CDS: Cross-Domain Self-supervised Pre-training

Donghyun Kim[1], Kuniaki Saito[1], Tae-Hyun Oh[2], Bryan A. Plummer[1], Stan Sclaroff[1], Kate Saenko[1,3]
[1]Boston University, [2]POSTECH, [3]MIT-IBM Watson AI Lab
{donhk, keisaito, bplum, sclaroff, saenko}@bu.edu, taehyun@postech.ac.kr

Abstract

*We present a two-stage pre-training approach that improves the generalization ability of standard single-domain pre-training. While standard pre-training on a single large dataset (such as ImageNet) can provide a good initial representation for transfer learning tasks, this approach may result in biased representations that impact the success of learning with new multi-domain data (*e.g.*, different artistic styles) via methods like domain adaptation. We propose a novel pre-training approach called Cross-Domain Self-supervision (CDS), which directly employs unlabeled multi-domain data for downstream domain transfer tasks. Our approach uses self-supervision not only within a single domain but also across domains. In-domain instance discrimination is used to learn discriminative features on new data in a domain-adaptive manner, while cross-domain matching is used to learn domain-invariant features. We apply our method as a second pre-training step (after ImageNet pre-training), resulting in a significant target accuracy boost to diverse domain transfer tasks compared to standard one-stage pre-training.*

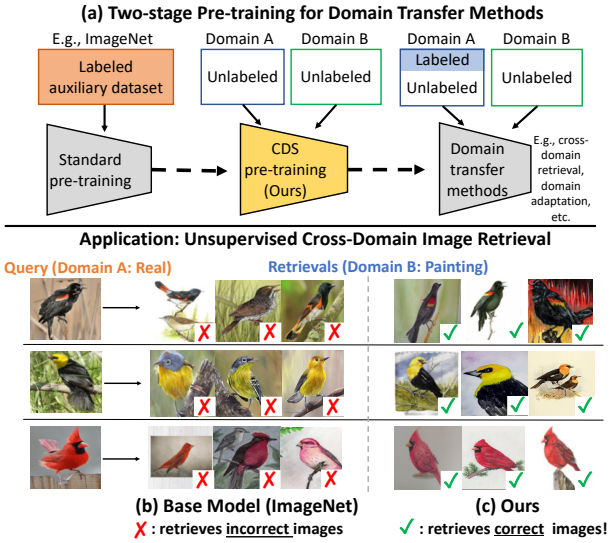


Figure 1: **Top:** Two-stage pre-training for domain transfer methods. To learn discriminative and domain-invariant features on downstream domains, we propose Cross-Domain Self-supervised pre-training (CDS) by leveraging unlabeled data from multiple domains. **Bottom:** An application of CDS to unsupervised cross-domain image retrieval. CDS learns a better semantic relationship across domains compared to ImageNet pre-training.

1. Introduction

Real-world image data can come from many sources: different weather, viewpoints, lighting, artistic styles, etc. Therefore, many tasks require visual representations that generalize across multiple domains. For example, domain adaptation aims to transfer knowledge from a labeled source domain to an unlabeled target domain [32, 12]. Cross-domain image retrieval aims to match semantically related images regardless of domain shift (*e.g.*, see Fig. 1-(b, c)).

Pre-training has been very effective for deep neural networks across many visual tasks, providing strong initial representations [21, 7]. Typically, prior work pre-trains a model on a large-scale supervised auxiliary domain (mostly on ImageNet [31]) and assumes the learned features are a good starting point for downstream tasks. However, ImageNet pre-training learns biased representations [13] and suffers from domain shift caused by changes in background, rotation, and viewpoints [1]. This suggests that pre-training on a single domain does not encourage domain-invariant features and may not be a good match for downstream tasks encountering new domains (*e.g.*, Fig. 1-(b), domain adaptation). In this paper, we address the problem of pre-training representations that are robust to domain shift and useful for downstream methods that operate on multiple domains.

We propose a two-stage pre-training approach to improve standard ImageNet pre-training with respect to generalization to new domains for downstream tasks. After the standard pre-training on a generic supervised dataset (*e.g.*, ImageNet), we add a second self-supervised pre-training stage that uses unlabeled downstream data from multiple domains as illustrated in Fig. 1-(a). Our second pre-training stage ensures that the representation gains discriminative power on the new domains and invariance to domain shift.

After this two-stage pre-training, the representation can be used directly for tasks like cross-domain image retrieval or used to initialize a model for existing transfer methods (*e.g.*, train with both labeled source and unlabeled target data). We compare our two-stage pre-training with standard one-stage pre-training and show significant gains across multiple tasks and methods. For example, Fig. 1-(c) shows that our method is better at learning class-semantic similarity across new domains compared to ImageNet pre-training (Fig. 1-(b)) and improves cross-domain image retrieval.

Self-supervised learning (SSL) has been shown to be very effective for pre-training on unlabeled data. SSL solves pre-text tasks such as predicting rotation [14] or instance discrimination [42, 7]. However, state-of-the-art SSL (*e.g.*, [42, 7, 16]) focuses on learning from a single domain. Naively adapting SSL to multiple domains cannot learn domain-invariant representations as the images of the same class across domains can have different visual characteristics, as we will show in our experiments.

To address the issue, we propose a new pre-training method called Cross-Domain Self-supervision (CDS) that overcomes the limitations of the prior single-domain SSL methods. CDS effectively learns the relationship between domains using unlabeled data (*i.e.*, unsupervised). Specifically, we devise two types of self-supervision to extract discriminative[1] and domain-invariant features across domains. First, we propose *in-domain instance discrimination*. This is motivated by recent SSL [42, 7], but we apply it in a domain adaptive manner to learn discriminative features in each domain. Second, we propose *cross-domain matching*. This objective matches each sample to a neighbor in the other domain while forcing it to be far from unmatched samples. While *in-domain instance discrimination* encourages a model to learn discriminative features by separating every instance within a domain, the *cross-domain matching* enables better knowledge transfer across domains by performing domain alignment. We hypothesize that such pre-training optimized for downstream multi-domain data can gain domain-invariance and discriminability to new domains.

