# SEMANTIC CATEGORY DISCOVERY WITH VISION-LANGUAGE REPRESENTATIONS

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

## ABSTRACT

Object recognition is the task of identifying the category of an object in an image. While current models report excellent performance on existing benchmarks, most fall short of the task accomplished by the human perceptual system. For instance, traditional classifiers (e.g those trained on ImageNet) only learn to map an image to a predefined class index, without revealing the actual semantic meaning of the object in the image. Meanwhile, vision-language models like CLIP are able to assign semantic class names to unseen objects in a 'zero-shot' manner, though they are once again provided a predefined set of candidate names at test-time. In this paper, we reconsider the recognition problem and bring it closer to a practical setting. Specifically, given only a large (essentially unconstrained) taxonomy of categories as prior information, we task a vision-language model with assigning class names to all images in a dataset. We first use non-parametric methods to establish relationships between images, which allow the model to automatically narrow down the set of possible candidate names. We then propose iteratively clustering the data and voting on class names within clusters, showing that this enables a roughly 50% improvement over the baseline on ImageNet. We demonstrate the efficacy of our method in a number of settings: using different taxonomies as the semantic search space; in unsupervised and partially supervised settings; as well as with coarse-grained and fine-grained evaluation datasets.

## 1 INTRODUCTION

Image recognition has emerged as a fundamental task for demonstrating progress in computer vision and machine learning (He et al., 2016; Simonyan & Zisserman, 2015; Caron et al., 2020; 2021; Radford et al., 2021). However, in this work, we propose that current image recognition settings fall short of the task accomplished by human visual systems. Specifically, given a set of images, humans are able to directly map them to semantic object names — *e.g.*, 'that is a bird', 'that is a fish', 'that is a lion' if one were visiting a zoo. In a 'zero-shot' fashion, a human can leverage relations between images to narrow down a large list of possible nouns to a small yet precise set of semantic labels. This is the problem we tackle in this paper: given a collection of images and a large (essentially 'open') vocabulary, assigning class names to each image. We term this task *Semantic Category Discovery*, or SCD.

In contrast, consider the conventional image recognition task, which involves mapping a set of images to a static set of class indices (Russakovsky et al., 2015; Krizhevsky & Hinton, 2009), where these indices represent a predefined set of class names. The limitation here is the finite set of class-labels provided to the model as prior information for the recognition task. On the flip side, recent vision-language models trained on internet scale data are endowed with 'open' vocabularies (Radford et al., 2021; Jia et al., 2021a). These models have the ability to map images into a representation space where they can be directly compared with semantic labels. However, again, during evaluation, the models are always given a limited set of nouns from which to select the best match for an image.

Our solution is to this is to provide a vision-language model with a large, 'open' vocabulary of possible concepts (e.g WordNet with $68k$ nouns), and use relations between images to find the most relevant set of concepts for the task. Specifically, we propose a solution based on non-parametric clustering and iterative refinement. We first cluster the images using off-the-shelf non-parametric clustering methods on top of self-supervised features, before employing a vision-language model to infer an initial set of candidates names from the open vocabulary. Using a voting strategy within

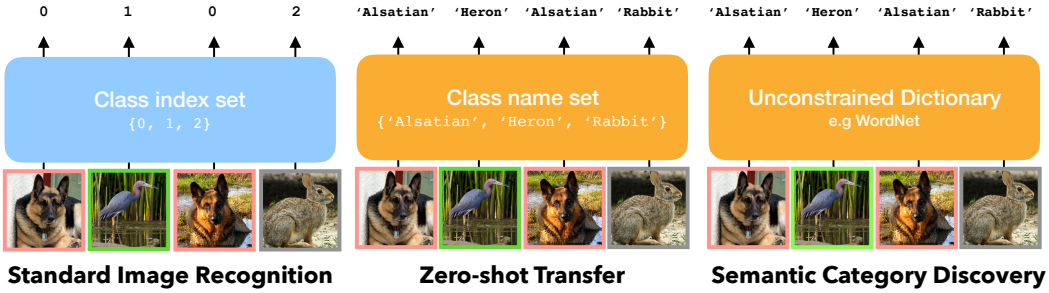


Figure 1: An illustration of how our proposed tasks extends existing image recognition settings. Left: A model is trained to predict class indices for a pre-defined set of categories (*e.g.*, supervised recognition, unsupervised clustering, category discovery etc.) Middle: A vision-language model is given a pre-defined set of class names to recognize images in a 'zero-shot' manner. Right: Our proposed setting, a model must predict an image's class name given only a large, unconstrained vocabulary. Note that the leftmost setting (blue model) uses only a visual representation, while the middle and right settings (orange models) use vision-language representations.

clusters to narrow down the initial set of candidate names, followed by an iterative refinement process, we can build a zero-shot classifier which can reliably predict the class names of images.

We evaluate our solution in a number of experimental settings. We control for: the initial choice of vocabulary, using the generic WordNet vocabulary as well as a task-specific vocabulary of $10k$ bird species; and also dataset granularity, including a large-scale ImageNet experiment (Russakovsky et al., 2015) as well as a fine-grained CUB evaluation (Wah et al., 2011a). We further conduct experiments when only images are available (the unsupervised setting) as well as when we have labels for a subset of the images (partially supervised (Vaze et al., 2022a)). For the latter, we further introduce a constrained semi-supervised clustering algorithm based on a Minimum Cost Flow (MCF) optimization problem (Bradley et al., 2000). In all cases, we find that our method substantially improves performance over baselines (e.g 50% relative improvement on ImageNet). In fact, surprisingly, we find that our method not only allows models to *name* the images, but in many cases improves *clustering performance* over prior state-of-the-art.

In summary, we make the following contributions: (1) We propose a visual recognition task which is closely aligned to that tackled by humans, which involves assigning semantic label names to a collection of images, and which addresses limitations in current tasks; (2) We propose novel solutions to this task using non-parametric clustering and iterative refinement with pre-trained vision-language models; (3) In a range of evaluation settings, we find that our method achieves substantial improvement of existing baselines and prior state-of-the-art. The task we tackle, and its distinctions with prior problems, is illustrated in fig. 1.

