than doubling the ACC in this case).

Semi-supervised k-means Further performance gains can be realized with semi-supervised clustering. Across 'All' classes, we observe a 2% and 7% increase in ACC on CIFAR100 and Herbarium19 respectively. On Herbarium19, ACC on the 'Old' classes is improved by 19%. Interestingly, it appears that semi-supervised k-means slightly hurts performance on the 'Old' classes on CIFAR100. We suggest that this is an artefact of the Hungarian algorithm, which opts to assign some 'clean' clusters to the 'New' ground truth categories, to maximize overall ACC. This can be observed in the 10% boost provided by the semi-supervised method on the 'New' classes in CIFAR100. Furthermore, we found that if we performed the Hungarian algorithm on 'Old' and 'New' instances independently (allowing the reuse of clean clusters during evaluation), semi-supervised k-means improved ACC on all data subsets. We refer to Appendix E for more details on the interaction between the reported ACC and the Hungarian assignment.

Summary Overall, we find that none of the components of our method are individually sufficient to achieve good performance across our benchmark datasets. Specifically, the *combination* of a vision transformer backbone and contrastive fine-tuning facilitates strong k-means clustering directly in the feature space of the model. The semi-supervised k-means algorithm further allows us to guide the clustering process with labels and achieve better ACC, especially on the 'New' classes in the fine-grained datasets.

We further illustrate this point in the TSNE visualizations in Fig. 2, performed on the CIFAR10 dataset. We show TSNE projections of raw ResNet-50 and ViT DINO features, as well as those of our model. For the ResNet-50 features, points from the same class are generally projected close to each other, indicating that they are likely to be separable given a simple transformation (*e.g.* a linear probe). However, they do not form clear clusters, hinting at these features' poor down-stream clustering performance. In contrast, the ViT features form far clearer clusters, which are further distinguished when trained with our approach.

4.5. Qualitative results

Finally, we visualize the attention mechanism of our model to better understand its performance. Specifically, in Fig. 3 we look at how the final multi-head attention layer attends to different spatial locations when supporting the output `[CLS]` token (which we use as our feature representation). We show this both for the pre-trained DINO model and after training with our method. We visualize the atten-

