# PSS: Progressive Sample Selection for Open-World Visual Representation Learning

Tianyue Cao*, Yongxin Wang†, Yifan Xing, Tianjun Xiao, Tong He, Zheng Zhang, Hao Zhou, and Joseph Tighe

Amazon Web Services
vanessa_@sjtu.edu.cn, {yongxinw, yifax, tianjux, htong, zhnzhe, zhouho, tighej}@amazon.com

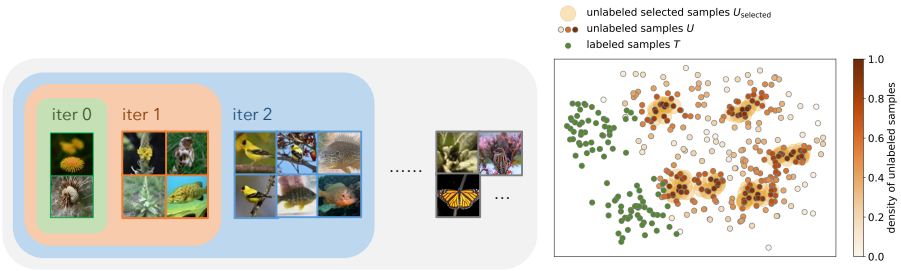
**Abstract.** We propose a practical open-world representation learning setting where the objective is to learn the representations for unseen categories without prior knowledge or access to images associated with these novel categories during training. Existing open-world representation learning methods make assumptions, which are often violated in practice and thus fail to generalize to the proposed setting. We propose a novel progressive approach which does not depend on such assumptions. At each iteration our approach selects unlabeled samples that attain a high homogeneity while belonging to classes that are distant to the current set of known classes in the feature space. Then we use the high-quality pseudo-labels generated via clustering over these selected samples to improve the feature generalization iteratively. Experiments demonstrate that the proposed method consistently outperforms state-of-the-art open-world semi-supervised learning methods and novel class discovery methods over nature species image retrieval and face verification benchmarks. Our training and inference code are released.[1]

**Keywords:** Open-World Representation Learning, Semi-Supervised Learning, Sample Selection, Iterative Methods

## 1 Introduction

Great progress has been made in the past decade to improve the accuracy of computer vision models and they are starting to be used in real-world applications. But one thing that holds back the wide adoption of such models is their restrictive closed universe requirements. Many real-world applications for computer vision do not operate in a fixed set of categories known a priori. Take the task of building a fine-grain species recognition system for example. One would start with a large set of annotated images, perhaps with a focus on mammals, for a set of known species and deploy such a system. Its users will expect it to recognize all fine-grain categories of, not only mammals, for which there might

*Currently at The Shanghai Jiao Tong University. Work conducted while at AWS.
†Corresponding author.
[1] https://github.com/dmlc/dgl/tree/master/examples/pytorch/hilander/PSS




(a) Progressively grown training set. (b) Sample selection.

Fig. 1: *Progressive Sample Selection* (PSS) approach. (a) shows the progressively grown training set. The training set gradually expands during iterations. (b) shows a toy example of sample selection in one iteration. We select samples that are densely clustered together with high local homogeneity

already be good coverage, but also birds, reptiles, etc. An effective way to expand to these new user specified fine-grain categories is to pose the problem as a metric learning task that leverages a single visual representation to retrieve examples from an ever expanding pool of labeled data. But learning such a visual representation without knowledge of the complete set of target labels is challenging. To this end, we present a practical open-world representation learning setting to realistically reflect real-world applications. Here, the training procedure has access to both labeled and unlabeled data, with the unlabeled data containing images of both the labeled classes as well as a set of unseen categories. We aim to train a visual representation that can generalize to the open-world setting where new unseen categories are encountered. Thus, the test data comes from labels that are disjoint from both the labeled and unlabeled training sets.

Our formulation is different from semi-supervised learning (SSL) or open-set semi-supervised learning as ours requires learning representations that cover both known and novel classes. Although there have been existing works aiming to discover novel classes and learn representations for them with a partially labeled dataset [10,37], they often assume constraints which are impractical in the real-world. Those include the assumptions that the unlabeled data is only comprised of samples from novel classes and the number test-time unseen classes is known a priori. The closest setting to our is [37]. However, they directly test unseen class recognition performance over the unlabeled dataset that is already accessible in training. They also assume an up-to a 2:1 ratio between the unlabeled and labeled data where in practice the ratio is much larger. Our setting does not assume any of the before-mentioned constraints and thus is more practical: 1) we test over a disjoint set of classes to the known classes in the labeled training data and novel classes in the unlabeled; 2) the unlabeled to labeled data ratio (up-to 10:1) is much higher, thus better approximates the real-world scenario where unlabeled data is far more abundant; 3) the unlabeled data can contain

both seen and unseen classes and 4) we do not know a priori the number of novel classes in the unlabeled training data.

The new proposed setting is more challenging than previous works due to the high unlabeled to labeled data ratio, introducing a large distribution gap in between, and an absence of prior knowledge over novel classes. Additionally, since we test on a class-disjoint set, the representation needs to be highly generalized. We test existing semi-supervised learning and novel class discovery methods in our proposed setting and found they fail to generalize.

One observation we make is that addressing such a challenging setting with one-step training is difficult. Thus, we propose a novel *Progressive Sample Selection* (PSS) approach as illustrated by Figure 1. Our method, partially inspired by [6], recurrently clusters a *selected* set of unlabeled data with representation learnt at the current iteration and adopts the cluster assignments as pseudo-labels to refine subsequent representations. However, PSS differs in that, within each iteration, we propose a novel sample selection method to gather samples which are closely clustered together via a density criterion.