## 2 Related work

Our work concerns assigning semantic label names to a collection of unlabelled images. Here, we review the two most closely related fields to this task: *clustering/category discovery*, which aims to group sets of semantically related images together; and *semantic representation learning*, which concerns joint learning of visual and semantic (text) representations.

**Clustering and category discovery** *Unsupervised clustering* is a canonical problem in machine learning, where the aim is to find natural groupings of data without any labels (Aggarwal & Reddy, 2013; Van Gansbeke et al., 2020; MacQueen, 1967; Lloyd, 1982; Yang et al., 2017). With the advent of deep learning, many works have tackled the problem of simultaneously learning an embedding space and clustering data (Ji et al., 2019; Caron et al., 2018; Ronen et al., 2022). For instance, Invariant Information Clustering, IIC (Ji et al., 2019), seeks to learn a representation which maximizes the mutual information between class assignments of two augmentations of the same image, while DeepCluster (Caron et al., 2018) iteratively performs $k$-means (MacQueen, 1967; Lloyd, 1982) and gradient descent, where the cluster assignments are used as pseudo-labels for learning. Recent work has shown that, on top of well-trained features, the classic $k$-means algorithm can perform comparably to more sophisticated methods (Ronen et al., 2022).

Meanwhile, a rich vein of research considers 'category discovery', where labelled data is leveraged to discover and group images from new categories in unlabelled data. *Novel Category Discovery* (NCD) considers the setting in which the categories in the labelled and unlabelled data are disjoint (Han et al., 2019; Zhao & Han, 2021; Han et al., 2020; 2021; Jia et al., 2021b). DTC (Han et al., 2019) approaches this problem by first learning an embedding from the labelled data and using the representation to cluster unlabelled data, while (Fini et al., 2021) proposes a SWaV-like (Caron et al., 2020) clustering approach to partition the unlabelled data. Very recently, NCD has been extended to a more generalized setting as *'Generalized Category Discovery'* (GCD) (Vaze et al., 2022a) and concurrently as *'Open-World Semi-supervised Learning'* (Cao et al., 2022), wherein unlabelled data can include instances from both labelled and new categories.

All methods and tasks described here involve only grouping similar images together or, equivalently, predicting a cluster index for each image. In contrast, we extend these works by directly predicting an interpretable, semantic *class name*, which we suggest results in a more useful and practical perception system.

**Semantic representation learning** Most image recognition models are trained on fully annotated data with a predefined set of class indices (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; He et al., 2016). Here, semantics are only recovered by mapping indices to names, which is limited by the set of mappings defined a-priori. Meanwhile, multi-modal settings with *vision-language models* involve the learning of an embedding space where images and semantic text are represented jointly (Frome et al., 2013; Karpathy & Fei-Fei, 2015; Faghri et al., 2018; Desai & Johnson, 2021; Chen et al., 2021; Radford et al., 2021; Jia et al., 2021a). Notably, contrastive learning with image-text pairs on internet-scale data (*e.g.*, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021a)) results in a robust representation which embeds an 'open' vocabulary of semantic concepts. LiT (Zhai et al., 2022) also introduced a more data efficient image-text embedding approach by freezing a pre-trained image encoder and tuning only the text embedding. These large-scale models have been applied to a range of downstream tasks such as image captioning (Lu et al., 2019), VQA (Goyal et al., 2017), visual commonsense reasoning (Zellers et al., 2019), etc. They further show encouraging *zero-shot transfer* performance, where the models are repurposed for image recognition without any labelled data in the target domain. In this case, at test-time, the models are given a finite set of candidate class names from which to select the best match for a given image.

In this work, instead of tasking the vision-language models to match images with a pre-defined set of ground-truth names or text descriptions, we leverage such models to facilitate automatic semantic category discovery using only a large, unconstrained vocabulary.

## 3 SEMANTIC CATEGORY DISCOVERY

In this paper, we consider the problem of Semantic Category Discovery. Given a collection of images, the objective is to automatically assign a class name to each image, with only a large ('open') vocabulary as prior information. In general, we assume datasets with $K$ distinct semantic categories, and the aim is to find the optimal set of $K$ category names and assign each image to one of them.

In this section, we first describe our baseline for this task in sec. 3.1, before describing our solutions to SCD in two settings: a fully unsupervised setting where we are have no labelled data (sec. 3.2); and a partially supervised setting, similar to Generalized Category Discovery (Vaze et al., 2022a), where we have some labelled data (though not from all categories, sec. 3.3).

### 3.1 BASELINE: ZERO-SHOT TRANSFER

Recent large-scale vision-language (VL) models, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021a), have demonstrated the ability to match images with arbitrary textual (semantic) inputs by modelling in a joint embedding space. These models are pre-trained with weak supervision on large-scale image-text pairs. In current evaluation settings, models are fed a finite set of $K$ possible nouns and tasked with selecting the best match for a given image. For instance, when evaluating on ImageNet, VL models are given the class names corresponding to the $K = 1000$ categories in the standard ILSVRC dataset, and the models are expected to select one out of the 1000 predefined true names describing all classes of ImageNet. *We argue that this is unrealistic prior knowledge to obtain for many real-world tasks, and that humans are able to narrow down a large, open vocabulary to assign semantic class names.*

Our baseline for SCD respects this intuition. Specifically, instead of giving the VL model $K$ possible names for an image, we give it all nouns from a large dictionary, of size $N \gg K$, which is a proxy for all possible nouns. This baseline is termed *zero-shot transfer* in this work. Unsurprisingly, if we relax the semantic space from the true set of categories to a universal space, we find the model's performance degrades substantially. The goal in this work can be seen as finding a way to narrow the large dictionary of size $N$ to the ground truth set of $K$ class names *automatically*.

### 3.2 UNSUPERVISED SETTING

We first consider the situation where we only have unlabelled images. Different from conventional clustering, we wish to recognize the images by their semantic names rather than simply a cluster index. Specifically, we consider a collection of unlabelled images $\mathcal{D}_{\mathcal{U}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{U}}$, where $y_i \in \mathcal{Y}_{\mathcal{U}} = \{1, \dots, K\}$ are the ground truth class indices associated with a unique noun.

