OwMatch: Conditional Self-Labeling with Consistency for Open-World Semi-Supervised Learning

Shengjie Niu[1*], **Lifan Lin**[2*], **Jian Huang**[1], **Chao Wang**[2†]
[1]Hong Kong Polytechnic University, [2]Southern University of Science and Technology
shengjie.niu@connect.polyu.hk, 12012816@mail.sustech.edu.cn,
j.huang@polyu.edu.hk, wangc6@sustech.edu.cn

Abstract

Semi-supervised learning (SSL) offers a robust framework for harnessing the potential of unannotated data. Traditionally, SSL mandates that all classes possess labeled instances. However, the emergence of open-world SSL (OwSSL) introduces a more practical challenge, wherein unlabeled data may encompass samples from unseen classes. This scenario leads to the misclassification of unseen classes as known ones, consequently undermining classification accuracy. To overcome this challenge, this study revisits two methodologies from self-supervised and semi-supervised learning, self-labeling and consistency, tailoring them to address the OwSSL problem. Specifically, we propose an effective framework called *OwMatch*, combining conditional self-labeling and open-world hierarchical thresholding. Theoretically, we analyze the estimation of class distribution on unlabeled data through rigorous statistical analysis, thus demonstrating that OwMatch can ensure the unbiasedness of the self-label assignment estimator with reliability. Comprehensive empirical analyses demonstrate that our method yields substantial performance enhancements across both known and unknown classes in comparison to previous studies. Code is available at https://github.com/niusj03/OwMatch.

1 Introduction

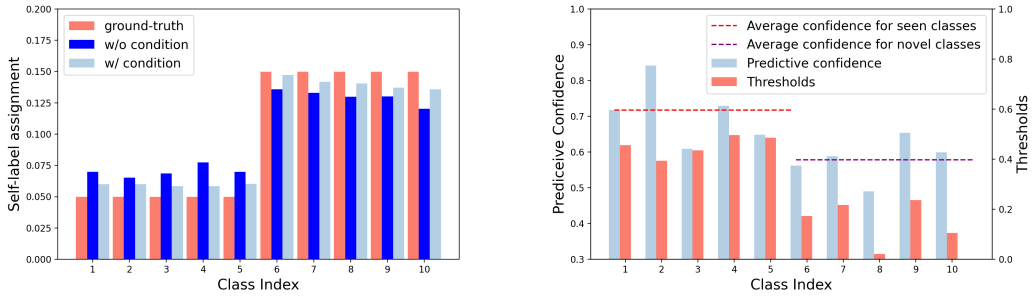
Deep learning has made remarkable success in various tasks by leveraging substantial labeled training data [22, 21, 13]. However, the costly and time-consuming labeling process limits their application in practical scenarios. Semi-supervised learning (SSL) significantly reduces the dependency on labeled data by exploring the inherent structure of unlabeled data [16]. Despite promising results, SSL methods assume a closed-world scenario where, though limited, all classes possess labeled instances. This assumption may be violated due to difficulties in data collection, such as in medical diagnostics, where it is common to encounter new symptoms or fail to annotate due to technical constraints. As a result, only a subset of the categories can be precisely labeled during the annotation process. Recently, numerous studies have sought to identify such novel classes effectively. Open-world SSL (OwSSL) is innovative in promoting dual objectives: classifying instances of seen classes and discovering instances of novel classes [5].

A notable challenge in OwSSL is the *confirmation bias* of model: model tends to predict instances as seen classes owing to the lack of ground-truth supervision of novel-class instances. To eliminate this bias, existing works utilize unsupervised clustering methods, including contrastive loss and binary cross-entropy (BCE) loss, to group pairs identified by similarity metrics [5, 15]. Among these unsupervised techniques, self-labeling [1, 6] has shown remarkable success, which involves assigning self-labels to unlabeled data, with the generation of high-quality self-labels being the key

[*]The first two authors contributed equally to this work.
[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3167



(a) (b)

Figure 1: Experimental results on the OwSSL problem. (a) Self-label assignment of seen classes (1-5) and novel classes (6-10) with or without conditional component in self-labeling. (b) Predictive confidence and hierarchical threshold for each class.

factor. Previous studies utilize optimal transport to align the self-labels for unlabeled data with a given distribution. However, this self-label generation fully relies on the accurate prior distribution and lack of consideration of the supervision of labeled data. In TRSSL [34], the unlabeled data are assigned with a soft self-label based on the inaccurate class distribution, which raises a biased estimation. Moreover, the confirmation bias still exists even if we use the ground-truth distribution to align the unlabeled data in the same process. In addition to confirmation bias, a new issue called *clustering misalignment* arises when self-labeling depends solely on unlabeled data: without proper guidance, the self-labeling process may adopt varying criteria for clustering. For example, it might cluster data based on superficial features like color rather than high-level semantic information. This misalignment can lead to results that deviate from expected outcomes and even contradict the classification criteria established by labeled data.

Consequently, we propose a new self-labeling scheme, conditional self-labeling, designed to address the challenges of OwSSL, particularly targeting issues related to confirmation bias and misalignment. This scheme limits self-labels for each class and incorporates labeled data to generate debiased and informative self-label assignments for all training data, further mitigating the confirmation bias as shown in Figure 1a. Additionally, as illustrated in Figure 1b, seen classes typically exhibit higher predictive confidence, while novel clusters demonstrate variability in their internal learning progresses. The disparities in learning paces between seen and novel classes, coupled with their distinct behaviors, necessitate the selection of appropriate thresholds to facilitate cluster learning. To address these challenges and ensure a balanced learning process across classes, we propose a hierarchical thresholding scheme.

We demonstrate our contributions as follows: **1)** We introduce a novel conditional self-labeling method to incorporate labeled data into the clustering process, reducing confirmation bias and misalignment. **2)** We design a hierarchical thresholding strategy that balances learning difficulties across different classes, helping unstable clusters gradually form. **3)** Our theoretical analysis rigorously discusses the unbiasedness and reliability of conditional self-labeling estimator from population-level statistics. To the best of our knowledge, this is the first work proposing the expectation of chi-square statistics (ECS) to evaluate the reliability of self-label assignment estimation. **4)** We conduct extensive experiments on various datasets, demonstrating the effectiveness of our approach, OwMatch, through detailed comparisons. On CIFAR-10, OwMatch significantly outperforms FixMatch [36] by up to 47.3% in all-class accuracy, while on CIFAR-100, it enhances TRSSL [34], the state-of-the-art model in OwSSL, by up to 14.6% in novel-class and 7.2% in all-class accuracy.

2 Related work

Traditional semi-supervised learning (SSL). Traditional SSL assumes that labeled and unlabeled data share an identical distribution. Extensive researches on SSL have spanned a considerable duration. The commonly employed strategies in SSL consist of entropy minimization [29, 14], consistency regularization [44, 30] and holistic methods [4, 3, 36]. The latest progresses in SSL include adaptive thresholding strategies [42, 45, 47], which enhance model performance by accounting for varying

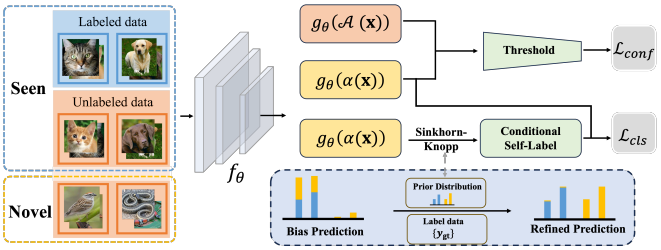


Figure 2: Overview of the OwMatch framework.

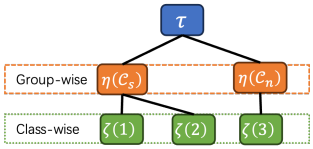


Figure 3: Illustration on the hierarchical thresholding scheme.

difficulties and learning conditions across classes, alongside other innovative techniques [19, 41, 2] that employ self-SL approaches to facilitate extracting the semantic information from unlabeled data. However, traditional SSL algorithms typically struggle to tackle the open-world problem in the presence of novel-class instances within unlabeled data.

Open-set semi-supervised learning (OSSL). OSSL expands the traditional SSL boundaries by allowing novel-class instances or outliers within unlabeled data. A variety of OSSL approaches have emerged in recent years [46, 16, 35, 8, 23]. A common solution among these methods is the optimization of the SSL objective exclusively for unlabeled samples deemed inliers. For instance, MTC [46] optimizes the network and estimates the anomaly score of unlabeled data alternately. OpenMatch [35] and T2T [23] train one-vs-all (OVA) classifiers for each known class to detect outliers. Subsequently, standard SSL objective [36] are applied to the remaining training data, excluding detected outliers. Furthermore, DS3L [16] leverages a bi-level optimization technique to train a weighting function, which mitigates the passive impact of out-of-distribution (OOD) samples. Nonetheless, these approaches are designed for classifying seen classes, thereby failing to learn from the novel class instances.

Open-world semi-supervised learning (OwSSL). OwSSL [5] has been proposed to address a practical challenge: enabling the model to effectively cluster novel-class instances while maintaining classification robustness on seen classes. One predominant research direction in this under-explored domain is BCE-based methods, including ORCA [5] and NACH [15]. Additionally, there exist methods to discover novel classes by employing various other clustering techniques: OpenLDN [33] employs bi-level optimization to train a pairwise similarity prediction network, which provides a supervisory signal to the similarity of all pairs; TRSSL [34] converts clustering into the self-labeling problem and applies Sinkhorn-Knopp algorithm to optimize self-label assignments. One subsequently proposed Generalized Category Discovery (GCD) setting is similar to the OwSSL [40, 49], with detailed discussion is provided in Section 5.3.

