MUTUAL INFORMATION-GUIDED KNOWLEDGE TRANSFER FOR OPEN-WORLD SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We tackle the open-world semi-supervised learning problem, aiming to cluster novel classes and classify seen classes in unlabeled data based on labeled data from seen classes. The main challenge is to transfer knowledge contained in seen class data to unseen ones. Previous methods mostly transfer knowledge through sharing representation space. However, they learn the seen and unseen classes classifier in a disjoint manner, neglecting the underlying relation between predictions on the seen and unseen classes. Therefore, the learned representations and classifiers are less effective for clustering unseen classes. In this paper, we propose a novel and general method to transfer knowledge between seen and unseen classes. Our insight is to utilize mutual information to measure the generic statistical dependency between seen and unseen classes in the classifier output space, which couple the learning of classifier and promote transferring knowledge between two data sets. To validate the effectiveness and generalization of our method, we conduct extensive experiments on several benchmarks, including CIFAR10/100, Imagenet100, Oxford-IIIT Pet and FGVC-Aicraft datasets. Our results show that the proposed method outperforms previous SOTA by a significant margin on almost all benchmarks.

1 INTRODUCTION

Recent development of deep learning has achieved remarkable success in a broad range of visual recognition tasks (He et al., 2016; 2017; Ren et al., 2015). However, most traditional models focus on the closed-world setting, in which all the visual classes are pre-defined. As a result, it is usually hard to deploy those models in realistic settings with novel classes. In contrast, human visual systems can automatically learn new classes unseen before without supervision. Inspired by such human ability, several studies (Han et al., 2019; Hsu et al., 2018a) propose the task of Novel Class Discovery (NCD) which aims to learn novel categories from their unlabeled data based on a set of seen classes. In order to cope with more realistic scenarios, recent works (Cao et al., 2022; Vaze et al., 2022) have extended the NCD problem into a more challenging setting, termed by Open-world Semi-Supervised learning (OWSSL), where unlabeled data contains both seen and unseen classes. Although NCD and OWSSL settings have somewhat different assumptions on unlabeled data, they both share a similar challenge that requires clustering unseen classes based on the knowledge of seen classes.

A key strategy for discovering novel classes is to learn a semantically meaningful feature representation by transferring knowledge contained in the classifier of seen classes to the unseen. Most existing NCD and OWSSL methods (Han et al., 2021; Zhong et al., 2021a; Cao et al., 2022) focus on exploiting the learned knowledge of seen classes through sharing feature representations and learn two classifier heads for the seen and unseen classes in a disjoint manner. While such strategies have achieved promising results, they often neglects the potential relation between the classifier outputs of the seen and unseen classes, which limits the scope of shared knowledge and may lead to inferior representation for unseen classes.

Nonetheless, it is difficult to model the semantic relationship between the seen and unseen classes in the NCD/OWSSL setting as the unseen classes are unknown to us beforehand. In this work, we instead consider capturing a generic statistical dependency between two groups of classes based on a information-theoretic measure defined in the classifier output space. To verify our hypothesis, we conduct an empirical study to estimate the mutual information (MI) between the class predictions on the labeled and unlabeled data in an OWSSL task during training, and to investigate its correlation

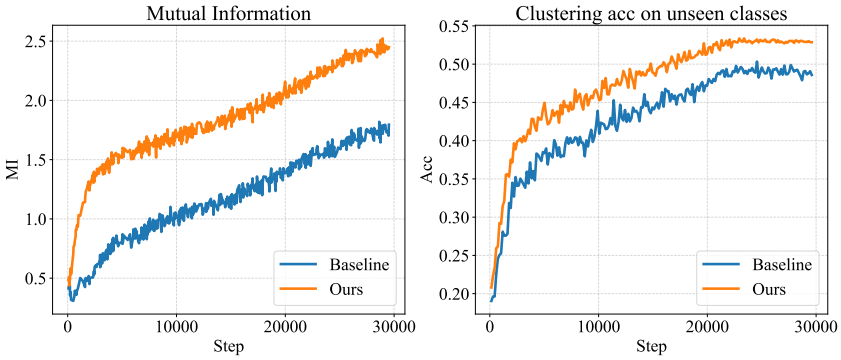

Figure 1: Mutual information and cluster acc on unseen classes. The left figure is the mutual information between labeled and unlabeled data in the projected model prediction space. The right figure is the clustering accuracy of unseen class in unlabeled data.

with the clustering accuracy of unseen classes. As shown in Fig.1, the mutual information estimation gradually increases during training, and the cluster accuracy also increases accordingly, indicating a positive correlation between those two quantities.

Motivated by the above observation, we propose a novel learning framework for the task of open-world semi-supervised learning, which augments the classifier training with an additional constraint based on the mutual information between labeled and unlabeled data in the model prediction space. This constraint couples the classifier learning of seen and unseen classes, enabling us to transfer additional knowledge on predictive distributions. Intuitively, maximizing mutual information enforces the model to reduce the predictive uncertainty on unlabeled data given the knowledge of seen classes.

Specifically, we adopt a two-head network architecture consisting of a ConvNet encoder and two classifier heads for the seen and unseen classes respectively. To learn the model under the OWSSL setting, we design a novel learning strategy that comprises three loss terms. The first loss term is a typical cross-entropy loss on the labeled data that extracts semantic knowledge from the seen classes. The second term is an unsupervised loss on the unlabeled data, which can take different forms and is agnostic to our learning strategy. The third term is the mutual information between the labeled and unlabeled data, aiming to facilitate transferring knowledge on classifiers from the seen classes to the unseen. To estimate the mutual information, we develop an approximate sampling method along with a projection operator for the OWSSL task. Given this loss, we train our classifier network for both seen and unseen classes in an end-to-end manner.

To validate the effectiveness of our method, we conduct extensive experiments on five datasets (Cao et al., 2022; Fini et al., 2021), including CIFAR10/100, ImageNet100, and two fine-grained datasets. Our results surpass the previous state of the art by a large margin in most cases, demonstrating the efficacy of our MI-based design. To summarize, our contributions are three-fold:

1. We propose a simple and effective learning framework to facilitate knowledge transfer from the seen to unseen classes, which provides a new perspective to solving OWSSL problems.
2. We design a new regularization strategy based on the mutual information between labeled and unlabeled data in a projected prediction space for coupling the classifier learning.
3. Our method significantly outperform previous works on four out of five public benchmarks under a broad range of learning settings.