Our key insight is that under such a selection criterion over the clustering density in the feature space, we choose samples signaling compact intra-class distance distributions and thus a higher homogeneity to reduce noise in the pseudo-labels. In traditional SSL methods, samples with high pseudo-label quality are often those represented confidently by known classes, or "close" to known classes semantically. However, we find that our selection method is also able to sample, with high quality, from dis-similar novel classes whose class centroids are far-apart in the feature space. These samples from distant novel classes help improve the model generalization to disjoint unseen classes at test-time. Compared to adding all unlabeled samples at once, our progressively selected samples improve generalization more effectively, as shown in Figure 4.

We test our method on two open-world metric learning tasks: image retrieval for natural species, where the task is given a query image, to find nearest neighbor images across animal species, and 1:1 face verification, which classifies a pair of faces as being from the same person or different. The proposed progressive sample selection and representation learning method outperforms state-of-the-art semi-supervised learning and open-world representation learning methods. Specifically, it improves the Recall@1 performance from 55% to 57% over the image retrieval benchmark for nature species, and reduces the False Non Match Rate (FNMR) @ 1e-4 False Match Rate (FMR), from 22% to 21%, for face verification, relative to SOTA methods, as shown in Table 3 and 4.

To summarize, the key contributions of our method are as follows: 1) we formalize a practical open-world representation learning setting that reduces the gap between existing settings in the literature to the real-world application; 2) we propose a novel iterative method that progressively selects, at each iteration, samples that are most effective in improving representation generalization over test-time unseen classes; 3) we outperform state-of-the-art semi-supervised and novel class discovery methods using labeled and unlabeled data under our

practical setting for open-world representation learning, over the nature species image retrieval and 1:1 face verification tasks.

## 2 Related Work

Our work is closely related to semi-supervised learning and novel class discovery, we review literature in these two fields in the following discussion. We also discuss different sample selections methods and literature that iterate between feature learning and pseudo labeling as we do in this work.

**Semi-Supervised Learning.** Traditional SSL follows a closed-set setting which assumes the same set of classes in labeled and unlabeled data. The goal is to improve in-distribution classification performance with the help of an unlabeled dataset. The core challenge is in on how to leverage unlabeled data, which is roughly categorized into the following spectrums: consistency regularization [28,32,35,42], pseudo-labeling [20,29,44], generative methods [25,27,8] and graph based methods [2,23]. However, when the unlabeled set contains out-of-distribution (OOD) samples, termed as open-set SSL [26], traditional approaches inevitablely suffer performance degradation [22,34,26]. To mitigate this adverse impact, recent works [49,11,31,16] proposed to detect and down play OOD samples for classification performance of in-distribution data. We encourage readers to refer [36,48,22] for more comprehensive review of SSL literature.

**Novel Class Discovery.** A related line of work is novel class discovery (NCD) [13]. Different from closed-set or open-set SSL which leverages an unlabeled image set to improve learning performance on seen classes, NCD aims at discovering new classes in an unlabeled image set, assuming disjoint classes for labeled and unlabeled images. In order to transfer the representation from the labeled set to the unlabeled one, prevailing works [12,18,50] opt to learn transformation invariant features from labeled set first then pairwise relationship among unlabeled samples. These two-step approaches are unified into one single objective in [10] with multi-view self-labeling strategy. NCD [13] makes the assumption that all unlabeled data comes from novel classes and the number of classes is known. To be less constrained, [37] introduced Generalized Category Discovery (GCD) and proposed contrastive training with semi-supervised k-means to cluster unlabelled data into seen and unseen classes. In contrast to novel class discovery, [5] considers an open-world setting, in which testing set contains both seen and unseen classes. It proposed to train a classifier with both supervised and pairwise unsupervised loss in a unified fashion. Different from [5], which use all the unlabeled data in training, we progressively select samples which are most informative, avoiding bring too many noisy labels into feature training.

**Sample Selection.** A good sample selection method [19] will make the curated dataset to contain less noise, class imbalance and redundancy. Such resultant

datasets will allow models trained over them to maximize the information gain from unlabeled data. When the labeling algorithm is not robust enough, it is beneficial to add those high confidence unlabeled samples to the labeled set. Confidence score [33] is one of the leading metrics to measure the quality of assigned labels. [30] applied confidence threshold to select a disjoint face identity set and assigned pseudo labels. Similarly, [44] proposed to use classification score on all unlabeled samples and selected top-K examples of each target class. However, samples with such high scores do not help close the distribution gap between labeled and unlabeled data in our open world setting. Therefore, adding them to the training set makes no guarantee of improved recognition rates on unseen classes. [45] proposed a progressive labeling algorithm similar to ours. It selected the most representative samples by ranking the in-degree of nodes on a directed k-nearest neighbor (kNN) graph. Different from this handcrafted node similarity metric, our approach adopts a density criterion that exploits rich semantics between graph nodes and we employ a GNN clustering model [43] which learns such a criterion with supervision from class labels.

**Iteration between clustering and feature learning.** Some existing research studies model the clustering (pseudo labeling) and feature learning into a unified framework [7,46,21,41,6]. These two tasks are usually solved in an alternative fashion under the same objective, leading to iterative methods similar to ours. [7,46,21,41] assign pseudo labels to all unlabeled samples and apply them for feature learning in the next step. However, due to the imperfect performance of clustering, this can easily bring noisy labels into feature learning. Deep clustering [6], on the other hand, proposed to sample data from a uniform distribution over the classes to circumvent the issue caused by class imbalance. Different from these methods, we proposed to sample unlabeled samples progressively based on their clustering density to avoid noisy labels being used in feature learning.