CDS is applicable to a variety of domain transfer tasks encountering new domains, where domain-invariant representations across downstream multi-domain should be considered. We present three tasks to evaluate SSL baselines: (1) unsupervised cross-domain image retrieval, (2) universal domain adaptation, and (3) few-shot domain adaptation. In our experiments, we show CDS improves various domain transfer methods by providing a better pre-training approach that outperforms the existing state-of-the-art SSL methods.

In summary, our work has the following contributions:

1. We present two-stage pre-training to improve the generalization ability of the standard single-stage pre-training for downstream multi-domain tasks.
2. We propose novel a Cross-Domain Self-supervised pre-training, which learns discriminative and domain-invariant features using unlabeled multi-domain data.
3. We show that CDS outperforms standard ImageNet pre-training and state-of-the-art SSL baselines on various domain transfer tasks.

[1]This term refers to instance-level discriminative representations [42].

2. Related Work

Domain Adaptation. Traditionally, unsupervised domain adaptation (UDA) addresses the problem of generalization to a different but related unlabeled target domain from a labeled source domain. Conventional UDA assumes a closed set where categories are fully shared between the source and target domains. With this assumption, the target features are aligned with the source features by minimizing domain distances [2] using: adversarial domain classifier based learning [12, 19, 24, 38], maximum discrepancy of domain distributions [35, 45], and entropy optimization [34, 24, 33]. Recently, DANCE [34] addressed the problem of universal domain adaptation, where arbitrary category shift exists between the source and target domains (*i.e.*, closed set, open set [22], partial [4], open-partial [44]). DANCE proposes neighborhood clustering via entropy optimization on the unlabeled target domain and entropy-based rejection to identify private classes in the target domain. In this paper, we focus on a pre-training approach to be used before applying UDA methods. While most prior work uses ImageNet pre-training for initialization, we aim to improve pre-training on multi-domain data via self-supervised learning which induces both domain-invariant and class-discriminative features on new domains without additional labels.

Self-supervised Learning. Self-supervised learning (SSL) [9, 14, 28, 42, 16, 6] introduces self-supervisory signals for solving pretext tasks. These pretext tasks enable a model to learn semantically meaningful features from unlabeled data for later use in downstream tasks. Prior work proposes pretext tasks such as: rotation prediction [14] or Instance Discrimination (ID) [20, 42]. Instance Discrimination and SimCLR [42, 7] achieve very powerful performance by classifying an image as its own unique class but treating all other instances as negative instances. However, these works focus on pre-training a model on a large-scale single domain dataset such as ImageNet. We later show that these methods are not very effective for a downstream dataset that has a domain shift.

Self-supervised Learning for Adaptation. Some unsupervised domain adaptation methods [5, 10, 36, 43] add existing SSL objectives (*e.g.*, [14, 28]) to improve performance by jointly training with source labels. These methods rely on a large amount of source supervision to guide

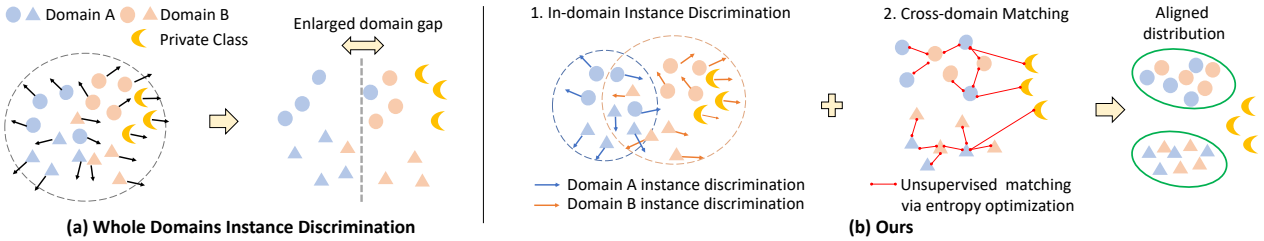


Figure 2: Comparison of Instance Discrimination (ID) [42] and ours (CDS): (a) ID distinguishes every feature from all the others without considering the domain gap, so that the domain gap between domains increases. (b) In order to reduce the domain gap, CDS jointly uses In-domain ID and cross-domain matching to learn features that are domain-invariant as well as discriminative (best viewed in color).

unlabeled target data and often assume the initial representation is already discriminative for the target domain. In contrast, our method explicitly finds an instance-to-instance matching across domains for domain alignment without any source supervision as shown in Fig. 2. Recently, a clustering method with unlabeled multiple domains [27] is proposed for a domain generalization task. However, it underperforms ImageNet pre-training while our method outperforms ImageNet pre-training by a large margin in our task.

3. Cross-Domain Self-supervision

We explore a pre-training approach called Cross-Domain Self-supervision (CDS) for multi-domain settings where we are given a domain A, $\mathcal{D}_A = \left\{\left(\mathbf{x}_i^A\right)\right\}_{i=1}^{N_A}$ and a different but related domain B, $\mathcal{D}_B = \left\{\left(\mathbf{x}_j^B\right)\right\}_{j=1}^{N_B}$. $\mathcal{D}_A$ and $\mathcal{D}_B$ contain the shared categories but there could be some category shift between $\mathcal{D}_A$ and $\mathcal{D}_B$ [44, 34]. For example, $\mathcal{D}_B$ may contain private classes which are not shared with $\mathcal{D}_A$. Our goal is to learn discriminative features on each $\mathcal{D}_A$ and $\mathcal{D}_B$, and domain-invariant features across the same categories in $\mathcal{D}_A$ and $\mathcal{D}_B$. We use a CNN architecture $F(\cdot)$ with L2 normalization [42], which outputs a feature vector $\mathbf{f} \in \mathbb{R}^d$.