Our solution here leverages a pre-trained VL model (*e.g.*, CLIP (Radford et al., 2021)) and non-parametric clustering. In a nutshell, we propose to first cluster images in an embedding space, before using the VL model to bridge the gap between cluster assignment and object semantics. Next, given an initial guess of the set of semantic category names, we perform iterative refinement to improve the predictions.

#### 3.2.1 INITIAL CLUSTERING

Image recognition with standard VL models is a form of *parametric classification*. Specifically, the embeddings of the candidate semantic names through a text encoder form a (parametric) linear classifier on top of visual features. In contrast, we propose to also leverage a non-parametric recognition signal through the form of clustering. (Ronen et al., 2022) shows that clustering of features is a good replacement for parametric classification in the case when we have limited labelled data; in our case, the clustering allows us to establish relationships *between data points* which the VL model would otherwise be blind to. In detail, given a feature extractor $f_\theta$, we extract a feature vector $\mathbf{z}_i$ as $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$ for each image $\mathbf{x}_i$ in $\mathcal{D}_{\mathcal{U}}$. We then run the standard $k$-means algorithm to partition the data into $K$ clusters.

Next, we use the cluster assignments to narrow down a large, open dictionary to an initial estimate of the set of $K$ class names by voting on class names within a cluster. In particular, provided with the cluster assignment $y_i'$ for each $\mathbf{x}_i$, we collect instances in a cluster with index $c$ as $\mathcal{D}_c = \{\mathbf{x}_i \mid y_i' = c, \mathbf{x}_i \in \mathcal{D}_{\mathcal{U}}\}$. Let $f_\omega^v$ and $f_\omega^t$ be the pre-trained visual and text encoders of a large VL model. We extract visual features $\mathcal{Z}_{\mathcal{V}}^c = \{\mathbf{z}_i^v = f_\omega^v(\mathbf{x}_i) \mid \mathbf{x}_i \in \mathcal{D}_c\}$, and text embeddings for all nouns in the large dictionary $\mathcal{N}$ as $\mathcal{S}_{\mathcal{T}} = \{\mathbf{s}_i^t = f_\omega^t(\mathbf{n}_i) \mid \mathbf{n}_i \in \mathcal{N}\}$. For each $\mathbf{z}_i^v$ in cluster $c$, a predicted semantic name $\mathbf{s}_{\alpha(i)}$ can be obtained by querying the nearest neighbour (NN) in $\mathcal{S}_{\mathcal{T}}$:

$$\alpha(i) = \arg\max_j \{\mathbf{z}_i^v \cdot \mathbf{s}_j^t \mid \mathbf{s}_j^t \in \mathcal{S}_{\mathcal{T}}\}. \tag{1}$$

After obtaining predicted the semantic name embedding vector $\mathbf{s}_{\alpha(i)}$ for each image $i$ in cluster $c$, all the unique text embeddings in $\{\mathbf{s}_{\alpha(i)}\}$ forms a subset of $\mathcal{S}_{\mathcal{T}}$, denoted as $\mathcal{S}_{\mathcal{T}}^c$. We simply choose the most common semantic name in the cluster to give an initial set of candidate names $\mathbf{S}_{\mathcal{T}}^{\mathcal{D}_{\mathcal{U}}}$ as:

$$\mathbf{S}_{\mathcal{T}}^{\mathcal{D}_{\mathcal{U}}} = \{\mathbf{s}_c = \arg\max_{\mathbf{s}_j} \mathcal{P}_c(\mathbf{s}_j) \mid \mathbf{s}_j \in \mathcal{S}_{\mathcal{T}}^c, c = 1, \cdots, K\}, \tag{2}$$

where $\mathcal{P}_c(\mathbf{a})$ counts the occurrences of the semantic vector $\mathbf{a}$ of a noun in cluster $c$.

#### 3.2.2 SEMANTIC REFINEMENT

Having obtained an initial estimate of the set of $K$ class names, we perform two *refinement* steps. Firstly, we note that by enforcing one semantic name to represent each cluster in the initial clustering step, any duplicated class names would necessarily give us a smaller set of candidate names than those actually present in $\mathcal{D}_{\mathcal{U}}$. To remedy this problem, instead of measuring the frequency of only the nearest semantic name in eq. (1) for each instance in a cluster $c$, we instead track the frequency of the $m$-NN. In this way, for each cluster $c$, there are $m$ candidate semantic vectors. This allows us to form a cost matrix of assigning a class name to a cluster, where the cost is equal to the proportion of instances in that cluster assigned that class name. Next, we solve the linear assignment problem

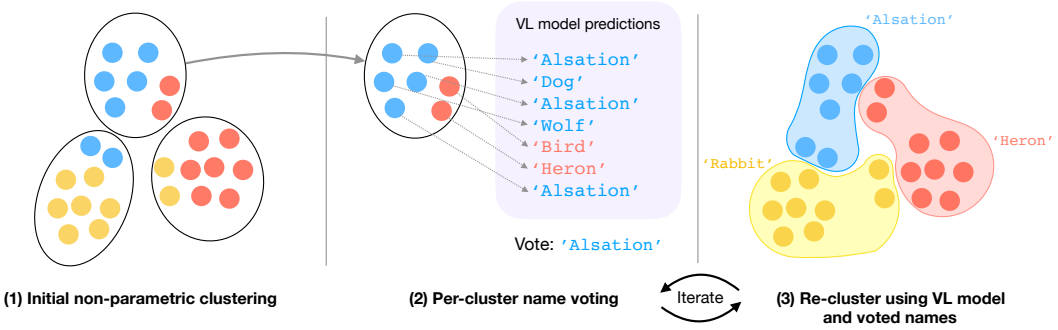


Figure 2: An illustration of our proposed method. Left: We first perform non-parametric clustering on deep features to get initial cluster assignments. Middle (Step 1): For each cluster, we use a VL model to assign a class names for each image from the *entire open vocabulary*. We select one class name for each cluster based on the most common occurrence. Right (Step 2): using the voted class names, we label each image as one of these, using these assignments to form new clusters. We then iterate steps one and two. Note: Here we have not illustrated refinement with top-$k$ voting (see sec. 3.2.2).

between class names and clusters, using the Hungarian algorithm to assign one semantic vector to each cluster without duplication.

