Self-Labeling Framework for Novel Category Discovery over Domains

Qing Yu,[1] Daiki Ikami,[1,2] Go Irie,[2] Kiyoharu Aizawa[1]

[1]The University of Tokyo, Japan
[2]NTT Corporation, Japan
[1]{yu,ikami,aizawa}@hal.t.u-tokyo.ac.jp, [2]goirie@ieee.org

Abstract

Unsupervised domain adaptation (UDA) has been highly successful in transferring knowledge acquired from a label-rich source domain to a label-scarce target domain. Open-set domain adaptation (open-set DA) and universal domain adaptation (UniDA) have been proposed as solutions to the problem concerning the presence of additional novel categories in the target domain. Existing open-set DA and UniDA approaches treat all novel categories as one unified unknown class and attempt to detect this unknown class during the training process. However, the features of the novel categories learned by these methods are not discriminative. This limits the applicability of UDA in the further classification of these novel categories into their original categories, rather than assigning them to a single unified class. In this paper, we propose a self-labeling framework to cluster all target samples, including those in the "unknown" categories. We train the network to learn the representations of target samples via self-supervised learning (SSL) and to identify the seen and unseen (novel) target-sample categories simultaneously by maximizing the mutual information between labels and input data. We evaluated our approach under different DA settings and concluded that our method generally outperformed existing ones by a wide margin.

Introduction

Deep neural networks (DNNs) have demonstrated excellent training capabilities in many large-scale annotation tasks. However, it has been observed that when the domain of the test data differs from that of the training data, the performance of the DNNs declines significantly. Unsupervised domain adaptation (UDA) has achieved reasonable performance in addressing the domain shift problem without additional annotations, by learning a recognition model in case of domain shifts between the source-domain training data and target-domain test data (Ghifary et al. 2016; Taigman, Polyak, and Wolf 2017; Tzeng et al. 2017; Saito et al. 2018a).

Although most UDA methods assume that the categories of the source domain are the same as those of the target domain, the target-domain sample classes are often unknown in real-world settings. For example, certain categories in the target domain may be absent in the source domain, i.e., open-set domain adaptation (open-set DA) (Panareda Busto and Gall 2017; Saito et al. 2018b); the source domain may contain categories that are absent in the target domain, i.e., partial domain adaptation (partial DA) (Cao et al. 2019), or a mixture of open-set DA and partial DA, i.e., universal domain adaptation (UniDA) may be observed (You et al. 2019; Saito et al. 2020). In open-set DA and UniDA, the target domain may contain samples belonging to unknown classes, i.e., classes that do not appear in the source domain.

Although certain samples in the target domain may belong to multiple novel categories, existing open-set DA and UniDA methods consider all unknown classes of the tar-

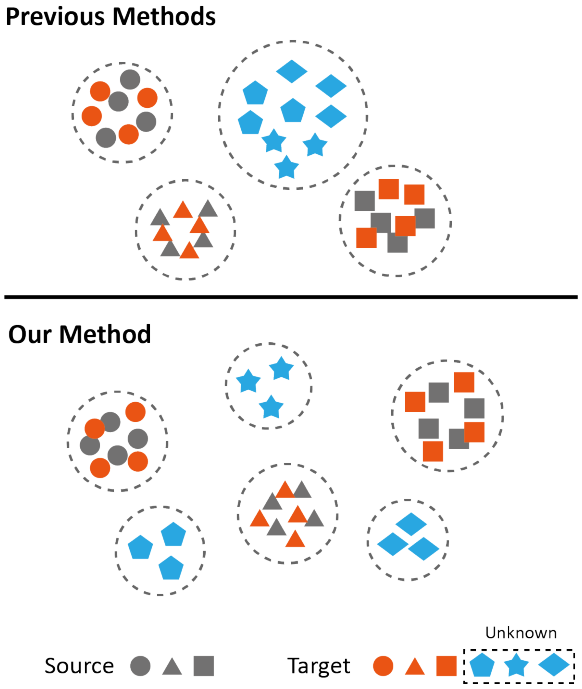


Figure 1: Novel category discovery goal in domain adaptation via the proposed method. While existing methods attempt to classify all "unknown" target samples into one category, our method aims to cluster all target samples according to their semantic categories.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

get domain as *one unified* "unknown" class and try to reject these unknown samples during the training process. Moreover, most existing methods focus on aligning the distributions of the source and target domains rather than on explicitly learning the representation of the target samples. Consequently, the features learned by these methods do not help classify "unknown" target domain samples into their original categories, which limits the applicability of UDA if the further classification of these novel categories is required (e.g., in one-shot learning).

Recent studies (Han, Vedaldi, and Zisserman 2019; Han et al. 2020) have proposed novel category discovery (NCD) methods; however, they assume that all samples in the unlabeled data belong to novel categories and there is no domain gap between the labeled and unlabeled data. Therefore, these methods cannot address NCD in the DA problem.

Hence, we propose a self-labeling framework for the simultaneous identification of seen categories (i.e., source-domain classes) and discovery of the novel categories of the target data. First, we utilize the target samples to improve the representation learning of the target domain through an existing self-supervised learning (SSL) technique, prototypical contrastive learning (PCL), which can learn the discriminative features useful for identifying novel categories.