3 Methodology

Problem setup. Given training data consisting of labeled data $\mathcal{D}_l = \{(\mathbf{x}^{(i)}, \mathbf{y}_{\text{gt}}^{(i)})\}_{i=1}^{N^l}$ and unlabeled data $\mathcal{D}_u = \{\mathbf{x}^{(i)}\}_{i=N^l+1}^{N^l+N^u}$, where $N = N^l + N^u$ and $N^u \gg N^l$. Here $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is the i-th instance with one-hot vector $\mathbf{y}_{\text{gt}}^{(i)} \in \{0,1\}^K$ as the corresponding label , where K is the number of all classes. We denote the set of classes in $\mathcal{D}_l$ as $\mathcal{C}_l$ and the set of classes in $\mathcal{D}_u$ as $\mathcal{C}_u$. Previous traditional SSL studies assume $\mathcal{C}_l = \mathcal{C}_u$. Here for OwSSL, we assume $\mathcal{C}_l \neq \mathcal{C}_u$ and $\mathcal{C}_u \setminus \mathcal{C}_l \neq \emptyset$. Denote $\mathcal{C}_s = \mathcal{C}_l \cap \mathcal{C}_u$ as a set of seen classes, $\mathcal{C}_n = \mathcal{C}_u \setminus \mathcal{C}_l$ as a set of novel classes, and $\mathcal{C} = \mathcal{C}_l \cup \mathcal{C}_u$ as a set of all considered classes. The desired OwSSL model is required to assign instances to either a previously seen class $c \in \mathcal{C}_s$, or a novel class $c \in \mathcal{C}_n$.

For labeled dataset $\mathcal{D}_l$, standard supervised objective is employed as shown in Equation 9. Additionally, OwMatch primarily incorporates two objectives: a) clustering objective, which leverages conditional self-labeling to refine the self-label assignment with the assistance of supervision; b) confidence objective, which applies consistency loss with open-world hierarchical thresholding strategy to enhance predictive confidence and balance the different learning difficulties across all classes. We will elaborate on them respectively in Section 3.1 and 3.2.

https://doi.org/10.52202/079017-3167

3.1 Conditional self-labeling

To effectively cluster the novel class instances, the self-labeling scheme [1] has been considered in OwSSL. Formally, consider a deep neural network (encoder) f_θ mapping input data $\mathbf{x}$ to representation $\mathbf{z} \in \mathbb{R}^D$, the representation is followed by a classification head $h : \mathbb{R}^D \to \mathbb{R}^K$, usually consisting of a single linear layer, converting the feature vectors into a vector of class scores. Denote $g_\theta = \sigma \circ h \circ f_\theta$ as a probability function, where σ refers to the SoftMax function. Moreover, denote $\mathbf{q}^{(i)} \in \mathbb{R}^K$ as the soft self-label for $\mathbf{x}^{(i)}$, and set $\mathbf{Q} = [\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \ldots, \mathbf{q}^{(N)}] \in \mathbb{R}^{K \times N}$ as the self-label assignment for $\{\mathbf{x}^{(i)}\}_{i=1}^N$. Asano et al. [1] utilize a constraint of desired partition of $\mathbf{Q}$ to construct the transportation polytope:

$$\mathcal{Q}_1 := \{\mathbf{Q} \in \mathbb{R}_+^{K \times N} | \mathbf{Q}\mathbf{1}_N = N\boldsymbol{\mathcal{P}}, \mathbf{Q}^T\mathbf{1}_K = \mathbf{1}_N\}, \tag{1}$$

where $\mathbf{1}_v$ is the v-dimensional vector of all ones, $\boldsymbol{\mathcal{P}}$ denotes the desired class distribution. On the other hand, we can obtain a probability output through $\mathbf{p}^{(i)} = g_\theta(\alpha(\mathbf{x}^{(i)}))$, where $\alpha(\cdot)$ refers to a specific weak augmentation, and denote $\mathbf{P} = [\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(N)}]$ as the matrix of probability outputs. This self-label assignment generation can be understood as solving an optimal transport problem [1]. It minimizes the cross-entropy loss and aligns the training data with the desired class distribution:

$$\max_{\mathbf{Q} \in \mathcal{Q}_1} \operatorname{Tr}(\mathbf{Q} \log(\mathbf{P}^T)), \tag{2}$$

where $\operatorname{Tr}(\cdot)$ is the trace of a given matrix. Obviously, clustering through self-labeling primarily relies on the quality of generated self-label assignments. However, optimizing self-label assignments through unsupervised self-labeling is unreliable owing to the lack of supervision. TRSSL [34] utilizes the above-unsupervised technique to optimize self-label assignment merely on unlabeled data with the uniform class distribution. Despite prominent results, this unconditional self-labeling process [34] has a notable flaw: it constructs transportation polytope based on an inaccurate class distribution.

Moreover, we consider conducting self-labeling across all training data and constructing a transportation polytope with a precise class distribution. To mitigate the confirmation bias, we propose a conditional self-labeling method to refine the self-label assignment under partial supervision. Specifically, we exploit the ground-truth labels from the labeled dataset and introduce another constraint $\mathcal{Q}_2$:

$$\mathcal{Q}_2 := \{\mathbf{Q} \in \mathbb{R}_+^{K \times N} | \mathbf{q}^{(i)} = \mathbf{y}_{\text{gt}}^{(i)}, i = 1, \ldots, N^l\}. \tag{3}$$

Now, the conditional self-label assignment generation with the above two constraints can be formulated as:

$$\max_{\mathbf{Q} \in \mathcal{Q}_1 \cap \mathcal{Q}_2} \operatorname{Tr}(\mathbf{Q} \log(\mathbf{P}^T)) + \epsilon E(\mathbf{Q}), \tag{4}$$

where $E(\cdot)$ is the entropy function, ϵ is a hyper-parameter controlling the smoothness of $\mathbf{Q}$. We adopt fast version [10] of *Sinkhorn-Knopp algorithm* to optimize Equation 4 efficiently and denote the optimal solution as $\tilde{\mathbf{Q}} = [\tilde{\mathbf{q}}^{(1)}, \tilde{\mathbf{q}}^{(2)}, \ldots, \tilde{\mathbf{q}}^{(N)}]$. Empirically, conditional self-labeling significantly alleviates the confirmation bias, resulting in self-label assignments that are much closer to the expected distribution, as shown in Figure 1a. Further theoretical analysis regarding estimators from unconditional and conditional self-labeling is provided in Section 4. Then, the clustering objective has the form of: $\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^N H(\tilde{\mathbf{q}}^{(i)}, \mathbf{p}^{(i)})$.

3.2 Open-world hierarchical thresholding

Beyond the clustering objective, prompting the predictive confidence has proven effective for classification. A similar goal arises in traditional SSL, wherein entropy minimization is employed to encourage low entropy (i.e., high confidence) in the prediction. FixMatch [36] leverages both consistency and pseudo-labeling to achieve exceptional performance with the following regularization: $\sum_{i=1}^N \mathbb{I}(\max(\mathbf{p}^{(i)}) \geq \tau) H(\hat{\mathbf{p}}^{(i)}, g_\theta(\mathcal{A}(\mathbf{x}^{(i)})))$, where $\hat{\mathbf{p}}^{(i)} := \arg\max(\mathbf{p}^{(i)})$ is predictive one-hot pseudo-label, with the $\hat{p}^{(i)}$-th element set to 1. α and $\mathcal{A}$ represent weak and strong augmentation respectively. Here, τ is a scalar hyperparameter denoting the threshold above which we retain a pseudo-label. The effectiveness of the aforementioned regularization depends on accurate and sufficient pseudo-labels, which are directly influenced by the thresholding scheme. Under the close-word assumption, extensive efforts [47, 45, 42] have been devoted to devising thresholding techniques based on the idea of balancing learning pace across classes with varying learning difficulties. However, these techniques do not fit with the open-world scenario due to a critical challenge: the learning

pace of novel classes tends to be much slower [5]. The predictive confidence of these two groups does not share the same behavior, as shown in Figure 1b.

We introduce an open-world hierarchical thresholding scheme to balance this inconsistent learning pace at the group level, leveraging these well-defined thresholds to retain high-quality and adequate pseudo-labels for learning. As shown in Figure 3, this scheme first estimates the learning conditions of the two groups and then hierarchically modulates the thresholds in a class-specific fashion within each group.

First, we split the dataset into seen ($\mathcal{C}_s$) and novel ($\mathcal{C}_n$) groups based on the pseudo-label and estimate their overall learning condition by predictive confidence. Motivated by FreeMatch [42], we define the group-wise learning status for a set of classes $\mathcal{C}_i = \mathcal{C}_s$ or $\mathcal{C}_n$ as

$$\eta(\mathcal{C}_i) = \frac{1}{N_{\mathcal{C}_i}} \sum_{i=1}^{N} \max(\mathbf{p}^{(i)}) \mathbb{I}(\hat{p}^{(i)} \in \mathcal{C}_i),\ \mathcal{C}_i = \mathcal{C}_s \text{ or } \mathcal{C}_n, \tag{5}$$

where $N_{\mathcal{C}_i} = \sum_{i=1}^{N} \mathbb{I}(\hat{p}^{(i)} \in \mathcal{C}_i)$ denotes the number of samples whose predictive pseudo-labels belong to the group $\mathcal{C}_i$. Similarly, the class-wise learning conditions can be defined as

$$\zeta_c = \frac{1}{N_c} \sum_{i=1}^{N} \max(\mathbf{p}^{(i)}) \mathbb{I}(\hat{p}^{(i)} = c),\ c = 1, \ldots, K, \tag{6}$$

where $N_c = \sum_{i=1}^{N} \mathbb{I}(\hat{p}^{(i)} = c)$ denotes the number of samples whose predicted labels belong to the c-th class. In practice, we utilize the exponential moving average (EMA) to update at each iteration. Then, we merge these two learning statuses and obtain the open-world hierarchical threshold as

$$\tau(c) = \frac{\zeta_c}{\max_{c \in \mathcal{C}_i} \zeta_c} \cdot \eta(\mathcal{C}_i),\ c = 1, \ldots, K, \tag{7}$$

where the c-th class belongs to the set $\mathcal{C}_i$ (i.e., $c \in \mathcal{C}_s$ or $\mathcal{C}_u$). The learning condition η distinguishes between seen and novel classes, while the class-wise condition ζ adjusts for class-wise differences. Ultimately, the confidence objective is:

$$\mathcal{L}_{conf} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\max(\mathbf{p}^{(i)}) > \tau(\hat{p}^{(i)})) \cdot H(\hat{\mathbf{p}}^{(i)}, g_\theta(\mathcal{A}(\mathbf{x}^{(i)}))). \tag{8}$$

Together with the supervised objective $\mathcal{L}_{sup} = \frac{1}{N^l} \sum_{i=1}^{N^l} H(\mathbf{y}_{gt}^{(i)}, \mathbf{p}^{(i)})$, the overall objective for OwMatch is

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{cls} + \mathcal{L}_{conf}. \tag{9}$$

4 Theoretical analysis of conditional self-labeling

To illustrate the superiority of conditional self-labeling over unconditional, we evaluate their estimators of the class distribution on unlabeled data through rigorous statistical analysis. This transformation is justified as both self-labeling methods produce corresponding self-label assignments, each representing their estimation of the class distribution on unlabeled data.