At this point, by stacking all the semantic vectors in $\mathbf{S}_{\mathcal{T}}^{\mathcal{D}_{\mathcal{U}}}$, we construct a $d \times K$ matrix forming a parametric, zero-shot linear classifier $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^K$. By applying this classifier on the the visual features as $j' = \arg\max_j \phi(\mathbf{z}_i^v)[j]$, we compute semantic assignments and can also construct a new clustering of the data, which is not necessarily the same as the initial clustering assignment achieved with the non-parametric algorithm. This new clustering of the data allows us to *iteratively update* the candidate semantic vectors and the cluster assignments until convergence while ensuring $K$ unique semantic class names are assigned.

Intuitively, our final iterative process can be considered as analogous to the classic $k$-means algorithm. Clusters are iteratively created and updated based on an energy, though the energy in our case is based on similarities of the visual features to a set of semantic text embeddings, rather than similarities only between visual features (as in standard $k$-means).

### 3.3 Partially Supervised Setting

We also consider a partially supervised setting where, in addition to the collection of unlabelled images $\mathcal{D}_{\mathcal{U}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{U}}$, we also have access to some labelled images $\mathcal{D}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{L}}$, where $\mathcal{Y}_{\mathcal{L}} \subset \mathcal{Y}_{\mathcal{U}}$. This setting has recently been formalized as 'Generalized Category Discovery' (Vaze et al., 2022a), which we now extend to require the prediction of semantic names rather than simply class indices.

Here, one could simply apply the method from sec. 3.2, and ignore the labelled data. However, we can also leverage the labelled data to help us better recognize the unlabelled images. Similarly to sec. 3.2, we first obtain cluster indices through a non-parametric clustering algorithm, followed by narrowing down the candidate semantic names from the universal semantic space to a set of elements with the same or more semantic names than the actual class number. However, differently to sec. 3.2, we constrain the clustering stage using the labelled set.

#### 3.3.1 Constraining K-means

With $\mathcal{D} = \mathcal{D}_{\mathcal{L}} \cup \mathcal{D}_{\mathcal{U}}$, we adopt a semi-supervised $k$-means algorithm (SS-$k$-means) (Han et al., 2019; Vaze et al., 2022a) to obtain the cluster indices for each image. SS-$k$-means extends $k$-means to cluster partially labelled data $\mathcal{D}$ by enforcing labelled instances (in $\mathcal{D}_{\mathcal{L}}$) of the same class to fall into the same cluster. Though it performs well in the GCD setting, we find that SS-$k$-means often produces clusters that contain very few instances. This is the result of the supervision, as we find that the forced assignment on the labelled set results in a few 'attractor' clusters which are bloated in size. Here, we improve SS-$k$-means by introducing a loose cluster size constraint. Specifically, we constrain the minimum size of the clusters during the SS-$k$-means clustering. Drawing inspiration

from (Bradley et al., 2000), we formulate the cluster assignment step of SS-$k$-means as a Minimum Cost Flow (MCF) linear optimisation problem. We call the resulting algorithm constrained SS-$k$-means (CSS-$k$-means). After running CSS-$k$-means on $\mathcal{D}$, we obtain initial cluster assignments $y'_i$ for each $\mathbf{x}_i$ in the unlabelled data $\mathcal{D}_{\mathcal{U}}$. The refinement process remains largely the same as in sec. 3.2.2. However, the search space for the vision language model is reduced, as we know that categories in $\mathcal{D}_{\mathcal{L}}$ must be present.

# 4 EXPERIMENTS

## 4.1 VOCABULARIES AND DATASETS

Recall that our aim is to take a collection of images and, using a VL model, assign a semantic class name to every image, given only a large vocabulary as prior information. As such, a key component in our pipeline is the choice of vocabulary. In this work, we consider nouns from the generic WordNet taxonomy, which contains $68k$ concepts and can be considered to contain most objects a human might typically encounter. We further test a domain-specific vocabulary for naming bird-species, for which we scrape $11k$ bird names from Wikipedia.

For evaluation data, we first demonstrate results on ImageNet (Russakovsky et al., 2015), where each class name is a node in the WordNet hierarchy. For comparison with GCD models (Vaze et al., 2022a), we demonstrate results on ImageNet-100, a 100 class subset of the standard ILSVRC benchmark (Russakovsky et al., 2015). We also show results on the full dataset in table 3. Furthermore, we show results on *Stanford Dogs* (Khosla et al., 2011), a fine-grained dataset of 120 dog breeds, whose classes also come from the WordNet hierarchy. Finally, for an evaluation with the bird species vocabulary, we experiment with CUB (Wah et al., 2011a). In the partially supervised setting, we use the data splits from GCD and refer to the original paper for details.

## 4.2 EVALUATION METRICS

The nature of the Semantic Category Discovery task can be captured by the metrics we choose for evaluation. The primary aim of this work is assigning *semantic names* to a given image, rather than a typical class index as in standard recognition. To this end, we first measure 'Semantic Accuracy' or 'sACC', which is analogous to standard classification accuracy. Specifically, given the ground truth semantic name, $s_i$, and the predicted name, $\hat{s}_i$:

$$sACC = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}\{s_i = \hat{s}_i\}, \tag{3}$$

Secondly, we introduce a soft evaluation metric which accounts for the continuous nature of semantic similarity. For instance, if a 'mushroom' is predicted as 'fungus', the prediction should not be considered as completely wrong. It would be preferable to account for semantic similarity between the ground-truth and predicted name during evaluation and, as such, we introduce the 'soft Semantic Accuracy (Soft-sACC) metric. We adopt the Leacock Chodorow Similarity (LCS) (Fellbaum, 1998), which measures lexical semantic similarity by finding the shortest path in the WordNet graph between two concepts, and scales that value by the maximum path length. The LCS between two concepts $i$ and $j$ is defined as:

$$s_{i,j} = -log(\frac{p(i,j)}{2d}), \tag{4}$$

where $p(i,j)$ denotes the shortest path length between $i$ and $j$ and $d$ denotes the taxonomy depth. We further re-scale the similarity score into the range $[0,1]$ by dividing with the maximum possible score, with '0' indicating no semantic similarity and '1' indicating a perfect prediction. We evaluate the Soft-sACC on datasets for which the concepts are determined in the WordNet hierarchy: ImageNet-100, Standord Dogs, and ImageNet-1K.

