Reliable Weighted Optimal Transport for Unsupervised Domain Adaptation

Renjun Xu*, Pelen Liu, Liyan Wang, Chao Chen
Zhejiang University, China

Jindong Wang
Microsoft Research

Abstract

Recently, extensive researches have been proposed to address the UDA problem, which aims to learn transferrable models for the unlabeled target domain. Among them, the optimal transport is a promising metric to align the representations of the source and target domains. However, most existing works based on optimal transport ignore the intra-domain structure, only achieving coarse pair-wise matching. The target samples distributed near the edge of the clusters, or far from their corresponding class centers are easily to be misclassified by the decision boundary learned from the source domain. In this paper, we present Reliable Weighted Optimal Transport (RWOT) for unsupervised domain adaptation, including novel Shrinking Subspace Reliability (SSR) and weighted optimal transport strategy. Specifically, SSR exploits spatial prototypical information and intra-domain structure to dynamically measure the sample-level domain discrepancy across domains. Besides, the weighted optimal transport strategy based on SSR is exploited to achieve the precise-pair-wise optimal transport procedure, which reduces negative transfer brought by the samples near decision boundaries in the target domain. RWOT also equips with the discriminative centroid clustering exploitation strategy to learn transfer features. A thorough evaluation shows that RWOT outperforms existing state-of-the-art method on standard domain adaptation benchmarks.

1. Introduction

Deep learning recently has achieved remarkable success in diverse computer vision tasks such as image classification, object detection and semantic segmentation with the help of large-scale labeled datasets [43]. Unfortunately, it is always expensive and time-consuming to collect extensive amounts of labeled data. To avoid labeling efforts, domain adaptation [26] serves as a promising solution to enhance the learning performance on a label-scarce domain (*i.e.*, target domain) by transferring knowledge from a label-rich domain (*i.e.*, source domain). The target domain may contain data collected from different perspectives or by different sensors, leading to a large domain gap.

For unsupervised domain adaptation, a major line of work reduces the domain gap by learning domain invariant feature representation, such as Maximum Mean Discrepancy (MMD) [38, 30], Correlation Alignment (CORAL) [36, 24] and Kullback-Leiber divergence (KL) [51]. Besides, another promising direction is based on the adversarial training [1, 37], where a discriminator (domain classifier) is trained to distinguish between the source and target representations. Meanwhile, duo to its ability of encoding class-structure in distributions, the optimal transport that minimizes a global transportation effort or cost between distributions, has been widely used to reduce the domain shift under complex distributions [5, 33, 47, 40]. However, existing optimal transport approaches in domain adaptation do not consider intra-domain structure [44] of both domains, only achieving coarse pair-wise matching. Moreover, some images that are significantly dissimilar across domains in the feature space may cause a gross mismatch during optimal transport procedure, leading to negative transfer.

To address the aforementioned issues, we propose a Reliable Weighted Optimal Transport (**RWOT**) for domain adaptation. Inspired by the prototypical networks which has achieved significant performance in domain adaptation [27], we exploit the prototypical information in the feature space to mitigate the wrong transport procedure. Besides, we also take into account the discriminative feature learning to benefit the adaptation performance, which has been widely used to learn the deep invariant features [49, 52]. Our proposed RWOT, therefore, consists of Shrinking Subspace Reliability (SSR) and weighted optimal transport strategy, which can be seen in Figure 1. Moreover, we demonstrate the necessity of shrinking subspace reliability and the procedure of weighted optimal transport strategy. RWOT also equips with a discriminative centroid exploitation strategy to improve transfer performance. The main contributions of this paper are:

(1) We propose shrinking subspace reliability to measure the sample-level domain discrepancy across domains by exploiting spatial prototypical information and intra-domain structure dynamically. The method can be used as a pre-