2 Related Work

Novel class discovery: The idea of novel class discovery was initially explored in Hsu et al. (2018a;b), which perform transfer learning across domains and tasks, and utilize predictive pairwise similarity as the knowledge for clustering. The problem was formalized by Han et al. (2019), aiming to cluster unlabeled data with the help of labeled data. Most NCD methods attempt to transfer knowledge from labeled to unlabeled data by learning a shared representation. Han et al. (2019) utilize deep embedding clustering to learn a representation from labeled and unlabeled simultaneously. Han et al. (2021) propose robust rank statistics to measure the similarity of two data

in its representation space. Zhong et al. (2021a) apply contrastive learning to learn a discriminative representation. Zhong et al. (2021b) utilize the mixup strategy to combine the labeled and unlabeled data. Fini et al. (2021) utilize the Sinkhorn-Knopp algorithm to generate pseudo labels and apply a unified classification loss for both labeled and unlabeled data. Chi et al. (2021) solve novel class discovery from a meta-learning perspective. Although these methods have achieved some success, few of them consider the potential relationship between the seen and unseen classes in the prediction space during model learning.

Semi-supervised learning: There have been a large body of literature on semi-supervised learning (Van Engelen & Hoos, 2020). For semi-supervised learning of deep neural networks, most of recent works are based on the idea of enforcing certain types of consistency on the labeled and/or unlabeled data (Rasmus et al., 2015; Samuli Laine, 2017; Tarvainen & Valpola, 2017; Berthelot et al., 2019; Sohn et al., 2020; Zhang et al., 2021; Zheng et al., 2022; Li et al., 2021). However, those methods cannot tackle the scenarios where novel classes exist. To deal with novel classes, Cao et al. (2022) proposes the open-world semi-supervised learning, which bridges the gap between semi-supervised learning and novel class discovery. And it proposes an uncertainty adaptive margin mechanism to learn from labeled data gradually. Meanwhile, Vaze et al. (2022) tackle the same problem and present a baseline that uses vision transformers with contrastive learning. Concurrently, Rizve et al. (2022a;b) propose more effective pseudo labeling strategies for unlabeled data. However, as in NCD, most of the previous works focus on learning a shared feature representation for transferring knowledge between classes, while ignoring the relationship in the classifier output space.

Knowledge transfer: Transferring knowledge between classes has been explored in many different learning paradigms. For instance, transfer learning (Zamir et al., 2018; Zhuang et al., 2020; Weiss et al., 2016) typically aims to transfer knowledge from the source domain to the target domain, and knowledge distillation (Hinton et al., 2015; Gou et al., 2021; Park et al., 2019) aims to transfer knowledge from a large model to a small model. Among them, the idea of Ahn et al. (2019) is most relevant to our method, in which they transfer knowledge by retaining high mutual information between the layers of the teacher and student networks. In contrast, we transfer knowledge by maximizing mutual information between labeled and unlabeled data.

Mutual information: Mutual information is widely used in representation learning (Van den Oord et al., 2018; Poole et al., 2019; Sordoni et al., 2021; Hjelm et al., 2018; Bachman et al., 2019; Ji et al., 2019). Van den Oord et al. (2018) propose InfoNCE, which learns the underlying shared information between different parts of the signal. Later works (Poole et al., 2019; Sordoni et al., 2021) analyze the InfoNCE estimation process and proposed a better strategy to maximize the mutual information between input and its representation. Ji et al. (2019) propose invariant information clustering, which maximizes the mutual information between the class assignments of input data and its transformation, aimed to learn a robust representation. By contrast, we maximize the mutual information between labeled and unlabeled data in model prediction space and transfer knowledge from labeled to unlabeled.

3 METHOD

In this section, we first introduce the problem setup in Sec. 3.1 and the overview of our method in Sec. 3.2. Then we describe the losses on the labeled data for the seen classes and unlabeled data with unseen classes in Sec. 3.3. Finally, we present our mutual information constraint loss in Sec. 3.4, which is the core design of our method.

3.1 PROBLEM SETUP

The training dataset consists of two parts: a labeled set $\mathcal{D}^l = \{x_i^l, y_i^l\}_{i=0}^{|\mathcal{D}^l|}$ and a unlabeled set $\mathcal{D}^u = \{x_j^u\}_{j=0}^{|\mathcal{D}^u|}$. Here x, y represent an input data and the corresponding label, respectively. We use $Y^l = \{1, 2, ..., m\}$ to represent the category space of labeled data and Y^u to represent the category space of unlabeled data. While the typical semi-supervised learning setting assumes $Y^l = Y^u$, and the standard NCD problem requires $Y^l \cap Y^u = \emptyset$, we follow the open-world semi-supervised learning setting (Cao et al., 2022; Vaze et al., 2022), which assumes unlabeled data contains both known and unknown categories, i.e. $Y^l \subset Y^u$ and $Y^u = \{1, 2, ..., m + n\}$. The goal is to cluster unknown class while classify seen class in $\mathcal{D}^u$ with the presence of $\mathcal{D}^l$.

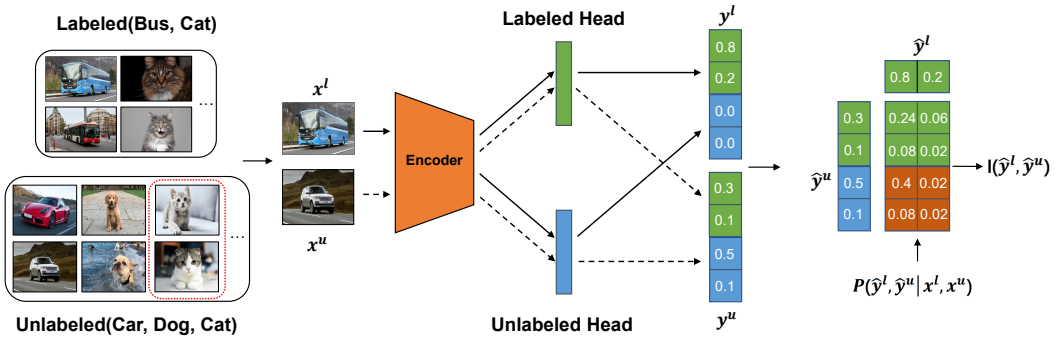

Figure 2: The diagram of our method. For simplicity, the figure mainly shows the detail of mutual information. First, we select a labeled data and unlabeled data from dataset (x^l, x^u), gets their prediction (y^l, y^u), and project y^l, y^u to $\hat{y}^l, \hat{y}^u$ by function π (See sec 3.4). Then we estimates joint distribution $p(\hat{y}^l, \hat{y}^u | x^l, x^u)$. After that, we integral $p(\hat{y}^l, \hat{y}^u | x^l, x^u)$ over many (x^l, x^u) pairs to estimate $p(\hat{y}^l, \hat{y}^u)$, subsequently compute the mutual information $I(\hat{y}^l, \hat{y}^u)$. By maximizing mutual information, the model are enforced to minimize the uncertainty of unlabeled data, leading to better representation of unlabeled data.

