

Rethinking Open-World Semi-Supervised Learning: Distribution Mismatch and Inductive Inference

Seongheon Park* Hyuk Kwon* Kwanghoon Sohn Kibok Lee
Yonsei University

{sam121796, kh12043, khsohn, kibok}@yonsei.ac.kr

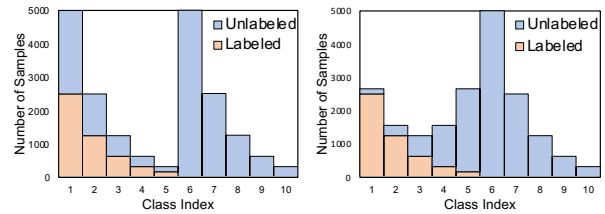
Abstract

Open-world semi-supervised learning (OWSSL) extends conventional semi-supervised learning to open-world scenarios by taking account of novel categories in unlabeled datasets. Despite the recent advancements in OWSSL, the success often relies on the assumptions that 1) labeled and unlabeled datasets share the same balanced class prior distribution, which does not generally hold in real-world applications, and 2) unlabeled training datasets are utilized for evaluation, where such transductive inference might not adequately address challenges in the wild. In this paper, we aim to generalize OWSSL by addressing them. Our work suggests that practical OWSSL may require different training settings, evaluation methods, and learning strategies compared to those prevalent in the existing literature.

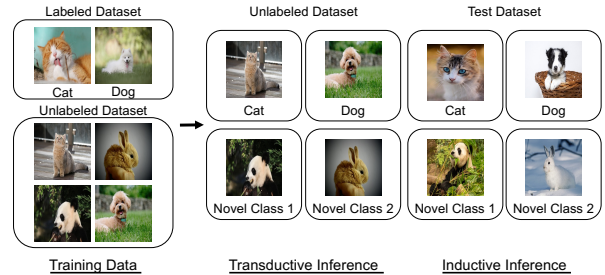
1. Introduction

OWSSL has been introduced to discover novel classes within an unlabeled dataset while accurately classifying known classes. However, we argue that OWSSL may not reflect real-world scenarios for the following reasons: 1) recent works on OWSSL assume balanced and identical class prior distribution between labeled and unlabeled datasets during the learning process, and 2) they only consider a transductive learning setting, which focuses on categorizing instances from the unlabeled training datasets.

Indeed, in-the-wild data naturally follow a long-tailed distribution and are exposed to label distribution shifts [25, 36], *i.e.*, labels are missing not at random (MNAR; Fig. 1a right) rather than missing completely at random (MCAR; Fig. 1a left). Class prior distribution mismatch between labeled and unlabeled datasets happens for multiple reasons, *e.g.*, the data distribution itself could change over time, or annotators might prefer to annotate relatively easy classes or they could miss difficult classes. However, most OWSSL methods assume a balanced class prior for training, which often hampers performance when the assumption does not hold. Also, most OWSSL methods assume a transductive



(a) In ROWSSL, we consider the cases when the class prior of labeled and unlabeled datasets are matched (left) and mismatched (right).



(b) Examples of transductive and inductive inference in ROWSSL. Inductive inference is performed without looking at other test data.

Figure 1. Examples of scenarios considered in ROWSSL.

learning setting, which is specialization on given unlabeled training data rather than generalization on unseen test data as illustrated in Fig. 1b. While transductive learning is useful for category discovery, it does not guarantee reliable performance when classifying discovered categories from unseen test data. Instead, inductive learning is important in safety-critical applications such as medical diagnosis, *e.g.*, a model that can discover novel diseases in a specific patient cohort might still misclassify diseases in unseen patients.

To this end, we extend OWSSL by addressing such practical training and evaluation settings, coined **Realistic Open-World Semi-Supervised Learning (ROWSSL)**. In this task, we consider long-tailed distribution with class prior distribution mismatch between labeled and unlabeled datasets for training, and inductive and transductive inferences for evaluation. To address the aforementioned challenges, we introduce **Density-based Temperature scaling and Soft pseudo-labeling (DTS)** to learn class-balanced representations taking account of local densities and reduce

* Authors contributed equally to this work.

classifier bias toward the head and known classes simultaneously. To achieve this, we propose to measure the tailedness as a proxy for the unknown class prior via density estimation on the representation space. With these proxies, we introduce a dynamic temperature scaling approach for balanced contrastive learning, which dynamically adjusts the temperature parameter for each anchor to shape a representation space to have better linear separability between head and tail classes. Also, we address the classifier bias via class uncertainty-aware soft pseudo-labeling, considering a density variance as an uncertainty measure.

2. ROWSSL

2.1. Training setting

Suppose we have a partially labeled dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, where $\mathcal{D}_l = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_l} \in \mathcal{X} \times \mathcal{Y}_l$ is the labeled dataset with N_l samples which belongs to one of the C_{old} known classes, and $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=1}^{N_u} \in \mathcal{X}$ is the unlabeled dataset with N_u samples, with its underlying label space \mathcal{Y}_u containing both of C_{old} known classes and C_{new} novel classes. In GCDW, the labeled dataset \mathcal{D}_l has a long-tailed distribution with an imbalance ratio $\gamma_l > 1$, while the unlabeled dataset can have an arbitrary class prior, including MCAR and MNAR scenarios as depicted in Fig. 1a. Following [44, 51], the total number of classes $C = C_{\text{old}} + C_{\text{new}}$ is either assumed to be known a priori or estimated through off-the-shelf methods [20]. Our objective is to train a parametric classifier $f : \mathbb{R}^d \mapsto [0, 1]^C$ to correctly assign class labels to both known and novel classes.

2.2. Evaluation setting and metrics

In Table 1, we compare the balanced overall accuracy of previous methods with different evaluation strategies on CIFAR-100-LT, where ‡ indicates the result aligned with [3], with a maximum discrepancy of 1.3%. The ratio of the number of known and novel classes is 80:20 for Split 1 and 50:50 for Split 2.

Transductive inference. Prior works [6, 44] have performed transductive inference for their methods on the unlabeled training dataset (“Train” evaluation set in Table 1). Following them, we measure the clustering accuracy between the ground truth labels y_i and the model’s predictions \hat{y}_i through the Hungarian algorithm [29]:

$$\text{ACC} = \frac{1}{|\mathcal{D}_u|} \sum_{i=1}^{|\mathcal{D}_u|} \mathbb{1}\{y_i = p^*(\hat{y}_i)\}, \quad (1)$$

where p^* is the optimal permutation that matches the predicted cluster assignments to the ground truth class labels. While transductive learning is useful for category discovery, it does not guarantee the reliable performance of the learned model to classify discovered categories in unseen test data.

Table 1. Comparison of different evaluation strategies.

Data split	Split 1 [3]				Split 2			
	Train	Test			Train	Test		
Recluster	-	✓	✗	✗	-	✓	✗	✗
Rematch	-	✓	✓	✗	-	✓	✓	✗
k -means	37.8	55.0	38.7	37.7	34.2	55.0	37.0	30.3
ORCA	38.9	51.2 [‡]	49.1	42.4	25.0	34.9	32.0	29.5
GCD	49.5	63.5 [‡]	50.3	48.5	42.3	54.4	40.5	38.1
SimGCD	49.2	61.3	52.5 [‡]	44.9	46.5	50.2	42.9	37.4
BaCon	50.8	67.9 [‡]	50.7	47.5	38.0	59.2	45.2	35.9
Ours	54.1	61.7	55.8	52.1	53.7	51.9	53.7	48.1

Transductive inference on the test set. BaCon [3] utilizes a balanced test dataset following the common practice in long-tailed recognition. However, they perform k -means clustering on the entire test dataset for evaluation, *i.e.*, the classification result depends on other test data, which corresponds to transductive inference (“Recluster ✓ and Rematch ✓” in Table 1). Also, they ignore the clusters found during training and match the clusters of the test set with the classes, resulting in unintentional concept shifts, *e.g.*, the cat class during training might be matched with the lion class at test time. Hence, their evaluation results do not properly reflect the generalizability of the models to online inference, which is often required in real-world scenarios, and they cannot identify the semantics of classes. Furthermore, k -means assumes the presence of the uniform cluster prior, which leads to biased results in relation to the balanced test set statistics [2]. We found that the high performance of BaCon might be due to the uniform prior assumption of k -means and concept shifts by rematching for the best performance, *e.g.*, when evaluated on the imbalanced training dataset, BaCon is on par with other methods in Split 1, and outperformed by other methods in Split 2.