We then assign labels to the target samples via a new self-labeling technique. An extended target-sample classifier is built, which integrates the samples belonging to seen and novel categories into a joint label distribution. We update the weights of the extended classifier for the seen categories based on prototypes from the source samples to classify the samples in the target data belonging to seen categories. To discover novel categories concurrently, we maximize the mutual information between the data indices and labels to encourage diversified classifier outputs in terms of the label distribution. Because the network is trained via SSL, samples belonging to a particular novel category tend to be classified into the same class by the extended classifier. Consequently, multiple novel categories can be discovered using the proposed method.

We evaluated the proposed method using a diverse set of DA settings and observed that our method clustered novel categories correctly and improved DA performance through SSL. Furthermore, our method outperformed existing UDA and NCD methods by a wide margin in many settings. The contributions of this study are summarized as follows.

- A new problem setting for NCD in DA is proposed.
- A novel self-labeling framework with PCL and mutual-information maximization is proposed for clustering target samples and learning discriminative features for domain-adaptation scenarios.
- The proposed method achieves high performance in several real-world domain adaptation tasks.

Related Work

Domain Adaptation

Several existing UDA approaches have demonstrated significant performance in learning a good target-domain classifier, given labeled source and unlabeled target data. Let us assume that C_s and C_t denote the label sets of the source and target domains, respectively. A closed-set DA ($C_s = C_t$) is a popular task in UDA, and distribution alignment approaches have been proposed (Ganin et al. 2016; Long et al. 2018) to solve this task. Partial DA (presence of private source classes, $C_t \subset C_s$) (Cao et al. 2018), open-set DA (presence of novel target classes, $C_s \subset C_t$) (Saito et al. 2018b), and UniDA (a mixture of open-set DA and partial DA) (You et al. 2019; Saito et al. 2020) have been proposed to handle the category-mismatch problem in the real world.

A universal adaptation network (UAN) (You et al. 2019) was proposed to manage UniDA by employing importance weighting of the source and target samples. Domain adaptive neighborhood clustering via entropy optimization (DANCE) (Saito et al. 2020) achieved high performance by applying neighborhood clustering and entropy separation to obtain weak domain alignment. Finally, the most advanced UniDA method in existence for these tasks is the one-vs-all network (OVANet) (Saito and Saenko 2021), which trained a one-vs-all classifier for each class using labeled source data and adapted the open-set classifier to the target domain by minimizing class entropy. However, most of these methods attempt to detect and group all novel target samples into *one unified* unknown class, which limits the representations that can be learned from these samples. The method proposed herein utilizes samples from the novel categories to improve feature learning in the target domain.

Self-Supervised Learning

SSL has been proposed to learn representative features for various image recognition tasks using a large-scale unlabeled dataset (Doersch, Gupta, and Efros 2015; Hjelm et al. 2019). One popular approach is to train a model to solve a pretext task, for example, to solve a jigsaw puzzle (Noroozi and Favaro 2016) or perform instance discrimination (Wu et al. 2018). Another popular approach is unsupervised clustering. Deep clustering (Caron et al. 2018) attempts to iteratively group the features using K-means clustering and uses the cluster index as the label to train the model. Online deep clustering (Zhan et al. 2020) improves upon the deep clustering method (Caron et al. 2018) by continuously updating the cluster labels at each iteration using memory. However, the most recent and advanced methods are based on Siamese networks (Bromley et al. 1994). Contrastive learning approaches, such as simple frameworks for contrastive learning of visual representations (SimCLR) (Chen et al. 2020), attempt to repel views of different images and attract the two views of the same image. Bootstrap your own latent (BYOL) (Grill et al. 2020), simple Siamese representation learning (SimSiam) (Chen and He 2020) and momentum contrast (MoCo) (He et al. 2020) predicted the output of one view from another view of the same sample using Siamese networks. Self-paced contrastive learning (Ge et al. 2020) introduced contrastive learning in domain adaptation.

Novel Category Discovery

The existing NCD problem setting aims to cluster an unlabeled dataset using the prior knowledge gained from the labeled dataset, whereby the unlabeled dataset consists of