## 3 Methods

### 3.1 Problem Formalization

We formalize a practical setting for open-world representation learning. Given a partially labeled dataset $\mathcal{D}$, we define the seen-class set $\mathcal{C}_{\text{in}}$, the unseen-classes set $\mathcal{C}_{\text{out}}$, and test class set $\mathcal{C}_{\text{test}}$. These three sets do not intersect with each other, and $|\mathcal{C}_{\text{in}}| \ll |\mathcal{C}_{\text{out}}|$. The training set consists of a labeled set $\mathcal{L}_{\text{in}}$ with $\mathcal{C}_{\text{in}}$ labels, and an unlabeled set $\mathcal{U} = \mathcal{U}_{\text{in}} \cup \mathcal{U}_{\text{out}}$, where $\mathcal{U}_{\text{in}}$ has $\mathcal{C}_{\text{in}}$ labels and $\mathcal{U}_{\text{out}}$ has $\mathcal{C}_{\text{out}}$ labels. The split of $\mathcal{U}_{\text{in}}$ and $\mathcal{U}_{\text{out}}$ is not known. We have $|\mathcal{L}_{\text{in}}| \ll |\mathcal{U}|$, and $|\mathcal{L}_{\text{in}}| : |\mathcal{U}| \approx 1 : 10$ in our setting. Figure 2 illustrates the dataset split. We aim to train a feature extractor $f$ to obtain generalized representations.

Compared with existing open-world semi-supervised learning and novel class discovery settings [13,10], ours differs in the aspects: (1) the unlabeled data is from both seen and unseen classes, instead of from only seen classes; (2) the unseen class number is not provided. Compared with the most similar existing

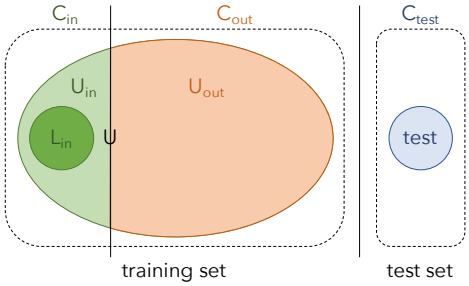



Fig. 2: Illustration of our practical open-world representation learning setting. We have a labeled training set $L_{\text{in}}$ with seen classes $\mathcal{C}_{\text{in}}$, and an unlabeled training set $\mathcal{U} = \mathcal{U}_{\text{in}} \cup \mathcal{U}_{\text{out}}$ with both seen classes $\mathcal{C}_{\text{in}}$ and unseen classes $\mathcal{C}_{\text{out}}$, only the images are accessible and the discrimination of $\mathcal{U}_{\text{in}}$ and $\mathcal{U}_{\text{out}}$ is unknown. The test set is from disjoint test classes $C_{\text{test}}$. We train on the labeled data $\mathcal{L}_{\text{in}}$ with ground truth labels; and the unlabeled data $\mathcal{U}$ with generated pseudo labels. The unlabeled data are selectively added during iterations

setting in [37]: (1) We have much larger $|\mathcal{U}| : |\mathcal{L}_{\text{in}}|$ and $|\mathcal{C}_{\text{out}}| : |\mathcal{C}_{\text{in}}|$ ratios; and (2) test on a hold out test set. The test images and classes are not accessible during training. In terms of the above differences, our setting is more challenging and of practical significance. Specifically, the high unlabeled to labeled data ratio leads to large distribution gap between the unlabeled and labeled data. Testing on a class-disjoint set requires a highly generalized model. The setting is also closer to the real world use scenarios, such as face verification and image retrieval.

### 3.2 Progressive Sample Selection (PSS) Pipeline

We propose a novel *Progressive Sample Selection* (PSS) approach to tackle the challenges in our practical open-world representation learning setting. We design a progressive pipeline to gradually expand the set of images used to train our model with the goals of being more robust to out-of-distribution unseen-class data. We select samples based on their cluster density, selecting samples in clusters with high local homogeneity. This selects points with less noise at both small and large distances from seen classes. We show that by continuing to add samples at each training iteration with our selection method, we expand the feature space and improve the model's generalization ability.

**PSS pipeline.** The overall PSS pipeline for open-set representation learning is shown in Figure 3. The training set $\mathcal{T}_0$ is initialized as the labeled set $\mathcal{L}_{\text{in}}$ and expands during iterations. Each iteration contains three steps, in iteration $i$:

1. **Representation Learning.** Train a feature extractor $f_i$ to learn representations on the training set $\mathcal{T}_i$. Note $f_i$ is retrained in each iteration instead of finetuning on top of $f_{i-1}$ to alleviate overfitting and avoid local optima.