We initialize the feature extractor $F(\cdot)$ with ImageNet pre-training, which is generally useful for many visual tasks. Then, we perform the second pre-training stage with CDS using downstream data $\mathcal{D}_A$ and $\mathcal{D}_B$ to provide more discriminative and domain-invariant representations for downstream multi-domain tasks. As shown in Fig. 2-(b), CDS consists of two objectives: (1) learning visual similarity with in-domain instance discrimination for each domain and (2) cross-domain matching for domain alignment. Then, this pre-trained model can be fine-tuned for downstream tasks including domain adaptation.

3.1. In-domain Instance Discrimination

The goal of this objective is to learn a discriminative feature extractor on the downstream data. We aim to improve ImageNet pre-training considering two aspects: (1) downstream tasks can contain novel categories that do not appear in ImageNet (category shift); (2) domain shift can exist between ImageNet and downstream datasets. Therefore, representations learned only on ImageNet can be less effective for initialization for downstream tasks. We utilize in-domain instance discrimination to learn visual similarity for two new domains to improve discriminative power.

For a single-domain, Instance Discrimination [9, 42, 7] (ID) learns visual similarity by imposing a unique class on every image instance and by training a model such that each image is classified to its own instance identity by *treating all the other images as negative pairs*. ID hypothesizes that a model can discover the underlying class-discriminative semantic similarity from instance similarity, which is helpful for a recognition task as shown in [42, 7].

A naive deployment of ID to multi-domain data could increase the domain gap between the $\mathcal{D}_A$ and $\mathcal{D}_B$, because ID treats all other samples as negatives against a given query sample without distinguishing domains. Given a query from the $\mathcal{D}_A$, if we treat all other samples in both domains as negatives, the negatives may contain samples in $\mathcal{D}_B$ belonging to the same class with the query. Besides, the difference between domains (*i.e.*, differences in style, color) can be more easily identified than categorical difference, as illustrated in Fig. 2-(a). Thus, the naive deployment of ID enlarges the difference between domains, which we do not aim to do.

In order to alleviate these problems, we propose to use In-domain ID, where negative pairs are sampled only from the same domain. This aims to prevent learning features to discriminate the two domains, as illustrated in Fig. 2-(b).

We sample features from domain-specific memory banks. We first initialize the memory banks, $\boldsymbol{V^A}$ and $\boldsymbol{V^B}$, from D_A and D_B with the feature extractor $F(\cdot)$,

$$\boldsymbol{V^A} = [\mathbf{v}_1^A, \cdots, \mathbf{v}_{N_A}^A], \quad \boldsymbol{V^B} = [\mathbf{v}_1^B, \cdots, \mathbf{v}_{N_B}^B], \quad (1)$$

where $\mathbf{v}_i$ is the "cashed feature" vector of the image $\mathbf{x}_i$, *i.e.*, $\mathbf{v}_i^A = F(\mathbf{x}_i^A)$. After this initialization, the memory bank features are updated with a momentum in every batch (described in the later section); as the cached features do not need gradient computations, this is highly memory efficient.

Using the feature extractor $F(\cdot)$, we obtain "live" feature vectors $\mathbf{f}_i^A = F(\mathbf{x}_i^A)$ and $\mathbf{f}_j^B = F(\mathbf{x}_j^B)$ from an image $\mathbf{x}_i^A \in \mathbb{B}$ and an image $\mathbf{x}_j^B \in \mathbb{B}$ in a batch. To perform In-domain ID, we compute the similarity distributions P_i^A and P_j^B by measuring the pairwise similarities (dot product) between features and the corresponding memory bank as

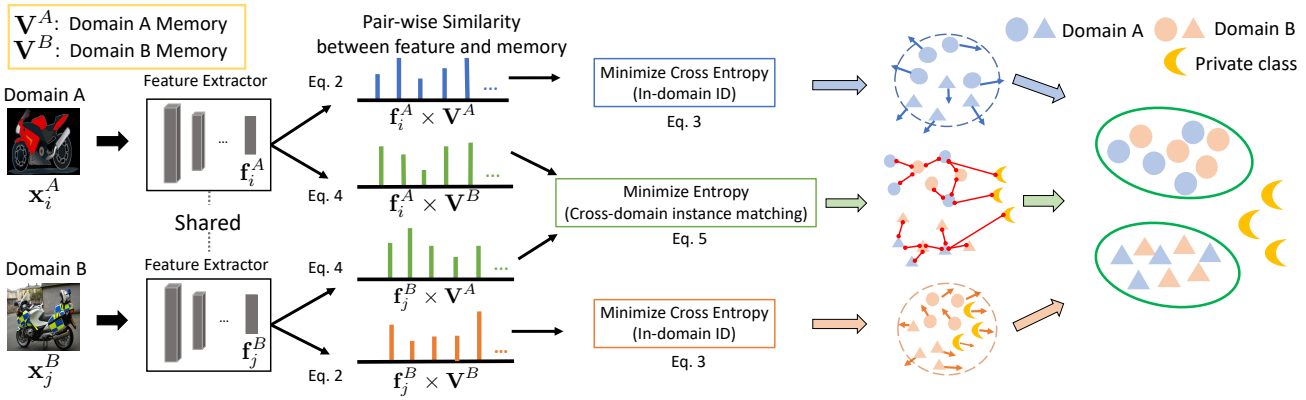


Figure 3: An overview of CDS. In *in-domain instance discrimination*, we measure the similarity of features within each domain, and then perform domain adaptive instance discrimination to learn discriminative features in each domain. In *cross-domain matching*, we measure similarity between a feature and cross-domain features from the cross-domain memory bank and then minimize the entropy for cross-domain matching (best viewed in color).

shown in Fig. 3,

$$P_i^A = \frac{\exp((\mathbf{v}_i^A)^\top \mathbf{f}_i^A/\tau)}{\sum_{k=1}^{N_A} \exp((\mathbf{v}_k^A)^\top \mathbf{f}_i^A/\tau)}, P_j^B = \frac{\exp((\mathbf{v}_j^B)^\top \mathbf{f}_j^B/\tau)}{\sum_{k=1}^{N_B} \exp((\mathbf{v}_k^B)^\top \mathbf{f}_j^B)/\tau)}, \tag{2}$$

where the temperature parameter τ determines the concentration level of the similarity distribution [18]. Finally, we perform In-domain ID by minimizing the averaged negative log-likelihood over a batch $\mathbb{B}$:

$$\mathcal{L}_{I\text{-}ID} = -\frac{1}{|\mathbb{B}|}(\sum_{i\in\mathbb{B}} \log P_i^A + \sum_{j\in\mathbb{B}} \log P_j^B)), \tag{3}$$

where i and j denote the unique index of x_i and x_j.