*Corresponding author: rux@zju.edu.cn

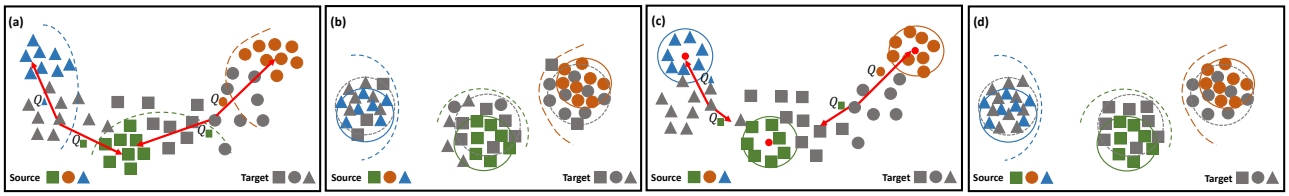


Figure 1. An overview of the proposed RWOT approach. Chromatic: Source samples; Gray: Target samples. Colorful imaginary lines: hyperplane learned from source domain. Red dots: shared class centers. Different kinds of shapes: different classes. (a) An example of unsupervised domain adaptation, where hard-aligned target samples distributed near the decision boundary, causing negative transfer. (b) The classification results of previous methods. (c) RWOT, with shrinking subspace reliability and discriminative centroid loss, exploits spatial prototypical information and intra-domain structure. (d) The final situation, our proposal achieves intra-class compactness and inter-class separability in the source and target domains. *Best viewed in color.*

processing step for existing domain adaptation techniques, improving efficiency significantly.

(2) We devise a weighted optimal transport strategy based on shrinking subspace reliability to achieve the *precise-pair-wise* optimal transport, reducing negative transfer brought by the samples near decision boundaries in the target domain. A discriminative centroid exploitation strategy is proposed to learn deep discriminative features.

(3) We analytically demonstrate that the combination of shrinking subspace reliability and optimal transport strategy can make deep features more distinguished and significantly enhance robustness and efficacy. Experimental results show that RWOT works stably in various datasets and outperforms existing methods.

2. Related work

Most of the discrepancy-based alignment methods are based on minimizing a divergence that measures the discrepancy between the source and target distributions. A representative work Maximum Mean Discrepancy (MMD) [38] aligns the source and target domains. Domain Adaptation Network (DAN) [21] utilizes multi-kernel strategies [11] in computing MMD to obtain a better performance. Another work Deep Correlation Alignment (Deep CORAL) [36] aligns the covariance of the source and target features. Weighted-MMD (WDAN) [46] introduces class-specific auxiliary weights into the original MMD to exploit the class prior probability on the source and target domains. Recent Joint Discriminative Domain Adaptation (JDDA) [4] claims that discriminative feature representations can enhance the performance of domain alignment. Contrastive Adaptation Network (CAN) [17] optimizes a new metric to explicitly model the intra-class domain discrepancy and the inter-class domain discrepancy. Deep Transfer Network (DTN) [49] utilizes pseudo label in conditional alignment. Easy-TL [43] programs intra-domain structures and Transferrable Prototypical Networks (TPN) [27] exploits prototypical distance [35] to avoid misclassification. Wang *et al.* [44, 42, 41] dynamic evaluate the importance of marginal and conditional distributions for distribution alignment.

Adversarial learning is another critical category to perform domain adaptation. GANs [12] have a generator that captures data distribution and a discriminator that predicts whether a sample is from the real data distribution or the generator. Based on this, to maximally confuse the domain classifier, Domain Adversarial Neural Network (DANN) [1] is proposed with a feature extractor. Deep Adversarial Metric Learning [9] considers distance metric in adversarial domain adaptation. Later, several extensions were proposed, such as Multi-Adversarial Domain Adaption (MADA) [28], and Conditional Domain Adversarial Networks (CDANs) [22] enable the alignment of multi-modal distributions. Domain-Symmetric Networks [50] learns invariant features in the domain adaptation. DANN [48] dynamically aligns the adversarial representations.

