Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering

Hui Tang, Ke Chen, and Kui Jia*
South China University of Technology
381 Wushan Road, Tianhe District, Guangzhou, Guangdong, China
eehuitang@mail.scut.edu.cn, {chenk,kuijia}@scut.edu.cn

Abstract

Unsupervised domain adaptation (UDA) is to make predictions for unlabeled data on a target domain, given labeled data on a source domain whose distribution shifts from the target one. Mainstream UDA methods learn aligned features between the two domains, such that a classifier trained on the source features can be readily applied to the target ones. However, such a transferring strategy has a potential risk of damaging the intrinsic *discrimination of target data. To alleviate this risk, we are motivated by the assumption of structural domain similarity, and propose to directly uncover the intrinsic target discrimination via discriminative clustering of target data. We constrain the clustering solutions using structural source regularization that hinges on our assumed structural domain similarity. Technically, we use a flexible framework of deep network based discriminative clustering that minimizes the KL divergence between predictive label distribution of the network and an introduced auxiliary one; replacing the auxiliary distribution with that formed by ground-truth labels of source data implements the structural source regularization via a simple strategy of joint network training. We term our proposed method as* Structurally Regularized Deep Clustering (SRDC)*, where we also enhance target discrimination with clustering of intermediate network features, and enhance structural regularization with soft selection of less divergent source examples. Careful ablation studies show the efficacy of our proposed SRDC. Notably, with no* explicit *domain alignment, SRDC outperforms all existing methods on three UDA benchmarks.*

1. Introduction

Given labeled data on a source domain, unsupervised domain adaptation (UDA) is to make predictions in the same label space for unlabeled data on a target domain, where there may exist divergence between the two domains. Mainstream methods are motivated by the classic UDA theories [2, 3, 40] that specify the learning bounds involving domain divergences, whose magnitudes depend on the feature space and the hypothesis space of classifier. Consequently, these methods (e.g., those recent ones based on adversarial training of deep networks [16, 48]) strive to learn aligned features between the two domains, such that classifiers trained on the source features can be readily applied to the target ones. In spite of impressive results achieved by these methods, they have a potential risk of damaging the *intrinsic* structures of target data discrimination, as discussed in [9, 50, 69]. Attempts are made in [9, 50] to alleviate this risk, however, explicit domain alignments are still pursued in their proposed solutions.

To address this issue, we first instantiate the general assumption of domain closeness in UDA problems [2, 50] as *structural domain similarity*, which spells as two notions of *domain-wise discrimination* and *class-wise closeness* — the former notion assumes the existence of intrinsic structures of discriminative data clusters in individual domains, and the later one assumes that clusters of the two domains corresponding to the same class label are geometrically close. This assumption motivates us to consider a UDA approach that directly uncovers the intrinsic data discrimination via discriminative clustering of target data, where we propose to constrain the clustering solutions using structural source regularization hinging on our assumed structural similarity.

Among various deep network based clustering algorithms [4, 8, 14, 61], we choose a simple but flexible non-generative framework [14], which performs discriminative clustering by minimizing the KL divergence between predictive label distribution of the network and an introduced auxiliary one. Structural source regularization is simply achieved via a simple strategy of joint network training, by replacing the auxiliary distribution with that formed by ground-truth labels of source data. We term our proposed method as *Structurally Regularized Deep Clustering (SRDC)*. In SRDC, we also enhance target discrimination with clustering of intermediate network features, and enhance structural regularization with soft selection of less

*Corresponding author.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00875

divergent source examples. We note that quite a few recent UDA methods [13, 27, 41, 51] consider clustering of target data as well; however, they still do explicit feature alignment between the two domains via alignment of cluster centers/samples, thus prone to the aforementioned risk of damaged intrinsic target discrimination. Experiments on benchmark UDA datasets show the efficacy of our proposed SRDC. We finally summarize our contributions as follows.

- To address a potential issue of damaging the *intrinsic* data discrimination by explicitly learning domain-aligned features, we propose in this work a source-regularized, deep discriminative clustering method in order to directly uncover the intrinsic discrimination among target data. The method is motivated by our assumption of structural similarity between the two domains, for which we term the proposed method as *Structurally Regularized Deep Clustering (SRDC)*.
- To technically achieve SRDC, we use a flexible deep clustering framework that first introduces an auxiliary distribution, and then minimizes the KL divergence between the introduced one and the predictive label distribution of the network; replacing the auxiliary distribution with that of ground-truth labels of source data implements the structural source regularization via a simple strategy of joint network training. In SRDC, we also design useful ingredients to enhance target discrimination with clustering of intermediate network features, and to enhance structural regularization with soft selection of less divergent source examples.
- We conduct careful ablation studies on benchmark UDA datasets, which verify the efficacy of individual components proposed in SRDC. Notably, with no *explicit* domain alignment, our proposed SRDC outperforms all existing methods on the benchmark datasets.

2. Related works

Alignment based domain adaptation. A typical line of works [16, 43, 53, 63] leverages a domain-adversarial task to align the source and target domains as a whole so that class labels can be transferred from the source domain to the unlabeled target one. Another typical line of works directly minimizes the domain shift measured by various metrics, e.g., maximum mean discrepancy (MMD) [34, 36, 37]. These methods are based on domain-level domain alignment. To achieve class-level domain alignment, the works of [35, 42] utilize the multiplicative interaction of feature representations and class predictions so that the domain discriminator can be aware of the classification boundary. Based on the integrated task and domain classifier, [52] encourages a mutually inhibitory relation between category and domain predictions for any input instance. The works of [7, 13, 41, 59] align the labeled source centroid and pseudo-labeled target centroid of each shared class in the feature space. Some works [31, 47, 48] use individual task classifiers for the two domains to detect non-discriminative features and reversely learn a discriminative feature extractor. Some works [30, 56, 57] focus attention on transferable regions to derive a domain-invariant classification model. To help achieve target-discriminative features, [28, 49] generate synthetic images from the raw input data of the two domains via GANs [19]. The recent work of [9] improves adversarial feature adaptation, where the discriminative structures of target data may be deteriorated [69]. The work of [60] adapts the feature norms of the two domains to a large range of values so that the learned features are not only task-discriminative but also domain-invariant.

