$DETECT$: A Deep Discriminative Clustering Baseline for Unsupervised and Universal Domain Adaptation

Bin Deng, Yabin Zhang and Kui Jia
School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
eebindeng@mail.scut.edu.cn

Abstract

Given a set of labeled instances on a source domain, unsupervised domain adaptation (UDA) aims to learn a prediction function to classify instances in a shifted, target domain. Depending on the degrees of overlap between the label spaces of the two domains, the problem variants of UDA range from the classical, closed set setting to the most general — arguably the most challenging — setting of universal domain adaptation. In this work, we argue that no matter what the degree of label space overlap, the problem nature of UDA remains unchanged when it comes to learning the intrinsic discrimination of target data in an unsupervised manner, regularized by the labeled discrimination of source data in an unknown but shared label space, and we argue that this regularization should not overwhelm the learning of a target prediction function. To this end, we propose a simple but strong baseline of neighborhooD-prEserved deep discriminaTivE ClusTering ($DETECT$) for UDA, whose design complies with the above learning principles. We conduct thorough experiments that verify the efficacy of constituent components in $DETECT$ across a range of label space overlaps. Such a simple baseline also outperforms all existing methods on four UDA benchmarks.

1 Introduction

Unsupervised domain adaptation (UDA) aims to learn a prediction function for the unlabeled data of a target domain, given labeled data in a shifted, source domain. Recently, problem variants of UDA have been introduced depending on the degree of overlap between the label spaces of the two domains. Specifically, the label space of the target data can be assumed as the same as that of the source data, a subset of the source data or a superset of the source data. Or it can be assumed to have a known overlap with the source data. These options respectively give rise to closed set [11, 27], partial [6, 48] and open set [38, 4] UDAs. You *et al.* [47] describe these UDA variants with a unified Jaccard distance, which measures the degree of overlap between the label spaces of both domains, and they introduce the task of universal domain adaptation.

The task of universal domain adaptation is challenging since we have no prior knowledge about the relationship between the label spaces of both domains, and this applies to general applications in practice. In universal domain adaptation, given labeled source data, for any related target data with different degrees of overlap in the label space, we aim to classify target data correctly if it belongs to overlapped classes across domains, else we reject it as an unknown class. As illustrated in [47], the seminal methods [11, 29, 48, 7, 4, 38] proposed for the closed set, partial and open set UDAs do not work well on universal domain adaptation, so You *et al.* [47] introduce a sample-level weighting mechanism to promote the feature alignment of overlapped classes across domains via the

seminal adversarial UDA method [11], and they reject a target sample as an unknown class using a pre-defined threshold.

We contribute in this paper a novel perspective to the challenging universal domain adaptation problem. We argue that no matter what the degree of the label space overlap, the problem nature of UDA remains unchanged when it comes to learning the intrinsic discrimination of target data in an unsupervised manner, regularized by the labeled discrimination of source data in an unknown but shared label space, and we argue that this regularization should not overwhelm the learning of a target prediction function. To this end, we propose a simple but strong baseline of neighborhooD-prEserved deep discriminaTivE ClusTering ($DETECT$) for UDA, whose design complies with the above problem nature. Concretely, we learn a model using deep discriminative clustering [45] on a data subset of the target domain whose labels (possibly) belong to the overlapped label space across domains. The target subset is achieved by filtering out target samples with high entropy as outliers (i.e. an unknown class), empowered by the existing technique of the out-of-distribution (OOD) detector [23]. Via deep discriminative clustering, the intrinsic discrimination of target data whose labels belong to the shared label space can be investigated and largely preserved. To further reduce contamination of the intrinsic discrimination of the entire target data, we adopt the neighborhood-preserved feature embedding [17, 5], which encourages samples close in the image space to maintain similar feature representations during the embedding learning. An additional regularization is imposed by training the model with labeled source data in a supervised manner, which provides the prior knowledge of category information. In Sec. 5, we investigate the efficacy of constituent components in $DETECT$ across a range of label space overlaps, and we present how such a simple baseline outperforms all existing methods in universal domain adaptation.

2 Related Work

In this section, we briefly review the UDA variants of the closed set, partial and open set, as well as universal domain adaptation, and their representative methods. We also review the technique of deep discriminative clustering, which sets up the technological basis of our $DETECT$.

Closed Set Domain Adaptation The label spaces of the source and target domains are assumed to be the same in the closed set domain adaptation. Classic UDA theories [3, 2, 31] suggest that the target risk can be minimized by bounding the source risk and the distribution discrepancy across domains, which motivates many methods [28, 44, 27, 39, 22, 11, 43, 37, 49] targeted at learning domain-invariant feature representations. Apart from these methods, others based on pseudo labels [36, 46, 21] or other semi-supervised learning techniques [15, 40] have also been widely studied. Recently, state-of-the-art results have been achieved by clustering-based UDA methods [9, 42], which adopt technique approaches that are similar to ours. The unrealistic assumption of a shared label space across domains, however, limits their practical applications.

Partial Domain Adaptation The task of partial domain adaptation assumes that the label space of the source domain subsumes that of the target domain, which partially relaxes the shared label space assumption of closed set domain adaptation. Current methods of partial domain adaptation typically adopt either a class-level [7], an instance-level [48, 6] or both a class- and instance-level [8] weighting mechanism for source data to filter the source samples whose labels are not present in the target data, after which they learn domain-invariant feature representations with samples of the domain-shared classes. Although the shared label space assumption is relaxed, the new assumption that the label space of target data is a subset of the source data also limits its application.