Optimal transport has been applied in domain adaptation to align the representations in the source and target domains with associated theoretical guarantees. Wasserstein Distance Guided Representation Learning (WDGRL) [34] uses Wasserstein distance as a core loss in promoting similarities between embedded representations by the dual formulation of the problem. The most famous attempt is Deep Joint Optimal Transport (DeepJDOT)[7], which applies the coupling matrix γ to transport the source samples to the target domain by an estimated mapping, achieving high accuracy on many transfer tasks. However, only adopting pure square root as a cost matrix for computing coupling γ will be quite weak, especially for hard-aligned samples in a more complicated task. Recently, researchers propose a feature selection procedure [10] that is implemented in optimal transport to leverage domain shift, but it is mainly based on the entropy regularization that cannot reduce negative transfer.

3. Reliable Weighted Optimal Transport

Most existing works resolve unsupervised domain adaptation by matching the shifted marginal distributions of source and target domains [22, 23]. A sophisticated approach is to formally define a statistical distance in the prob-

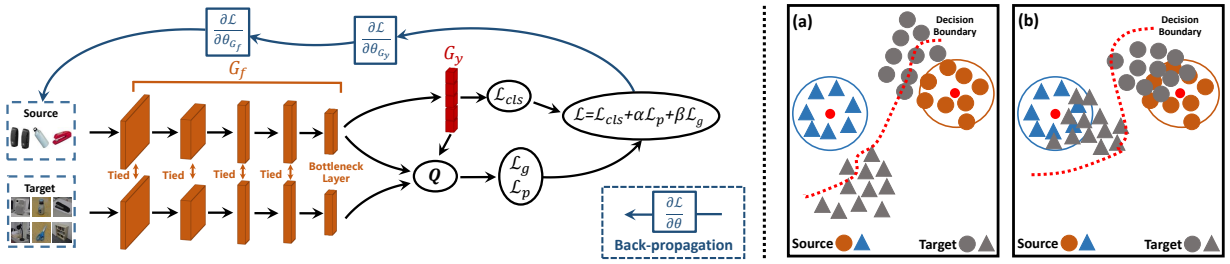


Figure 2. The architectures of Reliable Weighted Optimal Transport (RWOT), where G_f is the feature generator, G_y is the adaptive classifier; $\mathcal{L}_g$ is the proposed weighted optimal transport loss based on shrinking subspace realibility, $\mathcal{L}_p$ is the discriminative centroid loss, $\mathcal{L}_{cls}$ is the standard cross-entropy loss; α and β are hyper-parameters. The shrinking subspace realibility cost matrix $\mathbf{Q}$ is designed to balance the contribution of spatial prototypical information and intra-domain structures dynamicly during the training: (a) The decision boundary learned from source domain is not reliable to classify target samples, and the source samples are pushed to the spatial prototypes of corresponding classes correctly. (b) The decision boundary obtain reliable intra-domain structure of target samples, achieving better performance. *Best viewed in color.*

abilistic metric space and learn the optimal transport coupling to minimize that distance [7, 5, 33]. However, despite the great success, there is a common intrinsic limitation of this line of work: optimal transport is coarse pair-wise matching. Each image is reasoned as a whole to be transferred or not, without exploiting its intra-domain structures.

In this work, following the settings of unsupervised domain adaptation, we define a *source* domain $\mathcal{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled samples and a *target* domain $\mathcal{D}^t = \{(\boldsymbol{x}_i^t)\}_{i=1}^{n_t}$ with n_t unlabeled samples. The source and target domains follow joint probability distributions p and q respectively, and note that $p \neq q$. The discrepancy between these two distributions raises the key technical challenge of domain adaptation. The goal of this paper is to design a deep neural network that enables end-to-end training of a transferable feature generator and an adaptive classifier to minimize the distribution discrepancy across domains.