Finally, as in the existing recognition literature, we are also interesting in *clustering performance*, or how often images from the same category are grouped together. A model could predict the wrong name for every image but still achieve high 'Clustering Accuracy' (ACC), as long as images from

the same category were predicted the same (incorrect) name. Clustering Accuracy is defined as:

$$ACC = \max_{p \in \mathcal{P}(\mathcal{Y}_\mathcal{U})} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{y_i = p(\hat{y}_i)\}, \tag{5}$$

where $y_i$ represent the ground truth label and $\hat{y}_i$ the cluster assignment and $\mathcal{P}(\mathcal{Y}_\mathcal{U})$ denotes all possible permutations of the class labels in the unlabelled data. Note that 'sACC' is generally a more difficult metric than 'ACC', with random guessing performance being roughly $1/N$ (*e.g.*, $1/68k$ for WordNet), while random guessing for 'ACC' is roughly $1/K$ (*e.g.*, $1/100$ for ImageNet-100).

### 4.3 Implementation details

**Models.** For the visual feature extractor $f_\theta$ used for initial cluster assignment (sec. 3.2.1 and sec. 3.3.1) we could, in principle, use any feature extractor. For the unsupervised setting, we use a self-supervised vision transformer (DINO ViT-B-16 weights (Dosovitskiy et al., 2021; Caron et al., 2021)), while in the partially supervised setting we fine-tune the feature extractor with supervised and self-supervised contrastive losses as in (Vaze et al., 2022a). For the vision-language component ($f^v_\omega$, $f^t_\omega$), we adopt the off-the-shelf pre-trained CLIP model with a ViT-B-16 backbone.

**Compared methods.** We can find no single existing method which can perform the Semantic Category Discover task and, as such, we compare against a number of methods to benchmark different aspects of our model. Firstly, we use the *Zero-shot transfer (baseline)* described in sec. 3.1. Next, in the unsupervised setting, we employ *$k$-means* on top of DINO features as in (Vaze et al., 2022a) to compare clustering accuracy. It has been shown in (Ronen et al., 2022) that $k$-means is competitive with complex state-of-the-art methods when the underlying features are good. In the partially supervised setting, we compare against *GCD* (Vaze et al., 2022a). We also provide an upper-bound for performance as *Zero-shot transfer (UB)*. Here, we evaluate performance if a zero-shot classifier is evaluated on the unlabelled data using only the ground class names.

Finally, we report our results as *Ours (Semantic Naming)*. In the partially supervised case, we also report ACC after the initial non-parametric clustering step as *Ours (CSS-KM)* to demonstrate the efficacy of our proposed constrained semi-supervised $k$-means algorithm.

**Notes.** In all cases, we assume knowledge of the ground-truth number of categories, $k$. Though this is a limitation (see appendix D), we highlight that the estimation of the true number of clusters in a dataset is its own challenging research problem (Ronen et al., 2022). As such, we treat the problem of estimating the number of categories as orthogonal to our semantic naming question. We report numbers with an off-the-shelf $k$-estimation technique from (Vaze et al., 2022a) in appendix A.

Finally, no existing method can be evaluated in all settings. Specifically, clustering based methods cannot be evaluated for 'sACC'. Also, we found it computationally infeasible to compute the 'ACC' for the zero-shot transfer baseline. The baseline predicts too many unique class names (usually over 10k) to reasonably compute the Hungarian assignment with the ground truth classes. In table 3, when a method cannot be evaluated in a given setting, we fill the table entry with '-'.

### 4.4 Main results

We evaluate on ImageNet-100, Stanford Dogs and CUB in the unsupervised (table 1) and partially supervised (table 2) settings respectively, providing results for both semantic naming (left-hand tables) and clustering performance (righ-hand). We first note that the sACC of the baseline is surprisingly strong, given the random guessing performance of less that 0.1% in all cases. This speaks to the strength of the underlying large-scale VL model. However, our method provides improvements on this across the board. Specifically, our method roughly doubles sACC on ImageNet-100 in both cases, and improves performance on CUB by roughly 50% in the partially supervised setting for CUB. We note that gains are relatively modest for Stanford Dogs and in CUB's unsupervised evaluation. Next, we highlight a surprising finding that our method can usually also aid clustering accuracy. That is, in addition to our method giving the model the ability to assign a *semantic name* to an image, the resulting classifications can also be used to cluster images more reliably than prior methods.

Finally, we demonstrate results in both the unsupervised and partially supervised settings on a large-scale ImageNet evaluation in table 3. Particularly, in the partially supervised case, we are able to substantially improve upon the baseline here, with a 13% boost in sACC.