Clustering based domain adaptation. The cluster assumption states that the classification boundary should not pass through high-density regions, but instead lie in low-density regions [6]. To enforce the cluster assumption, conditional entropy minimization [20, 32] is widely used in the UDA community [11, 44, 45, 50, 51, 60, 64, 68]. The work of [27] adopts the spherical K-means to assign target labels. The recent work of [13] employs a Fisher-like criterion based deep clustering loss [38]. However, they use target clustering just as an incremental technique to improve explicit feature alignment. The previous work of [50] is based on the clustering criterion of mutual information maximization, which still explicitly forces domain alignment. In contrast, with no explicit domain alignment, SRDC aims to uncover the intrinsic target discrimination by discriminative target clustering with structural source regularization.

Latent domain discovery. Methods of latent domain discovery [10, 18, 22, 39] focus on capturing latent structures of the source, target data or a mixed one under the assumption that data may practically comprise multiple diverse distributions. Our proposed SRDC shares the same motivation with these methods, but differs in the aim to uncover the intrinsic discrimination among target classes by structurally source regularized deep discriminative target clustering, in a distinctive perspective of utilizing structural similarity between the source and target domains.

3. The strategies of transferring *versus* uncovering the intrinsic target discrimination

Consider a source domain $\mathcal{S}$ with n_s labeled examples $\{(\boldsymbol{x}_j^s, y_j^s)\}_{j=1}^{n_s}$, and a target domain $\mathcal{T}$ with n_t unlabeled examples $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$. Unsupervised domain adaptation (UDA) assumes a shared label space $\mathcal{Y}$ between $\mathcal{S}$ and $\mathcal{T}$. Let $|\mathcal{Y}| = K$ and we have $y^s \in \{1, 2, \ldots, K\}$ for any source instance $\boldsymbol{x}^s$. The objective of *transductive UDA* is to predict $\{\hat{y}_i^t\}_{i=1}^{n_t}$ of $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ by learning a feature embedding function $\varphi : \mathcal{X} \to \mathcal{Z}$ that lifts any input instance $\boldsymbol{x} \in \mathcal{X}$ to

the feature space $\mathcal{Z}$, and a classifier $f : \mathcal{Z} \to \mathbb{R}^K$. Subtly different from transductive UDA, *inductive UDA* is to measure performance of the learned $\varphi(\cdot)$ and $f(\cdot)$ on held-out instances sampled from the same $\mathcal{T}$. This subtle difference is in fact important since we expect to use the learned $\varphi(\cdot)$ and $f(\cdot)$ as off-the-shelf models, and we expect them to be consistent when learning with different source domains.

Domain closeness is generally assumed in UDA either theoretically [2, 40] or intuitively [50]. In this work, we summarize the assumptions in [50] as the *structural similarity* between the source and target domains, which include the following notions of domain-wise discrimination and class-wise closeness, as illustrated in Figure 1.

- *Domain-wise discrimination* assumes that there exist intrinsic structures of data discrimination in individual domains, i.e., data in either source or target domains are discriminatively clustered corresponding to the shared label space.
- *Class-wise closeness* assumes that clusters of the two domains corresponding to the same class label are geometrically close.

Based on these assumptions, many of exiting works [16, 35, 42, 48, 53, 66] take the *transferring* strategy of learning aligned feature representations between the two domains, such that classifiers trained on source features can be readily applied to the target ones. However, such a strategy has a potential risk of damaging the intrinsic data discrimination on the target domain, as discussed in recent works of [9, 50, 69]. An illustration of such damage is also given in Figure 1. We note that more importantly, classifiers adapting to the damaged discrimination of target data would be less effective for tasks of inductive UDA, since they deviate too much from the oracle target classifier, i.e. an ideal one trained on the target data with the ground-truth labels.

Based on the above analysis, we are motivated to directly *uncover* the intrinsic target discrimination via discriminative clustering of target data. To leverage the labeled source data, we propose to constrain the clustering solutions using *structural source regularization* that hinges on our assumed structural similarity across domains. Section 4 presents details of our method, with an illustration given in Figure 1. We note that quite a few recent methods [13, 27, 41, 51] consider clustering of target data as well; however, they still do explicit feature alignment across domains via alignment of cluster centers/samples, thus prone to the aforementioned risk of damaged intrinsic target discrimination.

4. Discriminative target clustering with structural source regularization

We parameterize the feature embedding function $\varphi(\cdot;\boldsymbol{\theta})$ and classifier $f(\cdot;\boldsymbol{\vartheta})$ as a deep network [21, 25, 26, 65],

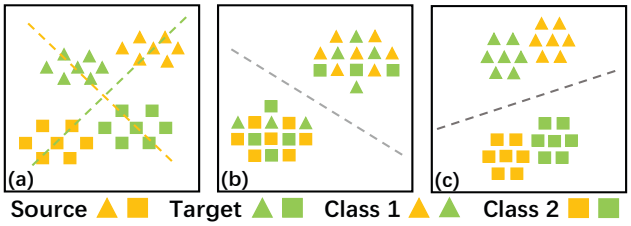


Figure 1. (Best viewed in color.) (a) Illustration of the assumption of structural domain similarity (cf. Section 3). The orange line denotes the classifier trained on the labeled source data and the green one denotes the classifier trained on the labeled target data, i.e. the oracle target classifier. (b) Illustration of damaging intrinsic structures of data discrimination on the target domain by the existing transferring strategy. The dashed line denotes the source classifier adapting to the damaged discrimination of target data, which has a sub-optimal generalization. (c) Illustration of our proposed uncovering strategy. Discriminative target clustering with structural source regularization uncovers intrinsic target discrimination.

where $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ collects the network parameters. We also write them as $\varphi(\cdot)$ and $f(\cdot)$ for simplicity, and use $f \circ \varphi$ to denote the whole network. For an input instance $\boldsymbol{x}$, the network computes feature representation $\boldsymbol{z} = \varphi(\boldsymbol{x})$, and outputs a probability vector $\boldsymbol{p} = \mathrm{softmax}(f(\boldsymbol{z})) \in [0,1]^K$ after the final softmax operation.