Following earlier works [7, 4], we choose two-stream siamese CNN architecture with shared weights. We propose Reliable Weighted Optimal Transport (RWOT) for unsupervised domain adaptation, which is an end-to-end method that learns a feature generator G_f and a classifier G_y, as illustrated in Figure 2. The main parts of RWOT are Shrinking Subspace Reliability (SSR) and weighted optimal transport strategy. The shrinking subspace reliability exploits spatial prototypical information and intra-domain structures, which is demonstrated to be of the great benefit of domain alignment and final classification. The weighted optimal transport strategy based on SSR achieves precise pair-wise matching to reduce negative transfer. Moreover, RWOT can learn discriminative transfer information by proposed centroid loss. In the following sections, we present the details of RWOT.

3.1. Shrinking Subspace Reliability

In a previous study [4], domain adaptation methods ignore the situation that target samples distributed near the edge of the clusters, or far from their corresponding class centers are easily to be misclassified by the hyperplane learned from the source domain. Inspired from self-labeling for domain adaptation [43], we exploit spatial prototypes of each class learned from labeled source data to assign the target samples a "pseudo" label. Considering negative transfer caused by target samples distributed near the edge of clusters, we propose Shrinking Subspace Reliability (SSR) to measure the sample-level domain discrepancy across domains, including spatial prototypical information to normalize prototypical distance and intra-domain structure computes the probability of target sample i belonging to the class k.

For quantifying spatial prototypical information in both domains, we define $\boldsymbol{c}^s$ as the class centers of deep features in the source domain, and $\boldsymbol{c}^s \in \mathbb{R}^{C\times d}$, where C denotes the number of classes in $\mathcal{D}^s$. d is the number of output neurons in the bottleneck layer. The spatial prototypical information is defined by matrix $\mathbf{D} \in \mathbb{R}^{n\times C}$ as:

$$D(i,k) = \frac{e^{-d(G_f(\boldsymbol{x}_i^t), \boldsymbol{c}_k^s))}}{\sum_{m=1}^{C} e^{-d(G_f(\boldsymbol{x}_i^t), \boldsymbol{c}_m^s))}}, \tag{1}$$

where $d(G_f(\boldsymbol{x}_i^t), \boldsymbol{c}_k^s))$ is the distance between the target sample $G_f(\boldsymbol{x}_i^t)$ and k-th source class center $\boldsymbol{c}_k^s$, where $k \in \{1, 2, ..., C\}$. n represents the batch-size for training. Compared with single kernel methods to measure the discrepancy of both domains monotonously, we focus on multiple kernels [13] to enhance the transferability of feature representation for deep domain adaptation comprehensively. Therefore, the multi-kernel formulation of $d(G_f(\boldsymbol{x}_i^t), \boldsymbol{c}_k^s))$ can be defined as:

$$\begin{aligned} d(G_f(\boldsymbol{x}_i^t), \boldsymbol{c}_k^s)) =& K(\boldsymbol{c}_k^s, \boldsymbol{c}_k^s) - 2K(G_f(\boldsymbol{x}_i^t), \boldsymbol{c}_k^s) + \\ & K(G_f(\boldsymbol{x}_i^t), G_f(\boldsymbol{x}_i^t)), \end{aligned} \tag{2}$$

The characteristic kernel associated with the feature map ϕ, the kernel $K(\boldsymbol{x}^s, \boldsymbol{x}^t) = \langle \phi(\boldsymbol{x}^s), \phi(\boldsymbol{x}^t) \rangle$, is defined as the

convex combination of m PSD kernels $\{K_u\}$:

$$\mathcal{K} = \left\{ K = \sum_{u=1}^{m} \beta_u K_u : \sum_{u=1}^{m} \beta_u = 1, \beta_u \geq 0, \forall u \right\}, \quad (3)$$

where $\mathcal{K}$ denotes the multi-prototypical kernel set. The constraints on coefficients $\{\beta_u\}$ are imposed to guarantee that the derived multi-kernel K is characteristic. As studied theoretically in Gretton *et al*. [13], the multi-kernel K can utilize different kernels to ensure lower test error, leading to a principled approach for optimal prototypical distance representations.