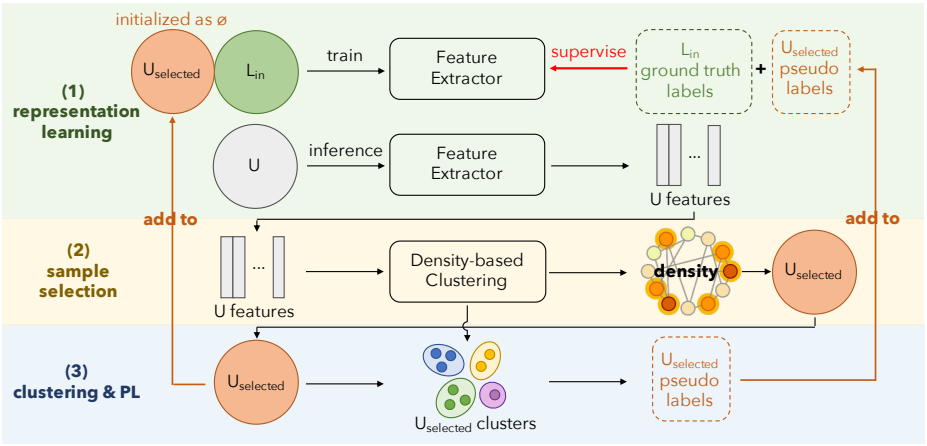



Fig. 3: The overall pipeline of Progressive Sample Selection (PSS). We progressively repeat three steps to train a generalized feature extractor with samples increasingly added to the training set in iterations. The feature extractor and density-based clustering model are shared in different steps. Best viewed in color

2. **Sample selection.** Estimate the density of the feature space defined by $f_i$ for each sample in the unlabeled training set $\mathcal{U}_i$. Select samples for inclusion in training the next iteration feature representation $f_{i+1}$ that have a density estimate above a threshold $\tau$ and pass the selection together with $\mathcal{U}_i$ to the clustering step for pseudo labeling as $\mathcal{U}_{\text{selected}}$. The details are in Section 3.3.
3. **Clustering and pseudo labeling.** Cluster samples in $\mathcal{U}$ and assign pseudo labels to points in $\mathcal{U}_{\text{selected}}$ corresponding to their cluster assignment. Then update the training set $\mathcal{T}_{i+1} = \mathcal{T}_i \cup \mathcal{U}_{\text{selected}}$ and the unlabeled set $\mathcal{U}_{i+1} = \mathcal{U}_i \setminus \mathcal{U}_{\text{selected}}$, preparing for iteration $i+1$ training. The details of clustering and pseudo labeling are in Section 3.4.

The above three steps are looped until the number of selected samples which are far away from the training set on the feature space is small. Finally, we re-train the feature extractor to learn better representations for retrieval or recognition.

### 3.3 Sample Selection

Due to the large unlabeled to labeled data ratio in our setting, a large class distribution gap exists between the unlabeled and labeled data. Thus, it is hard to get reasonable representations and clustering results for the subset of out-of-distribution data. In order to get high-quality training data and pseudo labels, we propose a novel sample selection method outlined above. We leverage Hi-LANDER [43] to both estimate the sample density and perform clustering but in principle these two systems could be independent estimators.

Hi-LANDER [43] is a hierarchical graph neural network model for image clustering which learns the grouping and model selection criteria in traditional

clustering methods. It recurrently builds k-nearest-neighbor graphs over nodes which are grouped from connected components at different level of the hierarchy. It defines *Density* as the proportion of same-class neighbors weighted by similarity. Specifically, the estimated density $\hat{d}_i$ for the $i$-th sample is:

$$\hat{d}_i = \frac{1}{k} \sum_{j \in \mathcal{N}(i)} \hat{e}_{ij} \cdot a_{ij}, \quad \hat{e}_{ij} = P(y_i = y_j) - P(y_i \neq y_j) \tag{1}$$

where $\mathcal{N}(i)$ refers to the neighbors of sample $i$, $\hat{e}_{ij}$ is the edge linkage probability of sample $i$ sharing the same class as its neighbour $j$, and $a_{ij}$ is the feature similarity between the two. Further details can be found in Equation (4) in [43]. A sample with high density therefore has a neighborhood that contains more consistent labels as itself, exhibiting a higher homogeneity and less noisy labels in the clustering pseudo-labels.

In each iteration, samples are selected through the above defined density metric with a threshold $\tau$. It is observed that using such as density selection criterion enforcing high homogeneity (thus high pseudo-label quality), we not only select samples that are close to known classes but also those that are far-apart from the centroids of existing classes in the feature space to close the distribution gap between labeled and unlabeled data, allowing improved generalization to unseen disjoint test-time classes (Figure 5).

### 3.4 Clustering and Pseudo Labeling

After obtaining the $\mathcal{U}_{\text{selected}}$ from $\mathcal{U}_i$ with the density metric, we then generate pseudo labels for them using the clustering results of Hi-LANDER. Given the clusters, we assign one new class label to all selected unlabeled samples in each cluster. Assume the current training set $\mathcal{T}_i$ has $m$ existing classes, samples in $\mathcal{U}_{\text{selected}}$ will then have their pseudo labels indexed from $m+1$. Note that for two same-class samples, this process might assign different pseudo labels to them. However, according to [17,10], over-clustering does not harm the downstream task performance, thus the class-split is acceptable.

With the selected samples and pseudo labels, the training set $\mathcal{T}_i$ is expanded to $\mathcal{T}_{i+1} = \mathcal{T}_i \cup \mathcal{U}_{\text{selected}}$. Finally, we re-train the feature extractor on $\mathcal{T}_{i+1}$ to learn more generalized representations.

## 4 Experiments

We evaluate PSS on fine-grained natural image retrieval and face verification benchmarks. First, we show ablation experiments over the design choices of PSS and demonstrate their significance in improving feature generalization. We then illustrate the performance comparison of PSS to state-of-the-art SSL and novel class discovery methods over the fine-grained nature species image-retrieval benchmark. Finally, we demonstrate the performance improvement of PSS in open-set face verification benchmarks.