Open Set Domain Adaptation There are two task settings for open set domain adaptation with one slight difference. The first one proposed by Busto *et al.* [4] assumes that both source and target domains enjoy their private classes, and that the overlapped classes are known in advance. To solve this problem, they propose an Assign-and-Transform-Iteratively (ATI) algorithm to assign target samples to source classes, and they make a final classification with SVM classifiers. Another task setting of open set domain adaptation proposed by Saito *et al.* [38] assumes that the data of source private classes are not required, and they reject the data of target private classes as an unknown class with adversarial training. Both task settings require prior knowledge about the overlapped classes, which usually does not hold in practice.

Universal Domain Adaptation The universal domain adaptation proposed by [47] assumes that the overlapped classes between source and target domains are unknown, which thoroughly relaxes the

unrealistic assumptions about the label spaces across domains. To solve this problem, You *et al.* [47] propose the universal adaptation network (UAN) by jointly training an adversarial domain adaptation network and a progressive instance-level weighting scheme, which quantifies the transferability of both source and target samples. More recently, Lifshitz *et al.* [26] propose a sample selection approach through the usage of pseudo-labels and a batch diversity loss, and in [35], a combination of self-supervised training and domain-specific batch normalization is adopted. Unlike these methods, we investigate the problem nature of UDA and correspondingly propose a deep discriminative clustering-based method. As illustrated in Sec. 5, empirical evidence on four benchmark datasets verifies the efficacy of our method across a range of label space overlaps.

Deep Discriminative Clustering Discriminative clustering aims to learn decision boundaries to represent distinctions between categories in an unsupervised manner [12]. By borrowing the powerful feature representation of a deep network, deep discriminative clustering models [20, 41, 45, 10] have been proposed and show promising performance. Among these approaches, a simple technique is proposed in [45] by introducing auxiliary target distributions and then minimizing the Kullback-Leibler (KL) divergence between these auxiliary targets and the posteriors of a discriminative deep networks classifier. Following [45], many methods [10, 14] have been proposed by using similar strategies. In this paper, we simply borrow the ideas of these works and propose a strong baseline of $DETECT$ to address the task for universal domain adaptation.

3 Problem Definition

Given a set of labeled data $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ on a source domain $\mathcal{X}^s \times \mathcal{Y}^s$, and unlabeled data $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ on a target domain $\mathcal{X}^t \times \mathcal{Y}^t$, unsupervised domain adaptation (UDA) aims to learn a prediction function $\boldsymbol{h} : \mathcal{X}^t \to \mathcal{Y}^t$ by utilizing the provided, labeled source data. For the problem of classification, we assume $\mathcal{Y}^s = \{1, ..., K\}$. Depending on different settings, the label space of $\mathcal{Y}^t$ can be assumed to be the same as $\mathcal{Y}^s$, a subset of $\mathcal{Y}^s$ or a superset of $\mathcal{Y}^s$. Or it could have a *known* overlap with $\mathcal{Y}^s$. These assumptions respectively give rise to the problem variants of closed set [11, 27], partial [6, 48] and open set [38, 4] UDAs. This spectrum of UDA variants can be more generally described using a measure of intersection-over-union (IOU) as $\xi = \frac{|\mathcal{Y}^s \cap \mathcal{Y}^t|}{|\mathcal{Y}^s \cup \mathcal{Y}^t|} \in [0, 1]$ [47]; the right extreme of $\xi = 1$ gives the closed set UDA, and when label spaces of the two domains are completely irrelevant, we have the left extreme of $\xi = 0$. In this work, we focus on the most general — arguably the most challenging — UDA setting where the IOU measure ξ is *unknown*. This setting is termed as *universal* [47] or *open-partial* [35] domain adaptation in existing literature. To make it more precise, we term the problem as ξ-UDA to show respect to both *Unsupervised* [11, 27] and *Universal* [47] Domain Adaptation.

Denote $\mathcal{Y}^{t/s} = \{y | y \in \mathcal{Y}^t, y \notin \mathcal{Y}^s\}$ as the label subset of $\mathcal{Y}^t$ that excludes labels in $\mathcal{Y}^s$, and denote the domain-shared label subset as $\mathcal{Y}^{st} = \mathcal{Y}^s \cap \mathcal{Y}^t$. Since nothing is known about the internal structure of $\mathcal{Y}^{t/s}$, the learning objective $\boldsymbol{h} : \mathcal{X}^t \to \mathcal{Y}^t$ of ξ-UDA can be relaxed as $\boldsymbol{h} : \mathcal{X}^t \to \mathcal{Y}^{st} \cup \{y^{t/s}\}$, where $y^{t/s}$ denotes the label of the (super-)class that contains all labels in $\mathcal{Y}^{t/s}$. Considering that $\mathcal{Y}^{st}$ is unknown but $|\mathcal{Y}^{st}| \leq |\mathcal{Y}^s| = K$, one can choose to implement the function $\boldsymbol{h}$ as a two-level, hierarchical classification of $h_1 : \mathcal{X}^t \to \{0, 1\}$ and $\boldsymbol{h}_2 : \mathcal{X}^t \to \{0, 1\}^K$; the binary $h_1(\boldsymbol{x}^t)$ classifies any $\boldsymbol{x}^t \in \mathcal{X}^t$ either as the label $y^{t/s}$ or any label in $\mathcal{Y}^s$, and when the latter is the case, $\boldsymbol{h}_2(\boldsymbol{x}^t)$ further classifies it potentially into a label index of $\mathcal{Y}^{st} \subseteq \mathcal{Y}^s$. One can alternatively learn $\boldsymbol{h}_2$ alone, and use it as an out-of-distribution (OOD) detector [18] to identify any $\boldsymbol{x}^t$ whose label is $y^{t/s}$. We choose to learn $\boldsymbol{h}_2$ alone in this work, and implement it as a K-way classification network.