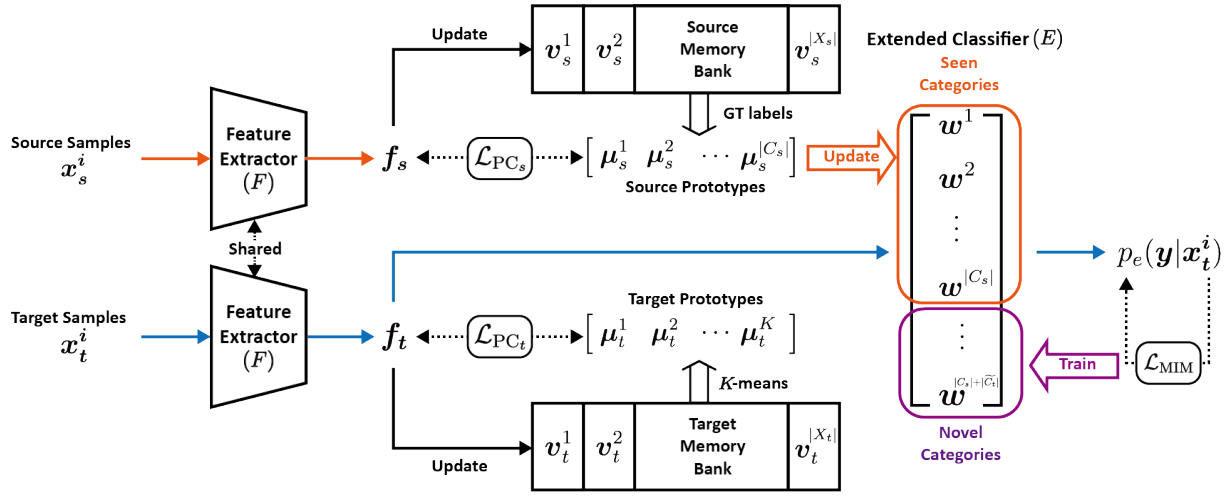


Figure 2: Overview of proposed framework. Our network has one shared feature extractor (F) and extended classifier (E). The former is trained via PCL, and the latter is trained via mutual information maximization.

classes disjoint to those present in the labeled dataset. Deep learning-based clustering methods have been proposed to solve this problem. Han et al. proposed deep transfer clustering (DTC) (Han, Vedaldi, and Zisserman 2019), which incorporated the information learned from known classes into a deep clustering framework. Later, Han et al. (Han et al. 2020) attempted to achieve unsupervised clustering by generating pairwise pseudo-labels of the unlabeled data using rank statistics. However, each method assumed that all the categories in the unlabeled dataset were novel (i.e., no seen categories were present in the unlabeled dataset), and no domain gap existed between the labeled and unlabeled datasets.

In this study, we developed an SSL framework for DA and demonstrated the better performance of our method in recognizing seen categories and discovering novel categories of target samples simultaneously.

Method

In this section, we present our problem statement and the proposed method for NCD in DA.

Problem Statement

We assume that a source image-label pair $\{\boldsymbol{x_s}, \boldsymbol{y_s}\}$ is drawn from a set of labeled source images, $\{X_s, Y_s\}$, while an unlabeled target image $\boldsymbol{x_t}$ is drawn from a set of unlabeled images X_t. The one-hot vector of the class label y_s is represented by $\boldsymbol{y_s}$. We use C_s and C_t to denote the label sets of the source and target domains, respectively, and $C_{com} = C_s \cap C_t$ to represent the set of labels common to, i.e., shared by both domains. The true labels for the source and target images are denoted by Y_s^{GT} and Y_t^{GT}, respectively (y_s^{GT} and y_t^{GT} represent the source and target image samples, respectively), which implies that $y_s^{GT} \in C_s$ and $y_t^{GT} \in C_t$. In the open-set DA and UniDA scenarios, certain novel classes exist in the target domain, i.e., $C_{com} \subset C_t$. These unknown target classes are denoted by $\widetilde{C_t} = C_t \setminus C_{com}$. Moreover, in the UniDA context, some classes of the source domain do not appear in the target domain, that is, $C_{com} \subset C_s$. These private source classes are denoted by $\widetilde{C_s} = C_s \setminus C_{com}$. Given a target sample, the goal of NCD in DA is to either classify it as one of the common seen classes C_{com} correctly or group it with similar unlabeled target samples to form one of the novel target classes amongst $\widetilde{C_t}$.

Approach Overview

Our model has four components: (1) a shared feature extractor F, which outputs an ℓ_2 normalized feature vector $\boldsymbol{f} \in \mathbb{R}^d$), and (2) an extended classifier E for target samples, which outputs a probability vector $\in \mathbb{R}^{|C_s|+|\widetilde{C_t}|}$ (note that the number of outputs is set to the sum of the number of seen classes and expected number of novel classes).

Prototypical Contrastive Learning

To learn the semantic structure of the unlabeled target samples, PCL (Li et al. 2020; Yue et al. 2021) was used to perform iterative clustering and representation learning. To discover novel classes, the features within the same cluster need to be in proximity, and the features within different clusters need to be separated further. We prepared a sample memory for storing features, expressed as:

$$\mathbf{V_t} = [\boldsymbol{v}_t^1, \cdots, \boldsymbol{v}_t^{|X_t|}], \tag{1}$$

where $\boldsymbol{v}_i$ denotes the feature vector of $\boldsymbol{x_t^i}$ stored in the memory. The memory is initialized via $F(\boldsymbol{x_t^i})$ and updated with momentum m after each batch, as follows.

$$\boldsymbol{v}_t^i \leftarrow m\boldsymbol{v}_t^i + (1-m)F(\boldsymbol{x_t^i}). \tag{2}$$

To perform PCL, we apply K-means clustering on $\mathbf{V_t}$ to obtain target clusters $\mathbf{C}_t = \{C_t^1, \cdots, C_t^K\}$, and the normalized target prototypes $\{\boldsymbol{\mu}_t^j\}_{j=1}^K$ can be derived through the

following equations:

$$\boldsymbol{u}_t^j = \frac{1}{|C_t^j|} \sum_{\boldsymbol{v}_t^i \in C_t^j} \boldsymbol{v}_t^i, \tag{3}$$

$$\boldsymbol{\mu}_t^j = \frac{\boldsymbol{u}_t^j}{\|\boldsymbol{u}_t^j\|}. \tag{4}$$