Formulation. Assuming that the class distribution of real-world data conforms to prior information $\boldsymbol{\mathcal{P}} = [p_1, p_2, \cdots, p_K]$. Suppose real-world data is composed of recognized labeled data and unrecognized unlabeled data, conforming to unknown class distribution $\boldsymbol{\mathcal{P}}^l = [p_1^l, p_2^l, \cdots, p_K^l]$ and $\boldsymbol{\mathcal{P}}^u = [p_1^u, p_2^u, \cdots, p_K^u]$ respectively. We independently sample $N = N^l + N^u$ instances from recognized and unrecognized data, respectively. Suppose $N_i = N_i^l + N_i^u$ is composed of two random variables that denote the number of recognized and unrecognized samples belonging to the i-th class. Obviously, we have $N^l = \sum_{i=1}^{K} N_i^l$ and $N^u = \sum_{i=1}^{K} N_i^u$.

Objective. We hope to estimate the unknown class distribution $\boldsymbol{\mathcal{P}}^u$ with $\hat{\boldsymbol{\mu}}$ based on prior information $\boldsymbol{\mathcal{P}}$ and observations of $N_1^l, N_2^l, \cdots, N_K^l$, then evaluate $\hat{\boldsymbol{\mu}}$ from unbiasness and ECS. Evaluation on both metrics requires estimating the number of samples in each class, denoted by

 https://doi.org/10.52202/079017-3167

$\mathbf{A} = (A_1, A_2, \ldots, A_K)$. Two self-labeling approaches (unconditional and conditional) can optimize self-label assignment, therefore obtaining two approximations of $\mathbf{A}$, denoted by $\mathbf{A}_{\text{uncon}}$ and $\mathbf{A}_{\text{con}}$. Denote the corresponding estimators as $\hat{\boldsymbol{\mu}}_{\text{uncon}}$ and $\hat{\boldsymbol{\mu}}_{\text{con}}$.

Assumption 4.1. Assume that all drawn samples with a static number of samples and class distribution follow the multinomial distribution as follows,

$$\begin{aligned} N_1, N_2, \cdots, N_K &\sim \text{Multinomial}(N, \boldsymbol{\mathcal{P}}) \\ N_1^l, N_2^l, \cdots, N_K^l &\sim \text{Multinomial}(N^l, \boldsymbol{\mathcal{P}}^l) \\ N_1^u, N_2^u, \cdots, N_K^u &\sim \text{Multinomial}(N^u, \boldsymbol{\mathcal{P}}^u). \end{aligned}$$

Given the independency between N_i^l and N_j^u. We basically have:

$$\begin{aligned} &\mathbb{E}[N_i] = \mathbb{E}[N_i^l] + \mathbb{E}[N_i^u] \quad \forall i, j \\ &Np_i = N^l p_i^l + N^u p_i^u. \end{aligned} \tag{10}$$

Lemma 4.2. *Suppose we want to test the null hypothesis (H_0) that categorical data $N_1, N_2, \cdots, N_{\mathcal{C}}$ come from a multinomial distribution with K classes and class probability of $\boldsymbol{\mathcal{P}}$. A chi-square statistic can be constructed to test the deviation between the observations $n_1, \cdots, n_K$ and expected outcomes for each class.*

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i]^2)}{\mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i]} \sim \chi^2_{K-1}, \tag{11}$$

where $\mathbb{E}_{\boldsymbol{\mathcal{P}}}[\cdot]$ denotes the population expectation of random variable. A lower chi-square value suggests that the observed data are consistent with H_0. Conversely, an exceedingly high chi-square value implies that either H_0 is incorrect or an event of low probability has happened.

Details of the above lemma are presented in Appendix E. Then, we define the following metric to evaluate the goodness of fit of estimation based on chi-square statistics.

Definition 4.3 (Expectation of chi-square statistics (ECS))**.** The expectation of chi-square statistics (ECS) for $\hat{\boldsymbol{\mu}}$ are defined as the population deviation between the estimator of unlabeled class distribution $\hat{\boldsymbol{\mu}}$ and its true distribution $\boldsymbol{\mathcal{P}}^u$:

$$\text{ECS}(\hat{\boldsymbol{\mu}}) := \mathbb{E}[\chi^2(\mathbf{A})] = \mathbb{E}\left[\sum_{i=1}^{K} \frac{(A_i - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^u])^2}{\mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^u]}\right], \tag{12}$$

where $\mathbf{A}$ are estimators based on $N_1^l, N_2^l, \cdots, N_K^l$, thus are still random variables.

Now, we introduce two main theorems and demonstrate the superiority of our conditional self-labeling.

Theorem 4.4. *Consider two estimators for class distribution on unlabeled data, $\boldsymbol{\mu}_{\text{uncon}}$ and $\boldsymbol{\mu}_{\text{con}}$, we have $\boldsymbol{\mu}_{\text{uncon}}$ is a biased estimator and $\boldsymbol{\mu}_{\text{con}}$ is an unbiased estimator.*

Theorem 4.5. *Suppose $r_i := \frac{N^l \cdot p_i^l}{N}$ denote the ratio of label samples of the i-th class to the whole samples, $r := \sum_i r_i$ denotes the ratio of labeled samples to the whole samples. For unlabeled sample size N^u, if $\sqrt{N^u} > \frac{1}{\max(|r_i - r \cdot p_i^u|, r \cdot p_j)}$ for $\forall i \in \mathcal{C}_l, \forall j \in \mathcal{C}_u$, then $\text{ECS}(\hat{\boldsymbol{\mu}}_{\text{con}}) \leq \text{ECS}(\hat{\boldsymbol{\mu}}_{\text{uncon}})$.*

Following rigorous statistical analysis, the generated self-label assignments from the conditional labeling method are closer to the true class distribution in the following scenarios:

- Estimation based on large unlabeled sample size (N^u);
- The difference between prior distribution $\boldsymbol{\mathcal{P}}$ and class distribution of unlabeled data $\boldsymbol{\mathcal{P}}^u$ is not negligible.

5 Experiments

This section presents a comprehensive evaluation of our approach. It includes experimental results and in-depth analysis, demonstrating the effectiveness of our approach.

Table 1: Average accuracy on the CIFAR-10/100 and ImageNet-100 with both novel class ratio and label ratio of 50%. We compare OwMatch with existing literature on OwSSL. Also compared with other related approaches of traditional SSL, OSSL, and NCD approaches following [5]. Proper modifications are made to make these approaches compatible with OwSSL; the details are in Appendix C. The results are averaged over three independent runs. The baseline figures are sourced from the respective papers.

Method	**CIFAR-10**			**CIFAR-100**			**ImageNet-100**		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
FixMatch [36]	71.5	50.4	49.5	39.6	23.5	20.3	65.8	36.7	34.9
DS^3L [16]	77.6	45.3	40.2	55.1	23.7	24.0	71.2	32.5	30.8
CGDL [37]	72.3	44.6	39.7	49.3	22.5	23.5	67.3	33.8	31.9
DTC [18]	53.9	39.5	38.3	31.3	22.9	18.3	25.6	20.8	21.3
RankStats [17]	86.6	81.0	82.9	36.4	28.4	23.1	47.3	28.7	40.3
SimCLR [7]	58.3	63.4	51.7	28.6	21.1	22.3	39.5	35.7	36.9
UNO [12]	91.6	69.3	80.5	68.3	36.5	51.5	-	-	-
ORCA [5]	88.2	90.4	89.7	66.9	43.0	48.1	89.1	72.1	77.8
NACH [15]	89.5	92.2	91.3	68.7	47.0	52.1	91.0	75.5	79.6
OpenLDN [33]	95.7	95.1	95.4	73.5	46.8	60.1	89.6	68.6	79.1
TRSSL [34]	**96.8**	92.8	94.8	80.0	49.3	64.7	-	-	-
OpenCon [38]	89.3	91.1	90.4	69.1	47.8	52.7	90.6	80.8	83.8
OwMatch	93.0	95.9	94.4	74.5	55.9	65.1	**91.7**	72.0	81.8
OwMatch+	96.5	**97.1**	**96.8**	**80.1**	**63.9**	**71.9**	91.5	79.6	**85.5**

5.1 Experimental setup

Datasets. We evaluate our approach on CIFAR-10/100 [25], ImageNet-100 [11] and Tiny ImageNet [28]. A detailed description of these datasets is provided in Appendix A. Specifically, ImageNet-100 dataset contains 100 classes sub-sampled from ImageNet-1k following [39]. On all datasets, we first split all classes into seen and novel classes with a *novel class ratio*. Subsequent experiments will adopt a novel class ratio of 50% unless otherwise specified. Then, we will randomly assign labels to a portion of the data from the seen classes according to the specified *label ratio*, while the remaining data, along with all samples from the novel classes, are assigned to the unlabeled set.

Implementation details. For a fair comparison, we apply ResNet-50 [22] as the backbone model for ImageNet-100 and ResNet-18 for other benchmarks. We train the model with a batch size of 256 for Tiny ImageNet and 512 for other benchmarks. Following [15], experiments across all benchmarks are implemented based on the pre-trained model from SimCLR [7]. We jointly optimize backbone and prototype parameters using the standard Stochastic Gradient Descent (SGD) with momentum. We apply the cosine annealing learning rate schedule for all experiments. Techniques including multi-crop and queue structure [6] are employed to enhance the clustering objective. Additionally, RandAugment [9] serves as the strong augmentation for confidence objective. Additional implementation details are available in the Appendix B.

Evaluation metric. In assessing the efficacy of OwMatch, we adopt a multifaceted approach to evaluate accuracy following [5]. Evaluation metrics include the standard accuracy for seen classes and the clustering accuracy for novel classes and all classes. Here, we leverage the Hungarian algorithm [26] to align the predicted class assignment for novel-class instances with their ground-truth labels to obtain clustering accuracy. We also report the joint clustering accuracy across all classes using the Hungarian algorithm.