As discussed in Section 3, in order to uncover the intrinsic discrimination of the target domain, we opt for direct clustering of target instances with structural regularization from the source domain. Among various clustering methods [4, 8, 14, 61], we choose a flexible framework of deep discriminative clustering [14], which minimizes the KL divergence between predictive label distribution of the network and an introduced auxiliary one; by replacing the auxiliary distribution with that of ground-truth labels of source data, we easily implement the structural source regularization via a simple strategy of network joint training, for which we term our proposed method as Structurally Regularized Deep Clustering (SRDC). In SRDC, we also enhance target discrimination with clustering of intermediate network features, and enhance structural regularization with soft selection of less divergent source examples.

4.1. Deep discriminative target clustering

For the unlabeled target data $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$, the network predicts, after softmax operation, the probability vectors $\{\boldsymbol{p}_i^t\}_{i=1}^{n_t}$ that we collectively write as $\boldsymbol{P}^t$. We also write as $p_{i,k}^t$ the k^{th} element of $\boldsymbol{p}_i^t$ for the target instance $\boldsymbol{x}_i^t$. $\boldsymbol{P}^t$ thus approximates the predictive label distribution of the network for samples of $\mathcal{T}$. Similar to [14, 24], we first introduce an auxiliary counterpart $\boldsymbol{Q}^t$, and the proposed SRDC then alternates in (1) updating $\boldsymbol{Q}^t$, and (2) using the updated $\boldsymbol{Q}^t$ as labels to train the network to update parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, which optimizes the following objective of deep

discriminative clustering

$$\min_{\boldsymbol{Q}^t,\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}} \mathcal{L}^t_{f\circ\varphi} = \mathrm{KL}(\boldsymbol{Q}^t||\boldsymbol{P}^t) + \sum_{k=1}^{K} \varrho^t_k \log \varrho^t_k, \tag{1}$$

where $\varrho^t_k = \frac{1}{n_t}\sum_{i=1}^{n_t} q^t_{i,k}$ and the second term in (1) is used to balance cluster assignments in $\{\boldsymbol{q}^t_i\}_{i=1}^{n_t}$ — otherwise degenerate solutions would be obtained that merge clusters by removing cluster boundaries [29]. In addition, it encourages entropy maximization of the label distribution on the target domain, i.e., encouraging cluster *size* balance. In aware of the lack of prior knowledge about target label distribution, we simply rely on the second term to account for a uniform one. The first term computes the KL divergence between discrete probability distributions $\boldsymbol{P}^t$ and $\boldsymbol{Q}^t$ as

$$\mathrm{KL}(\boldsymbol{Q}^t||\boldsymbol{P}^t) = \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{k=1}^{K} q^t_{i,k} \log \frac{q^t_{i,k}}{p^t_{i,k}}.$$

More specifically, the optimization of objective (1) takes the following alternating steps.

- **Auxiliary distribution update.** Fix network parameters $\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}$ (and $\{\boldsymbol{p}^t_i\}_{i=1}^{n_t}$ of target instances are fixed as well). By setting the approximate gradient of (1) as zero, we has the following closed-form solution [14]

$$q^t_{i,k} = \frac{p^t_{i,k}/(\sum_{i'=1}^{n_t} p^t_{i',k})^{\frac{1}{2}}}{\sum_{k'=1}^{K} p^t_{i,k'}/(\sum_{i'=1}^{n_t} p^t_{i',k'})^{\frac{1}{2}}}. \tag{2}$$

- **Network update.** By fixing $\boldsymbol{Q}^t$, this step is equivalent to training the network via a cross-entropy loss using $\boldsymbol{Q}^t$ as labels, giving rise to

$$\min_{\boldsymbol{\theta},\boldsymbol{\vartheta}} -\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{k=1}^{K} q^t_{i,k} \log p^t_{i,k}. \tag{3}$$

In this work, we also enhance uncovering of target discrimination via discriminative clustering in the feature space $\mathcal{Z}$. More specifically, let $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ be the learnable cluster centers of both the source and target data in the space $\mathcal{Z}$. We follow [58] and define a probability vector $\widetilde{\boldsymbol{p}}^t_i$ of soft cluster assignments of the instance feature $\boldsymbol{z}^t_i = \varphi(\boldsymbol{x}^t_i)$ based on instance-to-center distances in the space $\mathcal{Z}$, whose k^{th} element is defined as

$$\widetilde{p}^t_{i,k} = \frac{\exp((1+||\boldsymbol{z}^t_i - \boldsymbol{\mu}_k||^2)^{-1})}{\sum_{k'=1}^{K}\exp((1+||\boldsymbol{z}^t_i - \boldsymbol{\mu}_{k'}||^2)^{-1})}. \tag{4}$$

We write $\{\widetilde{\boldsymbol{p}}^t_i\}_{i=1}^{n_t}$ collectively as $\widetilde{\boldsymbol{P}}^t$. By introducing a corresponding auxiliary distribution $\widetilde{\boldsymbol{Q}}^t$, we have the following objective of deep discriminative clustering in the space $\mathcal{Z}$

$$\min_{\widetilde{\boldsymbol{Q}}^t,\boldsymbol{\theta},\{\boldsymbol{\mu}^t_k\}_{k=1}^{K}} \mathcal{L}^t_{\varphi} = \mathrm{KL}(\widetilde{\boldsymbol{Q}}^t||\widetilde{\boldsymbol{P}}^t) + \sum_{k=1}^{K} \widetilde{\varrho}^t_k \log \widetilde{\varrho}^t_k, \tag{5}$$

where $\widetilde{\varrho}^t_k = \frac{1}{n_t}\sum_{i=1}^{n_t} \widetilde{q}^t_{i,k}$. The objective (5) can be optimized in the same alternating fashion as for (1), by deriving formulations similar to (2) and (3), where we note that features $\{\boldsymbol{z}^t_i\}_{i=1}^{n_t}$ are computed with the updated network parameters $\boldsymbol{\theta}$, and we also re-initialize $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ at the start of each training epoch based on the current cluster assignments of $\{\boldsymbol{z}^t_i\}_{i=1}^{n_t}$ (together with labeled source $\{\boldsymbol{z}^s_j\}_{j=1}^{n_s}$). $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ are continuously updated during training iterations of each epoch via back-propagated gradients of (5).