Inductive inference. To evaluate the generalizability of models, we consider inductive inference. Specifically, we evaluate the models on the disjoint test dataset by nearest centroid classification, where the center of clusters found by optimal matching p^* from (1) on the training set are utilized as parametric class centers (“Recluster ✗ and Rematch ✗” in Table 1). To confirm that concept shifts are beneficial to maximize the performance, we also apply Hungarian matching between the parametric clustering results with the classes (“Recluster ✗ and Rematch ✓” in Table 1). While rematching results in better performance, this ignores the semantics of categories discovered during training. In fact, rematching corresponds to transductive inference, as it requires gathering the parametric clustering results. Throughout experiments, we focus on evaluation without reclustering and rematching for inductive inference.

3. Proposed Method

We propose an end-to-end approach that jointly learns the representation and parametric classifier, similar to Wen *et al.* [51]. The network architecture is composed of an en-

coder E followed by two heads f and g . The encoder E can be a pre-trained model, *e.g.*, a ViT pre-trained with DINO [7], $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$ is a feature vector representing the input image \mathbf{x} , f is an ℓ_2 -normalized linear classifier, and g is a multi-layer perceptron (MLP) projecting \mathbf{z} to a lower dimensional vector \mathbf{h} for representation learning.

3.1. Training objectives

Representation learning. We adopt contrastive learning (CL) loss for representation learning. From a mini-batch B , two views of an image are obtained through random augmentation, represented as \mathbf{x} , and \mathbf{x}' . These images are then fed into the query and key networks $E \circ g$ and $E' \circ g'$, yielding a pair of ℓ_2 -normalized embeddings $\mathbf{h} = (E \circ g)(\mathbf{x})$ and $\mathbf{b} = (E' \circ g')(\mathbf{x}')$, respectively, where the key network is updated by exponential moving average (EMA), following MoCo [22]. Self-supervised learning loss is defined as:

$$l_u(\mathbf{x}_i) = -\log \frac{\exp(\mathbf{h}_i \cdot \mathbf{b}_+ / \tau)}{\sum_{\mathbf{b}' \in \mathbf{Q}} \exp(\mathbf{h}_i \cdot \mathbf{b}' / \tau)}. \quad (2)$$

Here, \mathbf{b}_+ is a positive key, and the queue $\mathbf{Q} = \{\mathbf{b}_j\}_{j=1}^Q$ is updated sequentially with key embeddings \mathbf{b} following the first-in-first-out (FIFO) scheme, where Q is the predefined queue size. $\mathbf{Q} = \{\mathbf{b}_j\}_{j=1}^Q$ is a queue that contains the key embeddings \mathbf{b} of a predefined size Q . For effective utilization of label information, we adopt the variation of the supervised contrastive loss $l_{\text{sup}}(\mathbf{x}_i, \mathbf{y}_i)$ [27] which maintains multiple positive pairs on-the-fly by comparing the query label to a label queue [11]. Overall representation learning loss is defined as:

$$L_{\text{rep}} = (1 - \lambda_{\text{rep}}) \frac{1}{|B|} \sum_{i \in B} l_u(\mathbf{x}_i) + \lambda_{\text{rep}} \frac{1}{|B_l|} \sum_{i \in B_l} l_{\text{sup}}(\mathbf{x}_i, \mathbf{y}_i), \quad (3)$$

where B_l corresponds to the labeled subset of B and λ_{rep} is a balancing factor.

Classifier learning. Our parametric classification framework follows the self-distillation methods [7]. We employ a prototypical classifier where the weight parameters of linear classifier f are regarded as cluster centroids. To discover novel classes and allocate each sample to the optimal cluster, we condition the cluster centroids to contain class information through multi-tasking self-supervised and supervised objectives [17]. Classification loss is defined as:

$$l_{\text{cls}}(\mathbf{x}_i, \mathbf{y}_i) = -\sum_{k=1}^C \bar{\mathbf{y}}_i^k \log(\mathbf{p}_i^k), \quad \bar{\mathbf{y}}_i = \begin{cases} \mathbf{y}_i, & \mathbf{x}_i \in \mathcal{D}_l, \\ \mathbf{q}_i, & \mathbf{x}_i \in \mathcal{D}_u, \end{cases} \quad (4)$$

where $\mathbf{p} = \text{softmax}(f(\mathbf{z})/\tau_s)$ is the temperature-scaled softmax probability with τ_s , \mathbf{y} is a one-hot representation

of the ground-truth label, and the soft pseudo-label $\mathbf{q} = \text{softmax}(f(\text{sg}(\mathbf{z}'))/\tau_t)$ is produced by another augmented view of \mathbf{x} through sharpening, *i.e.*, $\tau_s > \tau_t$. Following [51], we also adopt a mean-entropy maximization regularizer $H(\bar{\mathbf{p}}) = \sum_{k=1}^C \bar{\mathbf{p}}^k \log(\bar{\mathbf{p}}^k)$, where $\bar{\mathbf{p}} = \frac{1}{2|B|} \sum_{i \in B} (\mathbf{p}_i + \mathbf{p}'_i)$, to avoid an inactivation of classifier heads. The classifier learning loss is defined as:

$$L_{\text{cls}} = \frac{1}{|B|} \sum_{i \in B} l_{\text{cls}}(\mathbf{x}_i, \mathbf{y}_i) - \varepsilon H(\bar{\mathbf{p}}), \quad (5)$$

where ε controls the weight of the regularizer. Overall training objective is defined as: $L_{\text{rep}} + L_{\text{cls}}$.

3.2. Constructing tailedness prototypes

Tailedness estimation. Different from prior OWSSL settings, the true class prior is unknown in ROWSSL, as MNAR is considered. To learn a model without knowing the true class prior, we define ‘‘tailedness’’ as a surrogate for the class prior based on density estimation within the representation space. Since tail classes often exhibit lower intra-class consistency than head classes, samples of tail classes tend to sparsely distribute on the representation space [4, 32]. Building on this, we learn tailedness prototypes, aiming to explore stable and efficient proxies to discover tail class samples. To begin with, we utilize the queue $\mathbf{Q} = \{\mathbf{b}_i\}_{i=1}^Q$ of the CL branch in Sec. 3.1, as the neighbors in the entire dataset cannot be captured by looking at only a mini-batch. We initialize ℓ_2 -normalized tailedness prototypes $\mathbf{M} = \{\mathbf{m}_i\}_{i=1}^M$ by k -means on the features of the queue, and estimate density d_i of a prototype \mathbf{m}_i based on the weighted average of the cosine similarity of its K -nearest neighbors:

$$d_j^K = \frac{1}{\sum_{k=1}^K w_k} \sum_{i \in \mathcal{N}_K(\mathbf{m}_j)} w_i (\mathbf{m}_j \cdot \mathbf{b}_i), \quad (6)$$

where $\mathcal{N}_K(\mathbf{m}_j)$ is the set of the indices of the K -nearest neighbors of \mathbf{m}_j , and the distance-based weighting $w_i = \text{argsort}_j(\mathbf{m}_j \cdot \mathbf{b}_i)$ to reflect the local density better, reducing the effect of noisy density estimation [15]. Tailedness score s_i of each sample \mathbf{x}_i is defined as:

$$s_i = d_{j^*(i)}, \quad \text{where } j^*(i) = \underset{j}{\text{argmax}} \mathbf{m}_j \cdot \mathbf{b}_i. \quad (7)$$

Prototype update. We update tailedness prototypes by EMA for stable learning. Specifically, the queue \mathbf{Q} is split into a disjoint set of key features $\{\mathbf{U}_j\}_{j=1}^M$, where each key feature is assigned to the nearest tailedness prototype:

$$\mathbf{m}_j \leftarrow \text{normalize} \left[\lambda_{\text{tail}} \mathbf{m}_j + (1 - \lambda_{\text{tail}}) \frac{1}{|\mathbf{U}_j|} \sum_{\mathbf{b}' \in \mathbf{U}_j} \mathbf{b}' \right], \quad (8)$$

where \mathbf{m}_j is ℓ_2 -normalized and λ_{tail} is a momentum coefficient.