During training, we first calculate the feature $\boldsymbol{f}_t^i = F(\boldsymbol{x_t^i})$, and subsequently compute the similarity distribution vector between $\boldsymbol{f}_t^i$ and $\{\boldsymbol{\mu}_t^j\}_{j=1}^K$, expressed as $P_t^i = [P_s^{i,1}, \cdots, P_s^{i,K}]$, through the following equation:

$$P_t^{i,j} = \frac{\exp(\boldsymbol{\mu}_t^j \cdot \boldsymbol{f}_t^i / \tau)}{\sum_{k=1}^K \exp(\boldsymbol{\mu}_t^k \cdot \boldsymbol{f}_t^i / \tau)}, \tag{5}$$

where the temperature parameter τ controls the distribution concentration degree. Let $D_s = \{(\boldsymbol{x_s^i}, \boldsymbol{y_s^i})\}_{i=1}^N$ and $D_t = \{(\boldsymbol{x_t^i})\}_{i=1}^N$ be mini-batches of size N, sampled from the source and target samples, respectively. The prototypical contrastive loss can be obtained using:

$$\mathcal{L}_{\mathrm{PC}_t}(D_t) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K c_t^{ik} \log P_t^{i,k}, \tag{6}$$

where c_t^{ik} denotes the cluster label of the samples.

To learn the representation of labeled source samples, the same operations are performed on the source samples. However, the ground-truth labels of the source samples are used instead of clustering indices, and $\mathbf{V_s}$ and $\{\boldsymbol{\mu}_s^j\}_{j=1}^{|C_s|}$ (i.e., the centroid of each source class) are obtained. Consequently, the overall loss for self-supervised clustering is denoted by:

$$\mathcal{L}_{\mathrm{CLU}}(D_s, D_t) = \mathcal{L}_{\mathrm{PC}_s}(D_s) + \mathcal{L}_{\mathrm{PC}_t}(D_t). \tag{7}$$

The loss is calculated at each iteration, and the memory, as well as class prototypes, are updated in each epoch.

Self-labeling via Mutual Information Maximization

Through the training of PCL, the network can learn the features for classifying seen categories and discover novel categories in the target data. However, the labels of the target samples cannot be obtained directly using these methods. To label the target samples into their original classes, we input the target samples into the extended classifier E and the output probability denoted by $p_e(\boldsymbol{y}|\boldsymbol{x_t^i}) = E(F(\boldsymbol{x_t^i})) \in \mathbb{R}^{|C_s|+|\overline{C_t}|}$ is obtained as a soft pseudo-label. Regarding $p_e(\boldsymbol{y}|\boldsymbol{x_t^i})$, the first $|C_s|$ dimensions denote the seen categories, and the remaining $|\widetilde{C_t}|$ dimensions denote the novel categories. The proposed self-labeling framework aims to optimize the pseudo-labels of the target samples via mutual information maximization.

To classify the seen categories correctly, the simplest method is to train the classifier using standard cross-entropy loss on labeled source data, expressed as:

$$\mathcal{L}_s(D_s) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{|C_s|} y_s^{ik} \log p_e(k|\boldsymbol{x_s^i}), \tag{8}$$

where $p_e(k|\boldsymbol{x^i})$ represents the probability that sample $\boldsymbol{x^i}$ belongs to class k predicted by the extended classifier.

However, this training method leads to an imbalance problem, wherein the gradient is updated for the seen source classes C_s, but not for the novel target classes $\widetilde{C_t}$. Consequently, the classifier will be biased toward the seen classes (Kang et al. 2019), which decreases the NCD performance.

To solve this problem, a cosine classifier consisting of weight vectors $\mathbf{W} = [\boldsymbol{w}^1, \cdots, \boldsymbol{w}^{|C_s|+|\widetilde{C_t}|}]$ is used as the extended classifier F. The probabilistic output can be obtained through the following:

$$p_e(\boldsymbol{y}|\boldsymbol{x_t^i}) = \mathrm{SoftMax}(\frac{1}{\tau}\mathbf{W}^{\mathrm{T}}\boldsymbol{f}), \tag{9}$$

where τ represents the temperature parameter. Instead of updating $\mathbf{W}$ via cross-entropy loss, we update it using the source sample prototypes $\{\boldsymbol{\mu}_s^j\}_{j=1}^{|C_s|}$ as follows.

$$\boldsymbol{w}^j = \boldsymbol{\mu}_s^j \ (1 \le j \le |C_s|). \tag{10}$$

While $\mathbf{W}$ is updated in each epoch based on the calculation of the prototypes, it continues to be updated based on other losses in each iteration.

At the same time, we attempt to discover novel classes by training the network to output diversified classes over the dataset. Because the dimension of the extended classifier is set to $|C_s| + |\widetilde{C_t}|$ and the seen classes are updated through Eq. (10), novel classes are automatically clustered in the $|\widetilde{C_t}|$ part. Consequently, the extended classifier tends to classify samples of the same novel category into the same class. To achieve this, we maximize the entropy of the expected network prediction $\mathcal{H}(\mathbb{E}_{\boldsymbol{x_t} \in X_t}[p_e(\boldsymbol{y}|\boldsymbol{x_t})])$.