Table 1: **Results in the unsupervised setting.** We use DINO features for the initial clustering step and report metrics for semantic accuracy (involving class naming, left) and clustering (right).

| | ImageNet-100 | | Stanford Dogs | | CUB |
|---|---|---|---|---|---|
| | sACC | Soft-sACC | sACC | Soft-sACC | sACC |
| Zero-shot transfer (UB) | 85.0 | 92.0 | 60.4 | 83.2 | 54.1 |
| Zero-shot transfer (baseline) | 22.7 | 57.7 | 51.7 | 77.4 | 20.2 |
| Ours (Semantic Naming) | **41.2** | **71.3** | **53.8** | **79.1** | **24.5** |

| | ImageNet-100 | Stanford Dogs | CUB |
|---|---|---|---|
| | ACC | ACC | ACC |
| Zero-shot transfer (UB) | 85.1 | 60.8 | 55.8 |
| $k$-means (baseline) | 73.2 | 47.2 | 34.4 |
| Ours (Semantic Naming) | **78.2** | **57.9** | **46.5** |

Table 2: **Results in the partially supervised setting.** We use GCD features for the initial clustering step and report metrics for semantic accuracy (involving class naming, left) and clustering (right).

| | ImageNet-100 | | Stanford Dogs | | CUB |
|---|---|---|---|---|---|
| | sACC | Soft-sACC | sACC | Soft-sACC | sACC |
| Zero-shot transfer (UB) | 85.0 | 92.0 | 60.4 | 83.2 | 54.1 |
| Zero-shot transfer (baseline) | 22.7 | 57.7 | 51.7 | 77.4 | 20.2 |
| Ours (Semantic Naming) | **54.8** | **77.5** | **53.7** | **79.6** | **28.2** |

| | ImageNet-100 | Stanford Dogs | CUB |
|---|---|---|---|
| | ACC | ACC | ACC |
| Zero-shot transfer (UB) | 85.1 | 60.8 | 55.8 |
| GCD Vaze et al. (2022a) (baseline) | 74.1 | 60.8 | **54.0** |
| Ours (CSS-$k$-means) | 78.7 | **62.1** | 52.9 |
| Ours (Semantic Naming) | **83.0** | 56.6 | 46.8 |

Table 3: **Results on the ImageNet-1k test set.** We evaluate on the standard ILSVRC Russakovsky et al. (2015) benchmark in both unsupervised and partially supervised settings. We use DINO features to provide initial cluster assignments in this case.

| | Unsupervised | | | Partially Supervised | | |
|---|---|---|---|---|---|---|
| | sACC | Soft-sACC | ACC | sACC | Soft-sACC | ACC |
| Zero-shot transfer (UB) | 63.4 | 81.3 | 64.1 | 63.4 | 81.3 | 64.1 |
| $k$-means | - | - | 50.2 | - | - | 50.2 |
| Zero-shot transfer (baseline) | 24.3 | 57.5 | - | 24.3 | 57.5 | - |
| Ours (Semantic Naming) | **31.1** | **63.5** | **55.5** | **37.7** | **66.7** | **54.9** |

### 4.5 Analysis

Here we investigate the various design choices in our solution to SCD. In table 4 we investigate both the choice of features for the initial clustering step, as well as the effects of different components of our semantic naming method, under the unsupervised setting.

Particularly, we gradually add our initial voting step, the iterative refinement, and the linear assignment, as introduced in sec. 3.2.1 and sec. 3.2.2, on top of the $k$-means non-parametric clustering. We experiment on the generic classification dataset ImageNet-100 and also the fine-grained Stanford Dogs dataset. We experiment with both CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) features for the initial $k$-means clustering.

As can be seen, each component of the semantic naming process improves sACC and ACC. The trend holds for different datasets and different features used for the initial clustering. Furthermore, DINO features provide a better starting point than CLIP in this setting. This is likely a result that DINO features form strong nearest-neighbour classifiers (Caron et al., 2021). Our full method gives an sACC of $41.2\%$ on ImageNet-100 and $53.8\%$ on Stanford Dogs using DINO based initial clustering. We also note the boost our methods can give in clustering accuracy on ImageNet-100 over $k$-means clustering with DINO features. This is surprising as, despite not being designed for this purpose, our semantic naming method is able to group similar images together more effectively than the strong DINO baseline. We also investigated the effects of using different initial clustering algorithms, and the results can be found at appendix B. We find our constrained semi-supervised $k$-means consistently leads to the best semantic naming results.

Table 4: **Effectiveness of different components of our semantic naming method.** Experiments under the unsupervised setting with CLIP and DINO features.

| Datasets | CLIP | | | | DINO | | | |
|---|---|---|---|---|---|---|---|---|
| | ImageNet-100 | | Stanford Dogs | | ImageNet-100 | | Stanford Dogs | |
| | sACC | ACC | sACC | ACC | sACC | ACC | sACC | ACC |
| $k$-means | - | 62.3 | - | 26.9 | - | 73.2 | - | 47.2 |
| + Initial Voting Step | 35.6 | 70.8 | 41.8 | 47.0 | 40.7 | 77.5 | 45.3 | 50.1 |
| + Iteration | 36.4 | 70.9 | 44.2 | 49.1 | 40.9 | 78.0 | 48.3 | 52.1 |
| + Linear assignment | **37.1** | **71.3** | **50.3** | **55.2** | **41.2** | **78.2** | **53.8** | **57.9** |

In addition, we show qualitative results of our method in fig. 3 on unlabelled images in ImageNet-100. We report correct, partially correct, and incorrect cases for our method, where the we use the