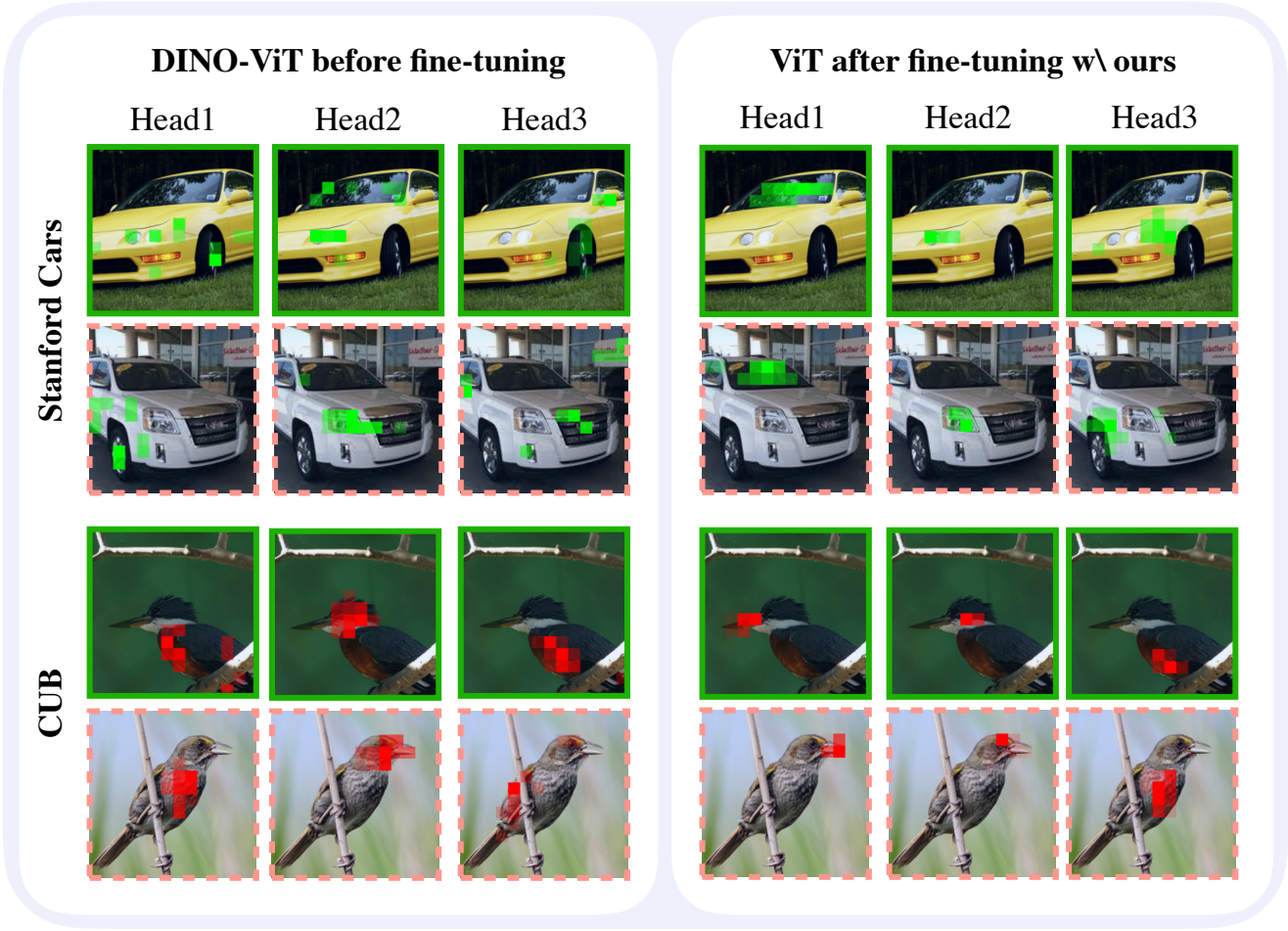


Figure 3. Attention visualizations for the DINO-ViT model before (left) and after (right) fine-tuning with our approach. For Stanford Cars and CUB, we show an image from the 'Old' (first row for each dataset) and 'New' classes (second row for each dataset). Our model learns to specialize attention heads (shown as columns) to different semantically meaningful parts, which can transfer between the labelled and unlabelled categories. The model's heads learn 'Windshield', 'Headlight' and 'Wheelhouse' for the cars, and 'Beak', 'Head' and 'Belly' for the birds. For both models, we select heads with as focused attention as possible. Recommended viewing in color with zoom.

tion maps for images from the 'Old' and 'New' classes for Stanford Cars and CUB.

It is demonstrated in [6] that different attention heads in the DINO model focus on different regions of an image, without the need for human annotation. We find this to be the case, with different heads attending to disjoint regions of the image and typically focusing on important parts. However, after training with our method, we find heads to be more specialized to semantic parts, displaying more concentrated and local attention. In this way, we suggest the model learns to attend to a set of parts which are transferable between the 'Old' and 'New' classes, which allows it to better generalize knowledge from the labelled data.

5. Conclusion

In this paper, we have proposed a new setting for image recognition, 'Generalized Category Discovery' (GCD). We highlight three take-home messages from this work: first, GCD is a challenging and realistic setting for image recognition; second, GCD removes limiting assumptions in existing image recognition sub-fields such as novel category discovery and open-set recognition; and third, while parametric classifiers tend to overfit to labelled classes in the generalized setting, direct clustering of features from contrastively trained ViTs proves to be a surprisingly good method for classification.

Acknowledgements We would like to thank Liliane Momeni for invaluable help with figures in this work. This research is funded by a Facebook AI Research Scholarship, a Royal Society Research Professorship RP\R1\191132, and the EPSRC Programme Grant VisualAI EP/T028572/1.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM symposium on Discrete algorithms*, 2007. 4

[2] Yuki M. Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. PASS: An ImageNet replacement for self-supervised pretraining without human. *NeurIPS Track on Datasets and Benchmarks*, 2021. 3

[3] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *CVPR*, 2016. 2

[4] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022. 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 5

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 4, 6, 7, 8

[7] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006. 2

[8] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021. 2

[9] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, 2020. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jua Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4, 5

[11] 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 *ICLR*, 2021. 4

[12] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *IJCAI*, 2020. 3

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

[14] Zongyuan Ge, Sergey Demyanov, and Rahil Garnavi. Generative openmax for multi-class open set classification. In *BMVC*, 2017. 2

[15] Sharath Girish, Saksham Suri, Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *ICCV*, 2021. 3

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010. 3

[17] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020. 3

[18] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE TPAMI*, 2021. 3, 4, 5, 6, 2

[19] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019. 2, 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[21] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018. 3

[22] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019. 3

[23] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single- and multi-modal data. In *ICCV*, 2021. 3

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3

[25] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. *ICCV*, 2021. 2

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013. 5

[27] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report*, 2009. 5

[28] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 4, 6, 2

[29] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2

[30] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967. 6, 2

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6, 1, 2

[32] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018. 2

[33] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 2

[34] Poojan Oza and Vishal M. Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *CVPR*, 2019. 2

[35] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? In *Proc. WACV*, 2020. 3

[36] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015. 2

[37] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Semi-supervised learning with scarce annotations. *arxiv*, 2019. 2

[38] Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Towards open set recognition. *IEEE TPAMI*, 2013. 2

[39] Yu Shu, Yemin Shi, Yaowei Wang, Tiejun Huang, and Yonghong Tian. P-odn: Prototype-based open deep network for open set recognition. *Scientific Reports*, 2020. 2

[40] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2

[41] Xin Sun, Zhenning Yang, Chi Zhang, Guohao Peng, and Keck-Voon Ling. Conditional gaussian distribution learning for open set recognition. In *CVPR*, 2020. 2

[42] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. In *Workshop on Fine-Grained Visual Categorization*, 2019. 5, 6

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[45] 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*, 2022. 2, 3, 5, 6, 1

[46] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011. 5

[47] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *CoRR*, abs/2103.06191, 2021. 3

[48] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, 2019. 2

[49] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. 2

[50] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *ECCV*, 2020. 2

[51] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021. 3

[52] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, 2021. 3

[53] 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 *CVPR*, 2021. 3