5.2 Main results

We consider and evaluate two versions of our method, called OwMatch and OwMatch+. OwMatch represents the standard version as illustrated in Figure 2, while OwMatch+ incorporates the multi-crop technique for additional augmentation. Detailed distinctions between the two versions are provided in the Appendix B. We evaluate our method on all benchmarks using a label ratio of 10% and 50% with the comprehensive experiment results provided in Table 1, 12, and 13. Results in Table 1 show that OwSSL approaches significantly outperform current state-of-the-art methods in traditional SSL, OSSL, and NCD by a considerable margin. On the other hand, OwMatch achieves

https://doi.org/10.52202/079017-3167

state-of-the-art across all benchmarks and evaluation metrics. It can not only classify novel classes accurately but also maintain robust performance on seen classes. On CIFAR-10, we observed OwMatch outperforms OpenLDN on novel and all classes by 2.0% and 1.4%, respectively. It is noteworthy that the enhancement brought about by OwMatch is more pronounced on the CIFAR-100 dataset, which presents a greater challenge due to the increasing number of classes. Regarding CIFAR-100, our method surpasses TRSSL by approximately 14.6% on novel classes and 7.2% on all classes. Subsequently, we extend to evaluate ImageNet-100 and observe a similar trend, with OwMatch+ showing significant improvement of **1.7%** on all-class accuracy compared to previous state-of-the-art approaches.

Principle analysis of conditional self-labeling. OwMatch primarily relies on high-quality self-label assignment to alleviate the model's confirmation bias. To clearly illuminate this progress during training, we employ the Manhattan distance $\sum_{i \in K} |c_i - c_i^{gt}|$ as a metric to evaluate the bias between the considered class distribution $\{c_i\}_{i=1}^K$ and the ground truth $\{c_i^{gt}\}_{i=1}^K$. Table 2 demonstrates the debiasing process: the model's confirmation bias is pronounced in the early epochs, whereas the bias of optimized self-label assignment is relatively minor. As training advances, the self-label assignment continues to guide the model, effectively mitigating the confirmation bias, as reflected in the decreasing B_m and the absolute difference between B_m and B_s.

Table 2: The Manhattan distance (MD) is used to evaluate the confirmation bias. The first row presents the bias between the model's predictive class distribution and the ground truth, denoted as B_m. The second row reflects the bias between the self-label assignment and the ground truth, denoted as B_s. The third row computes the absolute difference between B_m and B_s, highlighting the debiasing effect of high-quality self-label assignments.

Bias of	**Epoch 1**	**Epoch 2**	**Epoch 3**	**Epoch 6**	**Epoch 10**	**Epoch 30**	**Epoch 50**
B_m	0.4463	0.2939	0.2474	0.1753	0.1505	0.0904	0.0798
B_s	0.1004	0.0754	0.0893	0.0613	0.0407	0.0255	0.0219
$\lvert B_m - B_s \rvert$	0.3459	0.2185	0.1581	0.1140	0.1098	0.0649	0.0579

5.3 Ablations, analysis, and real-scenario applications

To investigate the impact of each component, we embark on comprehensive ablation studies with both novel class ratio and label ratio of 50%. The first row in Table 3 showcases the foundational model performance, whose objective consists of only *unconditional* clustering objective and supervised objective, already achieving impressive performance. We then analyze the effect of integrating a conditional self-labeling framework on CIFAR-100, which boosts novel-class accuracy by 1.0% on average. Additionally, the positive impact of consistency regularization is observed: roughly 0.9% enhancement across all evaluation metrics. Our ablation studies highlight the essential contribution of each component in OwMatch. Individually, each plays a significant part in the intended functionality, and together, these elements coalesce into a cohesive and robust framework. We also ablate other factors, including the number of local views for clustering objectives and iterations for the Sinkhorn-Knopp algorithm, with detailed statements provided in Table 15 and 16.

Table 3: Ablation studies on each component with both novel class ratio and label ratio of 50%. Here, **ConSL** refers to conditional self-labeling, **PLCR** refers to pseudo-label consistency regularization, and **OwHT** refers to an open-world hierarchical thresholding scheme.

Components			**CIFAR-10**			**CIFAR-100**			**Tiny-ImageNet**		
ConSL	PLCR	OwHT	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
×	×	×	96.5	90.2	93.3	78.8	56.7	67.7	66.5	38.1	52.0
✓	×	×	95.4	96.4	95.9	79.2	58.5	68.7	66.0	39.4	52.4
✓	✓	×	96.3	97.3	96.8	80.1	59.4	69.6	68.6	42.0	54.2
×	✓	✓	97.1	90.4	93.8	80.7	59.7	69.9	69.7	41.4	54.6
✓	✓	✓	96.5	97.1	96.8	80.1	63.9	71.9	68.8	42.4	55.0

Comparison study on varying thresholding strategies. We compare our proposed open-world hierarchical thresholding approach with two prominent techniques: static thresholding [36] and self-adaptive thresholding [42]. As illustrated in Table 4, our proposal achieves superior performance

in both novel- and all-class clustering accuracy across varying thresholding techniques. While self-adaptive has proven effective under closed-world scenarios, it encounters challenges in open-world settings. The pronounced disparity in overall learning conditions between seen and novel classes, as illustrated in Figure 1b, can lead to unstable global thresholds. The class-wise adaptive approach based on that may exaggerate this issue, resulting in suboptimal performance. We implement a hierarchical structure to mitigate the instability sourcing from distinct learning dynamics of seen and novel classes.

Table 4: Performance comparison of static, self-adaptive, and our OwHT thresholding techniques on CIFAR-100 with both novel class ratio and label ratio of 50%.

Thresholding	Seen Acc	Novel Acc	All Acc
Static - 0.7	80.1	59.4	69.6
Static - 0.8	79.8	63.9	71.7
Static - 0.9	80.2	62.8	71.3
Self-adaptive	**81.0**	60.5	70.6
OwHT	80.1	**63.9**	**71.9**

Table 5: Comparison with GCD-related works: average accuracy on the ImageNet-100 with both novel class ratio and label ratio of 50%.

Method	**Backbone**	Seen	Novel	All
GCD [40]	ViT-B/16	91.8	63.8	72.7
SimGCD [43]	ViT-B/16	93.1	77.9	83.9
InfoSieve [32]	ViT-B/16	84.9	78.3	80.5
CiPR [20]	ViT-B/16	84.9	78.3	80.5
PromptCAL [48]	ViT-B/16	92.7	78.3	83.1
OwMatch+	ResNet-50	91.5	**79.6**	**85.5**

On the comparison between OwSSL and Generalized Category Discovery. The OwSSL setting resembles the subsequently proposed Generalized Category Discovery (GCD) setting [40], with both assuming the existence of novel classes and that a portion of the data is labeled for seen classes. However, there are notable differences between these two groups of methods: 1) GCD-related methods leverage supervised contrastive learning [24] on labeled data and self-supervised contrastive learning [7] on all training data, whereas OwSSL typically employs pairwise similarity-based methods for clustering samples; 2) GCD-related works typically employ a pre-trained ViT-Base/16 backbone, which has significantly more parameters than the ResNet-18 or ResNet-50 models commonly used in OwSSL methods.

It is unfair to compare these two types of methods directly. Here, we still include a comparison with those GCD-related works to demonstrate the effectiveness of our method. Table 5 shows that our method outperforms existing approaches in novel-class and all-class accuracy on ImageNet-100 despite using a simpler model.

Ablation study on supervision components in the overall objective. The overall objective of OwMatch consists of a standard supervised objective, clustering objective, and confidence objective. Both the supervised and clustering objectives involve the use of labeled data, raising concerns about overlap in functionality. To investigate the significance of each component, we conduct an ablation study in which we modify the overall objective in two ways: 1) removing the supervised objective and 2) excluding labeled data from the online clustering process. The results are reported in Table 6.

In the first case, we observe a decrease in seen-class accuracy while maintaining novel-class clustering performance, while the latter case exhibits the opposite tendency: seen-class accuracy remains high, but novel-class clustering accuracy declines. In comparison to the previous cases, our overall objective integrates both components to strike a balance between clustering and confidence. The supervised objective enhances seen-class accuracy through one-hot supervision, while the clustering objective with conditional self-labeling improves novel-class clustering accuracy by incorporating labeled data. This harmonious approach yields the best all-class accuracy while roughly maintaining both seen- and novel-class performance.

Table 6: Model performance under varying applications of supervision in the overall objective.

Objective	Seen	Novel	All
w/o supervised objective	76.8	64.4	70.6
w/o supervision for clustering	80.3	61.2	70.7
overall objective	80.1	63.9	71.9

Table 7: Estimation of the number of classes across benchmarks.

	CIFAR10	CIFAR100	ImageNet100
Ground Truth	10	100	100
Estimation	10	104	111
Error	0%	4%	11%

The above evaluations are typically conducted under relatively ideal conditions: the datasets are class-balanced, and both the prior class distribution and the number of novel classes are available.

https://doi.org/10.52202/079017-3167

However, in real-world scenarios, it is crucial to address the dependency on these assumptions. We will elaborate on each of these aspects to demonstrate the practical effectiveness of OwMatch.

Estimating the number of novel classes. OwMatch and other baselines typically assume that the number of novel classes is pre-determined for clarity in evaluation. However, this prior knowledge is often unavailable in practice, necessitating a precise estimation of the number of novel classes in advance. We primarily follow the approaches of GCD [40] and TRSSL [34] to estimate the number of classes. Specifically, K-means clustering is performed on representations of the entire dataset from the pre-trained ViT-B/16 backbone. The optimal value of k is determined by evaluating the clustering accuracy on the labeled samples calculated by the Hungarian algorithm. This accuracy serves as a scoring function, optimized using Brent's algorithm to find the that maximizes performance on the labeled data. The estimation results across generic benchmarks are shown in Table 7, which illustrates that the estimated class numbers come close to the ground truth. We also evaluate OwMatch's sensitivity to varying extents of class number estimation error, with results reported in Appendix D.