Combining (1) and (5) gives our objective of deep discriminative target clustering, which will be used as the first term of our overall objective of SRDC algorithm

$$\min_{\boldsymbol{Q}^t,\widetilde{\boldsymbol{Q}}^t,\{\boldsymbol{\theta},\boldsymbol{\vartheta}\},\{\boldsymbol{\mu}_k\}_{k=1}^{K}} \mathcal{L}^t_{\mathrm{SRDC}} = \mathcal{L}^t_{f\circ\varphi} + \mathcal{L}^t_{\varphi}. \tag{6}$$

Remarks. Given unlabeled target data alone, the objective (1) itself is not guaranteed to has sensible solutions to uncover the intrinsic discrimination of target data, since the auxiliary distribution $\boldsymbol{Q}^t$ could be arbitrary whose optimization is subject to no proper constraints. Incorporation of (5) into the overall objective (6) would alleviate the issue by soft assignments of $\{\boldsymbol{z}^t_i\}_{i=1}^{n_t}$ to properly initialized cluster centers $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$. To guarantee sensible solutions, deep clustering methods [14, 58] usually employ an additional reconstruction loss as a data-dependent regularizer. In our proposed SRDC for domain adaptation, the following introduced structural source regularization serves a similar purpose as that of the reconstruction ones used in [14, 58].

4.2. Structural source regularization

Based on the UDA assumption made in Section 3 that specifies the structural similarity between the source and target domains, we propose to transfer the global, discriminative structure of labeled source data via a simple strategy of jointly training the same network $f\circ\varphi$. Note that the K-way classifier f defines hyperplanes that partition the feature space $\mathcal{Z}$ into regions, of which K ones are uniquely responsible for the K classes. Since the two domains share the same label space, joint training would *ideally* push instances of the two domains from same classes into same regions in $\mathcal{Z}$, thus *implicitly* achieving feature alignment between the two domains. Figure 1 gives an illustration.

Technically, for the labeled source data $\{(\boldsymbol{x}^s_j, y^s_j)\}_{j=1}^{n_s}$, we simply replace the auxiliary distribution in (1) with that formed by the ground-truth labels $\{y^s_j\}_{j=1}^{n_s}$, resulting in a supervised network training via cross-entropy minimization

$$\min_{\boldsymbol{\theta},\boldsymbol{\vartheta}} \mathcal{L}^s_{f\circ\varphi} = -\frac{1}{n_s}\sum_{j=1}^{n_s}\sum_{k=1}^{K} \mathrm{I}[k = y^s_j] \log p^s_{j,k}, \tag{7}$$

where $p^s_{j,k}$ is the k^{th} element of the predictive probability vector $\boldsymbol{p}^s_j$ of source instance $\boldsymbol{x}^s_j$, and $\mathrm{I}[\cdot]$ is the function of

indicator. We also enhance source discrimination in the feature space $\mathcal{Z}$, in parallel with (5), resulting in

$$\min_{\boldsymbol{\theta},\{\boldsymbol{\mu}_k\}_{k=1}^K} \mathcal{L}_{\varphi}^s = -\frac{1}{n_s}\sum_{j=1}^{n_s}\sum_{k=1}^{K} \mathrm{I}[k=y_j^s]\log \widetilde{p}_{j,k}^s, \tag{8}$$

where

$$\widetilde{p}_{j,k}^s = \frac{\exp((1+||\boldsymbol{z}_j^s-\boldsymbol{\mu}_k||^2)^{-1})}{\sum_{k'=1}^K \exp((1+||\boldsymbol{z}_j^s-\boldsymbol{\mu}_{k'}||^2)^{-1})}. \tag{9}$$

Combining (7) and (8) gives the training objective using labeled source data

$$\min_{\{\boldsymbol{\theta},\boldsymbol{\vartheta}\},\{\boldsymbol{\mu}_k\}_{k=1}^K} \mathcal{L}_{\text{SRDC}}^s = \mathcal{L}_{f\circ\varphi}^s + \mathcal{L}_{\varphi}^s. \tag{10}$$

Using (10) as the structural source regularizer, we have our final objective of SRDC algorithm

$$\min_{\boldsymbol{Q}^t,\widetilde{\boldsymbol{Q}}^t,\{\boldsymbol{\theta},\boldsymbol{\vartheta}\},\{\boldsymbol{\mu}_k\}_{k=1}^K} \mathcal{L}_{\text{SRDC}} = \mathcal{L}_{\text{SRDC}}^t + \lambda\mathcal{L}_{\text{SRDC}}^s, \tag{11}$$

where λ is a penalty parameter.

4.3. Enhancement via soft source sample selection

It is commonly hypothesized in transfer learning [23, 62] that importance of source samples varies for learning transferable models. A simple strategy to implement this hypothesis is to re-weight source instances based on their similarities to target ones [7, 17, 67]. In this work, we also employ this strategy into SRDC.

Specifically, let $\{\boldsymbol{c}_k^t \in \mathcal{Z}\}_{k=1}^K$ be the K target cluster centers in the feature space. For any labeled source example $(\boldsymbol{x}^s, y^s)$, we compute its similarity to $\boldsymbol{c}_{y^s}^t$, i.e., the target center of cluster y^s, based on the following cosine distance

$$w^s(\boldsymbol{x}^s) = \frac{1}{2}\left(1+\frac{\boldsymbol{c}_{y^s}^{t\top}\boldsymbol{x}^s}{||\boldsymbol{c}_{y^s}^t||\,||\boldsymbol{x}^s||}\right) \in [0,1]. \tag{12}$$