### 4.1 Evaluation Protocols

**Datasets.** For fine-grained natural image retrieval, we use iNaturalist [15] dataset, which contains a training set of 325,846 images across 5,690 classes and a disjoint test set of 136,093 images across 2,452 test classes. We randomly sample about 16% of the training classes as the seen classes $\mathcal{C}_{\text{in}}$, and take 60% of the images from each class in $\mathcal{C}_{\text{in}}$ as the labeled training set $\mathcal{L}_{\text{in}}$. The rest of the samples in the training set are used as the unlabeled training set $\mathcal{U}$. The test set has disjoint samples and classes to the training set. The attributes of three sets are shown in Table 1. The labeled training set $\mathcal{L}_{\text{in}}$ has about 9% samples over all the training samples. The ratios of both the labeled classes and the labeled samples are much smaller than previous open-set semi-supervised learning and novel class discovery settings.

For face verification training, we used the combined IMDB [39] and DeepGlint [1] datasets. The IMDB consists of 1.3 Million images with 49,990 identities and the DeepGlint dataset contains around 6.2 Million images with 180k identities. We randomly sampled 90% of the IMDB data as $\mathcal{L}_{\text{in}}$ and the rest are treated as $\mathcal{U}$. Similar to iNatualist experiments, we divide the dataset into $\mathcal{L}_{\text{in}}$ and $\mathcal{U}$ sets and the statistics of these sets can be found in Table 1. There is roughly a 1:6 ratio between the labeled and unlabeled data. We evaluate our method using the IJB-C [24] face verification benchmarks on the 1:1 face verification task, which contains about 3,531 identities and 140k images. The IJB-C benchmark is disjoint from the IMDB and the DeepGlint data that we use for training.

**Metrics.** For image retrieval and face verification, we respectively use *Recall@k* (higher the better) and False Non-Match Rate (FNMR) at False Match Rate (FMR) equaling 1e-4 (lower the better) as the metric to evaluate our method.

### 4.2 Implementation Details

For image retrieval, we use Smooth-AP loss [4] to finetune a ResNet-50 [14] backbone. The embedded dimension is 128. The learning rate is 1e-5. We train the feature extractor on a single machine with 8 NVIDIA T4 GPUs. The hidden dimension of Hi-LANDER is 512, and use GAT [38] as the base graph neural network model. The $k$ expansions in $k$-NN are 10,5,3.

For face verification, we use CosFace [40] for training our face embedding model. The embedding dimension is 128 and learning rate starts at 0.1 and decreases according to a cosine learning rate schedule. We train the embedding for 32 epochs on a distributed training system with 8 nodes, each with 8 NVIDIA Tesla V100 GPUs. We use the same Hi-LANDER setting as in image retrieval.

Section 4.3 describes the sample selection threshold and stopping criterion.

### 4.3 Ablation Experiments

We examine the effectiveness of our progressive system design by ablating our sample selection and progressive refinement. From Table 2, notice removing either our sample selection or progressive refinement results in a regression in

performance. Note that for progressive methods, the training time expands linearly with the number of iterations and thus the accuracy improvements do come at the cost of training time. The sample selection, however, not only brings performance gains but also reduces the number of samples being trained, in-turn reducing training time. We illustrate these components in more detail on the nature species image retrieval benchmark below.

Table 1: The labeled, unlabeled and class-disjoint test dataset attributes

| split | iNarturalist | | Face | |
|---|---|---|---|---|
| | #images | #classes | #images | #classes |
| $\mathcal{L}_{\text{in}}$ | 29,011 | 948 | 1,124,874 | 49990 |
| $\mathcal{U}$ | 296,835 | 5,690 | 6,323,702 | 209,551 |
| test | 136,093 | 2,452 | 141,139 | 3,531 |

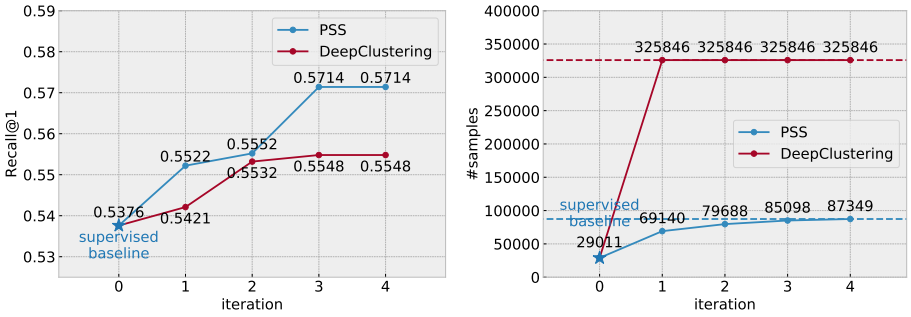
Table 2: Ablation of sample selection and iteration in our method. Both improve the performance. Sample selection also improves training efficiency

| sample selection | iterate | #training data | #iterations | Recall@1 | Recall@4 | Recall@16 | Recall@32 |
|---|---|---|---|---|---|---|---|
| × | × | 325,846 | 0 | 0.5421 | 0.7128 | 0.8318 | 0.8755 |
| ✓ | × | 69,140 | 0 | 0.5522 | 0.7224 | 0.8413 | 0.8848 |
| × | ✓ | 325,846 | 4 | 0.5548 | 0.7242 | 0.8407 | 0.8839 |
| ✓ | ✓ | 87,349 | 4 | **0.5714** | **0.7357** | **0.8501** | **0.8914** |

**Progressive pipeline.** The proposed open-world representation learning setting is challenging considering the large ratio between unlabeled data and labeled data. The base feature learned from the labeled data has limited representation ability on the unlabeled data and thus the pseudo-labels suffer more noise. Our progressive pipeline continually updates the feature and improves the pseudo-label quality at each iteration. This assumption is verified in Table 2 and Table 3, where our iterative method outperforms the one-step baselines. Figure 4 shows the retrieval performance at each iteration of PSS compared with the DeepClustering [6] iterative method. We notice DeepClustering stops improving after 2 iterations, while for PSS performance does not plateau until iteration 4.