Table 8: Performance on generic recognition benchmarks with varying imbalance factors (IF), with and without prior class distribution. These benchmarks come with both novel class ratio and label ratio of 50%.

Dataset	**Prior**	**Uniform (IF=1)**			**IF=10**			**IF=20**		
		Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
CIFAR-10	w/	96.5	97.1	96.8	93.7	72.1	82.5	92.9	70.1	80.9
	w/o	96.9	90.9	93.9	95.8	66.5	80.3	95.3	64.2	78.8
CIFAR-100	w/	80.1	63.9	71.9	76.8	42.0	57.3	76.1	35.2	51.9
	w/o	82.5	57.9	69.2	74.6	39.7	54.1	73.9	33.9	49.2
Tiny-ImageNet	w/	68.8	42.4	55.0	61.7	25.1	41.6	62.4	21.7	38.3
	w/o	69.6	40.6	54.8	61.0	24.9	40.1	61.3	20.3	36.9

Data imbalance. Most generic benchmarks feature class-balanced, whereas real-world data tend to exhibit long-tailed class distribution. Our approach accommodate to arbitrary class distribution by constraining the optimized self-label assignment to comply with the prior class distribution, thereby naturally mitigating performance degradation caused by data imbalance. We evaluate our approach on imbalanced benchmarks, constructed with varying imbalance factors following TRSSL [34]. Results in Table 8 demonstrate that OwMatch effectively addresses the challenge of data imbalance.

Training without prior. In scenarios where prior class distribution is unavailable, we propose an adaptive estimation scheme to make OwMatch still function *without relying on any prior assumptions*. Specifically, we initially adopt class-balanced prior if no prior information is available; then, the class distribution for conditional self-labeling is estimated and continuously updated based on model prediction. Next, standard training with estimated class distribution and distribution estimation are alternately conducted, with results reported in Table 8. We observe that the reduction in all-class accuracy achieved through the adaptive estimation scheme remains within 3% across almost all benchmarks and imbalance factors. These results reveal that the straightforward estimation technique performs robustly in the absence of prior knowledge.

6 Conclusion

This work integrates techniques from self-SL and SSL, refining them to present a new perspective on solving open-world SSL. We demonstrate that conditional self-labeling can achieve an unbiased estimation of the class distribution on unlabeled data with prior information, leading to high-quality self-label assignment with reduced confirmation bias. Our future endeavors will be directed toward developing solutions that are more aligned with realistic scenarios where such prior information might not be readily available or hard to be estimated. This will involve exploring methodologies that can effectively handle uncertainty and variability inherent in real-world data distributions.

Acknowledgements

Chao Wang would like to acknowledge the support from the National Natural Science Foundation of China under Grant 12201286, the Shenzhen Science and Technology Program 20231115165836001, Guangdong Basic and Applied Research Foundation 2024A1515012347. This research was conducted using the computing resources provided by the Research Center for the Mathematical Foundations of Generative AI in the Department of Applied Mathematics at The Hong Kong Polytechnic University.

References

[1] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 8443–8452, 2021.

[3] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[5] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML*, pages 1597–1607. PMLR, 2020.

[8] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3569–3576, 2020.

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 702–703, 2020.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[12] Enrico Fini, Enver Sangineto, Stéphane Lathuiliere, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR*, pages 9284–9292, 2021.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[14] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 17, 2004.

 https://doi.org/10.52202/079017-3167

[15] 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 *Advances in Neural Information Processing Systems*, May 2022.

[16] 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 *Proceedings of the 37th International Conference on Machine Learning*, pages 3897–3906. PMLR, November 2020.

[17] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations (ICLR)*, 2020.

[18] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 8401–8409, 2019.

[19] Tao Han, Junyu Gao, Yuan Yuan, and Qi Wang. Unsupervised semantic aggregation and deformable template matching for semi-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9972–9982, 2020.

[20] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*, 2023.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[23] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. 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 (CVPR)*, pages 8310–8319, 2021.

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

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

[26] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[27] P. Langley. Crafting papers on machine learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 1207–1216, 2000.

[28] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[29] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[30] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and machine intelligence*, 41(8):1979–1993, 2018.

[31] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[32] Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] 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 *European Conference on Computer Vision (ECCV)*, pages 382–401. Springer, 2022.

[34] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 437–455. Springer, 2022.

[35] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34:25956–25967, 2021.

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:596–608, 2020.

[37] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional Gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13480–13489, 2020.

[38] Yiyou Sun and Yixuan Li. OpenCon: open-world contrastive learning. *Transactions on Machine Learning Research*, 2023.

[39] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision (ECCV)*, pages 268–285. Springer, 2020.

[40] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.

[41] Xudong Wang, Long Lian, and Stella X Yu. Unsupervised selective labeling for more effective semi-supervised learning. In *European Conference on Computer Vision (ECCV*, pages 427–445. Springer, 2022.

[42] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[43] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16590–16600, 2023.

[44] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6256–6268, 2020.

[45] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021.

[46] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 438–454. Springer, 2020.

[47] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:18408–18419, 2021.

 https://doi.org/10.52202/079017-3167

[48] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023.

[49] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16623–16633, 2023.

Technical Appendices

Roadmap of technical appendices. These appendices are structured as follows: Appendix A introduces dataset details utilized in our experiments. Appendix B outlines the implementation specifics, including data augmentations and hyperparameters. Modification details for utilized baselines are provided in Appendix C. Additional experiment results consisting of supplementary main results, in-depth analysis, and ablation study of hyperparameters are reported in Appendix D. Complete and rigorous proof of theoretical results is represented in Appendix E. Appendix F discusses the social impacts of our work, and the limitations are considered in Appendix G

A Datasets

The details of the datasets utilized in our experiments are provided in Table 9, which includes dataset statements, as well as the corresponding backbones and batch sizes for training. The choice of backbone and batch size matches previous works [5, 15] for fair comparison. For CIFAR-10/100 datasets, we employ a simple pre-processing encompassing random crop with padding and horizontal flip. To make ResNet-18 compatible with CIFAR input data with a small resolution of 32×32, we refine the CNN by setting the kernel size to 3×3 and applying a stride of 1. We train the model with a batch size of 512 over 300 epochs. For ImageNet-100 and Tiny-ImageNet datasets, raw images are pre-processed with random resized crop and horizontal flip [22]. We use the standard version of ResNet-50 and ResNet-18 as the backbone, respectively. We leverage the standard SGD method with momentum and weight decay to optimize the network parameters; see hyperparameters in Table 11.

Table 9: Details of evaluation benchmarks, we show the number of classes, dataset statistics, selected backbone, and batch size for training.

Dataset	Num. Class	Train Samples	Test Samples	Backbone	Batch size
CIFAR-10 [25]	10	50,000	10,000	ResNet-18	512
CIFAR-100 [25]	100	50,000	10,000	ResNet-18	512
ImageNet-100 [11]	100	128,545	5,000	ResNet-50	512
Tiny-ImageNet [28]	200	100,000	10,000	ResNet-18	256

B Implementation details

Computational resources. The foundational algorithm of our study is constructed utilizing Python 3.8 and PyTorch 2.1 [31]. All experiments are carried out on NVIDIA's Tesla A100 GPU with 40G memory. All benchmarks are public and can be easily downloaded.

Strong augmentation. Furthermore, we apply the strong augmentation to input data for all experiments following FixMatch [36], including random resized crop, horizontal flip, and RandAugment [9]. It should be noted that the only additional enhancement in strong augmentation is RandAugment compared to basic pre-processing. Specifically, RandAugment randomly selects transformations from a collection of options for each sample in a mini-batch. We employed the same sets of image transformations as those used in RandAugment. A complete list of these transformations can be found in the original work [9].

OwMatch v.s. OwMatch+. We evaluate two versions of our approach in the main results, as illustrated in Table 1, 12 and 13. In short, OwMatch+ further incorporates the multi-crop technique as an additional augmentation to boost clustering capacity and, hence, improve model performance. The multi-crop strategy was proposed by SwAV [6] to additionally augment images by covering only small sections. The resulting low-resolution images, referred to as local views, allow more augmentations at a marginally increased computational cost. Compared to simple preprocessing (global views), local views generated through a multi-crop strategy involve resized cropping at smaller scales but in greater numbers, as illustrated in Table 10. And they experience additional color distortion [7] consisting of random color jitters, solarizing, and equalization in pursuit of model robustness. Here we apply the multi-crop technique to produce many low-resolution images (local views) and set the hyperparameters following PAWS [2], as detailed in Table 10.

https://doi.org/10.52202/079017-3167

Table 10: The details of crop augmentation hyperparameters.

Dataset	Global views		Local views			
	Crop scale	Resize	Crop scale	Resize	Numbers	Distortion
CIFAR-10	(0.75,1)	32	(0.3,0.75)	18	4	0.5
CIFAR-100	(0.75,1)	32	(0.3,0.75)	18	4	0.5
ImageNet-100	(0.2,1)	224	(0.14,0.3)	18	4	1
Tiny-ImageNet	(0.75,1)	64	(0.3,0.75)	36	4	1

Except for additional multi-crop augmentations, OwMatch and OwMatch+ differ slightly in the form of clustering objective. Specifically, we augment the input image by taking 2 full-resolution crops (normally and strongly augmented global views) and V low-resolution crops (local views). Note that we optimize the self-label $\tilde{\mathbf{q}}$ with only global views, since local views can only capture localized semantic information and are unable to provide a comprehensive overview of the entire image. We promote the model consistency by encouraging the prediction of different local views to be close to the optimized self-labels. Specifically, the clustering objective of OwMatch+ is formulated by

$$\mathcal{L}_{cls}^{+} = \frac{1}{(V+1)N} \sum_{i=1}^{N} \left[\sum_{v=1}^{V} H(\tilde{\mathbf{q}}^{(i)}, g_\theta(\alpha_v(\mathbf{x}^{(i)}))) + H(\tilde{\mathbf{q}}^{(i)}, \mathbf{p}^{(i)}) \right], \tag{13}$$

where $\tilde{\mathbf{q}}^{(i)}$ correspond to the optimized self-label of global view, and $g_\theta(\alpha_1(\mathbf{x}^{(i)})), \cdots, g_\theta(\alpha_V(\mathbf{x}^{(i)}))$ stand for predictions of V local views. Increasing the number of random low-resolution crops encourages the model to learn global-to-local information [2], which reflects in performance gain across all benchmarks; see main results in Table 1, 12 and 13. The utilization of low-resolution images boosts the model's efficiency with only a marginal rise in computational costs.