To convey the likelihood of intra-domain information by the pseudo classification probability of target samples, we define the sharpen probability annotation matrix $\mathbf{M}$ as:

$$M(i,k) = P\left(y = k | Softmax\left(\frac{G_y(G_f(\boldsymbol{x}_i^t))}{\tau}\right)\right), \quad (4)$$

where $\mathbf{M} \in \mathbb{R}^{n \times C}$, and $M(i,k)$ denotes the probability of the target samples i belongs to the label class k. τ is the *temperature* hyper-parameter [15] to obtain discriminative probability and reduce domain shift.

The purpose of shrinking subspace reliability is to quantitatively evaluate the importance of both spatial prototypical information, $D(i,k)$, and the intra-domain structure of target samples, $M(i,k)$. Formally, SSR is defined by $\mathbf{Q}$ as:

$$Q(i,k) = \frac{d_{\mathcal{A}(k)} D(i,k) + (2 - d_{\mathcal{A}(k)}) M(i,k)}{\sum_{m=1}^{C} \left(d_{\mathcal{A}(m)} D(i,m) + (2 - d_{\mathcal{A}(m)}) M(i,m)\right)}, \quad (5)$$

where $Q(i,k)$ weights the uncertainty of a target sample i belonging to class k. **The motivation for the numerator of Eq 5**: while both $D(i,k)$ and $M(i,k)$ measure the likelihood of target sample i having a label k, $D(i,k)$ is a distance measuring target sample i in the deep feature space to the class center c_k^s defined in the source domain, and $M(i,k)$ is measured by the classifier G_y. During the early training stage, $D(i,k)$ is much more reliable than the pseudo label given by classifier G_y. Fortunately, we can use the A-distance d_A to adjust the weights.The ***A*-distance**, $d_{\mathcal{A}(k)}(\mathcal{D}_k^s, \mathcal{D}_k^t) = 2(1 - 2\epsilon(h_k))$, followed from the definition of [3, 44], measures the discrepancy between the source and target domains. $\epsilon(h_k)$ is the error of a linear SVM classifier h_k discriminating the two domains. In the context of transfer learning, during the early training stage, the target domain is expected to be quite different from the source domain in the deep feature space, and one can have a near perfect classifier h_k, or $\epsilon(h_k) \to 0$ and $d_{\mathcal{A}(k)} \to 2$. The second term of numerator in Eq 5 vanishes, and the network is mainly trained through the first term. Finally, when the distributions of two domains coincide with each other, the classifier h_k cannot differentiate between two domains, and thus $\epsilon(h_k) = 0.5$ and $d_{\mathcal{A}(k)} = 0$. Now we can rely on G_y, and the second term $M(i,k)$ is the main contributor. The dynamic process is shown in Figure 2.

3.2. Weighted Optimal Transport

Optimal transport for domain adaptation performs the alignment of the sample representations in the source and target domains. However, existing optimal transport strategies fail to utilize the intra-domain structures, causing negative transfer arising from ambiguity coarse pair-wise matching. Therefore, to reduce the wrong pair-wise transport procedure, we devise the weighted optimal transport strategy by exploiting the proposed SSR. The optimization of weighted optimal transport is based on weighted Kantorovich problem [2] which seeks for a general coupling $\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)$ between $\mathcal{D}^s$ and $\mathcal{D}^t$:

$$\gamma^* = \underset{\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)}{\arg\min} \int_{\mathcal{D}^s \times \mathcal{D}^t} \mathcal{R}(\boldsymbol{x}^t, y(\boldsymbol{x}^s)) \mathcal{C}(\boldsymbol{x}^s, \boldsymbol{x}^t)\, d\gamma(\boldsymbol{x}^s, \boldsymbol{x}^t), \quad (6)$$