Furthermore, to obtain a high-confidence prediction for each sample, we apply entropy minimization to the network output, which is effective in semi-supervised learning (Grandvalet and Bengio 2005) and helps achieve DA in UDA tasks (Carlucci et al. 2017; Saito et al. 2019).

Maximizing $\mathcal{H}(\mathbb{E}_{\boldsymbol{x_t} \in X_t}[p_e(\boldsymbol{y}|\boldsymbol{x_t})])$ and minimizing the entropy of the network output are equivalent to maximizing the mutual information between the input and output (Asano, Rupprecht, and Vedaldi 2019; Cui et al. 2020; Yue et al. 2021), as follows.

$$\mathcal{I}(Y; X_t) = \mathcal{H}(\mathbb{E}_{\boldsymbol{x_t}}[p_e(\boldsymbol{y}|\boldsymbol{x_t})]) - \mathbb{E}_{\boldsymbol{x_t}}[\mathcal{H}(p_e(\boldsymbol{y}|\boldsymbol{x_t}))]. \tag{11}$$

The minimization of $\mathbb{E}_{\boldsymbol{x_t}}[\mathcal{H}(p_e(\boldsymbol{y}|\boldsymbol{x_t}))]$ can be achieved easily by reducing the output entropy of each sample in the mini-batch, as follows.

$$\mathcal{L}_e(D_t) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{|C_s|+|\widetilde{C_t}|} p_e(k|\boldsymbol{x_t^i}) \log p_e(k|\boldsymbol{x_t^i}). \tag{12}$$

To maximize $\mathcal{H}(\mathbb{E}_{\boldsymbol{x_t}}[p_e(\boldsymbol{y}|\boldsymbol{x_t})])$, it is estimated based on the following loss calculation:

$$\mathcal{L}_p(D_t) = -\sum_{k=1}^{|C_s|+|\widetilde{C_t}|} \bar{p}_e(k|D_t) \log \hat{p}_e(k|D_t), \tag{13}$$

where $\bar{p}_e(\boldsymbol{y}|D_t)$ represents the mean probability for each mini-batch D_t, expressed as:

$$\bar{p}_e(\boldsymbol{y}|D_t) = \frac{1}{N}\sum_{i=1}^{N} p_e(\boldsymbol{y}|\boldsymbol{x_t^i}), \tag{14}$$

and $\hat{p}_e(\boldsymbol{y}|D_t)$ denotes the moving average of $\bar{p}_e(\boldsymbol{y}|D_t)$, which is calculated in each iteration.

Finally, the objective of maximizing the mutual information is achieved by minimizing the following loss:

$$\mathcal{L}_{\mathrm{MIM}}(D_t) = -\mathcal{L}_p(D_t) + \mathcal{L}_e(D_t). \tag{15}$$

Distribution Weighting

In the case of UniDA, because some classes in the source domain do not appear in the target domain ($C_{com} \subset C_s$), if the mutual information is maximized over the extended classifier ($|C_s| + |\widetilde{C_t}|$ classes), some samples belonging to the common or the novel classes will be misclassified into the private source classes $\widetilde{C_s}$.

To solve this problem, we calculate the label distribution of the seen classes by combining the proposed method with the existing UniDA method, OVANet (Saito and Saenko 2021). We selected the target samples predicted to belong to seen classes and calculated the distribution of the target samples belonging to each seen class by averaging the label predictions based on the following:

$$\boldsymbol{\gamma} = \frac{1}{N_t}\sum_{i=1}^{N_t} \boldsymbol{y}^i_{\mathrm{ova}}, \tag{16}$$

where N_t is the number of target samples selected, and $\boldsymbol{y}^i_{\mathrm{ova}}$ denotes the predicted probabilities via OVANet for sample i. Because $\sum_{k=1}^{|C_s|}\boldsymbol{\gamma}_k = 1$, we consider the classes $\boldsymbol{\gamma}_k > \frac{1}{|C_s|}$ as common classes, and other classes as private source classes. The weight of each seen class is defined by:

$$\eta_k = \begin{cases} 1 & \text{if } \boldsymbol{\gamma}_k > \frac{1}{|C_s|} \\ \delta & \text{otherwise} \end{cases}, \tag{17}$$

where $\delta < 1$, and Eq. (13) can be rewritten as follows.

$$\begin{aligned}\mathcal{L}'_p(D_t) = &-\sum_{k=1}^{|C_s|}\eta_k \bar{p}_e(k|D_t)\log\hat{p}_e(k|D_t) \\ &-\sum_{k=|C_s|+1}^{|C_s|+|\widetilde{C_t}|}\bar{p}_e(k|D_t)\log\hat{p}_e(k|D_t).\end{aligned} \tag{18}$$

Because of the weights assigned to the seen classes, the classes with few samples will be considered as private source classes and no further samples will be classified into these classes via mutual information maximization. Consequently, the objective can be rewritten as

$$\mathcal{L}'_{\mathrm{MIM}}(D_t) = -\mathcal{L}'_p(D_t) + \mathcal{L}_e(D_t). \tag{19}$$