4 The Proposed Method

We present our proposed method for ξ-UDA in this section. We first note that no matter what the values of ξ are, the problem nature remains unchanged, i.e.:

- to learn the *intrinsic* discrimination of $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ in an unsupervised manner, regularized by labeled discrimination of $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ in the unknown $\mathcal{Y}^{st}$;
- considering that distribution shift is deemed to exist between the two domains, care should be taken when imposing regularization so that the intrinsic structure of target discrimination is *not* overwhelmed by that of the labeled source discrimination.

To this end, we propose a simple but strong baseline for ξ-UDA, based on the above principles that comply with the problem nature. Simply put, our method learns $\boldsymbol{h}_2$ for deep discriminative clustering on those instances of $\{\boldsymbol{x}^t\}$ whose labels belong to $\mathcal{Y}^{st}$, where neighborhood-preserved embedding [17] is used to maintain the intrinsic structure of the target data; regularization from the source data is imposed simply by training the same $\boldsymbol{h}_2$ using $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ in a supervised manner, where we borrow the existing technique of OOD detection [23] to improve the identification of $\{\boldsymbol{x}^t\}$, whose labels are $y^{t/s}$. We thus term this method as *neighborhooD-prEserved deep discriminaTivE ClusTering ($DETECT$)*. Experiments in Sec. 5 show that this simple baseline outperforms all existing methods for ξ-UDA.

We practically implement $\boldsymbol{h}_2$ as two cascaded functions of $\varphi_\vartheta : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{f}_\theta : \mathbb{R}^d \to [0,1]^K$ in a deep network, where φ_ϑ learns the feature embedding, $\boldsymbol{f}_\theta$ is a softmax classifier and (ϑ, θ) denotes the network parameters, collectively. For any instance $\boldsymbol{x}$, we write its feature embedding as $\boldsymbol{z} = \varphi_\vartheta(\boldsymbol{x}) \in \mathbb{R}^d$ and its network output as $\boldsymbol{p} = \boldsymbol{f}_\theta(\boldsymbol{z}) \in [0,1]^K$. We also write the k^{th} element of $\boldsymbol{p}$ as p_k. We omit the superscript s or t in this notation, since the network processes source or target instances equally. We present the specifics of $DETECT$ as follows.

Neighborhood-Preserved Deep Discriminative Clustering

Given the relaxed ξ-UDA task of $\boldsymbol{h} : \mathcal{X}^t \to \mathcal{Y}^{st} \cup \{y^{t/s}\}$, which is practically implemented as the prediction function $\boldsymbol{h}_2 : \mathcal{X}^t \to \{0,1\}^K$, our discovery of the intrinsic target structure focuses on those $\{\boldsymbol{x}^t\}$ whose labels are in $\mathcal{Y}^{st} \subseteq \mathcal{Y}^s$, while treating those with the label $y^{t/s}$ simply as OOD outliers. Assume for now that our choice of deep model $\boldsymbol{h}_2 = \boldsymbol{f}_\theta \circ \varphi_\vartheta$ serves as a good OOD detector; we then identify any $\boldsymbol{x}^t \in \mathcal{X}^t$ either as inliers or outliers based on the following indicator

$$I(\boldsymbol{x}^t) = \begin{cases} 1, & H(\boldsymbol{x}^t) < \tau \log(K), \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $H(\boldsymbol{x}^t) = -\sum_{k=1}^K p_k^t \log(p_k^t)$ measures the entropy of the probability vector $\boldsymbol{p}^t = \boldsymbol{f}_\theta \circ \varphi_\vartheta(\boldsymbol{x}^t)$, and $\tau \in [0,1]$ is a threshold parameter. We discuss how to train $\boldsymbol{f}_\theta \circ \varphi_\vartheta$ as a good OOD detector shortly. To cluster the target instances $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$, we follow the framework of [45, 10], and introduce for the collection $\boldsymbol{P}^t = \{I(\boldsymbol{x}_i^t) \cdot \boldsymbol{p}_i^t\}_{i=1}^{n_t}$ an auxiliary $\boldsymbol{Q}^t = \{I(\boldsymbol{x}_i^t) \cdot \boldsymbol{q}_i^t\}_{i=1}^{n_t}$. Given that the KL divergence between $\boldsymbol{P}^t$ and $\boldsymbol{Q}^t$ is defined as