where$\mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)$ denotes the probability distribution between $\mathcal{D}^s$ and $\mathcal{D}^t$. $y(\boldsymbol{x}^s)$ is the label of source data $\boldsymbol{x}^s$ and $\mathcal{R}(\boldsymbol{x}^t, y(\boldsymbol{x}^s))$ represents the adaptive matrix based on deep reliable prior knowledge according to the intra-domain structures. The cost function matrix $\mathcal{C}(\boldsymbol{x}^s, \boldsymbol{x}^t) = ||\boldsymbol{x}^s - \boldsymbol{x}^t||^k$ denotes the cost to move probability mass from $\boldsymbol{x}^s$ to $\boldsymbol{x}^t$, where $k = 2$ [7, 6]. In our optimal transport problem, the weighted optimal transport strategy needs to estimate the adaptive transport coupling γ^* between two distributions and achieve feature transformation by minimizing the cost of γ^*. Here is the discrete reformulation:

$$\gamma^* = \underset{\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)}{\arg\min} \langle \gamma, \boldsymbol{Z} \rangle_F = \underset{\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)}{\arg\min} \langle \gamma, \mathcal{R} \cdot \mathcal{C} \rangle_F, \quad (7)$$

where $\gamma^* \in \mathbb{R}^{n \times n}$ is the weighted ideal coupling matrix between the source and target domains, representing as a joint probability measure. $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product and $\boldsymbol{Z} \in \mathbb{R}^{n \times n}$ is the adaptive cost function matrix. Choosing different weighted cost matrix will bring totally different pairwise matching [8]. It is crucial to utilize deep reliable prior knowledge $\mathcal{R}(x, y)$.

Considering SSR cost matrix $\mathbf{Q}$ which evaluates the spatial prototypical information and the intra-domain structure of target samples, we first propose a *precise-pair-wise* optimal transport mechanism by exploiting SSR. The discrete formulation of adaptive cost matrix $\mathbf{Z}$ can be defined as:

$$Z(i,j) = \left|\left| G_f(\boldsymbol{x}_i^s) - G_f(\boldsymbol{x}_j^t) \right|\right|^2 \cdot (1 - Q(j, y_i^s)), \quad (8)$$

The further constraints of the $(1 - Q(j, y_i^s))$ help to resolve the pairing ambiguity of traditional optimal transport strategy. As the bottleneck layer encodes both semantic and spatial information, discriminative representations allow a speedy computation of a transportation coupling with significant performance gains. With the above analysis, weighted optimal transport optimizes jointly in this feature space by reducing sample-wise distance of the same classes.

Then, the solution to this problem can be achieved by minimizing the following objective function:

$$\mathcal{L}_g = \sum_{i,j} \gamma^*_{i,j} (||G_f(\boldsymbol{x}_i^t) - G_f(\boldsymbol{x}_j^s)||^2 + \mathcal{F}_1(Softmax(G_y(G_f(\boldsymbol{x}_i^t)), y_j^s)), \tag{9}$$

where $\mathcal{F}_1$ is the cross-entropy function.

3.3. Discriminative Centroid Exploitation

The motivation of discriminative domain alignment is that samples belonging to the same class should be as closer as possible in the feature space. Inspired by the Center Loss [45], we propose discriminative centroid loss $\mathcal{L}_p$ for unsupervised domain adaptation as below:

$$\begin{aligned}\mathcal{L}_p = &\sum_{i=1}^{n} \left|\left|G_f(\boldsymbol{x}_i^s) - \boldsymbol{c}^s_{y_i^s}\right|\right|_2^2 \\ &+ \sum_{k=1}^{C}\sum_{i=1}^{n} Q(i,k) \left|\left|G_f(\boldsymbol{x}_i^t) - \boldsymbol{c}_k^s\right|\right|_2^2 \\ &+ \lambda \sum_{k_1,k_2=1,k_1\neq k_2}^{C} \max(0, \nu - \left|\left|\boldsymbol{c}^s_{k_1} - \boldsymbol{c}^s_{k_2}\right|\right|_2^2),\end{aligned} \tag{10}$$