Overall Objective Function

In summary, our self-labeling framework performs the contrastive learning of prototypes and the learning of an extended classifier. The overall learning objective is:

$$\mathcal{L}_{total} = \lambda_{\mathrm{CLU}}\mathcal{L}_{\mathrm{CLU}}(D_s, D_t) + \lambda_{\mathrm{MIM}}\mathcal{L}'_{\mathrm{MIM}}. \tag{20}$$

Open-set DA			
Method	Office	OfficeHome	VisDA
UAN	60.29	32.01	42.23
DANCE	73.48	47.00	42.38
OVANet	90.10	69.04	60.05
Ours	**91.84**	**78.02**	72.35
Ours (OC)	91.70	77.93	**73.53**
UniDA			
Method	Office	OfficeHome	VisDA
UAN	64.69	56.24	36.82
DANCE	81.18	49.25	17.26
OVANet	84.78	70.65	50.10
Ours	**89.85**	78.25	64.45
Ours (OC)	88.99	**78.50**	**65.86**

Table 1: H-scores (%) for known classes and one unified unknown class under different settings. The average scores of all tasks for each dataset are reported. The bold values represent the highest scores for each row. (OC stands for over-clustering.)

Experiment

Experimental Setup

Datasets. Based on existing studies (You et al. 2019), we used three datasets to validate our approach. Office (Saenko et al. 2010), which consisted of three domains (Amazon, DSLR, Webcam), and 31 classes, was used as the first dataset. The second dataset was OfficeHome (Venkateswara et al. 2017), which contained four domains (art, clipart, product, and real) and 65 classes. The final dataset was VisDA (Peng et al. 2017), which contained two domains (synthetic and real) and 12 classes. To create the open-set DA conditions, we split the classes of each dataset according to (Saito et al. 2020), as $|C_{com}|/|\widetilde{C_s}|/|\widetilde{C_t}| = 10/0/11$ for Office, $15/0/50$ for OfficeHome, and $6/0/6$ for VisDA. To construct the UniDA setting, we split the classes of each dataset based on $|C_{com}|/|\widetilde{C_s}|/|\widetilde{C_t}| = 10/10/11$ for Office, $10/5/50$ for OfficeHome, and $6/3/3$ for VisDA.

Comparison of Methods. We compared the proposed method with three UniDA methods: (1) UAN (You et al. 2019), (2) DANCE (Saito et al. 2020), and (3) OVANet (Saito and Saenko 2021), and two NCD methods: (1) deep transfer clustering (DTC) (Han, Vedaldi, and Zisserman 2019), and (2) ranking statistics (RS) (Han et al. 2020). Because these methods achieved state-of-the-art performance in their respective settings, it would be valuable to analyze their performance in the NCD setting and compare the performance of the proposed method against those of the other methods.

Evaluation Protocols. To evaluate the performance of UniDA, we used the H-score metric (Saito and Saenko 2021; Bucci, Loghmani, and Tommasi 2020). When the unknown target classes are regarded as a unified unknown class, the H-score is the harmonic mean of the accuracy of common classes (acc_c) and that of the unified unknown class (acc_t), expressed as follows.

Open-set DA

Method	Office			OfficeHome			VisDA		
	Seen	Novel	H-score	Seen	Novel	H-score	Seen	Novel	H-score
UAN	95.24	30.21	45.85	81.98	11.21	19.63	78.81	26.84	40.06
DANCE	**97.67**	43.36	60.04	**83.41**	21.91	34.61	**82.71**	29.78	43.58
OVANet	89.12	54.82	67.98	66.67	38.24	48.69	56.02	38.10	45.30
DTC	83.11	53.81	64.70	47.85	20.58	27.83	23.91	20.37	22.00
RS	54.88	38.45	41.85	23.75	18.16	19.58	42.54	**47.44**	44.85
Ours	89.84	66.79	76.37	69.55	**50.54**	**58.34**	62.62	39.78	48.65
Ours (OC)	89.43	**67.39**	**76.56**	68.64	49.53	57.44	62.37	42.28	**50.40**

UniDA

Method	Office			OfficeHome			VisDA		
	Seen	Novel	H-score	Seen	Novel	H-score	Seen	Novel	H-score
UAN	68.32	51.89	58.87	82.85	13.58	23.12	58.00	28.85	38.36
DANCE	**95.23**	55.62	69.72	**86.26**	25.61	39.26	**83.79**	9.79	17.44
OVANet	81.31	60.51	69.29	69.13	40.51	50.92	35.56	33.73	34.59
DTC	85.96	50.15	63.23	50.61	24.10	31.95	31.11	26.52	28.63
RS	56.19	40.69	44.31	52.79	37.62	43.69	35.89	0	0
Ours	89.51	**69.59**	**77.75**	73.80	**51.77**	**60.72**	63.06	40.87	49.60
Ours (OC)	89.85	66.62	76.24	71.47	51.55	59.84	59.74	**49.97**	**54.42**

Table 2: Clustering accuracy for seen and novel classes and H-scores of the two previous accuracies under different settings.