3.2. Cross-domain Matching

With In-domain ID, we assume a model learns to extract class-discriminative features in each domain. However, it does not explicitly promote domain-invariant features between D_A and D_B. To encourage domain aligned yet discriminative features across the two related domains, we perform cross-domain feature matching as shown in Fig. 2-(b). This is done by making relatively nearby cross-domain points closer, while keeping dissimilar points further.

The key difference is that prior alignment methods using an adversarial domain classifier [12], MMD [25], or optimal transport [8, 3] focus on minimizing the domain gap between distributions of the two domains D_A and D_B. These do not consider class-class semantic similarity between the two domains and may lose class-discriminative power [23]. We propose to use the knowledge that samples of the same class are closer than other samples of different classes in the feature space across different domains (*e.g.*, a chair image in D_A share more similar property (shape, pattern) to a chair image in D_B than other classes such as bike or computer. We discover both positive and negative cross-domain pairs in an unsupervised way. Then, we maximize distances of negative matchings while minimizing a distance of a positive matching to enhance class-discriminative features in different domains. In comparison, optimal transport [3] scales poorly and is limited to finding a match in a batch, while we find a match globally in all cross-domain samples by using "cached" features in the memory bank instead of "live" features in the batch.

To discover positive and negative pairs, we minimize the entropy of the pairwise similarity distribution between a feature in one domain and features in the other domain memory bank. Since entropy minimization encourages a model to make a confident prediction, the model chooses a sample to match and enforces the query feature (*i.e.*, $\mathbf{f}_i^A$ or $\mathbf{f}_j^B$) to be closer to the matched sample. At the same time, the model enforces the query feature to be far from all the other unmatched examples in another domain, which makes it learn class-discriminative features across domains.

To be specific, given the query vectors, $\mathbf{f}_i^A{=}F(\mathbf{x}_i^A)$ and $\mathbf{f}_j^B{=}F(\mathbf{x}_j^B)$ from $\mathbf{x}_i^A$ and $\mathbf{x}_j^B$ in a batch $\mathbb{B}$, we first measure cross-domain pairwise similarities between the live features and the cross-domain memory bank features (*i.e.*, $\mathbf{v}^\mathbf{B}$, $\mathbf{v}^\mathbf{A}$) in Fig. 3:

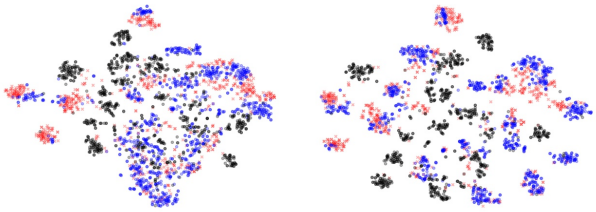
$$P_{j',i}^{A\to B} = \frac{\exp((\mathbf{v}_{j'}^B)^\top \mathbf{f}_i^A/\tau)}{\sum_{k=1}^{N_B} \exp((\mathbf{v}_k^B)^\top \mathbf{f}_i^A/\tau)}, P_{i',j}^{B\to A} = \frac{\exp((\mathbf{v}_{i'}^A)^\top \mathbf{f}_j^B/\tau)}{\sum_{k=1}^{N_A} \exp((\mathbf{v}_k^A)^\top \mathbf{f}_j^B/\tau)}. \tag{4}$$

Then we minimize the averaged entropy of the similarity distributions in a batch:

$$\mathcal{L}_{CDM} = \frac{1}{|\mathbb{B}|}(\sum_{i\in\mathbb{B}} H(P_i^{A\to B}) + \sum_{j\in\mathbb{B}} H(P_j^{B\to A})), \tag{5}$$

$$H(P_i^{A\to B}) = -\sum_{j'}^{N_A} P_{j',i}^{A\to B} \log P_{j',i}^{A\to B},$$

$$H(P_j^{B\to A}) = -\sum_{i'}^{N_B} P_{i',j}^{B\to A} \log P_{i',j}^{B\to A},$$



(a) CDS w/o I-ID (Eq. 5 only) (b) CDS (Eq. 3 + Eq. 5)

Figure 4: t-SNE visualization from CDS w/o In-domain ID and CDS. **Red** / **blue** / **black** dots represent shared classes in D_A and D_B and private classes in D_B, respectively. We observe that In-domain ID distinguishes the private classes (**black**) from shared classes in D_B (**blue**).

where $H(\cdot)$ represents the entropy measured from the probability in Eq. 4.

The overall objective function for CDS is to minimize:

$$\mathcal{L}_{CDS} = \mathcal{L}_{I\text{-}ID} + \mathcal{L}_{CDM}. \tag{6}$$

We also update the memory banks with the features in the batch with a momentum η to encourage smoothness of training following [42]:

$$\begin{aligned} \forall i \in \mathbb{B}, \mathbf{v}_i^A &= (1-\eta)\mathbf{v}_i^A + \eta \mathbf{f}_i^A, \\ \forall j \in \mathbb{B}, \mathbf{v}_j^B &= (1-\eta)\mathbf{v}_j^B + \eta \mathbf{f}_j^B. \end{aligned} \tag{7}$$

After we pre-train a model with CDS, we fine-tune the pre-trained model with existing domain transfer methods and evaluate performance gains.