3.2 Method Overview

We address the open-world semi-supervised learning problem by facilitating knowledge transfer from labeled data to unlabeled ones. To this end, we propose a generic method that regularizes the classifier learning by maximizing mutual information between the labeled and unlabeled data in a projected model prediction space. This enables us to couple the feature and classifier learning of seen and unseen classes, which leads to more discriminative representations for both groups of classes.

As shown in the Fig.2, our model consists of an encoder and two heads, corresponding to classification head for labeled data and clustering head for unlabeled data. We select a batch of labeled and unlabeled data pairs (x^l, x^u), and project them into the embedding space through the shared encoder. Then we feed them to the labeled and unlabeled head. Note that no matter whether an input is labeled or unlabeled, it will go through two heads to generate two outputs. Finally, we concatenates two outputs as the final prediction. The forward prediction process can be formulated as follows:

$$p(y|x) = \text{Softmax}((h_l \cdot f(x) \oplus h_u \cdot f(x))/\tau) \tag{1}$$

where $p(\cdot)$ is the model predictive distribution, f is the encoder, h_l and h_u are labeled and unlabeled head, and τ is the temperature of the softmax function. Our objective function consists of three terms: 1) a supervised loss for the labeled data, 2) an unsupervised loss for the unlabeled data, and 3) a mutual information loss for the labeled and unlabeled data. The overall loss can be written as:

$$\mathcal{L} = \mathcal{L}_l + \alpha \mathcal{L}_u + \beta \mathcal{L}_{MI} \tag{2}$$

where $\mathcal{L}_l$ is the standard supervised loss on labeled data, $\mathcal{L}_u$ is the unsupervised loss for unlabeled data, and $\mathcal{L}_{MI}$ is our novel mutual information loss, which aims to enforce coupling classifier learning of the labeled and unlabeled data. Here α, β are the weighting factors.

3.3 Losses for labeled and unlabeled data

We now present the first two loss terms for the labeled and unlabeled data, respectively. We note that the choice of those two losses is orthogonal to our method and here we summarize some widely-used loss functions. For the supervised loss on the labeled data, we adopt the standard cross-entropy loss. For the unsupervised loss on the unlabeled data, we evaluate our method with two options: 1) a pairwise similarity loss (Hsu et al., 2018a; Han et al., 2021; Cao et al., 2022) and 2) a self-labeling loss (Asano et al., 2020; Fini et al., 2021).

Pairwise similarity loss (Chang et al., 2017; Hsu et al., 2018a): The pairwise similarity loss encourages grouping a pair of similar data, thus learning compact representation for unlabeled data. Specifically, given a batch of B unlabeled data, we compute the embedding $z^u = f(x^u)$ and

prediction $\mathbf{y}^u = p(y^u|x^u)$. For each unlabeled data, we find its nearest neighbor in the embedding space from the B unlabeled data, and denote the nearest neighbor of z_i^u as $\hat{z}_i^u$. The pairwise loss (Cao et al., 2022) can be written as:

$$\mathcal{L}_u = \frac{1}{|\mathcal{D}^u|} \sum_{i=0}^{|\mathcal{D}^u|} -\log(\mathbf{y}_i^u)^T \hat{\mathbf{y}}_i^u \tag{3}$$

To prevent all the unseen classes from collapsing to a single cluster, Cao et al. (2022) also introduce a simple entropy regularization term to regularize the size of cluster.

Self-labeling loss (Asano et al., 2020): The self-labeling loss first generates pseudo label for unlabeled data, then utilizes the generated pseudo label to self-train the model. It assumes unlabeled data are equally partitioned into each cluster and utilizes Sinkhorn-knopp algorithm to find an approximate assignment. We denote $\mathbf{y}^q = q(y^u|x^u), \mathbf{y}^p = p(y^u|x^u)$, and $\mathbf{y}^p, \mathbf{y}^q \in \mathbb{R}^{(m+n)\times 1}$. Let $\mathbf{Q} = [\mathbf{y}_1^q, \mathbf{y}_2^q, , , \mathbf{y}_B^q]\frac{1}{B}$, $\mathbf{P} = [\mathbf{y}_1^p, \mathbf{y}_2^p, , , \mathbf{y}_B^p]\frac{1}{B}$ be the joint distribution of B sampled data. We estimate $\mathbf{Q}$ by solving an optimal transport problem. We refer readers to (Cuturi, 2013; Asano et al., 2020) for details. The optimal $\mathbf{Q}$ is the pseudo label of unlabeled data and we denote the optimal pseudo label as $q^*(y^u|x^u)$. The self-labeling loss is formulated as:

$$\mathcal{L}_u = \frac{1}{|\mathcal{D}^u|} \sum_{i=0}^{|\mathcal{D}^u|} -q^*(y_i^u|x_i^u) \log p(y_i^u|x_i^u) \tag{4}$$

To transfer knowledge between labeled and unlabeled data, previous methods (Han et al., 2019; Zhao & Han, 2021; Cao et al., 2022) couple the learning of labeled and unlabeled data by sharing the encoder f. The models are typically learned by optimizing the supervised cross-entropy loss on labeled data and the pairwise similarity loss or self-labeling loss on unlabeled data. Such parameter sharing and the joint optimized model allows them to learn representations helpful for unseen classes clustering. However, due to the noise contained in the pseudo label for unlabeled data, those methods tend to be biased toward the seen classes, leading to an inferior representation for the unseen classes. What's more, the classifier learning of seen and unseen classes is disjoint, neglecting the potential relation constraint via the model output space. Below, we introduce a regularization term to constrain the model learning and facilitate the representation learning of the unseen classes.

3.4 Mutual Information Loss

We first introduce a graphical model for the data distributions. As shown in Fig. 3, the feature extractor parameterized by θ encode $\mathbf{X}^l, \mathbf{X}^u$ to z^l, z^u, which lie in the same feature space, and the knowledge are transferred through the shared feature space. The classifier parameters ω^l and ω^u are typically learned via the two loss terms mentioned above in the prior works. To better transfer knowledge, we propose a novel mutual information regularization term between Y^l and Y^u. Therefore, the learning of ω^l, ω^u is coupled, which can facilitate transferring the knowledge contained in the seen classes classifier to the unseen one. According to the graphical model, our assumption is that maximizing mutual information between Y^l and Y^u enforces the model to learn underlying relation between $\mathbf{X}^l$ and $\mathbf{X}^u$, leading to a better representation of unseen classes.