**Sample selection.** The difference between our method and DeepClustering's [6] ability to continue to improve over iterations, lies in the process of sample selection. As mentioned in the progressive analysis, a large portion of the unlabeled data will be tagged with noisy pseudo-label since the generalization ability of the base feature is limited. Noisy pseudo-labels hurt model performance and diminishes the gain from correctly pseudo-labeled samples.

Usually, samples close to the existing labeled classes are easier to get high-quality pseudo-labels, since the learned feature are easier to generalize to those similar samples. Several existing sample selection works are adopting this intuition or its variants to keep pseudo-label quality, such as FixMatch [3], which only keeps samples assigned a high probability to a known class. In PSS, we use density defined in Hi-LANDER [43] as the selection metric. Intuitively, high density for a sample is an indicator of high-quality pseudo-label since the local intra-class homogeneity is kept. We can consider current feature "works" for that high-density sample by collecting same-class samples into its closest




(a) Retrieval performance. (b) Training data size.

Fig. 4: The retrieval performance (Recall@1) and training data size on iNaturalist in different iterations. The performance improves during iterations with more training data. DeepClustering is almost converged after iteration 2 while PSS still witnesses a performance boost at iteration 3. Meanwhile, PSS uses fewer samples to train the feature representation benefiting from sample selection

neighborhood. High density does not necessarily mean close distance to labeled classes. In Figure 5, we indeed found the density and the distance-to-closest-labeled-class are not negatively correlated, making it possible to select some high-quality pseudo-labeled samples which differ from the labeled classes. This is vital to generalize to unseen data for the challenging open-world representation learning setting and a major differentiator with existing sample selection methods. To verify this statement, we split the selected samples to equal-sized two parts based on the distance-to-closest-labeled-class. Samples with distance higher than 0.15 (mean distance for all the selected samples) are collected in the "far" part, vice versa. After feature learning with these two parts separately, we found that the samples far-away from existing labeled classes bring more gain on the test set. The *Recall@1* on the feature learnt with "far" part is 0.5524, while the one from "close" part is 0.5490 in iteration 1.

**Density thresholding criterion.** The density threshold is one of the hyper-parameters of PSS. The threshold we use is the density inflection point of the approximated upper-bound density-distance curve, as illustrated in Figure 5 over the nature species image retrieval benchmark (the first iteration). It shows that the density of the inflection point is around 0.8 and thus we pick it as the sample selection threshold and keep it for subsequent iterations til convergence. For face verification, the same rule is applied and we select threshold of 0.9.

**Stopping criterion.** The stopping criterion is usually empirical. We define our stopping criterion as the portion of samples who's distance to the nearest training class centroid is large (greater than 0.15). This portion drops iteration by

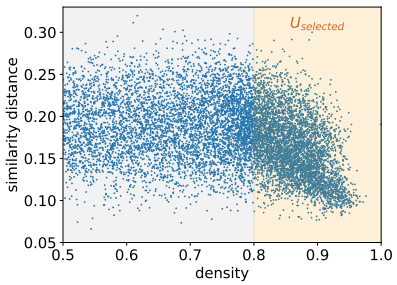



Fig. 5: The non-negative correlation between density and the cosine distance to the nearest training class centroid after iteration 1. Threshold 0.8 is selected for species retrieval

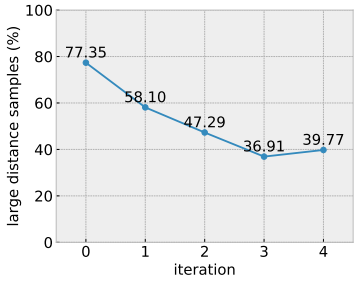



Fig. 6: The percentage of distances to the nearest training class centroid considered to be large (greater than 0.15) in $\mathcal{U}$ over different iterations

iteration. When it drops to a certain level, the benefit effect from the informative large distance samples will be negated by the negative effect from noisy pseudo-labels and thus we stop the first iteration we see this indicator rise. Figure 6 shows such stopping indicator value over the nature species retrieval task, which leads to stopping after the 4th iteration. For face verification, the same stopping criteria is applied and we stop at iteration 4.

### 4.4 Image Retrieval Results

We validate the effectiveness of PSS on natural images using fine-grained image retrieval as the downstream task.

We compare PSS with the supervised baseline trained with only labeled data $\mathcal{L}_{\text{in}}$, oracle trained with all training data $\mathcal{L}_{\text{in}} \cup \mathcal{U}$ with ground truth labels, traditional semi-supervised learning method Pseudo Labeling (PL) [20], open-set semi-supervised learning method UNO [10], one-step clustering method Hi-LANDER [43], and iterative clustering method DeepClustering [6]. UNO has classification heads for the training set but not for the disjoint-class test set, so we train a UNO model to generate pseudo labels for the unlabeled data $\mathcal{U}$, and finetune a pretrained feature extractor with Smooth-AP [4] loss. For Hi-LANDER, we apply a trained Hi-LANDER clustering model on $\mathcal{U}$ to generate pseudo labels and re-train the feature extractor with both $\mathcal{L}_{\text{in}}$ and $\mathcal{U}$. Here we do not compare with some other semi-supervised learning methods such as self-training [34] because they rely on the design of classification heads, which have difficulty expanding to the open-world retrieval task with unseen classes.