3.3. Density-based learning strategy

Dynamic temperature scaling. We aim to handle long-tailed data through self-supervised representation learning by controlling temperature parameter τ , which has been shown to play a significant role in learning good representations [48]. Specifically, we view the contrastive loss through the average distance maximization perspective [30]. From this view, a large τ allows the model to maximize the average distance across a wide range of neighbors, which is advantageous for preserving local semantic structures. On the other hand, a small τ helps to learn instance-specific features by encouraging a uniform distribution of embeddings across the representation space. Based on this perspective, we present a novel representation learning method, the dynamic temperature scaling for CL. Specifically, we adjust the temperature parameter τ in (2) as a function of the anchor’s tailedness score s_i :

$$\tau(\mathbf{x}_i) = \tau_{\min} + \frac{s_i - \min_t(d_t)}{\max_t(d_t) - \min_t(d_t)}(\tau_{\max} - \tau_{\min}), \quad (9)$$

where τ_{\min} and τ_{\max} are hyperparameters, denoting the minimum and maximum values of temperature, respectively. As tail classes benefit from learning instance-specific features while head classes are required to preserve their local semantic structure [30], our approach dynamically assigns smaller τ to tail classes and larger values to head classes. This allows the model to learn class-balanced representations, achieving better linear separability between the long-tailed classes without knowing the true class prior.

Class uncertainty-aware soft pseudo-labeling. For pseudo-labeling in long-tailed recognition, the distribution of pseudo-labels on unlabeled data tends to be biased toward head classes [1]. For conventional long-tailed recognition, the effect of bias can be mitigated by giving more weight to tail classes inversely proportional to their class prior [35]. However, this approach might not work well in ROWSSL, as pseudo-labels tend to be biased toward known classes, such that they are often more biased toward known-tail classes than novel-head classes [46]. To this end, we propose to adjust the soft pseudo-label \mathbf{q}_i in (4) with regard to the class uncertainty. Intuitively, for classes that are easy to learn, their samples will consistently be assigned to a specific tailedness prototype. Conversely, samples from more difficult and uncertain classes will be arbitrarily distributed across various prototypes. Based on this idea, we propose to use the standard deviation of tailedness scores among samples within each class as a measure of the relative learning uncertainty of each class as the additive class uncertainty. At each training iteration, we gather the tailedness scores in the dataset with respect to each sample’s

Table 2. Results on CIFAR-100-LT. **Tr**: transductive, **In**: inductive, **ACC**: average accuracy in Eq. (1), **bACC**: average of per-class accuracy. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Distribution Match ($\gamma_l = \gamma_u$)								
	Tr-ACC			Tr-bACC			In-bACC		
	All	Old	New	All	Old	New	All	Old	New
k -means	40.1	39.6	40.6	34.2	35.0	33.4	30.3	32.9	27.6
ORCA [†] [6]	51.2	64.9	43.9	25.0	31.5	18.6	29.5	39.1	19.9
GCD [44]	<u>55.0</u>	52.1	57.7	42.3	45.9	<u>38.6</u>	38.1	42.8	<u>33.4</u>
TRSSL [†] [41]	41.3	73.3	25.4	33.7	46.7	20.6	37.9	<u>53.5</u>	22.4
OpenCon [†] [42]	53.5	79.9	39.9	<u>48.5</u>	62.8	35.2	<u>47.7</u>	62.3	33.2
PromptCAL [54]	52.3	72.6	32.1	46.0	<u>62.9</u>	29.1	38.5	52.6	24.4
SimGCD [51]	51.7	54.3	49.2	46.5	59.8	33.2	37.4	44.1	30.8
BaCon [3]	45.8	40.0	51.5	38.0	41.9	34.2	35.9	40.5	31.2
Ours	65.3	<u>77.4</u>	<u>53.3</u>	53.7	68.4	39.1	48.1	52.9	43.2

Method	Distribution Mismatch ($\gamma_l \neq \gamma_u$)								
	Tr-ACC			Tr-bACC			In-bACC		
	All	Old	New	All	Old	New	All	Old	New
k -means	46.0	48.4	43.6	41.8	48.4	35.2	36.9	36.9	37.0
ORCA [†] [6]	48.8	35.5	55.5	23.8	25.5	22.2	27.2	30.5	23.8
GCD [44]	52.8	56.8	48.9	44.3	59.7	28.9	44.6	54.0	35.1
TRSSL [†] [41]	34.5	39.0	32.3	31.7	36.6	26.8	35.4	39.6	31.2
OpenCon [†] [42]	49.6	50.7	49.0	46.3	51.1	<u>41.5</u>	47.4	54.3	<u>40.4</u>
PromptCAL [54]	56.6	76.0	37.3	54.2	78.0	30.4	48.1	67.4	28.8
SimGCD [51]	<u>65.8</u>	<u>75.2</u>	<u>56.4</u>	<u>55.2</u>	<u>77.0</u>	33.4	<u>50.3</u>	<u>65.3</u>	35.4
BaCon [3]	56.0	56.5	55.6	46.4	61.2	31.7	42.8	50.9	34.8
Ours	66.6	74.2	59.0	57.3	68.7	45.9	53.1	64.3	41.8

pseudo-label into the class-wise tailedness queue S^c . We define the class uncertainty vector of the e -th training iteration $\mathbf{u}_e = [u^1, \dots, u^C]$, $e = 1, \dots, E$, as a collection of the standard deviation of tailedness scores per class:

$$u^c = \text{std}(S^c) \text{ where } S^c = \{s_i \mid x_i \in \mathcal{D}, \arg\max_k(\bar{y}_i^k) = c\}. \quad (10)$$

Note that $u^c = 0$ when $S^c = \emptyset$ and $\mathbf{u}_0 = \mathbf{0}$. Then, we adjust the output of the classifier with the class uncertainty:

$$\mathbf{q}_i = \text{softmax}[(f(\text{sg}(\mathbf{z}'_i)) + \lambda_{\text{var}}\mathbf{u}_{e-1})/\tau_t], \quad (11)$$

where λ_{var} is a hyperparameter. Our approach can be considered as a variation of the uncertainty-adaptive margin loss in [6], which mitigates classifier bias towards the head and known classes in a unified way.

4. Experimental Results

We compare our method with the state-of-the-art OWSSL methods [3, 6, 41, 42, 44, 51, 54]. We report the results on CIFAR-100-LT with an imbalance ratio $\gamma = 100$ in Table 2. We explore two scenarios with different class priors of the unlabeled dataset D_u : 1) the class prior of D_u is consistent with D_l , *i.e.*, MCAR ($\gamma_l = \gamma_u$; Fig. 1a left), and 2) the class prior of D_u is reversed from D_l , leading to a discrepancy in class prior distribution between them, *i.e.*, MNAR ($\gamma_l \neq \gamma_u$; Fig. 1a right). In most cases, our method outperforms others in terms of overall accuracy for both transductive and inductive inferences. Specifically, our method shows superior novel class accuracy, demonstrating that the density-based approach is effective in compensating for the difficulty of learning novel classes.