where λ is a hyper-parameter and ν is a constraint margin to control the distance between the paired inter-class samples. And $\boldsymbol{c}^s_{y_i^s}$, as the y_i^s-th class center in the source domain, can be approximately evaluated by averaging deep features of several batch-size samples as

$$\boldsymbol{c}_k^s = \frac{1}{S}\sum_{i=1}^{N_b} G_f(\boldsymbol{x}_i^s)\phi(y_i^s, k), \tag{11}$$

where $\phi(y_i^s, k) = 1$ if $y_i^s = k$, otherwise $\phi(y_i^s, k) = 0$. $S = \sum_{i=1}^{N_b} \phi(y_i^s, k)$ and $k \in \{1, 2, \cdots, C\}$ is the class indicator. Ideally the class centers should be calculated based on all the samples while the procedure is time-consuming. Herein, we compute the class centers using N_b samples, where $N_b = m_b \times n$, and $m_b \in \{3, 4, 5\}$ is recommended.

3.4. Training

In this section, we introduce the training process of RWOT. We first define the standard classification loss of the source domain to train a classifier as follow:

$$\mathcal{L}_{cls} = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathcal{F}_1(G_y(G_f(\boldsymbol{x}_i^s)), y_i^s), \tag{12}$$

Considering weighted optimal transport based on shrinking subspace reliability and discriminative centroid loss, the total training objective of RWOT can be described as:

$$\min_{G_y, G_f} \mathcal{L}_{cls} + \alpha\mathcal{L}_p + \beta\mathcal{L}_g, \tag{13}$$

where α, β denote hyper-parameters that respectively trade-off the contribution of weighted optimal transport strategy and discriminative domain alignment under different datasets. The training process is shown in Algorithm 1.

Algorithm 1 The optimization strategy of RWOT

Require: source data as $\mathcal{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$, target data as $\mathcal{D}^t = \{(\boldsymbol{x}_i^t)\}_{i=1}^{n_t}$. T is set as the total number of training iterations, and n represents the batch-size for training. N_b is the number of samples to compute the source class centers in Eq 11.

1: Initialize two-stream CNN architectures.
2: **for** $i = 1$ to T **do**
3: Randomly choose N_b source samples $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N_b}$.
4: Calculate class centers $\boldsymbol{c}_j^s$ in the source domain according to the Eq.(11).
5: Randomly choose source samples $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n} \in \mathcal{D}^s$ and target samples $\{(\boldsymbol{x}_i^t)\}_{i=1}^{n} \in \mathcal{D}^t$.
6: Calculate spatial prototypical matrix **D** following Eq.(1) and sharpen probability annotation matrix **M** following Eq.(4). Update shrinking subspace reliability cost matrix **Q** following Eq.(5).
7: Calculate $\mathcal{L}_g$ following Eq.(9), $\mathcal{L}_p$ according to Eq.(10), and $\mathcal{L}_{cls}$ according to Eq.(12).
8: Update parameters of G_y and G_f following Eq.(13).
9: **end for**

4. Experiment

We conduct experiments to evaluate our approach with state-of-the-art domain adaptation methods.

4.1. Setup

Digits contains three standard digit classification datasets: **MNIST** [19], **USPS**[16] and **SVHN** [25]. Each dataset consists of 10 classes of digits, ranging from 0 to 9. We follow previous work [4] to construct three transfer tasks: **USPS→MNIST**, **MNIST→USPS** and **SVHN→MNIST**.