Table 3 shows the retrieval performance of PSS and the state-of-the-art SSL and novel class discovery methods on the hold out iNaturalist test set. All the previous methods, except UNO, boost the supervised baseline by training with more data. Though UNO performs excellently in terms of classification accuracy

Table 3: Retrieval performance on iNaturalist. PSS improves Recall@1 from 55.48% (DeepClustering) to 57.14%

| method | Recall@1 | Recall@4 | Recall@16 | Recall@32 |
|---|---|---|---|---|
| sup. baseline | 0.5376 | 0.7135 | 0.8359 | 0.8817 |
| oracle | 0.6554 | 0.8074 | 0.8966 | 0.9261 |
| PL[20] | 0.5447 | 0.7188 | 0.8398 | 0.8832 |
| Hi-LANDER[43] | 0.5421 | 0.7128 | 0.8318 | 0.8755 |
| UNO[10] | 0.5372 | 0.7138 | 0.8367 | 0.8808 |
| DeepClustering[6] | 0.5548 | 0.7242 | 0.8407 | 0.8839 |
| PSS | **0.5714** | **0.7357** | **0.8501** | **0.8914** |

Table 4: IJBC Face Verification. PSS reduces FNMR@FMR 1e-4 from best prior (DeepClustering)

| method | FNMR@FMR1e-4 |
|---|---|
| sup. baseline | 0.3007 |
| oracle | 0.0672 |
| DB-SCAN[9] | 0.4203 |
| GCN-V[47] | 0.2508 |
| Hi-LANDER[43] | 0.2472 |
| RoyChowdhury et al.[30] | 0.2706 |
| DeepClustering[6] | 0.2234 |
| PSS | **0.2165** |

on the unlabeled training set, it does not work well in our setting since the large unlabeled to labeled distribution gap and it over-fits on the labeled data. The one-step methods PL, Hi-LANDER, and UNO add all the unlabeled data at once thus introduce more noises to the pseudo labels. Training with the noisy pseudo labels regress the performance boost. DeepClustering iteratively train with all the unlabeled data in each iteration. Though the progressive refinement brings performance boost by increasing quality of both features and pseudo labels, the pseudo labels are still noisy thus the gain is limited and stops increasing after two iterations (see Figure 4). Benefiting from our progressive sample selection pipeline, PSS is able to steadily improve feature generalization to test-time disjoint novel classes over iterations due to its capability to choose distant samples which also exhibit high homogeneity and high pseudo-label quality. Although DeepClustering also achieves relatively high Recall@1 performance, it's much more costly than PSS in the feature training step. Even if we consider DeepClustering to have converged at iteration 2 while PSS stops at iteration 4, PSS improves training efficiency by 60% via training on less than 40% of the samples summed-up across all iterations compared to DeepClustering.

### 4.5 Face Verification Results

Figure 7 shows the performance gain of PSS over iterations and Figure 8 shows the progressive number of samples selected. From iteration 1 to 3, the face verification performance steadily improves, with convergence at iteration 4.

We compare PSS with state-of-the-art semi-supervised and open-world representation learning methods [43,9,30,6,47] in Table 4. For DB-SCAN [9], GCN-V [47], Hi-LANDER [43], and RoyChowdhury et al. [30] which use clustering and pseudo-labeling, we perform a one-step feature training, where all unlabeled samples are gathered to generate pseudo-labels and used for training at once. For DeepClustering [6], we recurrently run the feature training with the same backbone architecture as PSS and pseudo-labeling via Hi-LANDER clustering in an alternating manner, however, it lacks the progressive sample selection. Due to its high training cost with full unlabeled and labeled data within each iter-

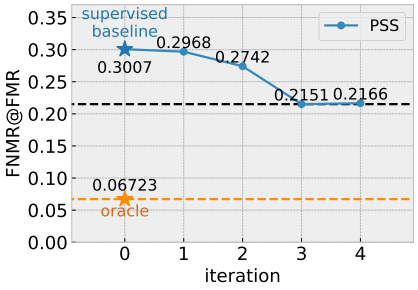



Fig. 7: IJBC Face Verification FNMR @ FMR 1e-4 (the lower, the better)

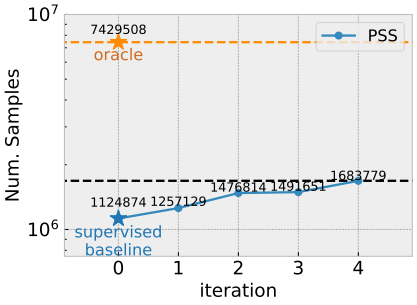



Fig. 8: Number of sampled selected by PSS for face verification

ation (>7 million samples), we stop DeepClustering after iteration 2. PSS, on the other hand, uses a much more efficient number of samples across all iterations (<7 million in total). PSS significantly improves feature generalization over methods that directly use all samples. In addition, PSS achieves a training efficiency boost over DeepClustering, where it reduces FNMR@FMR=1e-4 from 22.34% to 21.66% with training over only $< 50\%$ of the samples summed-up across iterations.