Appendix

A. Related Works

Table A.1. Comparison of ROWSSL with other related settings.

Setting	Known Classes	Novel Classes	Data Distribution	Distribution Mismatch	Evaluation
SSL	Classify	Not present	Balanced	✗	Inductive
Robust SSL [18]	Classify	Reject	Balanced	✗	Inductive
LT-SSL [49]	Classify	Not present	Imbalanced	✗	Inductive
RLT-SSL [50]	Classify	Not present	Imbalanced	✓	Inductive
NCD [24]	Not present	Discover	Balanced	-	Transductive
DA-NCD [53]	Not present	Discover	Imbalanced	-	Transductive
OWSSL/GCD [6, 44]	Classify	Discover	Balanced	✗	Transductive
DA-GCD [3]	Classify	Discover	Imbalanced	✗	Transductive *
ROWSSL	Classify	Discover & Classify **	Imbalanced	✓	Transductive & Inductive

* Evaluated on the disjoint test dataset, but it requires to see the entire test dataset for inference, i.e., transductive inference.

** Discover novel classes on the unlabeled training dataset and classify them on the disjoint test dataset.

Open-world semi-supervised learning (OWSSL) or generalized category discovery (GCD) is a transductive learning setting which extends semi-supervised learning (SSL) and novel category discovery (NCD) [23] by classifying known classes as well as discovering novel classes in the unlabeled training dataset. Vaze et al. [44] addresses this task via contrastive learning (CL) on a pre-trained vision transformer (ViT) [7, 13] followed by constrained k -means clustering [34]. Since then, a plethora of works have explored CL to achieve robust representations in OWSSL. XCon [16] learns fine-grained discriminative features by dataset partitioning. PromptCAL [54], DCCL [38], OpenNCD [33], and CiPR [21] construct an affinity graph, and OpenCon [42] utilizes a prototype-based novelty detection strategy to mine reliable positive pairs for the contrastive loss. GPC [55] introduces a novel representation learning strategy based on a semi-supervised variant of the Gaussian mixture model. SPTNet [47] proposes an iterative optimization method which optimizes both model and data parameters. In parallel with them, ORCA [6], NACH [19], and OpenLDN [40] utilize pairwise learning, generating pseudo-labels for unlabeled data by ranking distances in the feature space. ORCA and NACH also propose uncertainty-based loss to alleviate known class bias caused by different learning speeds between known and novel classes.

However, these advances are mostly based on the assumption that the class prior of the training dataset is balanced; indeed, data imbalance poses further challenges in OWSSL. For example, while a majority of methods proposed for OWSSL employed CL, it has been known that CL is not immune to data imbalance, such that representations learned on long-tailed distribution might be biased toward head classes [26]. Also, they mostly rely on k -means clustering, which assumes the presence of isotropic data clusters [31, 52], such that the uniform cluster prior assumption often hampers representation learning [2]. In the case of pairwise learning-based methods, the classifier is learned to be biased toward head classes due to the lack of positive pairs in tail classes [9]. Learning pace-based methods only take account of the uncertainty of known and novel classes, such that it might be difficult to distinguish between known-tail classes and novel-head classes [46]. Lastly, non-parametric methods [3, 44, 54] apply k -means clustering at inference time, which requires access to the entire test dataset for inference, often unattainable in real-world scenarios and hinders online inference, i.e., inductive learning.

In this paper, we advance OWSSL to a more practical setting, considering long-tailed distribution with class prior distribution mismatch, and inductive inference. Also, we address the aforementioned problems by density estimation on the latent feature space to achieve balanced CL and reduce the classifier bias toward the head and known classes.

Long-tailed recognition considers imbalanced class prior, which is natural in real-world scenarios. Early approaches to combat the imbalance include data re-sampling [8], re-weighting [10], and margin-based approach [5] with respect to given class-wise sample sizes. Based on this, DARP [28] and CReST [49] introduce long-tailed semi-supervised learning methods utilizing distribution alignment. Recently, ACR [50] and PRG [14] suggest realistic long-tailed semi-supervised learning setting considering class prior distribution mismatch between labeled and unlabeled datasets, i.e., MNAR. However, their closed-world assumption hinders direct application to ROWSSL. On the other hand, self-supervised learning under long-tailed distribution has also been investigated [26, 30]. As the temperature parameter plays a significant role in shaping the

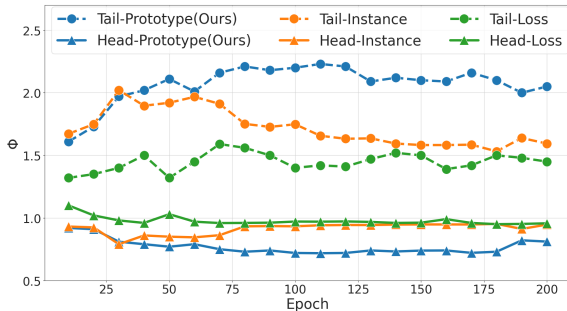


Figure D.1. Comparison of tail discovery methods.

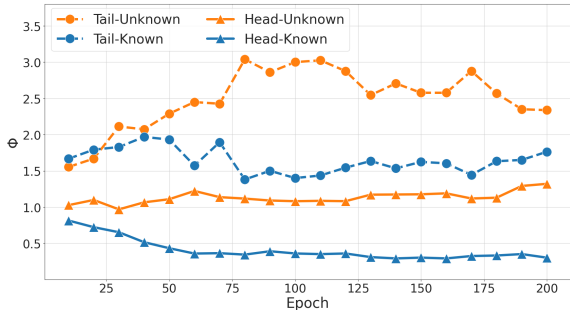


Figure D.2. Comparison between known and novel classes of our method.

serves as an indicator of the ability to identify tail samples: when the target group is head (tail), lower (higher) ϕ indicates that the method effectively localizes tail samples. In Fig. D.1, we compare our prototype-based estimation (“Prototype”) with the instance-wise estimation baseline (“Instance”) and the loss-based estimation (“Loss”) [56]. As shown in Fig. D.1, our method discovers tail samples more effectively than others. To observe the effect of our method on discriminating known and novel classes, we further divided head and tail groups into known and novel classes. In Fig. D.2, we observe that the difference between known and novel classes is not captured well at the beginning of training, but samples from novel classes begin to exhibit larger tailedness in the representation space than those from known classes through training. This implies that the learning difficulty could be captured by our density-based approach, which is further discussed in the following.

D.2. Design choices for the temperature in CL

To validate the effectiveness of the dynamic temperature for the contrastive loss, we experiment with different choices of the temperature. In addition to the constant temperature ($\tau = 0.07$), we compare our density-based approach with the estimated class prior by hard pseudo-labels for the classifier following [53] as baselines, and the true class prior as an oracle, where we apply the same min-max normalization as our method. We observe that our proposed method achieves better overall accuracy compared to baselines. Interestingly, our method even outperforms the oracle with the true class prior, implying that the learning difficulty of classes is not strictly proportional to the class prior, and our density-based approach can be more effective in addressing it.

D.3. Design choices for the class uncertainty

In this experiment, we validate the effectiveness of the choice of \mathbf{u} , which is the standard deviation of class-wise tailedness scores. We compare the variance of the maximum softmax probability as confidence for each class and the estimated distribution [53] as baselines and the ground-truth class prior as an oracle. For both estimated and ground-truth class prior, we convert the class frequency into a normalized probability distribution. As shown in Table D.2, our method achieves comparable performance to the oracle performance. Notably, our method boosts performance by 6.2% in novel classes and 3.8% in tail classes. This result confirms that focusing on class uncertainty is more effective than using class prior for mitigating the bias of the classifier in the ROWSSL setting.

Table D.1. Ablation study on τ .