$$KL(\boldsymbol{Q}^t || \boldsymbol{P}^t) = \frac{1}{N} \sum_{i=1}^{n_t} \sum_{k=1}^{K} I(\boldsymbol{x}_i^t) \cdot q_{i,k}^t \log \frac{q_{i,k}^t}{p_{i,k}^t} \;\text{ with }\; N = \sum_{i=1}^{n_t} I(\boldsymbol{x}_i^t), \tag{2}$$

the minimization of (2) is potentially able to cluster $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ via distribution matching when the auxiliary $\boldsymbol{Q}^t$ captures prior knowledge about the intrinsic, and ideally discriminative, structure of $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$. We inject such prior knowledge into $\boldsymbol{Q}^t$ based on two classical ideas. The first one directly leverages the UDA setting, and pre-trains $\boldsymbol{f}_\theta \circ \varphi_\vartheta$ using the labeled $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$, which enables the trained model to produce pseudo labels for $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$; we use these pseudo labels to initialize individual $\{\boldsymbol{q}^t \in \boldsymbol{Q}^t\}$. Due to the data and label mismatch between $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$, pseudo labels are not guaranteed to be reliable. As compensation, our second idea follows classical clustering and semi-supervised learning [12], and it promotes the balance of assignments among the K clusters for individual $\{\boldsymbol{q}^t \in \boldsymbol{Q}^t\}$; technically, we minimize $KL(\boldsymbol{\varrho}^t || \boldsymbol{u}^t)$, where $\boldsymbol{\varrho}^t = \frac{1}{N} \sum_{i=1}^{n_t} I(\boldsymbol{x}_i^t) \cdot \boldsymbol{q}_i^t$ is the empirical probability of target assignments over the K clusters, and where $\boldsymbol{u}^t$ is a uniform distribution. Combining $KL(\boldsymbol{\varrho}^t || \boldsymbol{u}^t)$ with (2) gives the first component of our learning objective of $DETECT$ for the ClusTering of target data

$$\mathcal{L}_{CT}(\vartheta, \theta; \{\boldsymbol{x}_i^t\}_{i=1}^{n_t}) = KL(\boldsymbol{Q}^t || \boldsymbol{P}^t) + KL(\boldsymbol{\varrho}^t || \boldsymbol{u}^t). \tag{3}$$

To reduce contamination of the intrinsic target discrimination, $DETECT$ implements the second learning principle of ξ-UDA based on the classical idea of neighborhooD-prEserved feature embedding [17, 5]

$$\mathcal{L}_{DE}(\vartheta, \theta; \{\boldsymbol{x}_i^t\}_{i=1}^{n_t}) = \frac{1}{2} \sum_{i,j=1}^{n_t} w_{ij}^t \cdot \|\boldsymbol{z}_i^t - \boldsymbol{z}_j^t\|^2 = \text{Tr}(\boldsymbol{Z}^{t\top} \boldsymbol{L}^t \boldsymbol{Z}^t), \tag{4}$$

where $\boldsymbol{W}^t$ denotes the adjacency matrix of the 2-nearest neighbor graph whose edge weight w_{ij}^t is computed as $w_{ij}^t = e^{-\|\boldsymbol{z}_i^t - \boldsymbol{z}_j^t\|^2/d}$ in the initial training epoch, $\boldsymbol{L}^t$ is the graph Laplacian [5]

built from $\boldsymbol{W}^t$, and $\boldsymbol{Z}^t = [\boldsymbol{z}_1^t; ...; \boldsymbol{z}_{n_t}^t] = [\varphi_\vartheta(\boldsymbol{x}_1^t); \ldots; \varphi_\vartheta(\boldsymbol{x}_{n_t}^t)] \in \mathbb{R}^{n_t \times d}$ contains the feature embedding of target instances. By minimizing (4), features learned by the embedding function φ_ϑ preserve the intrinsic neighborhood structure in the ambient space of $\mathcal{X}^t$.

Source Regularization in the Shared Label Space

Given the choice of deep model $\boldsymbol{h}_2 = \boldsymbol{f}_\theta \circ \varphi_\vartheta$, regularization from the labeled $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ is straightforward to enforce using cross-entropy loss $CE(\boldsymbol{p}^s, \boldsymbol{\delta}_{y^s}) = -\log p_{y^s}^s$, where $\boldsymbol{\delta}_{y^s} \in [0,1]^K$ denotes a one-hot vector with the only entry at the index y^s. To make $\boldsymbol{h}_2$ as a good OOD detector as well, we use a common strategy in existing OOD research [23] that regularizes the training of the deep model using an auxiliary set of unlabeled data $\{(\boldsymbol{x}_i')\}_{i=1}^n$; note that $\{(\boldsymbol{x}_i')\}_{i=1}^n$ is irrelevant to both $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\{(\boldsymbol{x}_i^t)\}_{i=1}^{n_t}$. This gives the regularizer of $DETECT$ for deep discriminaTivE training

$$\mathcal{R}_{TE}(\vartheta, \theta; \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{n_s}, \{\boldsymbol{x}_i'\}_{i=1}^n) = \frac{1}{n_s} \sum_{i=1}^{n_s} CE(\boldsymbol{p}_i^s, \boldsymbol{\delta}_{y_i^s}) + \gamma \mathcal{R}_{OOD}(\vartheta, \theta; \{\boldsymbol{x}_i'\}_{i=1}^n), \tag{5}$$

with

$$\mathcal{R}_{OOD}(\vartheta, \theta; \{\boldsymbol{x}_i'\}_{i=1}^n) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{1}{K} \log p_{i,k}', \tag{6}$$

where γ is a penalty parameter, p_k' is the k^{th} entry of $\boldsymbol{p}'$, and $\boldsymbol{p}' = \boldsymbol{f}_\theta \circ \varphi_\vartheta(\boldsymbol{x}')$; the regularizer $\mathcal{R}_{OOD}$ encourages the learned $\boldsymbol{f}_\theta \circ \varphi_\vartheta$ to give high values of entropy $H(\boldsymbol{x}')$ for any OOD instance $\boldsymbol{x}'$. Note that the alternative OOD manner of temperature scaling [19, 25] is used in [47].