We compute $\{\boldsymbol{c}_k^t\}_{k=1}^K$ once every epoch during network training. Note that $\{\boldsymbol{c}_k^t\}_{k=1}^K$ are different from $\{\boldsymbol{\mu}_k\}_{k=1}^K$ in (4) and (9), which are cluster centers of both the source and target data that are continuously updated during training iterations of each epoch. We compute weights for all $\{(\boldsymbol{x}_j^s, y_j^s)\}_{j=1}^{n_s}$ using (12), and enhance (7) and (8) using the following weighted version of objectives

$$\mathcal{L}_{f\circ\varphi(\cdot;\{w_j^s\}_{j=1}^{n_s})}^s = -\frac{1}{n_s}\sum_{j=1}^{n_s} w_j^s \sum_{k=1}^{K} \mathrm{I}[k=y_j^s]\log p_{j,k}^s, \tag{13}$$

$$\mathcal{L}_{\varphi(\cdot;\{w_j^s\}_{j=1}^{n_s})}^s = -\frac{1}{n_s}\sum_{j=1}^{n_s} w_j^s \sum_{k=1}^{K} \mathrm{I}[k=y_j^s]\log \widetilde{p}_{j,k}^s. \tag{14}$$

Experiments in Section 5 show that SRDC based on the above weighted objectives achieves improved results.

5. Experiments

5.1. Setups

Office-31 [46] is the most popular real-world benchmark dataset for visual domain adaptation, which contains $4,110$ images of 31 classes shared by three distinct domains: Amazon (**A**), Webcam (**W**), and DSLR (**D**). We evaluate all methods on all the six transfer tasks.

ImageCLEF-DA [1] is a benchmark dataset with 12 classes shared by three domains: Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**), and Pascal VOC 2012 (**P**). There are 50 images in each class and 600 images in each domain. We evaluate all methods on all the six transfer tasks.

Office-Home [55] is a more challenging benchmark dataset, with $15,500$ images of 65 classes shared by four extremely distinct domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**), and Real-World images (**Rw**). We evaluate all methods on all the twelve transfer tasks.

Implementation details. We follow the standard protocol for UDA [16, 33, 35, 48, 60] to use all labeled source samples and all unlabeled target samples as the training data. For each transfer task, we use center-crop target domain images for reporting results and report the classification result of $\mathrm{mean}(\pm\mathrm{std})$ over three random trials. We use the ImageNet [12] pre-trained ResNet-50 [21] as the base network, where the last FC layer is replaced with the task-specific FC layer(s) to parameterize the classifier $f(\cdot)$. We implement our experiments in PyTorch. We fine-tune from the pre-trained layers and train the newly added layer(s), where the learning rate of the latter is 10 times that of the former. We adopt mini-batch SGD with the learning rate schedule as [16]: the learning rate is adjusted by $\eta_p = \eta_0(1+\alpha p)^{-\beta}$, where p is the process of training epochs normalized to be in $[0,1]$, and $\eta_0 = 0.001, \alpha = 10, \beta = 0.75$. We follow [16] to increase λ from 0 to 1 by $\lambda_p = 2(1+\exp(-\gamma p))^{-1} - 1$, where $\gamma = 10$. The other implementation details are provided in the supplemtary material. The code is available at `https://github.com/huitangtang/SRDC-CVPR2020`.

5.2. Ablation studies and analysis

Ablation study. To investigate the effects of individual components of our proposed SRDC, we conduct ablation studies using Office-31 based on ResNet-50 by evaluating several variants of SRDC: (1) **Source Model**, which fine-tunes the base network on labeled source samples; (2) **SRDC (w/o structural source regularization)**, which fine-tunes a source pre-trained model using (6), i.e. without structural source regularization; (3) **SRDC (w/o feature discrimination)**, which denotes training without source and target discrimination in the feature space $\mathcal{Z}$; (4) **SRDC (w/o soft source sample selection)**, which denotes training without enhancement via soft source sample selection. The re-

Method	A →W	A →D	D →A	W →A	Avg
Source Model	77.8±0.2	82.1±0.2	64.5±0.2	66.1±0.2	72.6
SRDC (w/o structural source regularization)	87.3±0.0	92.1±0.1	73.9±0.1	75.0±0.1	82.1
SRDC (w/o feature discrimination)	94.2±0.4	94.3±0.4	74.3±0.2	75.5±0.4	84.6
SRDC (w/o soft source sample selection)	94.8±0.2	94.6±0.3	74.6±0.3	75.7±0.3	84.9
SRDC	**95.7**±0.2	**95.8**±0.2	**76.7**±0.3	**77.1**±0.1	**86.3**

Table 1. Ablation studies using Office-31 based on ResNet-50. Please refer to the main text for how different methods are defined.

Figure 2. The images on the left are randomly sampled from the target domain **A** and those on the right are the top-ranked (the 3^{rd} column) and bottom-ranked (the 4^{th} column) samples from the source domain **W** for three classes. Note that the red numbers are the source weights computed by (12).

Method	A →W	A →D	D →A	W →A	Avg
Source Model	79.3	81.6	63.1	65.7	72.4
DANN [16]	80.8	82.4	66.0	64.6	73.5
MCD [48]	86.5	86.7	72.4	70.9	79.1
SRDC	**91.9**	**91.6**	**75.6**	**75.7**	**83.7**
Oracle Model	98.8	97.6	87.8	87.8	93.0

Table 2. Comparative experiments under inductive UDA setting.

sults are reported in Table 1. We can observe that when any one of our designed components is removed, the performance degrades, verifying that (1) both feature discrimination and structural source regularization are effective for improving target clustering; (2) the proposed soft source sample selection scheme leads to better regularization.

Source refinement. To affirm that our proposed soft source sample selection scheme can select more transferable source samples, we show the images randomly sampled from the target domain **A**, and the top-ranked and bottom-ranked samples from the source domain **W** in Figure 2. Here, the red numbers are the source weights computed by (12). We can observe that (1) the lowest weight is more than 0.5, which is reasonable since all source samples are related to the target domain in that the two domains share the same label space; (2) the highest weight is less than 1, which is reasonable since there exists distribution shift between the two domains; (3) the source images with a canonical viewpoint have the higher weights than those with top-down, bottom-up, and side viewpoints, which is intuitive since all target images are shown only from a canonical viewpoint [46]. The above observations affirm the rationality of our proposed soft source sample selection scheme.