Method	CIFAR-100-LT					
	All	Old	New	Many	Med.	Few
Constant	45.3	55.1	35.5	<u>62.1</u>	53.8	20.0
Estimated prior	46.8	59.2	34.4	61.9	<u>55.2</u>	23.3
True prior	<u>47.2</u>	<u>57.4</u>	<u>36.8</u>	62.2	54.8	<u>24.3</u>
Ours	48.1	52.9	43.2	59.2	58.0	27.7

Table D.2. Ablation study on \mathbf{u} .

Method	CIFAR-100-LT					
	All	Old	New	Many	Med.	Few
Confidence	43.3	48.4	<u>38.1</u>	58.8	50.2	20.7
Estimated prior	45.5	<u>54.9</u>	36.1	<u>60.1</u>	54.8	21.6
True prior	<u>47.9</u>	58.7	37.0	62.9	<u>56.7</u>	<u>23.9</u>
Ours	48.1	52.9	43.2	59.2	58.0	27.7

D.4. Contribution of each component

We examine the impact of each component in Table D.3. Specifically, starting from the baseline [51], we ablate the momentum encoder [22] and dynamic temperature scaling and class uncertainty-aware pseudo-labeling. Comparing experiments (b) and (c), the proposed dynamic temperature scaling improves performance by 2.7% and 1.4% for head classes, alongside 6.1% and 14% for tail classes on the CIFAR-100-LT and CUB-200-LT datasets, respectively. This indicates that our method learns discriminative semantic structures for both head and tail classes. From (b) and (d), the proposed class uncertainty-aware pseudo-labeling yields a notable improvement in all metrics. Specifically, introducing \mathbf{u} enhances performance by 7.2% and 8.4% in novel classes, with 5.6% and 11.0% in tail classes for each dataset, effectively mitigating classification bias towards known and head classes. The full version of our method (e) shows superior performance on all evaluation metrics, which experimentally demonstrates that our approach plays a crucial role in addressing ROWSSL.

Table D.3. Component analysis of DTS.

Index	Component			CIFAR-100-LT					
	Momentum	Dynamic τ	Uncertainty \mathbf{u}	All	Old	New	Many	Med.	Few
(a)	✗	✗	✗	38.8	50.9	26.7	55.9	49.9	9.3
(b)	✓	✗	✗	41.5	<u>56.5</u>	28.5	56.4	52.7	14.4
(c)	✓	✓	✗	45.7	55.5	<u>35.9</u>	59.1	<u>56.1</u>	<u>20.5</u>
(d)	✓	✗	✓	<u>47.6</u>	59.6	35.7	68.5	55.6	20.0
(e)	✓	✓	✓	48.1	52.9	43.2	<u>59.2</u>	58.0	27.7

D.5. Unknown class numbers

In real-world applications, we often do not have prior knowledge of the true number of classes C . In Table D.4, we estimate the number of classes \hat{C} and use it for evaluation depending on the type of methods: for non-parametric clustering-based methods [3, 44], we apply Brent’s algorithm to estimate \hat{C} as in [44], and for parametric classification methods [6, 51] and ours, we provide an arbitrarily large number, *e.g.*, $\hat{C}_{\text{init}} = 2C$, and estimate \hat{C} by eliminating inactivated classes, *i.e.*, classes without mappings from any training data. Notably, the uniform prior assumption in the k -means algorithm leads GCD and BaCon to significantly underestimate the total class number in long-tailed datasets, resulting in overall performance degradation. In the case of ORCA, its pairwise learning could be dominated by known and head classes as pairs mostly consist of data from known and head classes, and its binary uncertainty estimation would not be suitable for distinguishing known-tail and novel-head classes, resulting in significant inactivation of classification heads. Our method demonstrates comparable performance to scenarios where the number of classes is known, with only a 1.0% decrease in overall inductive accuracy.

Table D.4. Comparison results on CIFAR-100-LT ($\gamma_l = \gamma_u$) with an unknown number of classes.

Method	Param.	Est. \hat{C}	Tr-ACC			Tr-bACC			In-bACC		
			All	Old	New	All	Old	New	All	Old	New
ORCA	✓	59	46.9	50.7	<u>45.1</u>	25.2	31.5	19.0	28.1	37.7	18.6
GCD	✗	76	44.7	47.4	39.3	37.9	38.1	<u>37.7</u>	38.6	51.6	25.7
SimGCD	✓	145	<u>52.8</u>	73.2	42.6	<u>42.9</u>	<u>55.6</u>	30.3	<u>41.6</u>	<u>56.0</u>	27.2
BaCon	✗	79	48.4	63.1	36.0	42.4	52.3	32.5	33.1	33.5	<u>32.7</u>
Ours	✓	94	60.8	<u>72.0</u>	49.6	51.3	61.2	41.4	47.1	57.6	36.6

D.6. Number of tailedness prototypes

To evaluate the performance sensitivity in relation to the number of tailedness prototypes M , we conduct an ablation study on different prototype numbers. As shown in Table D.5, aligning the number of prototypes with the class number yields the best performance. In general, our method demonstrates robustness across various numbers of prototypes, yielding the best performance among compared methods in most cases. Note that matching the number of prototypes with the true number of classes might not always result in the best performance, because multiple fine-grained classes might form a single coarse-grained class or a class might consist of multiple local clusters [39].

Table D.5. Comparison results on CIFAR-100-LT ($\gamma_l = \gamma_u$) with various number of prototypes.

M	Tr-ACC			Tr-bACC			In-bACC		
	All	Old	New	All	Old	New	All	Old	New
50	60.2	74.3	46.1	49.0	59.8	38.2	46.2	53.0	39.4
200	<u>63.4</u>	74.7	52.1	<u>51.2</u>	<u>64.6</u>	37.8	<u>47.7</u>	<u>54.0</u>	<u>41.5</u>
300	63.0	<u>75.0</u>	<u>51.0</u>	48.5	57.5	39.6	47.4	54.8	40.1
100	65.3	77.4	53.3	53.7	68.4	<u>39.1</u>	48.1	52.9	43.2

D.7. Results with different imbalance ratios

In previous experiments, we use $\gamma = 100$ for CIFAR-100-LT. In Tables D.6 to D.7, we conduct an ablation study for different imbalance ratios ($\gamma = 1, 10$) on CIFAR-100-LT. Our method shows superior performance in overall accuracy for various imbalance ratios, showing its generalization ability for the different class priors. Bacon [3] often outperforms our method in novel class accuracy in transductive inference, however, its performance is degraded in inductive inference, while our method maintains good performance in inductive inference.

Table D.6. Results on balanced CIFAR-100 ($\gamma = 1$).

Method	Tr-ACC			Tr-bACC			In-bACC		
	All	Old	New	All	Old	New	All	Old	New
k -means	49.2	50.1	48.5	49.3	50.1	48.6	50.0	54.9	45.2
ORCA [†]	41.6	50.1	37.3	43.7	50.1	37.3	44.5	52.5	36.5
GCD	64.6	<u>72.7</u>	60.5	66.6	<u>72.8</u>	60.3	63.5	74.9	52.1
TRSSL [†]	50.3	71.0	40.0	55.5	71.0	40.1	66.3	83.1	49.5
SimGCD	<u>65.4</u>	71.9	<u>62.6</u>	<u>67.2</u>	71.9	<u>62.6</u>	<u>69.8</u>	77.3	<u>62.4</u>
BaCon	65.3	72.3	61.8	67.0	72.4	61.7	69.2	81.0	57.4
Ours	69.0	79.4	63.8	71.6	79.4	63.8	72.6	<u>81.9</u>	63.2

Table D.7. Results on CIFAR-100-LT with $\gamma = 10$.