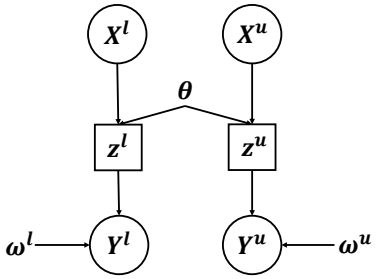

Figure 3: The graph of OWSSL problem. $\theta, \omega^l, \omega^u$ is the parameter of feature extractor, labeled and unlabeled classes classifier.

We intend to estimate the mutual information between seen and unseen classes, but they are mixed in the unlabeled data in our learning setting. Therefore, to simply our design, we directly estimate mutual information between labeled and unlabeled data, and maximize mutual information to transfer knowledge from labeled seen classes to unlabeled seen and unseen classes. We argue that maximizing the mutual information between labeled and unlabeled seen classes enforce the predictions of unlabeled seen classes to be more confident, while maximizing mutual information between labeled seen classes and unlabeled unseen classes can help model to transfer the knowledge contained in seen classes classifier to unseen classes. To transfer knowledge between classifier, we

approximate mutual information in the projected model prediction space. Specifically, we sample a batch of x_l, x_u pairs, and compute the model prediction distribution $p(y^l|x^l), p(y^u|x^u)$. Due to y^l, y^u are independent when given x^l and x^u, i.e. $p(y^l, y^u|x^l, x^u) = p(y^l|x^l) \cdot p(y^u|x^u)^T$. Finally, we marginalize over a batch of data pairs:

$$p(y^l, y^u) = \frac{1}{B}\sum_{b=1}^{B} p(y_b^l|x_b^l) \cdot p(y_b^u|x_b^u)^T \tag{5}$$

$$p(y^l) = \frac{1}{B}\sum_{b=1}^{B} p(y_b^l|x_b^l); \quad p(y^u) = \frac{1}{B}\sum_{b=1}^{B} p(y_b^u|x_b^u) \tag{6}$$

So, the mutual information can be estimated by:

$$I(y^l, y^u) = -\sum_{y^l}\sum_{y^u} p(y^l, y^u) \log \frac{p(y^l, y^u)}{p(y^l)p(y^u)} \tag{7}$$

Maximizing mutual information will reduce uncertainty of unlabeled data when given the knowledge of labeled data by utilizing underlying relations between labeled and unlabeled data. In our problem, all labeled data only contain seen classes, and unlabeled data contain both seen and unseen classes. The probability of labeled data belonging to the unseen class is less useful. Therefore we introduce a projection function π to project labeled data prediction y^l to $\hat{y}^l$, which only contains seen class probability. Then we re-normalize the projected class probability. The formulation of π function is:

$$\pi(y) = \begin{cases} \text{Softmax}(h_l \cdot f(x)/\tau), & x \in \mathcal{D}^l \\ y^u, & x \in \mathcal{D}^u \end{cases} \tag{8}$$

So, the mutual information term is estimated by:

$$\mathcal{L}_{MI} = -I(\hat{y}^l, \hat{y}^u) = \sum_{\hat{y}^l}\sum_{y^u} p(\hat{y}^l, y^u) \log \frac{p(\hat{y}^l, y^u)}{p(\hat{y}^l)p(y^u)} \tag{9}$$

We also provide another interpretation of our projection function. Without projection function, mutual information $I(y^l, y^u)$ can be decomposed into $H(y^l) - H(y^l|y^u)$. Maximizing mutual information will maximize $H(y^l)$ and minimize $H(y^l|y^u)$. $H(y^l)$ is maximized when $p(y^l)$ is partitioned equally to each class. So maximizing $H(y^l)$ will enforce the model to classify seen class data into unseen classes. Such an optimization direction is inconsistent with the cross-entropy loss. Minimizing $H(y^l|y^u)$ with the unseen class probability in y^l, on the other hand, involves unnecessary extra dimensions as stated above and hence leads to inefficient learning. For example, in the cifar100-50 setting, $y^l \in \mathbb{R}^{100\times 1}$, and $\hat{y}^l \in \mathbb{R}^{50\times 1}$. Therefore, we introduce the above projection functions to project y^l to seen classes prediction space to remove such undesirable effects.

4 Experiments

4.1 Experimental setup

Dataset: We evaluate the performance of our method on typical CIFAR10/100 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009), OxfordIIIT-Pet (Parkhi et al., 2012) and FGVC-Aircraft (Maji et al., 2013). Following Cao et al. (2022), we first divide all classes into 50% seen and 50% unseen classes, then select 50% of the seen classes data as labeled dataset. The rest data are treated as an unlabeled set, which contains all unseen class data and half of seen class data. The details of the dataset split are shown in Appendix Tab.9.

Metric: Similar to previous work (Cao et al., 2022), we evaluate our method on the unlabeled set, which contains both seen and unseen classes. First, we adopt the Accuracy and ClusterAcc to measure the performance of seen and unseen classes separately. However, we can not distinguish unlabeled data as seen or unseen in advance, and such a separate metric ignores the confusion between seen and unseen classes. Therefore, we utilize all the unlabeled data to calculate ClusterAcc. The formulation of ClusterAcc is:

$$ClusterAcc = \max_{perm\in P} \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i = perm(\hat{y}_i)\} \tag{10}$$

where $y_i, \hat{y}_i$ represent the ground-truth and predicted labels, respectively, and P represents the set of all permutations and combinations. We use the Hungarian algorithm to optimize permutations.

Table 1: Results on the unlabeled dataset. † denotes the results we obtained by running their code. * indicates the same augmentation is used.