Comparison under inductive UDA setting. To verify that our proposed strategy of uncovering the intrinsic target discrimination can derive the clustering solutions closer to the oracle target classifier than the existing transferring strategy of learning aligned feature representations between the two domains [16, 48], we design comparative experiments under the setting of inductive UDA. We follow a $50\%/50\%$ split scheme to divide each domain of Office-31 into the training and test sets. We use the both labeled sets of the source domain and the unlabeled training set of the target domain as the training data. In Table 2, we report results on the test set of the target domain using the best-performing model on the target training set. Here, **Oracle Model** finetunes the base network on the labeled target training set. We can see that our proposed uncovering strategy SRDC achieves closer results to Oracle Model, verifying the motivation of this work and the efficacy of our proposed SRDC.

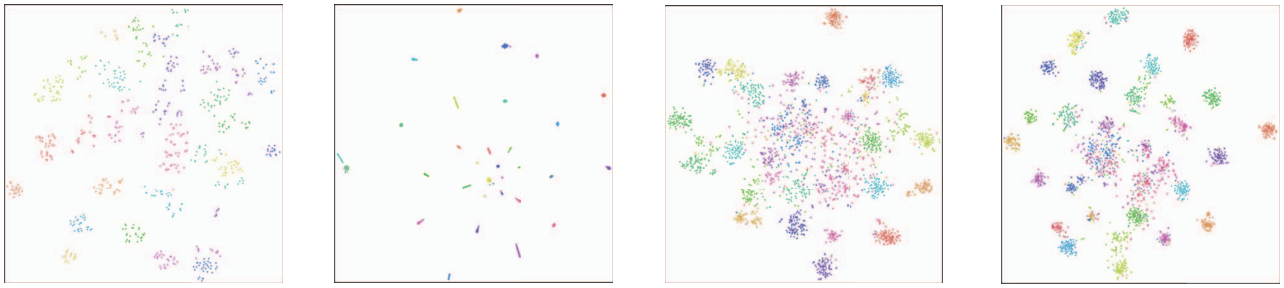
Feature visualization. We utilize t-SNE [54] to visualize embedded features on the target domain by Source Model and SRDC for two reverse transfer tasks of **A**→**W** and **W**→**A** in Figure 3. We can qualitatively observe that compared to Source Model, the target domain features can be much better discriminated by SRDC, which is based on data clustering to uncover the discriminative data structures.

Confusion matrix. We give confusion matrixes in terms of accuracy achieved by Source Model and SRDC on two reverse transfer tasks of **A**→**W** and **W**→**A** in Figure 4. Similar to the qualitative result of Figure 3, we can observe quantitative improvements from Source Model to SRDC, further confirming the advantages of SRDC.

Convergence performance. We verify the convergence performance of Source Model and SRDC with the test errors on two reverse transfer tasks of **A**→**W** and **W**→**A** in Figure 5. We can observe that SRDC enjoys faster and smoother convergence performance than Source Model.

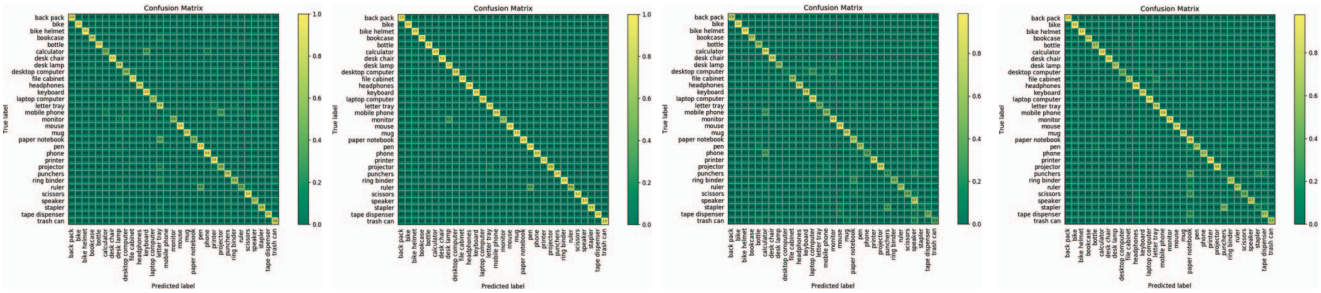
5.3. Comparisons with the state of the art

Results on Office-31 based on ResNet-50 are reported in Table 3, where results of existing methods are quoted from



(a) Source Model: **A**→**W** (b) SRDC: **A**→**W** (c) Source Model: **W**→**A** (d) SRDC: **W**→**A**

Figure 3. The t-SNE visualization of embedded features on the target domain. Note that different classes are denoted by different colors.



(a) Source Model: **A**→**W** (b) SRDC: **A**→**W** (c) Source Model: **W**→**A** (d) SRDC: **W**→**A**

Figure 4. The confusion matrix on the target domain. (Zoom in to see the exact class names!)

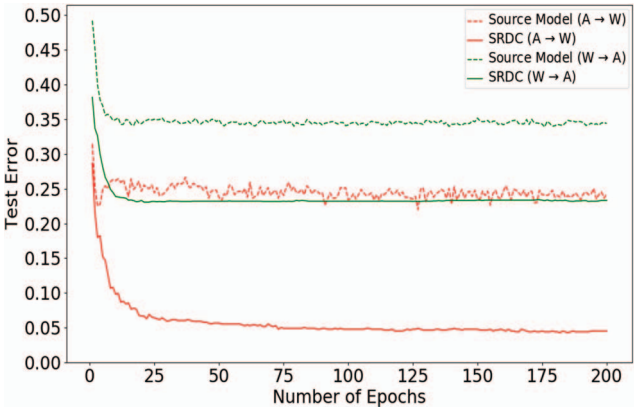


Figure 5. Convergence.

their respective papers or the works of [5, 33, 35]. We can see that SRDC outperforms all compared methods on almost all transfer tasks. It is noteworthy that SRDC significantly enhances the classification results on difficult transfer tasks, e.g. **A**→**W** and **W**→**A**, where the two domains are quite different. SRDC exceeds the latest work of BSP aiming to improve the discriminability for adversarial feature adaptation, showing that data clustering could be a more promising direction for target discrimination.