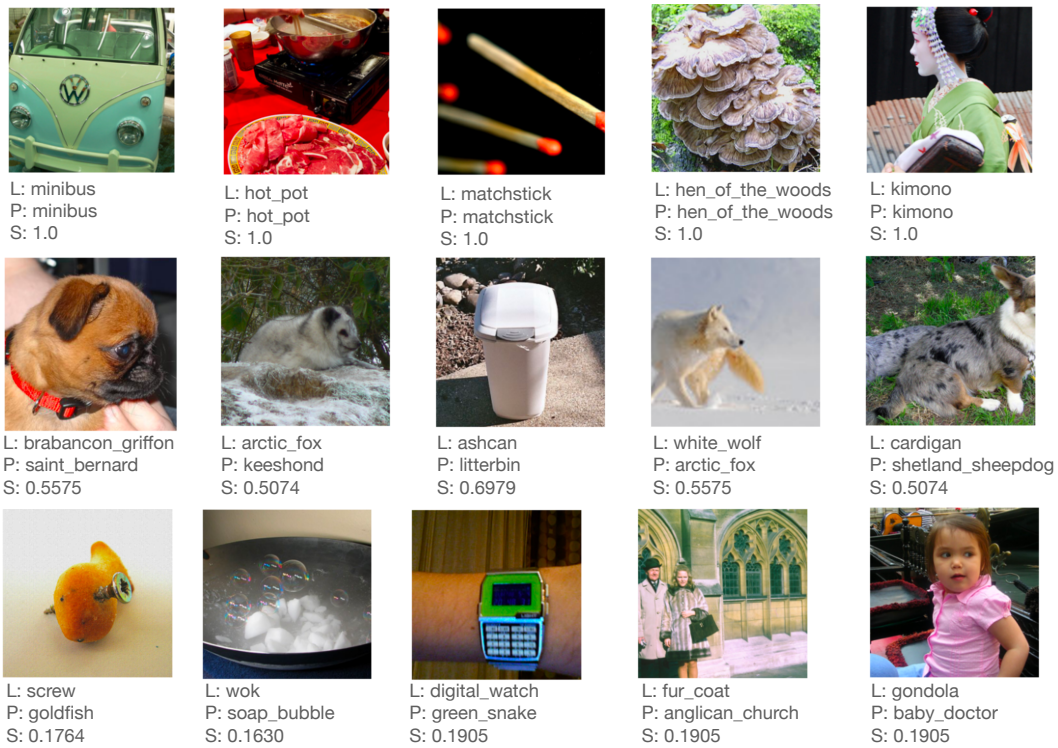


Figure 3: **Qualitative results on unlabelled instances from unknown classes.** Top row: correct predictions; Middle row: partially correct predictions; Bottom row: incorrect predictions. P: prediction; L: label; S: Soft semantic similarity score.

soft semantic similarity score (see sec. 4.2) to bin examples. For the partially correct predictions (middle row), we can see that the predictions are actually highly semantically relevant to the ground-truth names. The errors can be partially attributed to the pose (*e.g.*, 'brabancon griffon' vs 'saint bernard'), occlusion (*e.g.*, 'cardigan'), and background clutter (*e.g.*, 'arctic fox') in the content. For the incorrect predictions (bottom row), we can see that the predictions are mostly semantically relevant to the content. Incorrect predictions can also be caused by spurious objects appearing in the image. For example, in the 'screw' vs 'goldfish' image, a goldfish-shaped object actually occupies a higher region of the image than the screw, leading to an incorrect prediction.

When multiple dominant objects appear in the images, the model may get confused on which one to predict. For example, in the 'fur coat' vs 'anglican church' image, both items appear in the image but the latter occupies a larger region. Similar for the 'bondola' vs 'baby doctor' image. In these cases, multi-label evaluations of ImageNet may also be beneficial (Ridnik et al., 2021). Through these qualitative results, we can see that our method can produce reasonably good results which reflect the true semantics of the images, with failures generally occurring for understandable reasons.

## 5 Conclusions

In this work we have proposed and tackled the task of *Semantic Category Discovery* (SCD), where a model must predict the *semantic category name* of an image given only a large, open dictionary as prior information. This extends standard image recognition settings which only require mapping images to a pre-defined set of class indices. It also extends zero-shot evaluations of large-scale vision-language models, which assume a finite set of candidate class names will be given at test-time. We propose a solution based on parametric clustering followed by iterative semantic refinement, and show that this method substantially outperforms existing baselines for SCD on coarse and fine-grained datasets, including a full ImageNet evaluation.

## REFERENCES


Charu C. Aggarwal and Chandan K. Reddy. *Data Clustering: Algorithms and Applications*. CRC Press, 2013.

Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. Constrained k-means clustering, 2000.

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

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018.

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.

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.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, 2021.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021.

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.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

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

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, , and Tomasü Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

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

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

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

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

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021a.

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

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPRW*, 2011.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 1982.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *ArXiv e-prints*, 2021.

Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. Deepdpm: Deep clustering with an unknown number of clusters. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

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

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020.

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

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*, 2022b.

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, 2011a.

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*, 2011b.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, 2017.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022.

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

# A RESULTS WHEN ESTIMATING THE NUMBER OF CATEGORIES (K)

We use the method proposed in (Vaze et al., 2022a) to estimate the number of categories in the unlabeled data, and we get 109 classes for ImageNet-100, 114 classes for Stanford Dogs and 231 classes for CUB, respectively. We find a reduction in peformance on a number of datasets, including on Stanford Dogs. However, on some datasets, we surprisingly find improved performance, which we attribute to a quirk of the specific interaction of the initial clustering and semantic naming process. Note that, overall, our method still substantially outperforms baselines and prior art.

Table 5 shows the results in the partially supervised setting.

Table 5: **Results in the partially supervised setting with estimated class numbers.** We use GCD features for the initial clustering step and report metrics for semantic accuracy (involving class naming, left) and clustering (right).

| | ImageNet-100 | | Stanford Dogs | | CUB |
|---|---|---|---|---|---|
| | sACC | Soft-sACC | sACC | Soft-sACC | sACC |
| Zero-shot transfer (UB) | 85.0 | 92.0 | 60.4 | 83.2 | 54.1 |
| Zero-shot transfer (baseline) | 22.7 | 57.7 | **51.7** | 77.4 | 20.2 |
| Ours (Semantic Naming) | **54.5** | **75.4** | 49.7 | **78.1** | **29.8** |

| | ImageNet-100 | Stanford Dogs | CUB |
|---|---|---|---|
| | ACC | ACC | ACC |
| Zero-shot transfer (UB) | 85.1 | 60.8 | 55.8 |
| GCD (Vaze et al., 2022a) (baseline) | 74.1 | 60.8 | **54.0** |
| Ours (CSS-$k$-means) | 75.49 | **60.9** | 52.2 |
| Ours (Semantic Naming) | **79.1** | 54.8 | 49.71 |

## A.1 EVALUATION ON THE PREDICTED SEMANTIC NAMES

One perspective of the Semantic Category Discovery task is narrowing an unconstrained dictionary of $N$ possible names to the optimal set of $K$ categories for a given dataset (see sec. 3.1). As such, a further possible metric is the overlap between the set of discovered class names and the ground truth set of semantic labels. We evaluate the intersection over union (IoU) between the two sets and report results in table 6. We see that the reported IoUs are reasonably correlated with the sACC performance, substantiating our assumption that *discovering* the true set of class names is a key challenge in unconstrained semantic labelling.