Method	CIFAR10			CIFAR100-50			ImageNet100-50		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
FixMatch	71.5	50.4	49.5	39.6	23.5	20.3	65.8	36.7	34.9
DS^{3}L	77.6	45.3	40.2	55.1	23.7	24.0	71.2	32.5	30.8
CGDL	72.3	44.6	39.7	49.3	22.5	23.5	67.3	33.8	31.9
DTC	53.9	39.5	38.3	31.3	22.9	18.3	25.6	20.8	21.3
RankStats	86.6	81.0	82.9	36.4	28.4	23.1	47.3	28.7	40.3
SimCLR	58.3	63.4	51.7	28.6	21.1	22.3	39.5	35.7	36.9
ORCA†	$88.0_{\pm 0.2}$	$\mathbf{88.9}_{\pm 0.6}$	$88.8_{\pm 0.5}$	$66.3_{\pm 0.0}$	$41.1_{\pm 0.4}$	$46.2_{\pm 0.1}$	$89.1_{\pm 0.1}$	$69.9_{\pm 0.8}$	$76.1_{\pm 0.5}$
ORCA*	$88.0_{\pm 0.2}$	$\mathbf{88.9}_{\pm 1.6}$	$88.8_{\pm 1.1}$	$67.6_{\pm 0.0}$	$47.0_{\pm 1.4}$	$49.0_{\pm 0.6}$	$89.6_{\pm 0.2}$	$71.8_{\pm 1.3}$	$76.8_{\pm 1.0}$
ORCA* + **MI**	$90.1_{\pm 0.3}$	$87.9_{\pm 0.3}$	$88.6_{\pm 0.1}$	$69.0_{\pm 0.2}$	$48.9_{\pm 0.2}$	$53.4_{\pm 0.7}$	$90.4_{\pm 0.1}$	$74.8_{\pm 0.8}$	$79.3_{\pm 0.7}$
UNO*	$\mathbf{93.6}_{\pm 0.3}$	$88.1_{\pm 1.0}$	$89.9_{\pm 0.6}$	$75.4_{\pm 0.1}$	$48.3_{\pm 0.2}$	$55.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$59.4_{\pm 0.5}$	$67.8_{\pm 0.5}$
UNO* + **MI**	$93.0_{\pm 0.2}$	$88.4_{\pm 1.0}$	$\mathbf{89.9}_{\pm 0.6}$	$\mathbf{78.3}_{\pm 0.1}$	$\mathbf{53.4}_{\pm 0.6}$	$\mathbf{60.6}_{\pm 0.6}$	$\mathbf{93.7}_{\pm 0.0}$	$\mathbf{88.2}_{\pm 0.3}$	$\mathbf{88.3}_{\pm 0.0}$

Table 2: Results with 10% labeled data.

Method	Imagenet100-50		
	Seen	Novel	All
FixMatch	60.9	33.7	30.2
RankStats	41.2	26.8	37.4
ORCA	85.2	64.6	71.6
UNO	79.5	61.4	67.9
Ours	**88.7**	**88.6**	**84.4**

Table 3: Results with ViT Backbone.

Method	Imagenet100-50		
	Seen	Novel	All
K-means	75.5	71.3	72.7
UNO	93.8	62.4	68.4
ORCA	89.3	64.6	73.5
GCD	89.8	66.3	74.1
Ours	**97.6**	**92.8**	**92.5**

Implementation Details: We use ResNet18 as our backbone network for all dataset except Imagenet100, which we follow Cao et al. (2022) to use ResNet50. The training consists of two stages. First, we utilize label data to supervise train 100 epochs and then train 200 or 90 epochs with unlabeled data. In the second stage, we adopt four views, two strong augmentation views and two weak augmentation views. And, we set the batch size to 512 for CIFAR datasets, and 256 for other datasets. The optimizer is SGD, and the learning rate first grows linearly and then cosine decays. Balance factor α, β is set to 1 in most of the experiments, and temperature τ is set to 0.1. *By default, we adopt the self-labeling loss in the ablation study.* To reduce the effect of random factors, all results are the average of three repeated experiments.

4.2 Results

Comparison with SOTA: Our novel mutual information term is general and can be applied to existing methods directly. Therefore, we select two recent methods to validate the effectiveness of our approach. ORCA (Cao et al., 2022) and UNO (Fini et al., 2021) are simple and effective methods that adopt pairwise similarity and self-labeling loss to learn from unlabeled data, respectively. For fair comparisons, we implement them with the same augmentation.

Tab.1 shows that our method achieves sizeable improvements on CIFAR100-50 and ImageNet100-50 datasets with different losses on unlabeled data. On CIFAR10, due to limited base classes, the learned maximal mutual information is nearly zero. Therefore there is no apparent improvement. On CIFAR100-50, our method improved significantly. Specially, we improve **4.4%** and **5.0%** by ORCA and UNO on the All metric, respectively. And, on unseen classes, we also have 1.9% and 5.1% improvement. In addition, we largely decrease the confusion error between seen and unseen class on the ORCA method, because the improvement on the all metric is 4.4%, which is larger than the improvement on the seen and unseen class(1.4%, 1.9%), indicating our method learn a more discriminative representation for seen and unseen classes. On the challenging ImageNet100-50 dataset, based on UNO, we achieve **20.5%** and **28.8%** improvement on All and Novel metrics. Equipped with our MI term, the results of the ORCA method can also improve a lot. The results above indicating our method is general and effective, and can improve the existing methods significantly. To compare with concurrent work (Rizve et al., 2022a), which reports results on the test set, we provide the test set results in Appendix B.

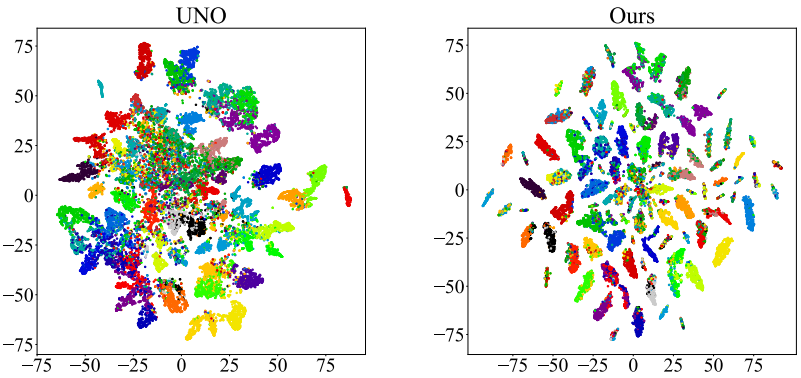

Figure 4: t-SNE visualization. Our method learns a more compact and discriminative representation.

Table 4: Results on fine-grained datasets.