Learning and Inference

Combining the terms (3), (4), and (5) gives the learning objective of our proposed $DETECT$

$$\mathcal{L}_{DETECT}(\vartheta, \theta) = \alpha \mathcal{L}_{DE} + \mathcal{R}_{TE} + \beta \mathcal{L}_{CT}, \tag{7}$$

where α and β are penalty parameters. The training of the deep model $\boldsymbol{f}_\theta \circ \varphi_\vartheta$ is easy to implement using stochastic gradient descent (SGD), where the graph Laplacian in (4) is also computed for instances in each mini-batch. Given a learned $\boldsymbol{f}_\theta \circ \varphi_\vartheta$, we label a target instance $\boldsymbol{x}^t$ by

$$\hat{y}^t = \begin{cases} \arg\max_k [\boldsymbol{f}(\varphi(\boldsymbol{x}^t))]_k, & H(\boldsymbol{x}^t) < \tau \log(K), \\ y^{t/s}, & \text{otherwise.} \end{cases} \tag{8}$$

5 Experiment

We evaluate our $DETECT$ on four benchmark datasets under the ξ-UDA setting with different values of ξ, and we investigate its components thoroughly. We begin by introducing the datasets and the learning setups that are used in our experiments in the following.

The **Office-31** dataset [34] contains 31 classes from three visually distinct domains: Amazon (**A**), DSLR (**D**) and Webcam (**W**). We follow the class split setting of [47] by adopting the 10 common classes between Office-31 and Caltech-256 [13] as $\mathcal{Y}^{st}$, and the next sets of 10 and 11 classes, in alphabetical order, as $\mathcal{Y}^{s/t}$ and $\mathcal{Y}^{t/s}$, respectively. **ImageNet-Caltech** is built from ImageNet-1K (**Im**) [33] with 1000 classes and Caltech-256 (**Cal**) [13] with 256 classes. Following [47, 26], we adopt the 84 classes shared by both domains as $\mathcal{Y}^{st}$ and use their private classes as their private label sets, respectively. **VisDA2017** [32] includes a source domain of synthetic images and a target domain of real-world images, focusing on a special transfer learning setting of simulation to the real world. There are 12 classes in this dataset. Following [47], in alphabetical order, we adopt the first 6 classes as $\mathcal{Y}^{st}$, the next 3 classes as $\mathcal{Y}^{s/t}$ and the rest as $\mathcal{Y}^{t/s}$. **ImageCLEF-DA** [1] consists of four domains, which are randomly selected from Caltech-256 (**C**), ImageNet ILSVRC2012 (**I**), Pascal VOC 2012 (**P**), and Bing (**B**). Each domain equally contains 600 images of 12 classes. In alphabetical order, we adopt the first 6 classes as $\mathcal{Y}^{st}$, the next 3 classes as $\mathcal{Y}^{s/t}$ and the rest as $\mathcal{Y}^{t/s}$.

Following [47], we adopt the average class accuracy (AA) as the evaluation metric for ξ-UDA, where the results are calculated by averaging the accuracy over all classes, including the (super-)class $y^{t/s}$. In our ablation studies, we also report the overall accuracy (OA), i.e. the averaged accuracy of all target samples, as a supplement. We empirically set α as 0.01 in all experiments. For the value of

$\mathcal{L}_{CT}$	$\mathcal{L}_{DE}$	$\mathcal{R}_{OOD}$	A2W	D2W	W2D	A2D	D2A	W2A	Avg.
-	-	-	80.34	95.38	97.02	85.11	82.92	82.25	87.17
✓	-	-	87.89	94.98	94.64	92.80	91.01	90.07	91.90
✓	✓	-	93.00	95.29	96.24	93.13	91.33	90.88	93.31
✓	✓	✓	94.43	96.87	96.31	94.69	91.08	92.06	94.24

Table 1: Ablation experiments on the Office-31 dataset with the evaluation metric of **AA** (%).

$\mathcal{L}_{CT}$	$\mathcal{L}_{DE}$	$\mathcal{R}_{OOD}$	A2W	D2W	W2D	A2D	D2A	W2A	Avg.
-	-	-	74.41	87.94	84.04	72.19	80.66	76.43	79.28
✓	-	-	72.28	82.74	79.82	75.20	76.29	72.21	76.42
✓	✓	-	82.56	86.76	85.14	89.76	77.68	75.33	82.87
✓	✓	✓	86.29	92.5	83.43	90.96	83.88	81.31	86.40

Table 2: Ablation experiments on the Office-31 dataset with the evaluation metric of **OA** (%).