Method	Distribution Match ($\gamma_l = \gamma_u$)									Distribution Mismatch ($\gamma_l \neq \gamma_u$)								
	Tr-ACC			Tr-bACC			In-bACC			Tr-ACC			Tr-bACC			In-bACC		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k -means	46.7	44.0	47.9	41.6	38.6	44.6	41.7	43.5	40.0	51.2	55.2	49.2	48.7	53.6	43.8	48.7	55.5	42.0
ORCA [†]	44.2	50.8	40.9	34.3	41.4	27.3	39.2	51.2	27.3	40.6	43.0	39.3	30.7	36.8	24.6	31.9	39.3	24.4
GCD	55.5	61.2	52.8	52.5	60.4	44.7	51.9	63.5	40.3	60.6	<u>75.1</u>	53.4	58.8	<u>71.8</u>	45.8	56.3	71.5	41.1
TRSSL [†]	42.9	<u>66.1</u>	31.4	42.4	56.1	28.7	49.0	<u>65.7</u>	32.3	43.3	59.6	35.2	43.7	54.7	32.7	51.4	63.4	39.4
SimGCD	54.2	59.1	51.8	53.4	<u>62.2</u>	44.6	52.8	61.5	44.2	62.1	73.7	<u>56.2</u>	<u>59.4</u>	69.6	49.2	<u>62.8</u>	<u>75.8</u>	<u>49.9</u>
BaCon	<u>59.7</u>	58.2	60.5	<u>54.5</u>	55.9	53.2	<u>55.1</u>	64.3	<u>45.9</u>	63.9	68.4	61.6	58.7	68.1	<u>49.3</u>	59.0	72.9	45.2
Ours	61.7	71.4	<u>55.4</u>	60.8	70.6	<u>51.0</u>	62.2	73.3	51.1	<u>63.8</u>	79.7	55.9	62.9	73.4	52.4	64.3	78.0	50.6

E. Hyperparameter Analysis

We conduct ablation experiments on critical hyperparameters of DTS, including (1) the number of neighbors for the K -NN, (2) τ_{max} for dynamic temperature scaling, and (3) λ_{var} for class-uncertainty aware pseudo labeling. We report overall inductive balanced accuracy performance on the CIFAR-100-LT dataset with distribution matched setting ($\gamma_l = \gamma_u$).

E.1. Number of neighbors for the K -NN

We consider $K = \{5, 10, 15, 20, 25\}$ for inspecting the impact of the number of nearest neighbors on tailedness estimation. As shown in Fig. E.1a, the optimal number of nearest neighbors is 15. When the neighborhood size is increased to include large neighbors ($K > 20$), we observe a slight degradation in performance, implying that the larger neighborhoods might accurately capture the local density that represents the class prior distribution.

E.2. Hyperparameter τ_{max}

In Fig. E.1b, we investigate the effect of the range of τ by considering $\tau_{max} = \{0.5, 0.7, 0.9, 1.0, 1.5\}$ with $\tau_{min} = 0.05$, where $\tau_{max} = 1.0$ shows the best performance. We argue that it optimally balances the uniformity and alignment of representation. A narrow range of tau ($\tau_{max} < 0.7$) may disrupt the semantic representation, while a wide range of tau ($\tau_{max} > 1.0$) could negatively impact learning instance-specific features.

E.3. Hyperparameter τ_{min}

In Fig. E.1c, we examine the effect of the range of τ by considering $\tau_{min} = \{0.01, 0.02, 0.05, 0.1, 0.3\}$ with $\tau_{max} = 1.0$, where $\tau_{min} = 0.05$ shows the best performance. We argue that it optimally balances the uniformity and alignment of representation. A high minimum value of tau ($\tau_{min} > 0.1$) may hinder the learning instance-specific features, while a low minimum value of tau ($\tau_{min} < 0.05$) may disrupt the semantic representation.

E.4. Hyperparameter λ_{var}

We examine the effect of the weight parameter λ_{var} as illustrated in Fig. E.1d, where we consider $\lambda_{var} = \{0.5, 1.0, 2.0, 3.0\}$. Among them, $\lambda_{var} = 1$ shows the best performance. Notably, a larger weight parameter appears to adversely affect the information contained in the original output logits of the cosine classifier.

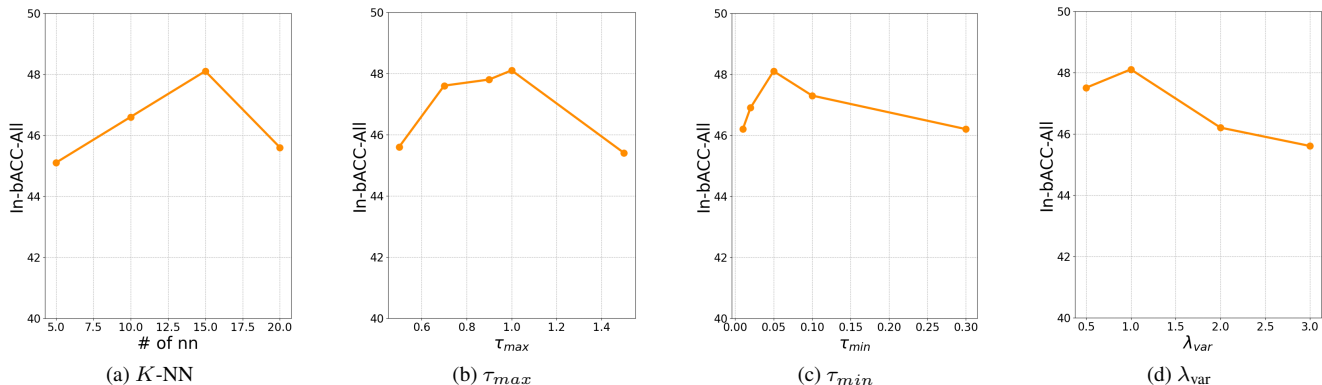


Figure E.1. Analysis of hyperparameters.

F. Detailed Results of Main Experiments

To better examine the impact of dataset imbalance, we conduct a detailed comparison in Tables F.1 to F.4. In Table F.4, we report the performance in Missing Not At Random (MNAR) scenarios for the in-nature long-tailed dataset Herbarium19. Our approach demonstrates a significant performance improvement for novel and tail classes, where the conventional open-world and long-tailed learning strategies do not take into account the importance of learning tail and novel classes, respectively. This validates that our method effectively addresses known and head class bias issues.

Table F.1. Results on CIFAR-100-LT ($\gamma_l = \gamma_u$).

CIFAR-100-LT ($\gamma_l = \gamma_u$)															
Method	Tr-ACC			Tr-bACC			In-bACC								
	All	Old	New	All	Old	New	All	Old	New	KMany	KMed	KFew	UMany	UMed	UFew
k -means	40.1	39.6	40.6	34.2	35.0	33.4	30.3	32.9	27.6	47.9	45.0	5.8	42.9	25.7	14.2
ORCA [†]	51.2	64.9	43.9	25.0	31.5	18.6	29.5	39.1	19.9	66.1	44.6	6.6	45.8	10.8	3.1
GCD	<u>55.0</u>	52.1	57.7	42.3	45.9	<u>38.6</u>	38.1	42.8	<u>33.4</u>	70.9	48.2	9.3	<u>50.6</u>	36.6	13.0
TRSSL [†]	41.3	73.3	25.4	33.7	46.7	20.6	37.9	<u>53.5</u>	22.4	80.6	51.3	<u>28.6</u>	35.2	25.6	6.4
OpenCon [†]	53.5	79.9	39.9	<u>48.5</u>	62.8	35.2	<u>47.7</u>	62.3	33.2	87.3	70.2	28.4	40.9	<u>46.4</u>	10.6
PromptCAL	52.3	72.6	32.1	46.0	<u>62.9</u>	29.1	38.5	52.6	24.4	75.3	59.7	22.8	35.5	24.2	13.5
SimGCD	51.7	54.3	49.2	46.5	59.8	33.2	37.4	44.1	30.8	67.1	43.9	21.3	44.6	35.2	12.6
BaCon	45.8	40.0	51.5	38.0	41.9	34.2	35.9	40.5	31.2	53.2	57.1	11.2	45.3	34.1	<u>14.2</u>
Ours	65.3	<u>77.4</u>	<u>53.3</u>	53.7	68.4	43.2	48.1	52.9	43.2	63.2	<u>61.1</u>	34.4	55.1	53.3	20.7

Table F.2. Results on CIFAR-100-LT ($\gamma_l \neq \gamma_u$).