## 5 Conclusion

We propose a practical open-world representation learning setting to learn unseen category representations with partially labeled data. Our setting has large unlabeled to labeled data ratio, no prior knowledge over the number of unseen classes during training, and no access to the images that have the same labels as the test set. Existing open-world representation learning methods fail to generalize to the proposed setting because of the large distribution gap between the unlabeled and labeled data. To tackle this challenging setting, we propose a novel *Progressive Sample Selection* (PSS) approach to improve representation generalization by iteratively training with increasingly effective samples selected during iterations. We use estimated density in Hi-LANDER clustering model to select samples that are densely clustered together with high local intra-class homogeneity. These samples can be from novel classes that are far from the existing categories, thus help improve the model generalization to test-time disjoint unseen classes. Experiments indicate that our method outperforms the state-of-the-art semi-supervised learning methods and novel class discovery methods in natural image retrieval and face verification.

## References


1. http://trillionpairs.deepglint.com/overview
2. Bengio, Y., Delalleau, O., Le Roux, N.: 11 label propagation and quadratic criterion (2006)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems **32** (2019)
4. Brown, A., Xie, W., Kalogeiton, V., Zisserman, A.: Smooth-ap: Smoothing the path towards large-scale image retrieval. In: European Conference on Computer Vision. pp. 677–694. Springer (2020)
5. Cao, K., Brbić, M., Leskovec, J.: Open-world semi-supervised learning. In: ICLR (2022)
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018)
7. Culp, M., Michailidis, G.: An iterative algorithm for extending learners to a semisupervised setting. In: Joint Statistical Meetings (2007)
8. Denton, E., Gross, S., Fergus, R.: Semi-supervised learning with context-conditional generative adversarial networks. arXiv preprint arXiv:1611.06430 (2016)
9. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
10. Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9284–9292 (October 2021)
11. Guo, L.Z., Zhang, Z.Y., Jiang, Y., Li, Y.F., Zhou, Z.H.: Safe deep semi-supervised learning for unseen-class unlabeled data. In: International Conference on Machine Learning. pp. 3897–3906. PMLR (2020)
12. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. arXiv preprint arXiv:2002.05714 (2020)
13. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8401–8409 (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778. IEEE Computer Society (2016)
15. Horn, G.V., Aodha, O.M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.J.: The inaturalist species classification and detection dataset. In: CVPR. pp. 8769–8778. Computer Vision Foundation / IEEE Computer Society (2018)
16. Huang, J., Fang, C., Chen, W., Chai, Z., Wei, X., Wei, P., Lin, L., Li, G.: Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8310–8319 (2021)
17. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

18. Jia, X., Han, K., Zhu, Y., Green, B.: Joint representation learning and novel category discovery on single-and multi-modal data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 610–619 (2021)
19. Killamsetty, K., Zhao, X., Chen, F., Iyer, R.: Retrieve: Coreset selection for efficient and robust semi-supervised learning. Advances in Neural Information Processing Systems **34** (2021)
20. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
21. Liao, R., Schwing, A., Zemel, R., Urtasun, R.: Learning deep parsimonious representations. In: NeurIPS (2016)
22. Luo, H., Cheng, H., Meng, F., Gao, Y., Li, K., Zhang, M., Sun, X.: An empirical study and analysis on open-set semi-supervised learning. arXiv preprint arXiv:2101.08237 (2021)
23. Luo, Y., Zhu, J., Li, M., Ren, Y., Zhang, B.: Smooth neighbors on teacher graphs for semi-supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8896–8905 (2018)
24. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P.: Iarpa janus benchmark - c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165 (2018). https://doi.org/10.1109/ICB2018.2018.00033
25. Odena, A.: Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583 (2016)
26. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. Advances in neural information processing systems **31** (2018)
27. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
28. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. Advances in neural information processing systems **28** (2015)
29. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models (2005)
30. RoyChowdhury, A., Yu, X., Sohn, K., Learned-Miller, E., Chandraker, M.: Improving face recognition by clustering unlabeled faces in the wild. In: European Conference on Computer Vision. pp. 119–136. Springer (2020)
31. Saito, K., Kim, D., Saenko, K.: Openmatch: Open-set semi-supervised learning with open-set consistency regularization. Advances in Neural Information Processing Systems **34** (2021)
32. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
33. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
34. Su, J.C., Cheng, Z., Maji, S.: A realistic evaluation of semi-supervised learning for fine-grained classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12966–12975 (2021)

35. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
36. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning **109**(2), 373–440 (2020)
37. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. arXiv preprint arXiv:2201.02609 (2022)
38. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJXMpikCZ
39. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Loy, C.C.: The devil of face recognition is in the noise. arXiv preprint arXiv:1807.11649 (2018)
40. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
41. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: ICML (2016)
42. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems **33**, 6256–6268 (2020)
43. Xing, Y., He, T., Xiao, T., Wang, Y., Xiong, Y., Xia, W., Wipf, D., Zhang, Z., Soatto, S.: Learning hierarchical graph neural networks for image clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3467–3477 (October 2021)
44. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019)
45. Yan, X., Chen, R., Feng, L., Yang, J., Zheng, H., Zhang, W.: Progressive representative labeling for deep semi-supervised learning. arXiv preprint arXiv:2108.06070 (2021)
46. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: CVPR (2016)
47. Yang, L., Chen, D., Zhan, X., Zhao, R., Loy, C.C., Lin, D.: Learning to cluster faces via confidence and connectivity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
48. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. arXiv preprint arXiv:2103.00550 (2021)
49. Yu, Q., Ikami, D., Irie, G., Aizawa, K.: Multi-task curriculum framework for open-set semi-supervised learning. In: European Conference on Computer Vision. pp. 438–454. Springer (2020)
50. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10867–10875 (2021)