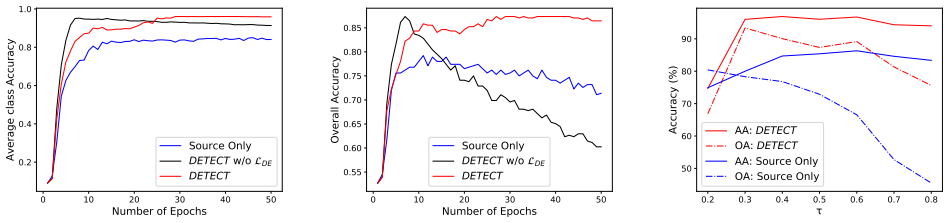
β, we set it to 0.5 in most cases and 1.0 to balance the target effect if the size of the target data is too small. We set γ to 0.5 and 0.01, respectively, for the small-scale datasets of Office-31 and ImageCLEF-DA and the large-scale datasets of VisDA2017 and ImageNet-Caltech. The threshold τ is fixed to 0.5 in all experiments, and this is investigated in Sec. 5.1. We adopt the photo domain images (containing 1670 natural images in total) in PACS [24] as the auxiliary dataset $\{(\boldsymbol{x}_i')\}_{i=1}^n$ in $\mathcal{R}_{OOD}$ for all our experiments. We implement our $DETECT$ based on PyTorch with a pre-trained ResNet-50 [16] as the backbone network. Other implementation details are the same as those in [47].

5.1 Ablation Studies and Analysis

Ablation Studies We conduct ablation studies on the Office-31 dataset to investigate the effects of components (i.e. the deep discriminative clustering object $\mathcal{L}_{CT}$ (3), neighborhood-preserved embedding term $\mathcal{L}_{DE}$ (4) and auxiliary regularizer $\mathcal{R}_{OOD}$ (6)) in detail. The results of AA and OA are illustrated in Tab. 1 and Tab. 2, respectively. We start with the baseline model trained with source data only, which is termed "Source Only". The AA increases and the OA decreases when we add $\mathcal{L}_{CT}$ to the baseline of Source Only, indicating that the objective of deep discriminative clustering benefits the recognition of target samples belonging to $\mathcal{Y}^{st}$ and degrades the detection of samples of the unknown categories (i.e. $\mathcal{Y}^{t/s}$). By taking $\mathcal{L}_{DE}$ into the overall objective, we improve the results significantly, especially the results of OA, verifying the importance and necessity of the neighborhood-preserved embedding objective in the clustering procedure, which is intuitively illustrated in Fig. 1(b) and Fig. 2. Specifically, the objective of $\mathcal{L}_{DE}$ successfully maintains the intrinsic discrimination of target data and prevents the target samples of unknown categories $\mathcal{Y}^{t/s}$ from aligning to the domain-shared samples of $\mathcal{Y}^{st}$ in the learned feature space. This contributes to category recognition, especially for unknown categories. The auxiliary regularizer $\mathcal{R}_{OOD}$ further improves the results of both OA and AA, leading to the best performance.

Convergence Performance We illustrate the convergence performance of our $DETECT$ with the evaluation metrics of AA and OA in Fig. 1(a) and Fig. 1(b), respectively. Compared to AA, the results of OA are more sensitive to the accuracy of target unknown categories, since samples of target unknown categories construct a large portion of the target domain in popular datasets. The OA of the Source Only baseline degrades significantly as the training proceeds, since the network becomes over-confident of its output, and therefore, increasingly more target samples of unknown classes are misclassified as domain shared categories. Our $DETECT$ without $\mathcal{L}_{DE}$ converges very fast and reaches a higher AA depending on deep discriminative clustering, but it converges to a lower OA than Source Only because the intrinsic target discrimination of unknown categories is contaminated. After including $\mathcal{L}_{DE}$, the proposed $DETECT$ significantly improves the convergence performance and converges stably to the highest accuracy levels for both AA and OA, certifying that it can be trained efficiently to learn the intrinsic discrimination of the entire target data.

Feature Visualization We visualize in Figs. 2(a)-2(d) the feature representations extracted by Source Only, UAN [47] and $DETECT$ without (w/o) $\mathcal{L}_{DE}$, as well as $DETECT$ on the A to W task of the Office-31 dataset with t-SNE [30]. Compared to Source Only and UAN [47], our $DETECT$ learns more discriminative representations for the target samples belonging to $\mathcal{Y}^{st}$, and

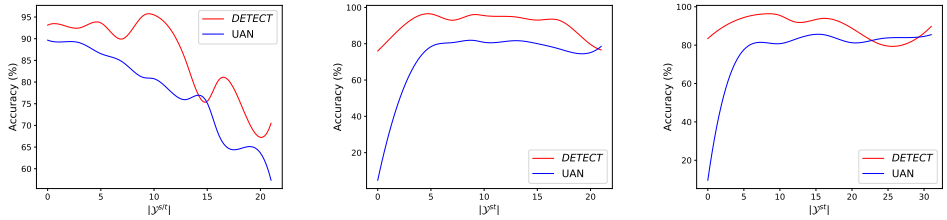


(a) (b) (c)

Figure 1: (a)-(b): Convergence performance on the task of A to D according to the evaluation metrics of AA and OA, respectively. (c): Performance w.r.t. the threshold τ in the task of A to D.