Method	Oxford-IIIT Pet			FGVC-Aircraft		
	Seen	Novel	All	Seen	Novel	All
DTC	20.7	16	13.5	16.3	16.5	11.8
RankStats	12.6	12.0	11.1	13.4	13.6	11.1
ORCA	52.0	33.1	28.2	48.7	28.5	24.2
UNO	59.1	29.1	39.2	40.5	24.4	27.2
Ours	**84.2**	**62.9**	**66.7**	**72.7**	**43.0**	**47.3**

Visualization: We use t-SNE (Van der Maaten & Hinton, 2008) to visualize the representation of unseen class on CIFAR100-50 setting. The visualization of seen classes are provided in Appendix C. And we choose the self-labeling loss as the $\mathcal{L}_u$. As shown in Fig.7, the cluster of our method is more discriminative and compact than UNO. We speculate that maximizing mutual information makes the model transfer the separate property of the labeled data in the representation space to the unlabeled data, thereby reducing the entropy of the unlabeled data.

Results with 10% labeled data: To verify the effectiveness of our method under a small amount of labeled data, we split 50% seen classes and 50% unseen classes, then select 10% seen classes data as the labeled data. As Tab.2 shows, with the same SimCLR pretrain model, we improve **12.8%** and **16.5%** compared to ORCA and UNO on the All metric, indicating our method can still transfer knowledge effectively even with limited labeled data.

Results with ViT backbone: Following Vaze et al. (2022), we also evaluate our method on Imagenet100-50 setting with ViT backbone, which is more powerful. The results in Tab.3 show that, we achieve **92.8%** on novel metric and are the only one outperforming K-means on novel metric. And on the All metric, our method outperforms the pervious method by $\sim$18%, and resnet50 backbone in Tab.1 by 4.2%, illustrating our method is independent with backbone and can achieve better results with a stronger backbone.

Comparison on Fine-grained datasets: We validate our method on two fine-grained datasets, where the relationship between seen and unseen classes is more prosperous than CIAFR and Imagenet datasets. Tab.4 shows, on Oxford-IIIT Pet and FGVC-Aircraft dataset, compared to UNO on the same experiment setting, we improve more than 20% on each metric. The excellent results indicate that our novel mutual information term did utilize the relation to transfer knowledge, boosting the representation learning of seen and unseen classes. In Appendix B, we also provide the test set results to compare with concurrent works (Rizve et al., 2022a;b)

Results on imbalanced dataset: To further validate the effectiveness of our method, we perform experiments on the imbalanced dataset. We set the imbalanced factor as 50. The implementation details and more results are illustrated in Appendix B. The results in Tab.5 shows, equipped with novel mutual information term: 1) the confusion error between seen and unseen classes of ORCA is largely improved, resulting in 5.1% improvement on "All" metric. 2) the clustering accuracy on Novel class of UNO method achieves 6.4% gains. The sizeable improvement demonstrates the effectiveness of our method in the imbalanced dataset with various methods.

Table 5: Results on imbalanced dataset.

Method	CIFAR100-50		
	Seen	Novel	All
ORCA	36.8	32.5	26.1
ORCA + **MI**	39.7	30.4	31.2
UNO	49.3	28.1	33.9
UNO + **MI**	**50.1**	**34.5**	**38.6**

Table 6: The number of clusters is unknown.

Method	CIFAR100-50		
	Seen	Novel	All
RankStats	33.7	22.1	20.3
ORCA	66.3	40.0	46.4
UNO	70.0	43.4	52.0
Ours	**75.9**	**45.9**	**55.6**

Table 7: Ablation of mutual information.

Method	CIFAR100-50		
	Seen	Novel	All
Baseline	75.4	48.3	55.6
Baseline + $I_{NWJ}(z^l, z^u)$	75.9	47.5	55.1
Baseline + $I(y^l; y^u)$	72.4	47.3	53.9
Baseline + $I(\hat{y}^l; \hat{y}^u)$	**78.3**	**53.4**	**60.6**

Table 8: Results on NCD setting.

Method	CIFAR100-50
	Novel
DTC	35.9
RankStats	44.1
UNO	61.3
Ours	**64.7**

The number of clusters is unknown: In reality, the number of clusters is unknown. Therefore, we use the method proposed by (Han et al., 2021) to estimate the number of unknown clusters. On CIFAR100-50 dataset, $k = 122$. Then we retest our method and compare it with UNO and ORCA. The results are shown in the Tab.6. It shows that our method improves 5.9% and 2.5% in seen and novel classes compared with UNO, indicating the robustness of our method in the wild.

Ablation of mutual information: To validate our learning objective, we ablate the mutual information term with comparisons to other variants and verify the efficacy of our projection function. As shown in Tab.7, the first row is our baseline, which adopts self-labeling loss. $I_{NWJ}(z^u, z^l)$(Poole et al., 2019) means we estimate mutual information in the feature space(pooling-5 feature). $I(y^u, y^l)$and $I(\hat{y}_u, \hat{y}_l)$ means we estimate mutual information on the model prediction space without and with the projection function, respectively. The results show (1) Compare Row1 with Row4: our mutual information term does improve the results by a large margin ($55.6\% \rightarrow 60.6\%$ on "All" metric); (2) Compare Row1 with Row2: Applying mutual information on the feature space has no apparent improvement. We conjecture that sharing a feature extractor is already enough to transfer knowledge contained in the feature extractor, indicating the importance of transferring knowledge in the classifier. (3) Compare Row1 and Row3: Directly using mutual information without projection function often hurts performance. The reason is that mutual information without projection function enforces the model to predict seen classes as unseen classes. Our analysis shows that 19.9% of seen class data are classified as unseen classes, while the baseline is 10%. This leads to a significant drop in the Seen class metric ($75.4\% \rightarrow 72.4\%$). Ablation of β is provided in Appendix B.

Results on Novel class discovery setting: With minimal modification, our method can be applied to the NCD setting. Specifically, due to all the unlabeled data belonging to the unseen class, we project the prediction of unseen classes only to contain unseen class probability. On CIFAR100-50, as shown in Tab.8, our method achieve 3.4% improvement by UNO. It demonstrates our method is still effective in the simplified NCD setting. More results are provided in the Appendix D.

5 Conclusion

In this paper, we propose a novel and general learning framework for the task of open-world semi-supervised learning, which transfer knowledge by coupling the learning of feature extractor and classifier. Specifically, we utilize mutual information to measure the statistical dependency of seen and unseen classes in model prediction space, and maximizing mutual information enable us to transfer additional knowledge on predictive distribution. Finally, we conduct plenty of experiments in various learning setting to demonstrate our method, and we improve significantly on several benchmarks, like CIFAR100, Imagenet100 and two fine-grained datasets. We hope our work shed light on future work to better transfer knowledge between seen and unseen classes, such that the learned model can discover novel class better.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9163–9171, 2019.