3.3. Category Shift between D_A and D_B

In the setting of universal domain adaptation [34], D_A and D_B may contain private classes not shared between them. Assuming we have private classes in D_B, a method should align shared classes in D_A and D_B and while separating these from private classes. As In-domain ID learns class-semantic similarity by separating visually dissimilar images in each domain, the private classes can be embedded far from the shared classes. Fig. 4 shows the feature visualization of shared classes in D_A (**red**) and D_B (**blue**) and private classes in D_B (**black**) on Art and Painting domains in Office-Home. In Fig. 4-(a), some private classes can be aligned with the shared classes without In-domain ID. However, Fig. 4-(b) shows that In-domain ID (Eq. 3) on D_B keeps the blue and black dots to be distinctive, which prevents aligning the private classes with the shared classes. Thus, In-domain ID serves as a good regularizer for separating the private classes from the shared classes.

4. Experiments

We evaluate Cross-Domain Self-supervision (CDS) in the variety of domain transfer applications: (1) unsupervised cross-domain image retrieval (Sec 4.2), (2) universal domain adaptation (Sec. 4.3), and (3) few-shot domain adaptation (Sec. 4.4). We summarize key findings: (1) existing SSL baselines do not work well under a domain shift, (2) In-domain ID tends to perform better than the naive ID, and (3) our cross-domain matching performs better than the adversarial domain alignment [12].

4.1. Experiment Setting

Datasets. We utilize three standard domain adaptation benchmarks: CUB which is a fine-grained bird classification dataset [40, 41]) with Real and Painting domains and 200 categories; Office-Home [39] with Art (Ar), Clipart (Cl), Real (Rw), and Product (Pr) domains, and 65 categories; Office [32] with Amazon (A), Dslr (D), and Webcam (W) domains and 31 categories; In the supplementary, we show the overall statistics of the datasets. While most of the categories in Office and Office-Home are shared with ImageNet, CUB contains many novel categories.

Implementation details. Our method is implemented in PyTorch [30] with a single GTX1080Ti. We use a ResNet-50 [17] pre-trained on ImageNet followed by a FC layer and a L_2 normalization layer as a feature extractor. In the pre-training with CDS, we use SGD with the moment parameter 0.9, a learning rate of 0.003, a batch size of 64, weight decay rate $5e^{-4}$. As for the parameters τ and η, we set $\tau = 0.1$ and $\eta = 0.5$ for all experiments. We apply standard data augmentation including random cropping and horizontal flipping. In the supplementary, we show additional details and sensitivity analysis.

Evaluation. Conventional unsupervised domain adaptation (DA) uses ImageNet pre-training as initialization. The goal is to evaluate whether self-supervised learning (SSL) on downstream multi-domain can provide better initialization to domain transfer methods. We choose one domain as D_A and one of the remaining domains as D_B in each dataset following domain adaptation settings [32, 25, 12]. We compare our CDS with ImageNet pre-training, which is a strong and widely used baseline and existing SSL baselines: Instance Discrimination (ID) [7], SimCLR [7], Jigsaw Puzzle [28], Predicting Rotation [14], MoCo [16], and SwAV [6]. We also integrate domain alignment with an adversarial domain classifier (DC) [12] to build a fair yet commonly used baseline. All the baselines start with ImageNet pre-training and each SSL is applied on the union set of D_A and D_B. Then each pre-trained model is finetuned on downstream tasks. We report average accuracy of three runs. We report mean accuracy on all settings in Office-Home (Ar, Cl, Rw, and Pr), CUB (Real and Painting), Office (A, D, and W). Detailed results are shown in the supplementary.

4.2. Unsupervised Cross-domain Image Retrieval

From the unsupervised pre-training, SSLs can be directly applied to an unsupervised cross-domain image retrieval task between D_A and D_B. We query an image from D_A

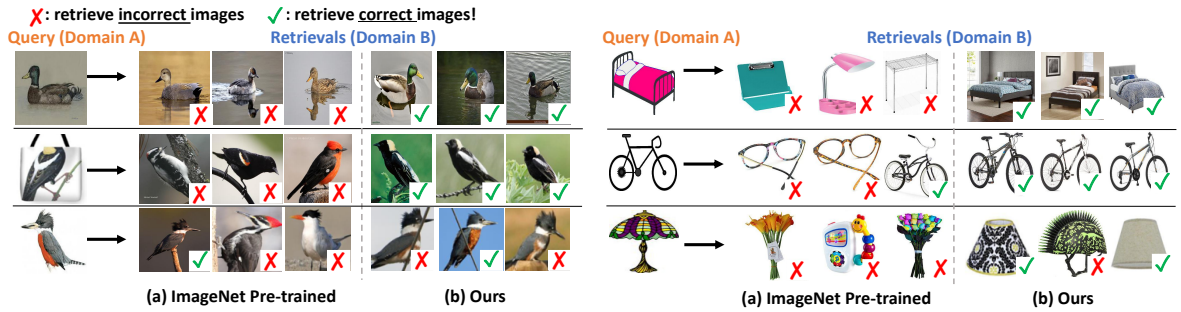


Figure 5: Retrieval of the cross-domain neighbors using (a) standard ImageNet pre-trained features and (b) ours (CDS). While ImageNet pre-trained features are biased to wrong textures and colors, our method learns better semantic similarity across domains.

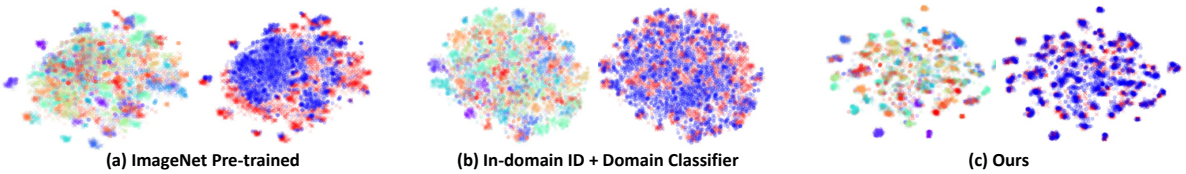


Figure 6: t-SNE visualization of ours and baselines. Each color represents different classes in left subfigures and red and blue represents D_A and D_B in right subfigures. Ours (CDS) extracts features that are clearly class-discriminative as well as domain-invariant.