Method	Open-set DA			UniDA		
	Office	OfficeHome	VisDA	Office	OfficeHome	VisDA
UAN	62.81	38.54	48.64	58.60	42.13	49.44
DANCE	67.65	44.50	70.48	63.35	48.90	56.53
OVANet	63.81	38.02	50.50	58.50	42.11	47.81
DTC	51.55	17.32	26.22	51.51	20.80	32.53
RS	48.09	21.15	33.68	43.28	24.57	45.07
Ours	**75.42**	51.09	**72.69**	**72.86**	54.18	**72.33**
Ours (OC)	74.12	**51.15**	71.75	71.87	**54.44**	72.15

Table 3: Average accuracy (%) of linear classification given one labeled target sample per novel class.

$$H_{score} = \frac{2\mathrm{acc}_c \cdot \mathrm{acc}_t}{\mathrm{acc}_c + \mathrm{acc}_t}. \tag{21}$$

A high H-score is obtained only when both, acc_c and acc_t are high, indicating that this metric accurately measures both accuracies.

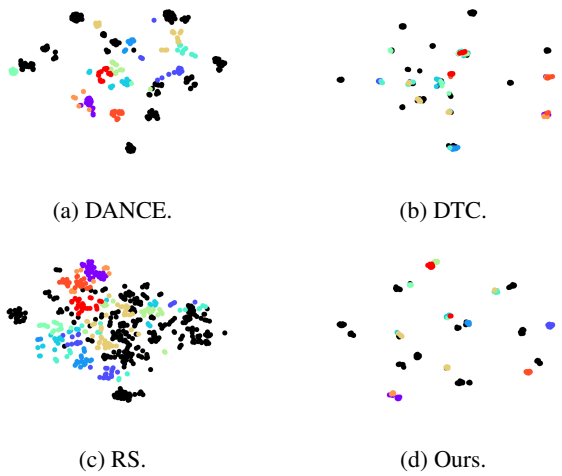
To evaluate the performance of NCD, we calculated the clustering accuracy between the labels obtained through each method and the ground-truth labels for all target samples belonging to the seen and novel classes, respectively. We also reported the H-scores of the common-class and novel-class accuracies. The clustering accuracy of UniDA methods are evaluated by further clustering the detected novel samples into clusters by K-means.

Implementation Details. In this experiment, we used the same network architecture and hyperparameters as in Saito et al. (2020). We implemented our network based on ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009). We used the modules of ResNet until the *global-average-pooling* layer as the feature extractor F, and one *fully connected* layer as the extended classifier E. We set the momentum m to 0.5, temperature τ to 0.1, class weight δ to 0.5, and loss weights λ_{CLU} and λ_{MIM} to 1 and 0.5, respectively, in all the experiments.

Note that the exact number of classes in the target domain $|C_t|$ is generally unknown. However, as in previous studies (Xie, Girshick, and Farhadi 2016; Van Gansbeke et al. 2020; Han et al. 2020), we used the cluster number K, equalling the number of ground-truth clusters $|C_t|$, for evaluation. In real-world applications, a rough estimation of the number of clusters is generally obtainable (e.g., the number of animal species observed in a national park). Based on the estimations obtained via prior domain knowledge, we can train our method with a larger number of clusters without compromising on performance. This is detailed in the following section.

Experimental Results

Table 1 compares the classification results of UniDA obtained via the proposed method with existing state-of-the-art UniDA methods. While OVANet outperformed other existing methods, the proposed method outperformed OVANet for all datasets. Furthermore, the proposed approach outper-



(a) DANCE. (b) DTC.

(c) RS. (d) Ours.

Figure 3: t-SNE plots of target samples (best viewed in color). Each color indicates a different class. Black dots represent "known" target samples while those of other colors represent "unknown" target samples.

formed existing methods by a considerable margin in the case of the complex VisDA dataset, wherein synthetic data comprises the source domain and real data comprises the target domain. However, the results in Table 1 do not capture the NCD performance.

The target-sample classification performances inclusive of novel categories, listed in Table 2, are of greater importance than the previous results. The accuracies for novel classes and corresponding H-scores demonstrate that NCD via the proposed approach is better than that via other methods in most settings. The accuracy for the seen classes observed via our method is also competitive with that of the OVANet method. Note that existing NCD methods perform poorly compared to the proposed method because they do not consider that the seen classes are present in the unlabeled data and cannot manage the domain shift problem.

Over-clustering. Thus far, our approach assumed that the number of ground-truth classes in the target domain when K-means is applied was known. To investigate the sensitivity to the number of clusters K, we overestimated the number of ground-truth classes used in the training process by a factor of two (e.g., we clustered the Office into 42 rather than 21 classes). The results obtained for each dataset are listed in Table 1 and Table 2 as "Ours (OC)." Although the accuracy was marginally lower than that of the non-OC method in some settings, it outperformed the other methods in most tasks.

Feature Visualization. The target features using t-SNE (Van der Maaten and Hinton 2008) are visualized in Fig. 3 for A→W task in the open-set DA setting. The "known" target-sample features are represented using black dots, and are tightly clustered via our method. Additionally, most "unknown" features, represented using dots of other colors, are tightly clustered and adequately distanced from the "known" features.