Results on ImageCLEF-DA based on ResNet-50 are reported in Table 4, where results of existing methods are quoted from their respective papers or the work of [35]. SRDC achieves much better results than all compared methods on all transfer tasks and substantially improves the results on hard transfer tasks, e.g. **C**→**P** and **P**→**C**, verifying the efficacy of SRDC on transfer tasks with the source and target domains of equal size and class balance.

Results on Office-Home based on ResNet-50 are reported in Table 5, where results of existing methods are quoted from their respective papers or the works of [35, 45]. We can observe that SRDC significantly exceeds all compared methods on most transfer tasks, with still a large room for improvement. This is reasonable since the four domains in Office-Home contain more categories, are visually more different from each other, and have much lower in-domain classification results [55]. It is inspiring that SRDC largely improves over the current state-of-the-art method MDD on such difficult tasks, which underlines the importance of discovering the discriminative structures by data clustering.

6. Conclusion

In this work, motivated by the assumption of structural domain similarity, we propose a source regularized, deep discriminative clustering method, termed as *Structurally Regularized Deep Clustering (SRDC)*. SRDC addresses a potential issue of damaging the intrinsic data discrimination by the existing alignment based UDA methods, via directly uncovering the intrinsic discrimination of target data. Technically, we use a flexible framework of deep network based discriminative clustering that minimizes the KL divergence between predictive label distribution of the network and an introduced auxiliary one; replacing the auxiliary distribution with that formed by ground-truth labels of

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
Source Model [21]	77.8±0.2	96.9±0.1	99.3±0.1	82.1±0.2	64.5±0.2	66.1±0.2	81.1
DAN [34]	81.3±0.3	97.2±0.0	99.8±0.0	83.1±0.2	66.3±0.0	66.3±0.1	82.3
DANN [16]	81.7±0.2	98.0±0.2	99.8±0.0	83.9±0.7	66.4±0.2	66.0±0.3	82.6
ADDA [53]	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
VADA [51]	86.5±0.5	98.2±0.4	99.7±0.2	86.7±0.4	70.1±0.4	70.5±0.4	85.4
SimNet [43]	88.6±0.5	98.2±0.2	99.7±0.2	85.3±0.3	73.4±0.8	71.8±0.6	86.2
MSTN [59]	91.3	98.9	**100.0**	90.4	72.7	65.6	86.5
GTA [49]	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
MCD [48]	88.6±0.2	98.5±0.1	**100.0**±0.0	92.2±0.2	69.5±0.1	69.7±0.3	86.5
SAFN+ENT [60]	90.1±0.8	98.6±0.2	99.8±0.0	90.7±0.5	73.0±0.2	70.2±0.3	87.1
DAAA [28]	86.8±0.2	**99.3**±0.1	**100.0**±0.0	88.8±0.4	74.3±0.2	73.9±0.2	87.2
iCAN [63]	92.5	98.8	**100.0**	90.1	72.1	69.9	87.2
CDAN+E [35]	94.1±0.1	98.6±0.1	**100.0**±0.0	92.9±0.2	71.0±0.3	69.3±0.3	87.7
MSTN+DSBN [5]	92.7	99.0	**100.0**	92.2	71.7	74.4	88.3
TADA [56]	94.3±0.3	98.7±0.1	99.8±0.2	91.6±0.3	72.9±0.2	73.0±0.3	88.4
TAT [33]	92.5±0.3	**99.3**±0.1	**100.0**±0.0	93.2±0.2	73.1±0.3	72.1±0.3	88.4
SymNets [68]	90.8±0.1	98.8±0.3	**100.0**±0.0	93.9±0.5	74.6±0.6	72.5±0.5	88.4
BSP+CDAN [9]	93.3±0.2	98.2±0.2	**100.0**±0.0	93.0±0.2	73.6±0.3	72.6±0.3	88.5
MDD [66]	94.5±0.3	98.4±0.1	**100.0**±0.0	93.5±0.2	74.6±0.3	72.2±0.1	88.9
CAN [27]	94.5±0.3	99.1±0.2	99.8±0.2	95.0±0.3	**78.0**±0.3	77.0±0.3	90.6
SRDC	**95.7**±0.2	99.2±0.1	**100.0**±0.0	**95.8**±0.2	76.7±0.3	**77.1**±0.1	**90.8**

Table 3. Results (%) on Office-31 (ResNet-50).

Methods	I → P	P → I	I → C	C → I	C → P	P → C	Avg
Source Model [21]	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
DAN [34]	74.5±0.4	82.2±0.2	92.8±0.2	86.3±0.4	69.2±0.4	89.8±0.4	82.5
DANN [16]	75.0±0.6	86.0±0.3	96.2±0.4	87.0±0.5	74.3±0.5	91.5±0.6	85.0
JAN [37]	76.8±0.4	88.0±0.2	94.7±0.2	89.5±0.3	74.2±0.3	91.7±0.3	85.8
CDAN+E [35]	77.7±0.3	90.7±0.2	97.7±0.3	91.3±0.3	74.2±0.2	94.3±0.3	87.7
TAT [33]	78.8±0.2	92.0±0.2	97.5±0.3	92.0±0.3	78.2±0.4	94.7±0.4	88.9
SAFN+ENT [60]	79.3±0.1	93.3±0.4	96.3±0.4	91.7±0.0	77.6±0.1	95.3±0.1	88.9
SymNets [68]	80.2±0.3	93.6±0.2	97.0±0.3	93.4±0.3	78.7±0.3	96.4±0.1	89.9
SRDC	**80.8**±0.3	**94.7**±0.2	**97.8**±0.2	**94.1**±0.2	**80.0**±0.3	**97.7**±0.1	**90.9**

Table 4. Results (%) on ImageCLEF-DA (ResNet-50).