Pre-train	Office-Home			CUB		
	P@1	P@5	P@15	P@1	P@5	P@15
ImageNet	49.9	44.9	39.5	22.6	18.8	16.2
ID [42]	42.2	36.2	31.8	22.4	18.0	14.9
SimCLR [7]	48.0	43.6	38.3	13.4	11.6	10.1
SimCLR+DC	48.0	43.5	38.4	13.5	11.7	10.1
In-domain ID	44.2	39.1	32.6	22.8	18.9	15.8
CDS	**56.3**	**53.9**	**50.2**	**40.9**	**37.5**	**35.2**

Table 1: Precision@k (P@k) comparison of different pre-training methods on the unsupervised cross-domain image retrieval task.

and retrieve images in D_B. If a retrieval is the same class as the query, we consider this as correct. Fig. 1 and 5 compare the retrieval results from the ImageNet pre-training and ours on CUB and Office-Home. We observe that the ImageNet weights are biased to wrong color or texture information, whereas CDS tends to capture better shape representation with proper shape and texture information. Table 1 reports precision@k (P@k) averaged on all cross-domain settings from cross-domain retrievals in Office-Home and CUB. We observe that the ImageNet pre-training is strong and outperforms the existing SSL baselines. CDS outperforms all the other baselines. Especially, CDS significantly improves the score on CUB. These show that CDS can adapt well to new datasets under domain shift. We show more visualization in the supplementary.

Feature visualization. Fig. 6 shows t-SNE visualization [26] of features from the ImageNet pre-training and ours on the setting Rw→Cl of Office-Home. Compared to (b) DC for feature alignment [12], it qualitatively shows that (c) CDS clusters examples in the same class in the feature space; thus, CDS favors more discriminative features. The red-blue dot plots represent the D_A and D_B domains, which illustrates that CDS can yield well-aligned features while preserving the class-discriminative power.

$\lvert C\rvert/\lvert\bar{C}_s\rvert/\lvert\bar{C}_t\rvert$	CUB	Office-Home	Office
Closed-set	200 / 0 / 0	65 / 0 / 0	35 / 0 / 0
Partial	100 / 100 / 0	25 / 45 / 0	10 / 21 / 0
Open set	100 / 0 / 100	15 / 0 / 50	10 / 0 / 11
Open-partial	100 / 50 / 50	10 / 5 / 50	10 / 10 / 11

Table 2: Statistics on category shift under different UDA settings.

4.3. Universal Domain Adaptation

Setup. Unsupervised domain adaptation is the task of transferring knowledge from a labeled source domain to an unlabeled target domain. D_A denotes the labeled source domain and D_B denotes the unlabeled target domain (*i.e.*, Adapting D_A to D_B: $D_A \rightarrow D_B$). The source and target domain may contain private classes: closed set [24], open set [22], partial [4], or open-partial [44]. $|C|$ denotes the numbers of the shared classes ($|C_s \cap C_t|$) and $\bar{C}_s, \bar{C}_t$ denotes the number of source private and target private classes. For Office and Office-Home, we use the same split in [34]. The overall statistics are in Table 2. We classify target private classes as the "unknown" class. DANCE [34] is the recently proposed method, which achieves state-of-the-art results including closed set, open set, partial, and open-partial DA. We employ SO (Source-only) and DANCE [34], which achieve higher performances than all the other DA baselines. DANCE utilizes target neighborhood clustering and entropy-based sample rejection. SO also uses the entropy-based rejection as DANCE to identify unknown classes. We report the overall target accuracy for all UDA scenarios. For open set DA, we additionally report the mean class accuracy and H-score [11], which is a harmonic mean of accuracy on known classes and accuracy on unknown classes. In open set DA, it is important to consider both metrics, as the number of unknown samples can overwhelm the number of known samples. After pre-training a model with SSL

Pre-train	CUB			Office-Home		
	Closed	Partial	Open	Closed	Partial	Open
ImageNet	54.5	58.1	54.6	69.1	**71.1**	78.1
ID [42]	55.6	54.7	40.5	66.3	67.3	71.0
ID+DC	56.0	54.8	38.1	66.1	67.0	71.0
SimCLR [7]	49.1	52.7	33.2	66.6	68.1	72.5
MoCo [16]	52.4	52.5	32.2	65.9	66.6	72.0
SwAV [6]	55.4	56.1	45.2	67.4	69.0	73.2
In-domain ID	56.6	56.2	39.5	66.4	66.8	71.9
CDS	**59.0**	**65.4**	**55.9**	**69.9**	69.7	**78.7**

Table 3: Comparison with SSL baselines using DANCE [34] on each setting. We report mean class accuracy for open set DA.

baselines, we compare the performances of SO and DANCE from different SSL baselines.

Results. Table 3 shows the comparison with ours and SSL baselines on closed set, partial, and open set DA using DANCE [34]. SSL baselines obtain mixed results compared to ImageNet pre-training while CDS outperforms ImageNet pre-training and SSL baselines except the Office-Home partial. These results show that representations learned from CDS are more discriminative, domain-aligned, and effective for DA. By comparing CDS with ID + DC, our cross-domain matching clearly performs better than DC. In Table 4, we show the results on open set and open-partial DA. DANCE obtains good performance on mean class accuracy but low H-scores, which means DANCE is less effective in classifying unknown classes. CDS significantly improves the H-scores and overall accuracy (Acc) of SO and DANCE while improving/maintaining mean class accuracy compared to ImageNet pre-training.

The impact of category shift between ImageNet and downstream datasets. We observe that CDS is more beneficial to CUB than Office-Home. This is because most of the categories in Office-Home are shared with ImageNet, but there are many novel classes in CUB. CDS can be more useful for downstream tasks, which has a bigger category shift from ImageNet.