(a) Source Only (b) UAN [47] (c) $DETECT$ w/o $\mathcal{L}_{DE}$ (d) $DETECT$

Figure 2: t-SNE visualization of target features in the task of A to W. Red dots represent target samples that enjoy shared labels with the source domain (i.e. $\mathcal{Y}^{st}$) while black dots are samples from unknown classes (i.e. $\mathcal{Y}^{t/s}$). (Best viewed in color)



(a) $|\mathcal{Y}^{st}| = 10$ (b) $|\mathcal{Y}^{s/t}| = 10$ (c) $|\mathcal{Y}^{t}| = |\mathcal{Y}^{s}| + 1$

Figure 3: Average class accuracy (%) of different ξ-UDA settings on the task of A to W. The results of UAN [47] are reproduced with its public code.

it distinguishes the target samples of $\mathcal{Y}^{st}$ clearly from those of $\mathcal{Y}^{t/s}$. In addition, we can observe from the Figs. 2(c) and 2(d) that the introduced neighborhood-preserved embedding objective $\mathcal{L}_{DE}$ contributes to distinguishing between target samples of $\mathcal{Y}^{st}$ and $\mathcal{Y}^{t/s}$ in the feature space, verifying its efficacy.

Threshold Sensitivity We explore the sensitivity of our $DETECT$ to the threshold τ on the task of A to D on the Office-31 dataset. As illustrated in Fig. 1(c), the results of our method are stable under a wide range of τ, i.e. $\tau \in [0.3, 0.6]$, and they show consistent improvement over the results of Source Only, justifying the efficacy and robustness of our proposed method.

Analysis on Different Settings of ξ-UDA We compare the results of $DETECT$ and UAN [47] under different degrees of label space overlap between source and target domains on the A to W task of the Office-31 dataset, as illustrated in Fig. 3. First, given a fixed $|\mathcal{Y}^{st}| = 10$ and ξ, we compare results of different methods on the ξ-UDA with various sizes of source private classes, i.e. $|\mathcal{Y}^{s/t}|$, where $|\mathcal{Y}^{t/s}|$ changes correspondingly. As illustrated in Fig. 3(a), our $DETECT$ outperforms UAN [47] on most of the sizes of $|\mathcal{Y}^{s/t}|$. In Figs. 3(b) and 3(c), we evaluate the

Method	Office-31							ImageNet-Caltech		VisDA
	A2W	D2W	W2D	A2D	D2A	W2A	Avg.	Im2Cal	Cal2Im	
Source Only [47]	75.94	89.60	90.91	80.45	78.83	81.42	82.86	70.28	65.14	52.80
DANN[11]	80.65	80.94	88.07	82.67	74.82	83.54	81.78	71.37	66.54	52.94
RTN[29]	85.70	87.80	88.91	82.69	74.64	83.26	84.18	71.94	66.15	53.92
IWAN[48]	85.25	90.09	90.00	84.27	84.22	86.25	86.68	72.19	66.48	58.72
PADA[7]	85.37	79.26	90.91	81.68	55.32	82.61	79.19	65.47	58.73	44.98
ATI[4]	79.38	92.60	90.08	84.40	78.85	81.57	84.48	71.59	67.36	54.81
OSBP[38]	66.13	73.57	85.62	72.92	47.35	60.48	67.68	62.08	55.48	30.26
UAN[47]	85.62	94.77	**97.99**	86.50	85.45	85.12	89.24	75.28	70.17	60.83
Method in [26]	90.25	95.25	96.96	88.84	90.19	89.30	91.80	76.13	74.67	64.31
DANCE [35]	92.8	**97.8**	97.7	91.6	**92.2**	91.4	93.9	-	-	69.2
$DETECT$	**94.43**	96.87	96.31	**94.69**	91.08	**92.06**	**94.24**	**78.52**	**76.81**	**71.38**

Table 3: Average class accuracy (%) on datasets of Office-31, ImageNet-Caltech, and VisDA2017.

	C2I	C2P	C2B	I2C	I2P	I2B	P2C	P2I	P2B	B2C	B2I	B2P	Avg.
Source Only	81.43	72.21	58.86	87.90	76.90	58.38	85.33	80.19	55.05	87.90	77.14	67.72	74.08
UAN [47]	79.81	70.38	57.33	84.00	72.97	58.00	81.33	77.52	54.10	80.77	72.57	64.27	71.09
$DETECT$	**89.05**	**75.34**	**61.24**	**92.00**	**78.20**	**60.57**	**90.38**	**86.10**	**60.67**	**94.00**	**88.00**	**74.21**	**79.15**

Table 4: Average class accuracy (%) on the ImageCLEF-DA dataset. The results of UAN [47] are reproduced with its public code.

results by changing the size of $|\mathcal{Y}^{st}|$ with the fixed conditions of $|\mathcal{Y}^{s/t}| = 10$ and $|\mathcal{Y}^{t}| = |\mathcal{Y}^{s}| + 1$, respectively. In most cases, our method outperforms UAN significantly, certifying the efficacy of our proposed $DETECT$. When $|\mathcal{Y}^{st}| = 0$, i.e. no classes are shared across domains, our $DETECT$ improves over UAN by a large margin, justifying the robustness of our $DETECT$ in this extreme case. Note that comparisons between our $DETECT$ and UAN on existing closed set, partial and open set UDA are included here. Specifically, the partial UDA [6] is met when $|\mathcal{Y}^{s/t}| = 21$ in Fig. 3(a) and $|\mathcal{Y}^{st}| = 21$ in Fig. 3(b). The open set UDA [38] is met when $|\mathcal{Y}^{s/t}| = 0$ in Fig. 3(a), and the closed set domain adaptation is a special case when $|\mathcal{Y}^{st}| = 31$ in Fig. 3(c).