Method	Seen	Novel	H-score
Ours	89.51	69.59	77.75
w/o $\mathcal{L}_{\mathrm{PC}_s}$	82.23	62.96	70.49
w/o $\mathcal{L}_{\mathrm{PC}_t}$	90.12	55.40	68.18
w/o $\mathcal{L}'_p$	93.20	50.79	64.81
w/o $\mathcal{L}_e$	90.59	64.54	74.53
w/ $\mathcal{L}_{\mathrm{MIM}}$	86.31	55.89	67.40
$\lambda_{\mathrm{MIM}} = 0.1$	93.75	61.66	73.64
$\lambda_{\mathrm{MIM}} = 0.2$	93.29	63.68	74.26
$\lambda_{\mathrm{MIM}} = 1$	87.40	65.71	74.87

Table 4: Clustering accuracies and H-scores of ablation study tasks on the Office dataset under the UniDA setting.

One-shot linear classification. We evaluate how well the learned features can contribute to sample clustering by training a new linear classifier based on the previously learned feature extractor using one labeled sample per novel category. For example, for the OfficeHome task under the open-set DA setting, we trained a classifier with 50 "unknown" classes using one labeled sample per category as training data and the remaining samples as test data. Table 3 summarizes the average linear classification accuracies obtained for each dataset. These results reveal that our method outperforms other methods by a wide margin.

Ablation Study. The performances of variants of the proposed method were evaluated using the Office dataset under the UniDA setting, for further exploration of the efficacy of the proposed method. The following variants were studied. (1) "Ours w/o $\mathcal{L}_{\mathrm{PC}_s}$" is a variant that does not use PCL on source samples in Eq. (7). (2) "Ours w/o $\mathcal{L}_{\mathrm{PC}_t}$" is a variant that does not use PCL on target samples in Eq. (7). (3) "Ours w/o $\mathcal{L}'_p$" is a variant that does not maximize the entropy of the expected network predicted using Eq. (19). (4) "Ours w/o $\mathcal{L}_e$" is a variant that does not minimize the entropy of the network output obtained via Eq. (19). (5) "Ours w/ $\mathcal{L}_{\mathrm{MIM}}$" is the variant that uses the original mutual information loss in Eq. (15) instead of the weighted loss in Eq. (19). Table 4 reveals that the version of our approach that utilizes all the losses, outperforms other variants in all settings. Specifically, $\mathcal{L}'_p$ is the most important component for our method, and $\mathcal{L}_e$ and $\mathcal{L}_{\mathrm{PC}_t}$ are also necessary to achieve a more complete and accurate clustering. The results of "Ours w/ $\mathcal{L}_{\mathrm{MIM}}$" also demonstrate the effectiveness of the distribution weighting. In Table 4, we also demonstrate the sensitivity of the loss in Eq. (20) to different weights.

Conclusion

In this paper, we proposed a self-labeling framework for NCD in DA. Our framework uses clustering to learn the semantic structure of target samples, including "unknown" categories, and labels them into their original classes via mutual information maximization. We evaluated the performance of the proposed method in a diverse set of tasks across various source- and target-domain pairs, and observed that it outperformed existing state-of-the-art DA methods by a considerable margin.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR19F4, JSPS KAKENHI Grant Number 20J22372, and NTT Communication Science Laboratories.

References

Asano, Y.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. In *ICLR*.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1994. Signature verification using a" siamese" time delay neural network. In *NeurIPS*.

Bucci, S.; Loghmani, M. R.; and Tommasi, T. 2020. On the effectiveness of image rotation for open set domain adaptation. In *ECCV*.

Cao, Z.; Ma, L.; Long, M.; and Wang, J. 2018. Partial adversarial domain adaptation. In *ECCV*.

Cao, Z.; You, K.; Long, M.; Wang, J.; and Yang, Q. 2019. Learning to transfer examples for partial domain adaptation. In *CVPR*.

Carlucci, F. M.; Porzi, L.; Caputo, B.; Ricci, E.; and Bulo, S. R. 2017. Autodial: Automatic domain alignment layers. In *ICCV*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Chen, X.; and He, K. 2020. Exploring Simple Siamese Representation Learning. *arXiv preprint arXiv:2011.10566*.

Cui, S.; Wang, S.; Zhuo, J.; Li, L.; Huang, Q.; and Tian, Q. 2020. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; and Li, H. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; Balduzzi, D.; and Li, W. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*.

Grandvalet, Y.; and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *NeurIPS*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*.

Han, K.; Rebuffi, S.-A.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2020. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. In *ICLR*.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *NeurIPS*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*.

Panareda Busto, P.; and Gall, J. 2017. Open set domain adaptation. In *ICCV*.

Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019. Semi-supervised domain adaptation via minimax entropy. In *ICCV*.

Saito, K.; Kim, D.; Sclaroff, S.; and Saenko, K. 2020. Universal Domain Adaptation through Self-Supervision. In *NeurIPS*.

Saito, K.; and Saenko, K. 2021. OVANet: One-vs-All Network for Universal Domain Adaptation. In *ICCV*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018a. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*.

Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018b. Open set domain adaptation by backpropagation. In *ECCV*.

Taigman, Y.; Polyak, A.; and Wolf, L. 2017. Unsupervised cross-domain image generation. In *ICLR*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *ECCV*.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *CVPR*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*.

You, K.; Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2019. Universal domain adaptation. In *CVPR*.

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021. Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation. In *CVPR*.

Zhan, X.; Xie, J.; Liu, Z.; Ong, Y.-S.; and Loy, C. C. 2020. Online deep clustering for unsupervised representation learning. In *CVPR*.