Comparison with other DA baselines. We show that DANCE with CDS can achieve the state-of-the-art results by comparing with other DA baselines in Table 5 on the open-partial setting. We report the results of DANN [12], Universal Adaptation Network (UAN) [44], and Calibrated Multiple Uncertainties (CMU) [11]. DANCE obtains the state-of-the-art mean class accuracy but achieves lower H-scores than CMU on Office-Home. When CDS is applied to DANCE (DANCE+CDS), it improves the mean class accuracy by 0.6% and H-score by 21.6% on Office-Home. In Office, CDS slightly decreases the class mean accuracy but improves H-score by a large margin.

4.4. Few-shot Domain Adaptation

Setup. In this section, we explore domain adaptation with a source domain with few-source labels and an unlabeled target domain with closed set DA, where categories between source and target are fully shared. Similarly, D_A denotes the source domain with few-labels and many unlabeled data and D_B denotes the unlabeled target domain (*i.e.*, Adapting D_A to D_B: $D_A \rightarrow D_B$). Conventional DA assumes many source labels are available, which may limit the wide-spread application of DA as highlighted in semi-supervised learning literature [29]. SSL is shown to be effective in semi-supervised learning [42, 7]. Following the semi-supervised learning evaluation protocols, we randomly select few-source labels (*i.e.* 1-shot / 3-shots) as labeled and treat others as unlabeled. To show the benefit of SSL to diverse DA methods, we consider DANN [12], CDAN [24] with entropy conditioning, SRDC [37], and MME [33]. DANN and CDAN are based on an adversarial domain classifier (DC). SRDC uses clustering and MME uses adversarial entropy optimization for domain alignment. We apply entropy minimization [15] on unlabeled source data for all baselines and ours. We report average accuracy over three different random splits.

Results. Table 6 shows the comparison of CDS with ImageNet pre-trained weights on Office-Home and CUB, where CDS improves the performance in all cases. CDS shows higher performance gains on 1-shot settings than 3-shots settings against the baselines, which shows the label-efficiency of CDS. Even with the source-only model (SO), CDS largely improves the performance by domain-aligned features. In CUB, which is a challenging fine-grained classification dataset with few labels, CDS significantly im-

Adapt.	Pre-train	Open-set			Open-partial		
		Acc	Class Acc	H-score	Acc	Class Acc	H-Score
(a) CUB							
SO	ImageNet	51.8	47.9	49.9	51.3	44.9	50.5
SO	CDS	**56.6**	**50.9**	**54.0**	**57.4**	**46.0**	**56.3**
DANCE	ImageNet	42.0	54.6	32.9	53.0	**51.6**	49.1
DANCE	CDS	**56.8**	**55.9**	**53.6**	**64.2**	50.7	**62.8**
(b) Office-Home							
SO	ImageNet	55.7	69.6	57.8	54.0	72.5	57.8
SO	CDS	**60.5**	**73.1**	**62.5**	**59.1**	**74.3**	**62.3**
DANCE	ImageNet	46.8	78.1	46.6	45.5	80.4	49.2
DANCE	CDS	**66.1**	**78.7**	**68.6**	**66.7**	**81.0**	**70.8**
(c) Office							
SO	ImageNet	76.4	89.1	73.5	72.6	85.5	73.5
SO	CDS	**79.8**	**89.4**	**77.9**	**75.5**	**86.9**	**77.9**
DANCE	ImageNet	79.3	94.1	74.5	82.4	**93.7**	80.3
DANCE	CDS	**91.8**	**94.7**	**92.1**	**87.3**	91.2	**87.3**

Table 4: Target accuracy (%) on open set and open-partial DA averaged on all settings in each dataset.

Method	Office-Home		Office	
	H-score	Class Acc	H-score	Class Acc
SO [11]	47.3	73.2	50.9	82.7
DANN [12]	46.2	73.2	50.6	81.8
UAN [44]	56.6	77.0	63.5	89.2
CMU [11]	61.6	78.0	73.1	91.1
DANCE [34]	49.2	80.4	80.3	**93.7**
DANCE+CDS	**70.8**	**81.0**	**87.3**	91.2

Table 5: Comparison with other DA methods on open-partial DA. We report the mean class accuracy and H-score averaged on all settings in Office-Home and Office.

Pre-train	CUB: Target Acc. (%) on 1-shot / 3-shots			
	SO	DANN	CDAN	MME
ImageNet	5.1 / 15.0	6.1 / 17.6	6.5 / 18.5	12.0 / 41.9
CDS	**20.8 / 33.4**	**20.2 / 34.6**	**23.2 / 38.5**	**28.7 / 47.4**
Pre-train	Office-Home: Target Acc. (%) on 1-shot / 3-shots			
	SO	CDAN	MME	SRDC
ImageNet	18.7 / 34.2	19.6 / 35.0	28.9/ 50.3	28.2 / 48.9
CDS	**33.8 / 45.7**	**35.0 / 51.1**	**36.3 / 55.2**	**41.3 / 55.9**
Pre-train	Office: Target Acc. (%) on 1-shot / 3-shots			
	SO	CDAN	MME	SRDC
ImageNet	37.3 / 61.9	46.0 / 74.0	59.1 / 74.6	60.5 / 75.7
CDS	**60.9 / 73.9**	**65.2 / 79.2**	**65.8 / 79.8**	**69.2 / 79.8**

Table 6: Target accuracy (%) on 1-shot and 3-shots averaged on all settings in the Office-Home and CUB datasets.

Pre-train	Feature Analysis		Adaptation
	Linear	kNN	CDAN 1-shot
ImageNet	61.9±1.6	53.5±0.0	46.6±4.3
Jigsaw [28]	50.7±0.4	32.3±4.5	48.9±5.2
Rotation [14]	41.1±5.3	36.4±3.0	44.7±4.1
ID	62.2±0.6	59.6±0.4	45.2±2.8
SimCLR	62.5±0.6	60.2±0.2	54.0±3.0
SimCLR+DC	62.7±0.6	60.9±0.2	53.8±3.3
In-domain ID (Eq. 3)	63.0±0.4	61.7±1.1	49.7±0.6
CDS (Eq. 5)	**71.3**±0.4	**68.5**±0.5	**66.8**±2.1

Table 7: Comparison with the baselines under each evaluation protocol on Office D→A. CDS outperforms the baselines.

proves accuracy compared to the baseline. These results show that CDS is more effective than just naive adaptation of ImageNet pre-trained weights, which is a strong and generally used baseline in prior DA works.