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

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, 2019.

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*, 32, 2019.

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

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887, 2017.

Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Meta discovery: Learning to discover novel classes given very limited data. In *International Conference on Learning Representations*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, 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*, pp. 9284–9292, 2021.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

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*, pp. 8401–8409, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*, 2018a.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*, 2018b.

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9865–9874, 2019.

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

Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

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

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28, 2015.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

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. *arXiv preprint arXiv:2207.02261*, 2022a.

Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. *arXiv preprint arXiv:2207.02269*, 2022b.

Timo Aila Samuli Laine. Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations (ICLR)*, 30, 2017.

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*, 33:596–608, 2020.

Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pp. 9859–9869. PMLR, 2021.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pp. arXiv–1807, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. *arXiv preprint arXiv:2201.02609*, 2022.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

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*, 34:18408–18419, 2021.

Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481, 2022.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10867–10875, 2021a.

Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9462–9470, 2021b.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

APPENDIX

In the appendix, we provide the details of datasets split and more results to support the main paper. Moreover, we plot the training curve and visualize the representation of our method to understand our novel mutual information better. Finally, we validate our approach in the Novel class discovery setting thoroughly.

A DATASET SPLITS

The split details are shown in Tab. 9. We will release the code as soon as possible.

Table 9: The details of dataset split.

Dataset	Labeled		Unlabeled	
	Images	Classes	Images	Classes
CIFAR10	12.5k	5	37.5k	10
CIFAR100-50	12.5k	50	37.5k	100
ImageNet100-50	≈32k	50	≈96k	100
Oxford-IIIT Pet	≈0.94k	19	≈2.74k	37
FGVC-Aircraft	1.65k	50	≈5.0k	100

B MORE RESULTS

Comparison with concurrent works on the test set: To compare with concurrent works (Rizve et al., 2022a;b), we evaluate our method on the Imagenet100 test set. The results show we outperform the concurrent work largely in two settings. Especially, on 50% labeled setting, we outperform OpenLDN-MixMatch by 9.3% on "All" metric, and on 10% labeled setting, we outperform TRSSL by 10.7%. Moreover, our method drops a little when the number of labeled samples is decreased. Furthermore, we report the results of the OxfordIIIT-Pet test set. As shown in Tab. 11, we outperform TRSSL and OpenLDN-UDA by 3.0% and 6.5%.

Results on Imbalanced dataset: We first create the CIFAR100-LT dataset with different imbalanced factor, like 10, 50. Then, we random select 50 classes as the seen classes. The rest classes are treated as the unseen classes. Finally, we follow the evaluation protocol in the paper to evaluate the method on the balanced test set. As shown in Tab. 12, when imbalanced factor is 10, equipped with our novel mutual information regularization, our method outperform previous method by a large margin($\sim 5\%$).

Ablation of weight factor β: As shown in Tab. 13, our novel mutual information is not too sensitive to β. $\beta = 0.5$ can also achieve satisfied results. Our method degenerate to the baseline when the $\beta = 0$. Large β will fail due to the regularization term is too strong.

C MORE VISUALIZATION

Mutual information and accuracy of unlabeled data: As shown in Fig. 5, the mutual information has positive correlation with seen classes classification accuracy, unseen classes clustering accuracy. At the beginning of training, our method's unseen classes clustering accuracy surpasses the baseline quickly. And our method's seen classes accuracy is lower than baseline, indicating the mutual information term can regularize the model to learn representation helpful for unseen classes instead of biased to seen classes.

t-SNE analysis of seen classes in unlabeled data: We approximate mutual information between labeled and unlabeled data containing both seen and unseen classes. Therefore, similar to the main paper, we utilize t-SNE to visualize the representation of seen classes in unlabeled data. The results show the representation we learned is more compact. Thus, the classification results of seen classes is largely improved($75.4 \rightarrow 78.3$).

Table 10: Results on Imagenet100 testset with 50% and 10% labeled set.

Method	50% labeled			10% labeled		
	Seen	Novel	All	Seen	Novel	All
OpenLDN-MixMatch	89.6	68.6	79.1	-	-	-
TRSSL	-	-	-	82.6	67.8	75.4
Ours	**92.1**	**86.5**	**88.4**	**89.4**	**82.8**	**86.1**

Table 11: Results on OxfordIIIT Pet testset.

Method	Seen	Novel	All
UNO	49.8	22.7	34.9
OpenLDN-UDA	66.8	33.1	50.4
TRSSL	**70.9**	36.1	53.9
Ours	70.7	**49.5**	**56.9**

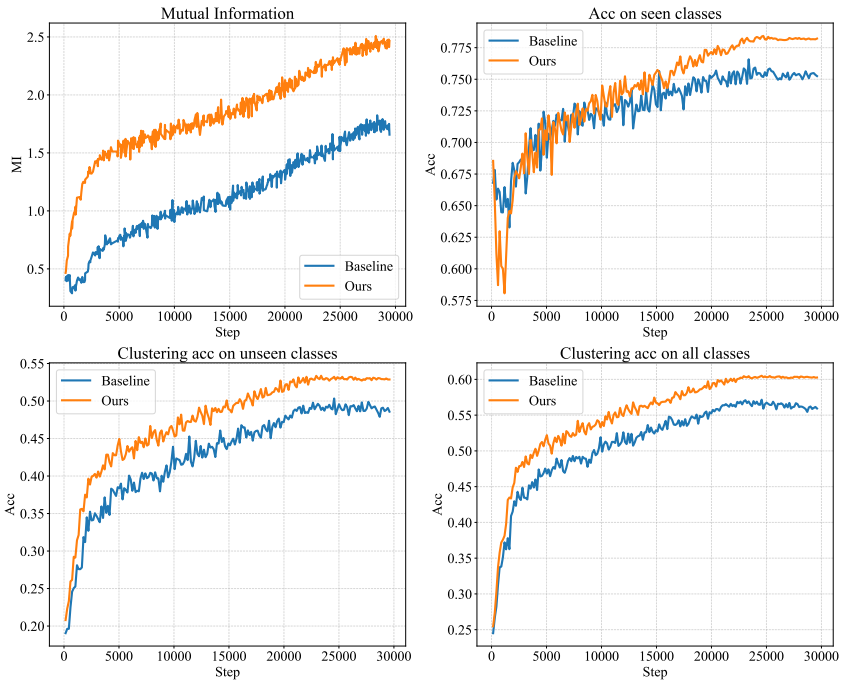

Figure 5: Mutual information and performance on unlabeled dataset.