5.2 Results

We compare our $DETECT$ with state-of-the-art UDA methods for the task of ξ-UDA on the datasets of Office-31, ImageNet-Caltech, VisDA2017 and the ImageCLEF dataset, as illustrated in Tab. 3 and Tab. 4, respectively. Results of other methods are quoted from [47], [26] and [35]. As illustrated, the methods proposed for the closed set UDA (DANN [11] and RTN [29]), the partial UDA (IWAN[48] and PADA[7]) and the open set UDA (ATI[4] and OSBP[38]) do not perform well on the task of ξ-UDA, since their prior assumptions about the label space are violated. Some of these methods, such as the PADA [7] and OSBP [38], present worse results than the baseline of Source Only, indicating the presence of negative transfer. This phenomenon also occurs in the method of UAN when evaluating on the ImageCLEF-DA dataset, as illustrated in Tab. 4. Our proposed $DETECT$ improves over the baseline of Source Only consistently, and it achieves better results than the specific methods of ξ-UDA [47, 26, 35] in all four datasets and most of the tasks, certifying its efficacy.

6 Conclusion

In this paper, we contribute a novel perspective to unsupervised and universal domain adaptation (ξ-UDA). We argue that no matter what the degrees of the label space overlap are, the problem nature of UDA remains unchanged in two principles: (1) to learn the intrinsic discrimination of target data in an unsupervised manner, regularized by the labeled discrimination of source data and (2) to reduce contamination of the intrinsic target discrimination during the source regularization. Based on that, we provide the simple but strong baseline of $DETECT$ for ξ-UDA. Experiments show that such a simple baseline can work effectively across a range of label space overlaps, and it outperforms all existing methods on four classic domain adaptation benchmarks.

Broader Impact

When applying an object recognition system in the wild, there usually exist both domain and category gaps between the training and target images. Our proposed method can seamlessly apply to this situation. Moreover, by preserving the intrinsic structure of target discrimination during the source regularization, our proposed method can better detect instances of unknown categories and predict more accurately the instances of training categories. This technology can be positively used in many applications. For example, it can be used to automatically label the target pictures as an unknown label or a specific label from the training data. When we found that if the recognition system labels most of the targets with unknown, then it can guide us to provide more labelling information of such data into the system. However, negative impact may arise if the technology is abused. We encourage future work to mitigate the risks arising from, e.g., medical recognition applications. Also, we expect that the policymakers to take active actions to penalize the misuse of such technologies.

References

[1] Imageclef-da dataset. `http://imageclef.org/2014/adaptation/`.

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

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144, 2007.

[4] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763, 2017.

[5] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1548–1560, 2011.

[6] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *CVPR*, pages 2724–2732, 2018.

[7] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *ECCV*, volume 11212, pages 139–155, 2018.

[8] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *CVPR*, pages 2985–2994, 2019.

[9] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *ICCV*, pages 9943–9952, 2019.

[10] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *ICCV*, pages 5747–5756, 2017.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Fran*cc*ois Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[12] Ryan Gomes, Andreas Krause, and Pietro Perona. Discriminative clustering by regularized information maximization. In *NIPS*, pages 775–783, 2010.

[13] Greg Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[14] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, pages 8400–8408, 2019.

[15] Philip Häusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *ICCV*, pages 2784–2792, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[17] Xiaofei He, Deng Cai, Shuicheng Yan, and HongJiang Zhang. Neighborhood preserving embedding. In *ICCV*, pages 1208–1213, 2005.

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

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

[20] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, volume 70, pages 1558–1567, 2017.

[21] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CPVR*, pages 4893–4902, 2019.

[22] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, pages 10285–10295, 2019.

[23] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5543–5551. IEEE Computer Society, 2017.

[25] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[26] Omri Lifshitz and Lior Wolf. A sample selection approach for universal domain adaptation. *CoRR*, abs/2001.05071, 2020.

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, volume 37, pages 97–105, 2015.

[28] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, pages 2200–2207, 2013.

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, pages 136–144, 2016.

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

[31] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

[32] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[34] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[35] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *CoRR*, abs/2002.07953, 2020.

[36] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, volume 70, pages 2988–2997, 2017.

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018.

[38] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, pages 153–168, 2018.

[39] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, pages 4058–4065. AAAI Press, 2018.

[40] Rui Shu, Hung H. Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T approach to unsupervised domain adaptation. In *ICLR*, 2018.

[41] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *ICLR*, 2016.

[42] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. *CoRR*, abs/2003.08607, 2020.

[43] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017.

[44] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[45] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, volume 48, pages 478–487, 2016.

[46] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, volume 80, pages 5419–5428, 2018.

[47] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *CVPR*, pages 2720–2729, 2019.

[48] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *CVPR*, pages 8156–8164, 2018.

[49] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, pages 5031–5040, 2019.