Table 6: **Evaluation on the predicted semantic names.** We measure the IoU between the ground-truth names and the predicted names for each dataset under both unsupervised and partially supervised settings.

| | Unsupervised | Partially supervised |
|---|---|---|
| ImageNet-100 | 0.290 | 0.575 |
| ImageNet-1K | 0.273 | 0.517 |
| Standord Dogs | 0.589 | 0.752 |
| CUB | 0.343 | 0.481 |

# B EFFECTS OF DIFFERENT INITIAL CLUSTERING ALGORITHMS

In table 7, we investigate the effects of using different non-parametric clustering algorithms, including $k$-means, semi-supervised $k$-means (SS-KM), and our constrained semi-supervised $k$-means (CSS-KM), on both CLIP and GCD features under the partially supervised setting, where the latter two clustering algorithms appears to be more demanding. Our semantic naming method can consistently improve the ACC, regardless of the non-parametric clustering algorithms nor the features used for initial clustering. Among these three different choices of algorithms, our CSS-KM gives the best semantic naming performance. Interestingly, we observe that the overall initial ACC of CSS-KM is not as good as SS-KM, but using our CSS-KM to provide initial clustering gives better ACC after semantic naming, validating the effectiveness of the size contstraint we introduced in CSS-KM for the name voting. As the GCD feature extractor is trained jointly using supervised contrastive learning on the unlabelled data and self-supervised contrastive learning on all data for better visual representation, the initial clustering results on GCD features are better and we obtain a better semantic naming result with GCD features. As a stronger training signal is used for the 'Old' classes, the ACC of $k$-means, SS-KM and CSS-KM are very high. Note the semantic naming is based on the CLIP feature, which is not biased to the labelled data, thus our semantic naming gives the best 'All' and 'New' classes and the gap between 'Old' and 'New' is much smaller than the initial clustering based on the GCD feature.

Table 7: **Effects of different initial clustering algorithms.** Experiments under the partially supervised setting with CLIP and GCD features.

| | CLIP | | | GCD | | |
|---|---|---|---|---|---|---|
| Classes | All | Old | New | All | Old | New |
| $k$-means | 62.3 | 57.2 | 64.8 | 78.2 | 89.0 | 72.8 |
| $k$-means + Semantic Naming | 71.3 | 64.6 | 74.7 | 80.9 | 85.1 | 78.9 |
| SS-KM | 68.1 | 74.1 | 65.0 | 73.2 | 87.8 | 65.8 |
| SS-KM + Semantic Naming | 71.2 | 80.7 | 66.4 | 74.0 | 81.9 | 70.0 |
| CSS-KM | 65.6 | 74.2 | 61.3 | 78.7 | **92.9** | 71.5 |
| CSS-KM + Semantic Naming | **79.9** | **83.7** | **78.0** | **83.0** | 84.9 | **82.1** |

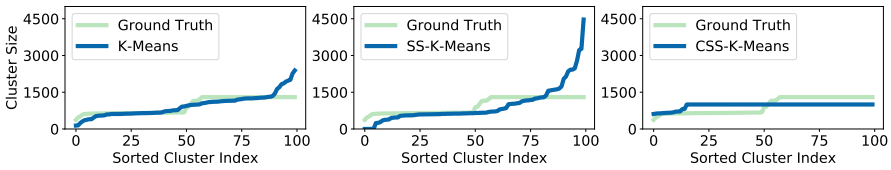


Figure 4: **Sorted cluster sizes obtained by different initial clustering algorithms.** Results are reported on ImageNet-100 dataset with DINO features.

## C CLUSTER SIZE COMPARISON

Here, we compare the cluster sizes under the partially supervised setting with different clustering methods, in order to illustrate the benefits of our constrained algorithm (sec. 3.3.1). As can be seen, both $k$-means and SS-$k$-means can have clusters with very few instances (which is more severe in SS-$k$-means), making the voting from these few-instance clusters less reliable. In contrast, our improved CSS-$k$-means can give cluster sizes that are better aligned with the ground truth, thus benefiting the subsequent semantic voting process. Note that in the partially supervised settings, rough cluster sizes can be estimated from the labelled data (and used to guide the CSS-$k$-means process).

## D LIMITATIONS AND ETHICAL CONSIDERATIONS

In this work we have tackled the task of assigning semantic names to images by automatically narrowing down an unconstrained vocabulary with a pre-trained vision language model. Here, we highlight a number of considerations when deploying such a method, in addition to the assumption of the number of ground truth classes (mentioned in the main paper).

Firstly, we note that though our method outperforms existing baselines for this task, its absolute accuracy (both ACC and sACC) is quite low in absolute terms; for instance we achieve roughly 30% sACC on the unsupervised ImageNet setting. This suggests that, even with internet-scale pre-training, current perception systems are not suitable for unconstrained deployment in the real-world. However, we hope that by highlighting this poor performance in the unconstrained setting, we can draw attention to an important research direction for the community.

Next, we note that our method hinges on a model trained on internet-scale data (CLIP (Radford et al., 2021)) for which the training data is private and unavailable for inspection. As such, we are unable to interrogate many sources of possible bias in the model, or else predict a-priori when its predictions may be unreliable.

Finally, we highlight a more subtle technical consideration for our setting. In this work, we assessed our model's ability to name arbitrary objects by categorically measuring its accuracy against a pre-defined vocabulary (*i.e.*its predictions were either considered correct or incorrect in a binary fashion). This raises a technical issue as it assumes that there is a single underlying taxonomy for the data, and does not permit other categorisation systems, which may be valid but do not align with the selected vocabulary. Though this is perhaps less of an issue for the fine-grained CUB evaluation (see (Vaze et al., 2022b) for a discussion on this point), there may be a number of equally valid partitions of a generic object recognition dataset like ImageNet.

## E LICENSE FOR DATASETS

We carefully follow the licenses of the datasets in our experiments. ImageNet (Russakovsky et al., 2015) and Stanford Dogs (Khosla et al., 2011) apply the same license for non-commercial research use. CUB (Wah et al., 2011b) dataset also allows non-commercial research use.