Table 12: Results on imbalanced CIFAR100-50 dataset. Imbalanced factor is 10.

Method	Seen	Novel	All
ORCA	49.8	38.6	35.9
ORCA + **MI**	51.5	37.9	40.8
UNO	63.0	35.9	44.3
UNO + **MI**	**62.8**	**49.0**	**51.4**

Table 13: Results with different β on CIFAR100-50.

β	Seen	Novel	All
0.0	75.4	48.3	55.6
0.5	78.8	50.9	59.0
1.0	78.1	54.1	61.3
3.0	76.6	4.1	54.6
5.0	6.3	9.4	5.3

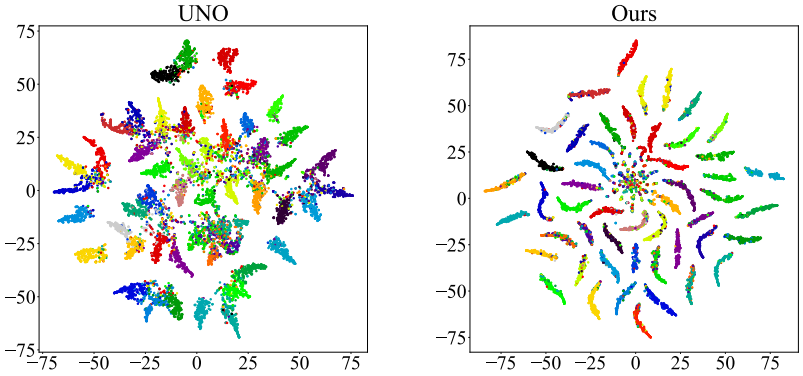

Figure 6: t-SNE visualization. Our method learns a more compact and discriminative representation.

D Novel Class Discovery

In this setting, most approaches assume the number of unseen class is given as a priori. To demonstrate efficiency of our method, we compare with previous SOTA when the number of unseen class is known and unknown respectively. Finally, we visualize representation with t-SNE(Van der Maaten & Hinton, 2008) for further analysis.

The number of clusters is known: The results are shown in Tab.14. Overall, our method surpass previous SOTA on CIFAR100-20, CIFAR100-50, and ImageNet1k. But our results is worse than UNO on CIFAR10. On CIFAR100-20 and CIFAR100-50, we have a 1.5% and 3.4% improvement over UNO, respectively. We think that when the ratio between unseen and seen classes increases, the knowledge of seen classes is relatively small, making the representation learning more difficult. Therefore, to transfer knowledge between them becomes particularly important. What's more, on the challenging ImageNet1k dataset, our method improves 1.3%, which means our method still effective on large-scale datasets.

Table 14: The number of clusters is known. All results are cluster accuracy on novel classes in the train set. * denotes the results we obtained by running their code.

Method	CIFAR10	CIFAR100-20	CIFAR100-50	ImageNet1k
K-means	72.5	56.3±1.7	28.3±0.7	71.9
KCL	72.3±0.2	42.1±1.8	-	73.8
MCL	70.9±0.1	21.5±2.3	-	74.4
DTC	88.7±0.3	67.3±1.2	35.9±1.0	78.3
RS	90.4±0.5	73.2±2.1	39.2±2.3	82.5
RS+	91.7±0.9	75.2±4.2	44.1±3.7	82.5
OpenMix	95.3	-	-	85.7
NCL	93.4±0.5	86.6±0.4	-	90.7
DRNCD	91.6±0.6	75.3±2.3	-	88.9
UNO*	93.6±0.2	90.2±0.7	61.3±0.7	91.1
Ours	93.8±0.2	**91.7±0.7**	**64.7±0.8**	**92.4**

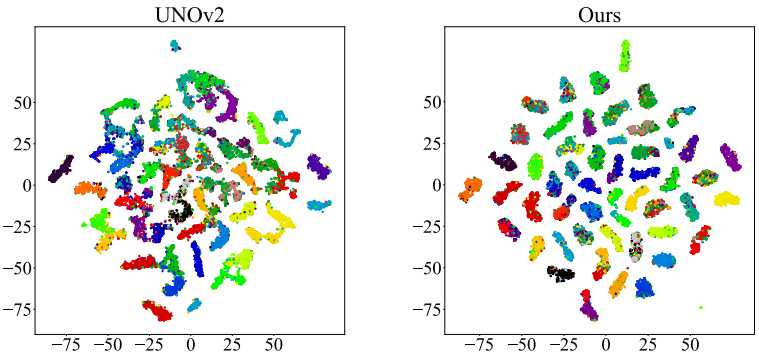

Figure 7: t-SNE visualization. Our method learns a more compact and discriminative representation.

The number of clusters is unknown: We first utilize the technique proposed by (Han et al., 2021) to estimate the number of clusters. On the CIFAR100-20 dataset, we estimate $k = 23$. Then we re-test our model. As shown in tab.15, our method still achieves the best result with a 3.6% improvement over UNOv2. This indicates that our approach can perform well even when the cluster is unknown. But this result is 11% lower than the setting in which the number of clusters is known as a priori.

Visualization Analysis: To further analysis our method, we use t-SNE to visualize the representation of unlabeled data on CIFAR100-50 setting. As shown in Fig.7, our representation is more compact and discriminative than UNO. It indicates that the representation our method learned is more helpful to cluster unlabeled data. Meanwhile, we also notice there are many error in the cluster.

Table 15: The number of clusters is unknown.

Method	CIFAR100-20
DTC	64.3
RS	70.5
RS+	71.2
UNO	77.1
Ours	**80.7**

Comparison on test dataset in NCD setting Most novel class discovery methods only test performance on the training dataset and do not compare the model's generalization performance on the test set. Therefore, we compare the model performance with previous methods on the test set, illustrating the generalization capability of our approach. Furthermore, based on Fini et al. (2021), we report results in task-agnostic evaluation metrics. The results are shown in Tab.16. Compared with the previous SOTA (UNOv2), our method is 2.9% higher in the unseen class and 1.9% lower in the seen classes. The results show that our method reduces the model's bias on the seen classes while transferring the knowledge to the unseen classes, thereby improving the performance of the unseen classes.

Table 16: The results show the performance of the test set in NCD setting. Seen and unseen represent the seen and unseen classes' performance on the test set.

Method	CIFAR100-50		
	Seen	Unseen	All
DTC	30.2	34.7	32.5
RS+	69.7	40.9	55.3
UNO	**75.4**	58.3	66.9
Ours	73.5	**61.2**	**67.4**