The effectiveness of conditional self-labeling is significantly compromised when the ratio of batch size (B) to class numbers (K) is relatively small. In a scenario where a class is disproportionately sampled in the labeled data of a batch, the conditional self-labeling mechanism might be unable to reassign the unlabeled data to that particular class. Notably, when this ratio falls below 1, assigning B samples to all K classes becomes unfeasible. Therefore, we leverage the queue structure to store data from previous batches by following SwAV [6]. In practice, a queue of 1024 logits are stored for the implementation of the Sinkhorn-Knopp algorithm, which is utilized to derive the self-label assignment. We then retain the logits from the last batch of the optimized self-labels to construct the clustering objective. Such queue length proves effective in our experiments with a large batch size (e.g., 512) and a relatively small number of categories (e.g., 100 classes for CIFAR-100). However, when dealing with high-resolution images that encompass a greater variety of categories, storing much more previous data information is essential.

Hyperparameters Here, we provide a comprehensive list of hyperparameters in Table 11. For hyperparameters related to the SGD optimizer, we adhere to the settings used in the previous works [5, 34] to ensure a fair comparison. Regarding hyperparameters introduced by our proposed method, we perform ablation studies to determine the most appropriate values, specifically for SK-iteration and the number of local views, as detailed in Appendix D. These hyperparameters are selected based on a balanced consideration of computational costs and model performance.

Table 11: List of hyperparameters for CIFAR-10/100, ImageNet-100, and Tiny-ImageNet.

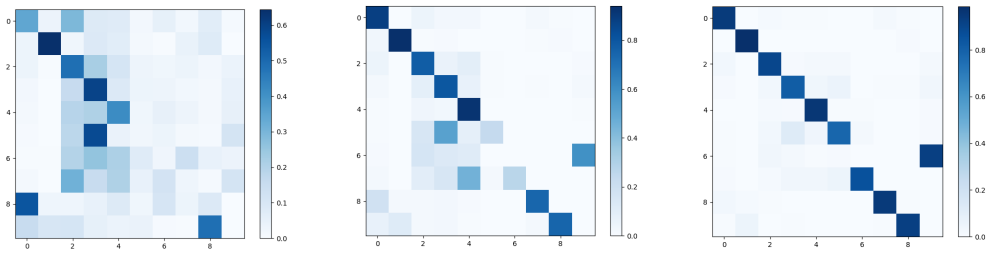
Hyper-parameter	CIFAR-10	CIFAR-100	ImageNet-100	Tiny-ImageNet
SGD-momentum	0.9			
SGD-learning rate	0.1			
SK-ϵ	10			
SK-iteration	10			
# of local views	4			
SGD-weight decay	0.0005	0.0005	0.0001	0.0001

C Baseline implementation details

We compare our proposal with baselines from other settings: traditional SSL, open-set semi-supervised learning (OSSL), and novel class discovery (NCD). We will elaborate on the modifications to these settings separately. Traditional SSL methods cannot deal with the novel-class instances and we extend it in the following manner: samples are firstly divided into seen-class and novel-class instances based on out-of-distribution (OOD) criteria, then we report the standard classification accuracy on seen-class instances and apply K-means algorithms to achieve clustering accuracy on the novel-class instances. Hungarian algorithm [26] is utilized to match the clustering result and their ground-truth labels, this result is reported as novel-class accuracy. For the traditional SSL method, FixMatch [36], we separate the OOD samples based on confidence scores produced by the Softmax function. Many OSSL approaches like CGDL [37] are naturally capable of detecting outliers, thus we directly cluster the considered outliers by the K-means algorithm and report the clustering accuracy without manually inspecting OOD samples. Since DS3L [16] applies the re-weighting technique to downsize the passive effect of OOD samples, we consider the samples with the lowest weight as outliers. Both traditional SSL and weighting-based OSSL approaches depend on the OOD likelihood score to partition the inliers and outliers. Here we follow ORCA [5] to determine the threshold for OOD samples by using ground-truth class information.

NCD methods are trained to discover novel classes in unlabeled data with totally novel-class instances, thus failing to recognize seen-class instances. For NCD approaches without seen-class classification heads, like DTC [18] and RankStats [17], we report the performance on novel classes and extend them to classify seen classes by assuming the seen-class instances in unlabeled data as novel. Then we extend the unlabeled classification head to include logits for seen classes by following [5] and leverage the Hungarian algorithm to match the discovered classes with ground-truth labels within labeled data, the best assignment is reported as seen-class accuracy. And for recent UNO [12] with explicitly labeled classification heads, we generate pseudo-labels for both seen- and novel-class instances based on model predictions from concatenated labeled and unlabeled classification heads. Therefore, both seen and novel class classification accuracy can be directly computed and reported. We apply the same pre-trained model on NCD methods to demonstrate that the enhanced performance is not attributable to the the application of pre-training. Additionally, we present the clustering outcomes based on representations from the pre-trained model.

D Additional results



(a) Epoch 1 (b) Epoch 10 (c) Epoch 100

Figure 4: Confusion matrices on CIFAR-10 with both novel class ratio and label ratio of 50%. The model needs to classify the initial five seen classes accurately (as reflected in the diagonal elements). While for the novel classes (6-10), the classes clustering are required to align with the ground-truth label (dark blue in one cell).

Training process. We plot the confusion matrices on CIFAR-10 with both novel class ratio and label ratio of 50% in Figure 4. This collection of images compellingly demonstrates that the bias derived from novel classes is progressively mitigated as the experiment advances, leading to continuous improvement in the model's prediction accuracy.

As depicted in Figure 4a, at the beginning of training, the model struggles to effectively distinguish between novel and seen class instances, although it can classify seen class instances normally. As training progresses, an increasing number of samples are assigned to the novel classes and prediction

https://doi.org/10.52202/079017-3167

accuracy for the seen-class instances advances, as illustrated in Figure 4b. In the later stages of training as shown in Figure 4c, the model becomes capable of accurately classifying the seen-class instances and clustering novel-class instances simultaneously.

Table 12: Average accuracy on the CIFAR-10/100 and ImageNet-100 with novel class ratio of 50% and labeled ratio of 10%.

Method	**CIFAR-10**			**CIFAR-100**			**ImageNet-100**		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
FixMatch [36]	64.3	49.4	47.3	30.9	18.5	15.3	-	-	-
DS^{3}L [16]	70.5	46.6	43.5	33.7	15.8	15.1	-	-	-
DTC [18]	42.7	31.8	32.4	22.1	10.5	13.7	-	-	-
RankStats [17]	71.4	63.9	66.7	20.4	16.7	17.8	-	-	-
UNO [12]	86.5	71.2	78.9	53.7	33.6	42.7	66.0	42.2	53.3
ORCA [5]	82.8	85.5	84.1	52.5	31.8	38.6	83.9	60.5	69.7
NACH [15]	91.8	89.4	90.6	65.8	37.5	49.2	-	-	-
OpenLDN [33]	92.4	93.2	92.8	55.0	40.0	47.7	-	-	-
TRSSL [34]	94.9	89.6	92.2	68.5	**52.1**	**60.3**	82.6	67.8	75.4
OpenCon [38]	-	-	-	62.5	44.4	48.2	-	-	-
OwMatch	89.3	92.2	90.7	59.5	43.7	51.2	86.4	69.2	77.8
OwMatch+	94.4	**96.2**	**95.3**	**69.9**	51.5	**60.3**	**87.8**	**72.7**	**80.2**

Main results with label ratio of 10%. We evaluate our approach on CIFAR-10/100 and ImageNet-100, closely similar to Table 1, but with the label ratio adjusted to 10%. The results are detailed in Table 12; OwMatch+ continues to maintain state-of-the-art performance across most benchmarks and evaluation metrics.

Performance sensitivity to varying extents of class number estimation error. As illustrated in Figure 5, OwMatch maintains robust performance over a range of errors.

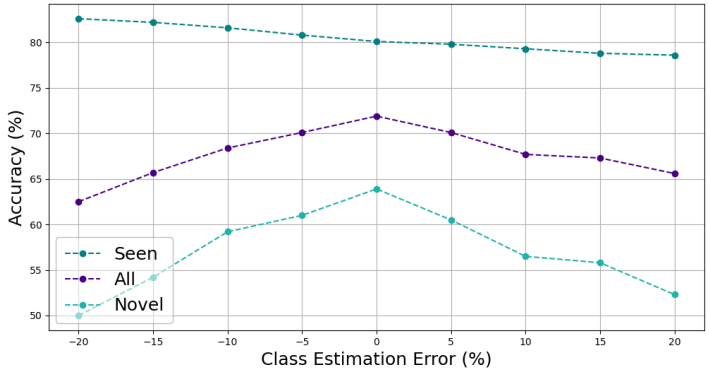


Figure 5: Accuracy as a function of class number estimation error on **CIFAR-100** dataset.

Main results on Tiny-ImageNet. In addition to the results on CIFAR and ImageNet datasets, we also evaluate OwMatch and OwMatch+ on Tiny-ImageNet with 50% novel classes in Table 13. We found that both OwMatch and OwMatch+ surpass previous state-of-the-art methods across benchmarks and evaluation metrics.

Different novel class ratio. Previously, we assessed model's performance with a constant novel class ratio of 50%, which is also variable in real-world scenarios. Here, we alter this value and fix the label ratio within seen classes to 50%; the results are reported in Table 14. The model's performance generally exhibits a declining trend across all benchmarks and evaluation metrics as the novel class ratio increases. It is important to note that as the number of novel classes increases, the total amount of labeled data decreases.

Table 13: Average accuracy on Tiny-ImageNet datasets with label ratio of 10% and 50%.

Method	**50% label ratio**			**10% label ratio**		
	Seen	Novel	All	Seen	Novel	All
DTC [18]	28.8	16.3	19.9	13.5	12.7	11.5
RankStats [17]	5.7	5.4	3.4	9.6	8.9	6.4
UNO [12]	46.5	15.7	30.3	28.4	14.4	20.4
OpenLDN [33]	58.3	25.5	41.9	-	-	-
TRSSL [34]	59.1	24.2	41.7	39.5	20.5	30.3
OwMatch	62.9	38.9	50.6	44.8	25.4	34.7
OwMatch+	**68.8**	**42.4**	**55.0**	**55.8**	**33.3**	**43.3**

Table 14: Model performance across benchmarks with varying novel class ratios.