Feature analysis. To see where this performance gain comes from, we conduct feature analysis in Table 7. Table 7 shows the comparison of ours with the SSL baselines on the Office D→A setting. First, following the SSL evaluation protocols in [42, 7], we evaluate the learned representations with a linear classifier and weighted k-Nearest Neighbor (kNN) classifier. We freeze the feature extractor but train a linear classifier or kNN classifier on top of the frozen features with full source labels and measure accuracy on the target domain. We observe that Jigsaw and Rotation hurt the performance of the ImageNet pre-training. Second, we fine-tune the whole network for the few-shot domain adaptation task in the column of *Adaptation* using CDAN on the 1-shot setting. CDS outperforms the baselines by a large margin in all cases.

Label efficiency comparison with SSL. Fig. 7 shows the results of ours and baselines with different fractions of labels using CDAN on Office-Home. Our method consistently outperforms the baselines and stably improves with the additional labels even with the full source labels, while the SSL baselines obtain similar accuracy as ImageNet. CDS greatly improves when there are a few source labels.

Consistency of SSL and downstream task objectives. A good SSL method should match SSL objectives with downstream performance [42]. We analyze this consistency by measuring downstream target accuracy according to the SSL learning training epoch in Fig. 8. In Fig. 8-(a), we

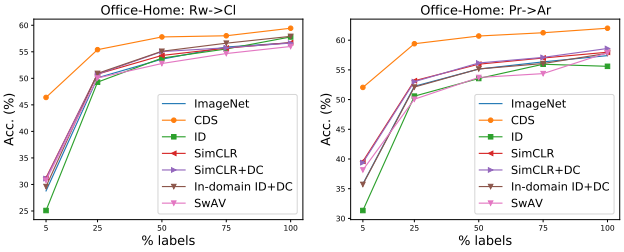
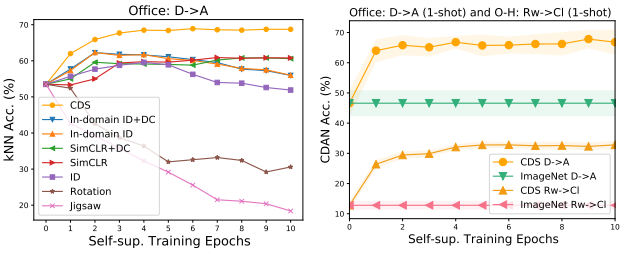


Figure 7: Target accuracy on different fractions of source labels. While other SSL baselines achieve similar results as ImageNet pre-training, ours consistently performs better.



(a) (b)

Figure 8: (a): Target accuracy using Weighted kNN according to training epochs of each self-supervised learning. (b): Target accuracy using CDAN according to training epochs of CDS.

measure the target accuracy using the same kNN classifier in Table 7 according to the SSL training epochs on the D→A setting on Office. The accuracy at epoch 0 reports the accuracy of ImageNet pre-training. Jigsaw and Rotation decrease in accuracy over training, *i.e.*, overfitting to the respective proxy tasks. Compared to the baselines, CDS improves the performance in early training epochs of SSL and converges. In Fig. 8-(b), we also show the target accuracy and standard deviation over three random splits from CDAN according to the training epochs on D→A in Office and Rw→Cl in Office-Home. We observe that CDS improves the accuracy compared to the ImageNet weights. The standard deviations show that the accuracy is not very sensitive to different random splits. Please refer to our supplementary for more detailed results of our experiments.

5. Conclusion

Conventional domain adaptation (DA) method uses ImageNet pre-training as a weight initialization. With two-stage pre-training with CDS, we aim to standard ImageNet pre-training by learning discriminative and domain-aligned features with SSL for downstream multi-domain data. We propose a novel Cross-Domain Self-supervised learning leveraging unlabeled data from multiple domains. CDS can be easily applied to boost performance of diverse domain transfer tasks and outperforms the standard pre-training and existing SSL baselines.

Acknowledgments: This work was supported by Honda, DARPA LwLL, NSF Award No. 1535797, and IITP funded by the Korea government (MSIT) No. 2020-0-00004, Development of Previsional Intelligence.

References

[1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Danny Gutfreund, Joshua Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 2019. 1

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 2

[3] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 447–463, 2018. 4

[4] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018. 2, 6

[5] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2, 5, 7

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 3, 5, 6, 7, 8

[8] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016. 4

[9] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 2, 3

[10] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning from multi-domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3245–3255, 2019. 2

[11] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*, pages 567–583. Springer, 2020. 6, 7

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer, 2017. 1, 2, 4, 5, 6, 7

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2, 5, 8

[15] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2005. 7

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 2, 5, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 2

[20] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. *arXiv preprint arXiv:1904.11567*, 2019. 2

[21] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019. 1

[22] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2019. 2, 6

[23] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022. PMLR, 2019. 4

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 2, 6, 7

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 4, 5

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 6

[27] Willi Menapace, Stéphane Lathuilière, and Elisa Ricci. Learning to cluster under domain shift. *arXiv preprint arXiv:2008.04646*, 2020. 3

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, pages 69–84. Springer, 2016. 2, 5, 8

[29] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018. 7

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Autodiff Workshop*, 2017. 5

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision*, pages 213–226. Springer, 2010. 1, 5

[33] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 7

[34] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *arXiv preprint arXiv:2002.07953*, 2020. 2, 3, 5, 6, 7

[35] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2

[36] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 2

[37] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8725–8735, 2020. 7

[38] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[39] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 5

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[41] Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Progressive adversarial networks for fine-grained domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9213–9222, 2020. 5

[42] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2, 3, 5, 6, 7, 8

[43] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019. 2

[44] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019. 2, 3, 6, 7

[45] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I Jordan. Bridging theory and algorithm for domain adaptation. *arXiv preprint arXiv:1904.05801*, 2019. 2