CIFAR-100-LT ($\gamma_l \neq \gamma_u$)															
Method	Tr-ACC			Tr-bACC			In-bACC								
	All	Old	New	All	Old	New	All	Old	New	KMany	KMed	KFew	UMany	UMed	UFew
k -means	46.0	48.4	43.6	41.8	48.4	35.2	36.9	36.9	37.0	39.2	41.9	29.6	63.0	29.3	18.7
ORCA [†]	48.8	35.5	55.5	23.8	25.5	22.2	27.2	30.5	23.8	37.0	35.1	19.4	54.2	10.9	6.3
GCD	52.8	56.8	48.9	44.3	59.7	28.9	44.6	54.0	35.1	57.9	54.9	49.3	<u>59.2</u>	38.2	7.9
TRSSL [†]	34.5	39.0	32.3	31.7	36.6	26.8	35.4	39.6	31.2	63.4	32.9	22.5	62.2	20.1	11.3
OpenCon [†]	49.6	50.7	49.0	46.3	51.1	<u>41.5</u>	47.4	54.3	<u>40.4</u>	<u>72.3</u>	63.0	27.6	54.9	<u>46.8</u>	<u>19.5</u>
PromptCal	56.6	76.0	37.3	54.2	78.0	30.4	48.1	67.4	28.8	80.1	74.3	47.8	39.5	28.2	18.7
SimGCD	<u>65.8</u>	75.2	<u>56.4</u>	<u>55.2</u>	77.0	33.4	<u>50.3</u>	65.3	35.4	69.4	63.9	<u>62.6</u>	51.0	42.4	12.8
BaCon	56.0	56.5	55.6	46.4	61.2	31.7	42.8	50.9	34.8	51.0	54.5	47.2	62.0	32.4	10.0
Ours	66.6	<u>74.2</u>	59.0	57.3	68.7	45.9	53.1	64.3	41.8	66.0	<u>64.1</u>	62.8	53.8	49.2	22.4

Table F.3. Results on Herbarium19 ($\gamma_l = \gamma_u$).

Herbarium19 ($\gamma_l = \gamma_u$)															
Method	Tr-ACC			Tr-bACC			In-bACC								
	All	Old	New	All	Old	New	All	Old	New	KMany	KMed	KFew	UMany	UMed	UFew
k -means	13.0	12.2	13.4	9.8	8.6	11.0	6.6	7.2	5.9	8.4	9.2	4.1	8.4	6.5	2.8
ORCA [†]	19.4	18.2	20.1	7.0	10.1	6.4	16.4	17.7	15.0	32.7	10.4	10.0	30.8	8.0	6.2
GCD	35.8	50.6	27.8	33.4	42.3	24.5	25.5	25.1	25.9	36.1	24.0	15.2	36.0	29.3	12.4
TRSSL [†]	40.2	67.2	16.4	32.0	54.0	10.0	33.3	33.4	33.3	56.3	31.9	12.0	49.7	36.0	14.2
OpenCon [†]	28.6	46.2	19.2	20.9	31.5	10.4	29.7	27.8	31.7	39.8	23.4	20.2	50.8	39.5	<u>29.1</u>
PromptCAL	34.1	49.7	25.7	<u>34.4</u>	44.3	24.5	32.0	33.1	30.9	42.6	33.1	<u>23.6</u>	41.6	30.7	20.4
SimGCD	<u>43.4</u>	<u>57.7</u>	<u>35.8</u>	33.9	<u>45.8</u>	22.1	<u>42.3</u>	<u>40.1</u>	<u>44.6</u>	55.8	<u>42.8</u>	21.7	<u>60.6</u>	<u>49.4</u>	23.8
BaCon	29.8	29.2	30.1	28.7	28.3	<u>29.2</u>	27.1	27.1	27.2	45.3	23.4	12.5	38.1	26.9	16.6
Ours	47.7	48.7	46.8	38.9	39.6	38.1	45.4	43.0	47.8	<u>56.1</u>	44.8	28.0	64.1	49.7	29.5

Table F.4. Results on Herbarium19 ($\gamma_l \neq \gamma_u$).

Method	Herbarium19 ($\gamma_l = \gamma_u$)														
	Tr-ACC			Tr-bACC			In-bACC								
	All	Old	New	All	Old	New	All	Old	New	KMany	KMed	KFew	UMany	UMed	UFew
GCD	27.3	32.8	24.3	28.9	34.2	23.6	18.2	23.0	13.4	25.4	29.1	14.3	19.2	11.8	9.1
SimGCD	<u>34.9</u>	<u>40.9</u>	<u>31.7</u>	31.2	<u>38.2</u>	24.3	<u>26.5</u>	<u>32.5</u>	<u>20.5</u>	39.2	<u>38.5</u>	<u>19.6</u>	<u>32.1</u>	<u>16.3</u>	<u>13.1</u>
BaCon	32.4	35.6	30.6	<u>31.6</u>	35.3	<u>27.9</u>	21.5	26.7	16.3	<u>39.6</u>	25.6	14.9	28.2	14.3	6.4
Ours	46.9	58.4	40.8	37.0	48.9	<u>25.1</u>	31.4	41.5	21.3	57.2	39.7	27.4	32.9	16.5	14.7

G. Results on ImageNet-100-LT

While ImageNet-100 has been often used for OWSSL in literature, we argue that ImageNet-100 might not be appropriate for the conventional OWSSL settings built on top of ImageNet-1K [12] pre-trained backbone, *e.g.*, DINO-ViT [7], as it already observed data from novel classes during pretraining. In other words, the performance could be boosted by preserving the pre-trained knowledge rather than learning to discover novel classes and classify all classes. Nevertheless, below we report the performance on ImageNet-100 with the data split from BaCon [4]. In Tables G.1 to G.2, our proposed method achieves significantly better performance on novel and tail classes, surpassing the baseline performance. Notably, the classic baseline GCD [44] often shows the best performance (mostly in transductive inference), implying that it preserves the pre-trained knowledge better.

Table G.1. Results on ImageNet-100-LT ($\gamma_l = \gamma_u$).

Method	ImageNet-100-LT ($\gamma_l = \gamma_u$)														
	Tr-ACC			Tr-bACC			In-bACC								
	All	Old	New	All	Old	New	All	Old	New	KMany	KMed	KFew	UMany	UMed	UFew
GCD	<u>63.8</u>	<u>69.5</u>	60.7	63.5	<u>69.4</u>	57.6	<u>59.2</u>	<u>67.1</u>	<u>51.3</u>	81.1	77.1	<u>41.8</u>	76.0	<u>56.8</u>	20.5
SimGCD	55.3	62.3	51.6	55.5	63.9	47.0	54.0	63.2	44.8	64.0	80.2	43.4	64.8	48.2	20.9
BaCon	60.7	68.8	56.4	58.6	66.9	50.4	54.8	65.8	43.7	<u>81.5</u>	<u>81.0</u>	33.0	67.8	41.3	<u>22.4</u>
Ours	65.6	82.5	<u>56.6</u>	<u>63.0</u>	71.2	<u>54.7</u>	61.4	69.6	53.3	85.2	81.2	40.9	<u>70.0</u>	64.6	23.6

Table G.2. Results on ImageNet-100-LT ($\gamma_l \neq \gamma_u$).