Novel class ratio	**CIFAR10**			**CIFAR100**			**Tiny-ImageNet**		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
50%	96.5	97.1	96.8	80.1	63.9	71.9	68.8	42.4	55.0
60%	96.5	92.1	93.9	80.3	60.4	68.1	68.5	41.1	51.7
70%	97.5	91.0	93.0	81.3	58.6	65.3	71.3	34.7	45.6
80%	98.4	92.9	94.0	78.9	58.0	61.8	69.5	32.9	40.0
90%	97.8	93.6	94.0	82.0	50.7	53.5	69.4	26.4	30.3

Efficient labeling strategy under a fixed budget. In the previously conducted experiments, we assessed the model's performance by altering the novel class ratio and label ratio, respectively. Merely altering a single factor does not yield highly convincing inferences, as it is impractical to fix the novel class ratio or the label ratio for some classes in real-world scenarios. Here, we consider a fixed budget, specifically the total number of labeled data, as illustrated in Figure 6. This comparison aims to shed light on how the balance between labeled data in seen classes and the proportion of novel classes influences model performance under a fixed level of supervision.

When the supervisory information is held constant, a configuration with a smaller portion of labeled data spread across a greater number of different classes results in higher accuracy for both all classes and novel classes. Additionally, it is observed that as the number of novel classes decreases, the accuracy of the seen classes improves. This improvement is attributed to the reduced complexity of the classification task when there are fewer categories to be classified. From these observations, it can be inferred that within a limited labeling budget, it is more effective to label a broader category of samples, thereby capturing as many representative points as possible within the feature space. This strategy appears to optimize the model's performance across both known and novel classes.

Hierarchical thresholding scheme with scarce supervision information. Previously, we observed that in the scenario where the label ratio on seen classes is 50%, the performance difference between adopting the hierarchical thresholding strategy and setting a high static threshold hyperparameter is not significant. Here, we maintain the novel class ratio as 50% while decreasing the label ratio to 10%. Figure 7a and Figure 7b demonstrate that a hierarchical thresholding scheme not only retains more pseudo-labels but also preserves the predictive accuracy of selected pseudo-labels.

Table 15: Model performance on CIFAR-10/100 and Tiny-ImageNet with different number of local views to contrastive learning.

# of Crops	**CIFAR10**			**CIFAR100**			**Tiny-ImageNet**		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
2	95.9	97.7	96.8	78.8	61.6	70.2	66.4	43.6	54.2
4	96.5	97.1	96.8	80.1	63.9	71.9	68.8	42.4	55.0
6	96.8	97.2	97.0	81.1	59.4	69.8	70.1	42.1	55.6

Different number of local views in contrastive learning. Contrastive learning boosts clustering by acquiring different views of the data and promoting consistency across these views at the representation level. The manner of data augmentation and the number of augmented views largely determine

https://doi.org/10.52202/079017-3167

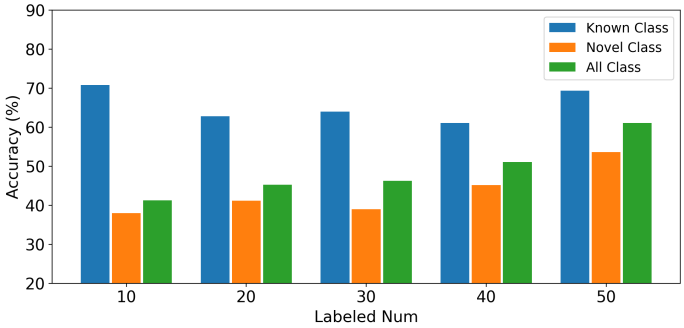
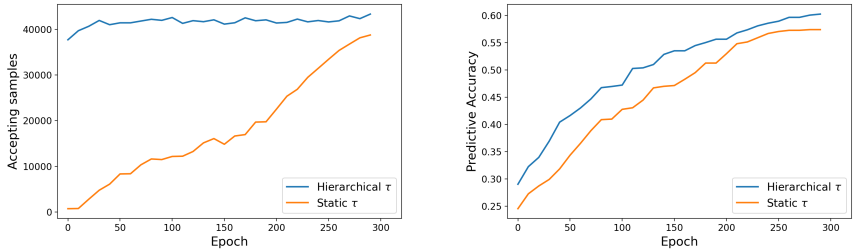


Figure 6: Model performance on CIFAR-100 with varying numbers of seen classes and a fixed amount of labeled data (5% of total data). Models trained with more seen classes generally perform better on both novel and all classes, despite having fewer labeled data per seen class.



(a) Number of selected pseudo-labels

(b) Pseudo-labels accuracy

Figure 7: Open-world hierarchical thresholding scheme generally selects more instances as pseudo-labels (a), and the quality of pseudo-labels is also enhanced (b).

the model performance. Here we take the combination of color distortion and random resized crop [7] as augmentation and apply a multi-crop technique [6] to produce many low-resolution images (local views). We compare the performance with varying number of local views, the results are reported in Table 15.

The incorporation of local views significantly enhances model performance. However, it's observed that as the number of local views increases, the incremental benefits to the model come to plateau, while the training and computational time considerably escalate. This aligns with the notion that local views assist in capturing local patterns, aiding in the development of robust representations. However, there is a threshold beyond which the addition of more local views contributes less to learning, overshadowed by the rise in computational demands.

Different number of iterations in the Sinkhorn-Knopp algorithm. Conditional self-labeling is proposed to optimize high-quality self-label assignments, which depends on a fast version of the Sinkhorn-Knopp algorithm [10] to solve this complex linear programs efficiently. Resolving this involves iterative processes to converge on the optimal solution. We evaluate the model performance across different iterations, with the results presented in Table 16.

Generally, model performance tends to increase with more iterations of the Sinkhorn-Knopp algorithm. Despite the positive correlation tendency, we observe that certain iterations (e.g., 6 for CIFAR-100 with both novel class ratio and label ratio of 50%) already achieve satisfactory outcomes, with only marginal gains from further iterations.

E Proof details

Here, we derive ECS for both unconditional and conditional self-labeling in the following two lemmas.

https://doi.org/10.52202/079017-3167

Table 16: Model performance on CIFAR-10/100 and Tiny-ImageNet with different iteration numbers in Sinkhorn-Knopp algorithm.

SK-iters	CIFAR10			CIFAR100			Tiny-ImageNet		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
3	96.8	91.6	94.2	81.5	55.0	67.6	68.6	41.7	54.1
6	96.7	91.0	93.9	81.4	62.4	71.3	68.0	42.8	54.5
10	96.5	97.1	96.8	80.1	63.9	71.9	68.8	42.4	55.0

Lemma E.1. *Estimators of* $\mathbf{A}$ *from unconditional self-labeling has the form of*

$$\mathbf{A}_{\text{uncon}} = [N^u p_1, N^u p_2, \cdots, N^u p_K], \tag{14}$$

and ECS for $\hat{\boldsymbol{\mu}}_{\text{uncon}} = \frac{1}{N^u}\mathbf{A}_{\text{uncon}}$ *can be derived as*

$$\text{ECS}(\hat{\boldsymbol{\mu}}_{\text{uncon}}) = \sum_{i=1}^{K} \frac{N^u (p_i - p_i^u)^2}{p_i^u}. \tag{15}$$

Lemma E.2. *Estimators of* $\mathbf{A}$ *from conditional self-labeling has the form of*

$$\mathbf{A}_{\text{con}} = [Np_1 - N_1^l, \cdots, Np_K - N_K^l], \tag{16}$$

and ECS for $\hat{\boldsymbol{\mu}}_{\text{con}} = \frac{1}{N^u}\mathbf{A}_{\text{con}}$ *can be derived as*

$$\text{ECS}(\hat{\boldsymbol{\mu}}_{\text{con}}) = \sum_{i=1}^{K} \frac{N^l p_i^l (1 - p_i^l)}{N^u p_i^u}. \tag{17}$$

E.1 Proof of Lemma 4.2

Proof. Under the null hypothesis H_0, the sample size of the i-th class follows the Binary distribution with parameters of N and p_i. Therefore, we have the expectation and standard error of N_i as $\mathbb{E}_{\boldsymbol{\mathcal{P}}}(N_i) = Np_i$ and $\mathbb{SD}(N_i) = Np_i(1 - p_i)$, respectively, where the standard deviation of N_i measures the average deviation of random variable N_i from its expected value. And for observation n_i for each class, the discrepancy can be denoted as $n_i - \mathbb{E}_{\boldsymbol{\mathcal{P}}}(N_i)$. To ensure discrepancies for each class are evaluated on a consistent basis, dividing them by their standard errors under H_0, which enables us to focus on the standardized variable $\frac{n_i - \mathbb{E}_{\boldsymbol{\mathcal{P}}}(N_i)}{(Np_i(1-p_i))^{1/2}}$.