Methods	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
Source Model [21]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [34]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [16]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [37]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
SE [15]	48.8	61.8	72.8	54.1	63.2	65.1	50.6	49.2	72.3	66.1	55.9	78.7	61.5
DWT-MEC [45]	50.3	72.1	77.0	59.6	69.3	70.2	58.3	48.1	77.3	69.3	53.6	82.0	65.6
CDAN+E [35]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
TAT [33]	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
BSP+CDAN [9]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
SAFN [60]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
TADA [56]	53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	60.0	82.9	67.6
SymNets [68]	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	74.5	52.6	82.7	67.6
MDD [66]	**54.9**	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	**60.2**	82.3	68.1
SRDC	52.3	**76.3**	**81.0**	**69.5**	**76.2**	**78.0**	**68.7**	**53.8**	**81.7**	**76.3**	57.1	**85.0**	**71.3**

Table 5. Results (%) on Office-Home (ResNet-50).

source data implements the structural source regularization via joint network training. In SRDC, we also enhance target discrimination with clustering of intermediate network features, and enhance structural regularization with soft selection of less divergent source examples. Experiments on benchmarks testify the efficacy of our method.

Acknowledgments. This work is supported in part by National Natural Science Foundation of China (Grant No.: 61771201), Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183), Guangdong R&D key project of China (Grant No.: 2019B010155001), and Microsoft Research Asia.

References

[1] The imageclef-da dataset is available at 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. *Machine Learning*, 79(1):151–175, May 2010.

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2007.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference On Computer Vision*, pages 1692–1700, 2018.

[5] W. Chang, T. You, S. Seo, S. Kwak, and B. Han. Domain-specific batch normalization for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7346–7354, June 2019.

[6] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *AISTATS 2005*, pages 57–64. Max-Planck-Gesellschaft, Jan. 2005.

[7] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, June 2019.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016.

[9] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1081–1090, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[10] Z. Chen, J. Zhuang, X. Liang, and L. Lin. Blending-target domain adaptation by adversarial meta-adaptation networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2243–2252, June 2019.

[11] S. Cicek and S. Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1416–1425, Oct 2019.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[13] Z. Deng, Y. Luo, and J. Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9943–9952, Oct 2019.

[14] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5747–5756, Oct 2017.

[15] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, Jan. 2016.

[17] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10–19, 2017.

[18] Boqing Gong, Kristen Grauman, and Fei Sha. Reshaping visual datasets for domain adaptation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1286–1294. Curran Associates, Inc., 2013.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[20] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 529–536, Cambridge, MA, USA, 2004. MIT Press.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[22] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 702–715, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[23] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 601–608. MIT Press, 2007.

[24] Mohammed Jabi, Marco Pedersoli, Amar Mitiche, and Ismail Ben Ayed. Deep clustering: On the link between discriminative models and k-means. *CoRR*, arXiv:1810.04246, 2018.

[25] K. Jia, S. Li, Y. Wen, T. Liu, and D. Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[26] K. Jia, J. Lin, M. Tan, and D. Tao. Deep multi-view learning using neuron-wise correlation-maximizing regularizers. *IEEE Transactions on Image Processing*, 28(10):5121–5134, Oct 2019.

[27] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, June 2019.

[28] Guoliang Kang, Liang Zheng, Yan Yan, and Yi Yang. Deep adversarial attention alignment for unsupervised domain adaptation: The benefit of target expectation maximization. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 420–436, Cham, 2018. Springer International Publishing.

[29] Andreas Krause, Pietro Perona, and Ryan G. Gomes. Discriminative clustering by regularized information maximization. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 775–783. 2010.

[30] V. K. Kurmi, S. Kumar, and V. P. Namboodiri. Attending to discriminative certainty for domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 491–500, June 2019.

[31] C. Lee, T. Batra, M. H. Baig, and D. Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10277–10287, June 2019.

[32] Haifeng Li, Keshu Zhang, and Tao Jiang. Minimum entropy clustering and applications to gene expression analysis. In *Proceedings of 2004 IEEE Computational Systems Bioinformatics Conference*, pages 142–151, 2004.

[33] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4013–4022, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 97–105. JMLR.org, 2015.

[35] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 1647–1657, USA, 2018. Curran Associates Inc.

[36] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 136–144, USA, 2016. Curran Associates Inc.

[37] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 2208–2217. JMLR.org, 2017.

[38] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] M. Mancini, L. Porzi, S. R. Bulò, B. Caputo, and E. Ricci. Boosting domain adaptation by discovering latent domains. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3780, June 2018.

[40] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[41] Y. Pan, T. Yao, Y. Li, Y. Wang, C. Ngo, and T. Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2234–2242, June 2019.

[42] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 3934–3941, 2018.

[43] P. O. Pinheiro. Unsupervised domain adaptation with similarity learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, June 2018.

[44] Ariya Rastrow, Frederick Jelinek, Abhinav Sethy, and Bhuvana Ramabhadran. Unsupervised model adaptation using information-theoretic criterion. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 190–197, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[45] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulò, N. Sebe, and E. Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9463–9472, June 2019.

[46] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 213–226, Berlin, Heidelberg, 2010. Springer-Verlag.

[47] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018.

[48] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, June 2018.

[49] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, June 2018.

[50] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1275–1282, USA, 2012. Omnipress.

[51] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018.

[52] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[53] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, July 2017.

[54] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9 (Nov):2579–2605, 2008.

[55] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394, 2017.

[56] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[57] Jun Wen, Risheng Liu, Nenggan Zheng, Qian Zheng, Zhefeng Gong, and Junsong Yuan. Exploiting local feature patterns for unsupervised domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[58] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning - Volume 48*, pages 478–487, 2016.

[59] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5423–5432, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[60] R. Xu, G. Li, J. Yang, and L. Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1426–1435, Oct 2019.

[61] Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[62] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 114–, New York, NY, USA, 2004. ACM.

[63] W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, June 2018.

[64] Yabin Zhang, Bin Deng, Hui Tang, Lefei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *ArXiv*, abs/2002.08681, 2020.

[65] Y. Zhang, K. Jia, and Z. Wang. Part-aware fine-grained object categorization using weakly supervised part detection network. *IEEE Transactions on Multimedia*, pages 1–1, 2019.

[66] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7404–7413, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[67] Yabin Zhang, Hui Tang, and Kui Jia. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[68] Y. Zhang, H. Tang, K. Jia, and M. Tan. Domain-symmetric networks for adversarial domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035, June 2019.

[69] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7523–7532, Long Beach, California, USA, 09–15 Jun 2019. PMLR.