Method	ImageNet-100-LT ($\gamma_l \neq \gamma_u$)														
	Tr-ACC			Tr-bACC			In-bACC								
	All	Old	New	All	Old	New	All	Old	New	KMany	KMed	KFew	UMany	UMed	UFew
GCD	66.2	75.9	61.1	62.1	67.2	57.0	60.6	66.2	55.0	87.1	66.8	44.8	73.4	62.2	28.5
SimGCD	60.4	<u>75.4</u>	52.4	56.5	66.9	46.2	56.0	66.6	45.4	75.6	<u>84.2</u>	37.6	69.2	53.1	12.8
BaCon	60.7	68.8	56.4	58.6	66.9	50.4	54.7	65.8	43.7	<u>81.5</u>	81.0	33.0	67.8	41.3	22.4
Ours	<u>63.2</u>	73.9	<u>57.4</u>	63.3	68.9	57.6	62.9	68.5	57.2	77.6	86.4	<u>41.5</u>	71.4	<u>55.3</u>	45.2

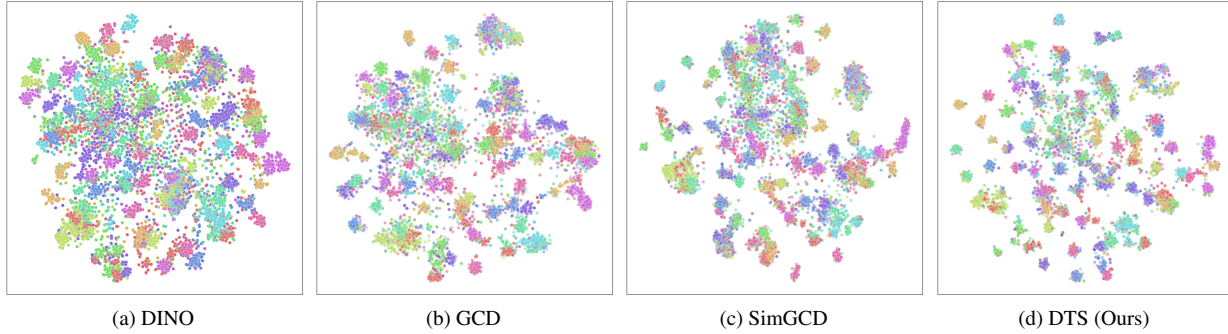


Figure H.1. t-SNE visualization on the test set of CIFAR-100-LT.

H. Visualizations

To inspect the learned semantic discriminativeness of the proposed DTS on the long-tailed dataset, we visualize embeddings by t-SNE [43] algorithm, trained on CIFAR-100-LT with distribution match. We show the feature embedding of pretrained DINO [7], GCD [44], SimGCD [51], and DTS (Ours), in Fig. H.1. Compared to other models, the model trained with our DTS learns less ambiguous features which exhibits a larger margin between different classes, with more compact clusters. This indicates that our method is more effective in learning a discriminative semantic structure, even under long-tailed datasets.

I. Conclusion and Limitations

In this paper, we formulate the practical ROWSSL setting, which considers the long-tailed distribution and the class prior distribution mismatch between labeled and unlabeled data for training, and inductive and transductive inferences for evaluation. To tackle ROWSSL, we introduce a novel method called Density-based Temperature scaling and Soft pseudo-labeling (DTS), which learns class-balanced representations and mitigates the classification bias based on local densities. Nevertheless, we acknowledge several limitations inherent in DTS and existing methods. First, the labeled and unlabeled data are sampled from the same dataset, which might not reflect domain shifts in real-world scenarios. Second, estimating the number of novel classes with off-the-shelf methods can result in inaccurate prediction due to the imbalanced class prior distribution. We believe that ROWSSL will establish a robust foundation for future research and contribute to the development of more reliable methods for practical applications of OWSSL.

J. Negative Societal Impact

While our work itself is not inherently harmful to society, there is a risk that it could be misused by those with malicious intent. For example, the proposed method could be used to unfairly single out and target certain groups, such as minorities. Consequently, we urge that this work must be utilized within ethical and legal boundaries.

References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020. 4
- [2] Mahmoud Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *ICLR*, 2023. 2, 1
- [3] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Ruizhe Chen, Lianrui Mu, Xiaomeng Li, Joey Tianyi Zhou, Yang Feng, Jian Wu, and Haoji Hu. Towards distribution-agnostic generalized category discovery. *arXiv preprint arXiv:2310.01376*, 2023. 2, 4, 1, 5
- [4] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. In *ICLR*, 2023. 3, 2, 8
- [5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1
- [6] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2021. 2, 4, 1
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 1, 2, 8, 9
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002. 1
- [9] Zhang Chuyu, Xu Ruijie, and He Xuming. Novel class discovery for long-tailed recognition. In *TMLR*, 2023. 1
- [10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1
- [11] Zhigang Dai, Bolun Cai, and Junying Chen. Unimoco: Unsupervised, semi-supervised and fully-supervised visual representation learning. *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2022. 3
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 8
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2
- [14] Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, and Yinghuan Shi. Towards semi-supervised learning with non-random missing labels. In *ICCV*, 2023. 1
- [15] Sahibsingh A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976. 3
- [16] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *BMVC*, 2022. 1
- [17] Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, and Elisa Ricci. Semi-supervised learning made simple with self-supervised clustering. In *CVPR*, 2023. 3
- [18] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, 2020. 1
- [19] Lan-Zhe Guo, Yi-Ge Zhang, Zhi-Fan Wu, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning when not all classes have labels. In *NeurIPS*, 2022. 1
- [20] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019. 2
- [21] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cidr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*, 2023. 1
- [22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 4
- [23] Yen-Chang Hsu and Zsolt Kira. Neural network-based clustering using pairwise constraints. In *ICLR Workshop*, 2016. 1
- [24] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019. 1
- [25] Xinting Hu, Yulei Niu, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. On non-random missing labels in semi-supervised learning. In *ICLR*, 2022. 1
- [26] Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *ICML*, 2021. 1
- [27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 3
- [28] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *NeurIPS*, 2020. 1
- [29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 2

- [30] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Ruppert. Temperature schedules for self-supervised contrastive methods on long-tail data. In *ICLR*, 2023. 4, 1, 2
- [31] Jiye Liang, Liang Bai, Chuangyin Dang, and Fuyuan Cao. The k -means-type algorithms versus imbalanced data distributions. *IEEE Transactions on Fuzzy Systems*, 2012. 1
- [32] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *ICLR*, 2022. 3
- [33] Jiaming Liu, Yangqiming Wang, Tongze Zhang, Yulu Fan, Qinli Yang, and Junming Shao. Open-world semi-supervised novel class discovery. In *IJCAI*, 2023. 1
- [34] James MacQueen et al. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967. 1
- [35] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 4
- [36] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 1
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 2
- [38] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, 2023. 1
- [39] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *ICCV*, 2019. 4
- [40] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *ECCV*, 2022. 1
- [41] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *ECCV*, 2022. 4
- [42] Yiyu Sun and Yixuan Li. Opencon: Open-world contrastive learning. In *TMLR*, 2023. 4, 1
- [43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. In *JMLR*, 2008. 9
- [44] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 2, 4, 1, 8, 9
- [45] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021. 2
- [46] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *ICML*, 2022. 4, 1
- [47] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, 2024. 1
- [48] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 4, 2
- [49] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, 2021. 1
- [50] Tong Wei and Kai Gan. Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In *CVPR*, 2023. 1
- [51] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. A simple parametric classification baseline for generalized category discovery. In *ICCV*, 2023. 2, 3, 4, 9
- [52] Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *KDD*, 2009. 1
- [53] Muli Yang, Liancheng Wang, Cheng Deng, and Hanwang Zhang. Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In *CVPR*, 2023. 1, 3
- [54] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guanyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, 2023. 4, 1
- [55] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *ICCV*, 2023. 1
- [56] Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, and Ya Zhang. Contrastive learning with boosted memorization. In *ICML*, 2022. 2, 3