Note that $\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mathcal{P}}}(N_i) = \sum_{i=1}^{K} Np_i = N$, therefore the discrepancies across the K classes cannot simultaneously be all positive or all negative. To measure the discrepancies of all K classes, we sum the squares of the discrepancies of each class. This formulation ensures that discrepancies are independent of sign, merely reflecting the deviation between observed and expected values under H_0:

$$\frac{(n_1 - Np_1)^2}{Np_1(1-p_1)} + \frac{(n_2 - Np_2)^2}{Np_2(1-p_2)} + \cdots + \frac{(n_K - Np_K)^2}{Np_i(1-p_K)}. \tag{18}$$

We then prefer to exclude the factors $(1 - p_i)$ from the denominators of the sum for two primary reasons. Firstly, if numerous classes exist and none of them has significant large probabilities, then $(1 - p_i)$ is almost negligible. Moreover, when the expectation $\mathbb{E}_{\boldsymbol{\mathcal{P}}}(N_i)$ is substantial, the approximation of each discrepancy adheres to a standard normal distribution, which leads to the construct of chi-square statistics with the degree of freedom of $K - 1$. Then the summary statistic can be obtained as follows

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - Np_i)^2}{Np_i} = \sum_{i=1}^{K} \frac{(n_i - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i]^2)}{\mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i]} \sim \chi^2_{K-1}. \tag{19}$$

□

https://doi.org/10.52202/079017-3167

E.2 Proof of Lemma E.1

Proof. TRSSL [34] posits that unlabeled and real data share the same class distribution, and optimize the self-label assignment on unlabeled data solely based on prior information $\boldsymbol{\mathcal{P}}$. Owing to the alignment constraint between the generative self-label assignment and prior information, as shown in (1). Then the estimated number of samples in each class from conditional self-labeling is

$$\mathbf{A}_{\text{uncon}} = [N^u p_1, N^u p_2, \cdots, N^u p_K], \tag{20}$$

where both N^u and $p_1, \cdots, p_K$ are known, therefore $\mathbf{A}_{\text{uncon}}$ is a constant vectors. Then the estimator $\hat{\boldsymbol{\mu}}_{\text{uncon}} = \frac{1}{N^u}\mathbf{A}_{\text{uncon}}$ is also a constant vectors. Thus, the ECS for $\hat{\boldsymbol{\mu}}_{\text{uncon}}$ can be calculated directly,

$$\begin{aligned}
\text{ECS}(\hat{\boldsymbol{\mu}}_{\text{uncon}}) =& \mathbb{E}\left[\sum_{i=1}^{K} \frac{(N^u p_i - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^u])^2}{\mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^u]}\right] && (21)\\
=& \sum_{i=1}^{K} \frac{(N^u p_i - N^u p_i^u))^2}{N^u p_i^u} && (22)\\
=& \sum_{i=1}^{K} \frac{N^u (p_i - p_i^u)^2}{p_i^u}. && (23)
\end{aligned}$$

□

E.3 Proof of Lemma E.2

Proof. Different from an unconditional setting, conditional self-labeling considers partial supervision. Specifically, except for the constraint brought from prior information, an additional constraint is constructed to realize the alignment between clustering results and existing labels in labeled data, as shown in (3). Then the estimated number of samples in each class from conditional self-labeling is

$$\mathbf{A}_{\text{con}} = [Np_1 - N_1^l, \cdots, Np_K - N_K^l], \tag{24}$$

where $N, p_1, p_2, \cdots, p_K$ are static values, and $N_1^l, \cdots, N_K^l$ are a set of random variables. Note that in a specific experiment, we can get a set of observations of $N_1^l, \cdots, N_K^l$, thus estimation based on conditional self-labeling is feasible. Then consider the estimator $\hat{\boldsymbol{\mu}}_{\text{con}} = \frac{1}{N^u}\mathbf{A}_{\text{con}}$, derivation steps of the ECS for $\hat{\boldsymbol{\mu}}_{\text{con}}$ are as follow,

$$\text{ECS}(\hat{\boldsymbol{\mu}}_{\text{uncon}}) = \mathbb{E}\left[\sum_{i=1}^{K} \frac{(Np_i - N_i^l - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^u])^2}{\mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^u]}\right] = \mathbb{E}\left[\sum_{i=1}^{K} \frac{(Np_i - N_i^l - N^u p_i^u)^2}{N^u p_i^u}\right]. \tag{25}$$

According to (10), we have

$$\mathbb{E}\left[\sum_{i=1}^{K} \frac{(Np_i - N_i^l - N^u p_i^u)^2}{N^u p_i^u}\right] = \mathbb{E}\left[\sum_{i=1}^{K} \frac{(N_i^l p_i^l - N_i^l)^2}{N^u p_i^u}\right] = \mathbb{E}\left[\sum_{i=1}^{K} \frac{(N_i^l - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^l])^2}{N^u p_i^u}\right]. \tag{26}$$

Since ECS is also defined at the population level, thus we have

$$\mathbb{E}\left[\sum_{i=1}^{K} \frac{(N_i^l - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^l])^2}{N^u p_i^u}\right] = \left[\sum_{i=1}^{K} \frac{\mathbb{E}(N_i^l - \mathbb{E}_{\boldsymbol{\mathcal{P}}}[N_i^l])^2}{N^u p_i^u}\right] = \sum_{i=1}^{K} \frac{\text{Var}(N_i^l)}{N^u p_i^u}. \tag{27}$$

Since $N_1^l, N_2^l, \cdots, N_K^l \sim \text{Multinomial}(N^u, \boldsymbol{\mathcal{P}}^l)$, we have

$$\sum_{i=1}^{K} \frac{\text{Var}(N_i^l)}{N^u p_i^u} = \sum_{i=1}^{K} \frac{N^l p_i^l (1 - p_i^l)}{N^u p_i^u}. \tag{28}$$

□

E.4 Proof of Theorem 4.4

Proof. From Theorem E.2 and Theorem E.3, we have that

$$\hat{\boldsymbol{\mu}}_{\text{uncon}} = [p_1, p_2, \cdots, p_K] = \boldsymbol{\mathcal{P}}, \tag{29}$$

$$\hat{\boldsymbol{\mu}}_{\text{con}} = \left[\frac{Np_1 - N_1^l}{N^u}, \cdots, \frac{Np_K - N_K^l}{N^u}\right]. \tag{30}$$

Note that $\hat{\boldsymbol{\mu}}_{\text{uncon}}$ is exactly the prior information, $\hat{\boldsymbol{\mu}}_{\text{con}}$ are function of a set of random variables $N_i^l, N_2^l, \cdots, N_K^l$, then we have

$$\mathbb{E}(\hat{\boldsymbol{\mu}}_{\text{con}}) = \left[\mathbb{E}(\frac{Np_1 - N_1^l}{N^u}), \cdots, \mathbb{E}(\frac{Np_K - N_K^l}{N^u})\right] \tag{31}$$

$$= \left[\frac{Np_1 - N^l p_1^l}{N^u}, \cdots, \frac{Np_1 - N^l p_K^l}{N^u}\right]. \tag{32}$$

According to (10), we have

$$\left[\frac{Np_1 - N^l p_1^l}{N^u}, \cdots, \frac{Np_1 - N^l p_K^l}{N^u}\right] = \left[\frac{N^u p_1^u}{N^u}, \cdots, \frac{N^u p_K^u}{N^u}\right] \tag{33}$$

$$= [p_1^u, p_2^u, \cdots, p_K^u] = \boldsymbol{\mathcal{P}}^u. \tag{34}$$

Thus, $\hat{\boldsymbol{\mu}}_{\text{con}}$ is a unbiased estimator of $\mathcal{C}^u$. □

E.5 Proof of Theorem 4.5

Proof. Note that $r_i := \frac{N^l \cdot p_i^l}{N}$ is non-negative and obtain zero if and only if the i-th class denote a novel class.
When the i-th class refers to seen classes, we have $p_i^u = \frac{p_i - r_i}{1-r}$, then

$$N^u(p_i - p_i^u)^2 \geq N^u(r_i - r \cdot p_i^u)^2 \geq 1.$$

When the i-th class refers to an novel class, we have $p_i^u = \frac{p_i}{1-r}$, then

$$N^u(p_i - p_i^u)^2 = N^u \cdot \left(p_i - \frac{p_i}{1-r}\right)^2 = N^u \cdot \left(\frac{rp_i}{1-r}\right)^2 \geq N^u(r \cdot p_i)^2 \geq 1.$$

Therefore, we get that

$$\text{ECS}(\hat{\boldsymbol{\mu}}_{\text{uncon}}) - \text{ECS}(\hat{\boldsymbol{\mu}}_{\text{con}}) = \sum_{i=1}^{K} \frac{N^u(p_i - p_i^u)^2 - \frac{N^l}{N^u} p_i^l(1 - p_i^l)}{p_i^u} \tag{35}$$

$$\geq \sum_{i=1}^{K} \frac{N^u(p_i - p_i^u)^2 - 1}{p_i^u} \geq 0. \tag{36}$$

□

F Broader impacts

This research delves into the issue of semi-supervised learning (SSL) in situations where not all classes possess labeled instances, an aspect that has received limited attention within the realm of SSL. We aim to draw increased focus towards examining the resilience of SSL in diverse real-world scenarios, thereby fostering a broader application of SSL in various contexts. However, the current accuracy is not very high for some challenging datasets. Therefore, the predictive results should be best used as references rather than treated as ground truth.

https://doi.org/10.52202/079017-3167

G Limitations

OwMatch, similar to existing methods in OwSSL, faces a significant challenge when applied to imbalanced datasets or unknown prior class distribution. Existing OwSSL methods are typically applied on class-balanced datasets where instances of each class share nearly the same frequency; the model performance would deteriorate when encountering imbalanced datasets. On the other hand, prior class distributions are not available in real-world applications. Addressing the dependency on prior class distribution and effectively handling datasets of arbitrary composition remain challenging for existing OwSSL algorithms, including OwMatch.

Recognizing this, we propose an adaptive estimation scheme for the OwMatch framework and show its feasibility in the experiments, with results reported in Table 8. Although a performance decline within 3% may seem acceptable, it is worth further consideration and exploration to determine whether further optimizations can enhance model performance without any prior. At the same time, we aim to prove the convergence of this adaptive algorithm in our future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim our scope and contributions clearly in abstract and Section 1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work and possible refinements in Appendix G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-3167

Justification: We provide all theoretical results with detailed assumptions in Section 4 and complete proof in Appendix E from the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all training implementation details in Section 5.1 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the datasets are public as illustrated in Appendix B. We will release the full code and commands once the manuscript gets accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We claim the experiment setting and details in Section 5.1 and Appendix B. Choosing hyper-parameters is based on the observations of the ablation study, as illustrated in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The reported results are averaged after running experiments multiple times with the same random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-3167

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information regarding compute resources in Appendix B from the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have meticulously reviewed the NeurIPS Code of Ethics to ensure compliance with the requirements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the social impact in Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work focuses on classification algorithms using public datasets and therefore poses no risks concerning safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite all the original papers involved. All utilized dataset and models are public, as detailed in Appendix B.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/079017-3167

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces newly developed code, which is well documented. The documentation includes detailed descriptions of the code's purpose, functionality, usage examples, and license details. The documentation will be publicly available.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3167