Office-31 [32] is a standard domain adaptation dataset which contains 4110 images from 31 categories with three domains: *Amazon* (**A**), with images collected from amazon.com, *Webcam* (**W**) and *DSLR* (**D**), with images shot by web camera and digital SLR camera respectively. By permuting the three domains, we obtain six transfer tasks: **A→W**, **A→D**, **D→W**, **W→D**, **D→A** and **W→A**.

ImageNet-Caltech is a large dataset built with **ImageNet-1K** [31] and **Caltech-256**. They share 84 classes, thus we select the same classes in both domains and form two transfer tasks: **ImageNet (84) → Caltech (84)** and **Caltech (84)→ ImageNet (84)**.

Office-Home [39] is a well organized, standard benchmark for visual domain adaptation, consisting of 15,500 images

Table 1. Classification accuracy (%) on Office-31 and ImageNet-Caltech dataset for unsupervised domain adaptation (ResNet)

Method	Office-31							ImageNet-Caltech		Avg
	A→W	A→D	D→W	W→D	D→A	W→A	Avg	I→C	C→I	
ResNet [14]	70.0±0.3	65.5±0.4	96.1±0.2	99.3±0.3	62.8±0.4	60.5±0.1	75.7	91.5±0.3	78.0±0.3	84.8
DeepCORAL [36]	83.0±0.1	71.5±0.2	97.9±0.2	98.0±0.2	63.7±0.3	64.5±0.2	79.8	92.0±0.4	85.5±0.2	88.8
DANN [1]	81.5±0.3	74.3±0.2	97.1±0.1	99.6±0.4	65.5±0.2	63.2±0.2	80.2	96.2±0.3	87.0±0.1	91.6
ADDA [37]	86.2±0.3	78.8±0.4	96.8±0.2	99.1±0.2	69.5±0.1	68.5±0.1	83.2	96.5±0.3	89.1±0.2	92.8
CDAN [22]	94.1±0.1	92.9±0.2	98.6±0.1	**100.0**±0	69.3±0.1	71.0±0.3	87.7	97.7±0.3	91.3±0.3	94.5
TPN [27]	91.2±0.3	89.9±0.2	97.7±0.2	99.5±0.1	70.5±0.2	73.5±0.1	87.1	96.1±0.2	90.8±0.3	93.5
DeepJDOT [7]	88.9±0.3	88.2±0.1	98.5±0.1	99.6±0.2	72.1±0.4	70.1±0.4	86.2	95.0±0.1	85.3±0.2	90.2
RWOT-M	92.5±0.1	89.5±0.2	99.2±0.2	**100.0**±0	75.1±0.2	74.5±0.2	88.5	95.9±0.1	90.1±0.2	93.0
RWOT-D	93.9±0.1	92.4±0.3	99.4±0.3	**100.0**±0	76.2±0.2	76.1±0.3	89.6	97.4±0.3	92.4±0.2	94.9
RWOT-C	94.7±0.2	94.0±0.2	99.4±0.2	**100.0**±0	77.1±0.2	77.4±0.3	90.4	97.7±0.1	**92.7**±0.2	95.2
RWOT	**95.1**±0.2	**94.5**±0.2	**99.5**±0.2	**100.0**±0	**77.5**±0.1	**77.9**±0.3	**90.8**	**97.9**±0.1	**92.7**±0.2	**95.3**

Table 2. Classification accuracy (%) on VisDA-2017 dataset for unsupervised domain adaptation (ResNet)

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
ResNet [14]	70.6	51.8	55.8	8.9	67.9	7.6	48.3	54.5	71.1	27.9	64.6	5.6	54.5
DANN [1]	75.9	70.5	65.3	17.3	72.8	38.6	58.0	77.2	72.5	40.4	70.4	44.7	58.6
SWD [20]	90.8	82.5	81.7	70.5	91.7	**69.5**	86.3	77.5	87.4	63.6	85.6	29.2	76.4
TPN [27]	93.7	**85.1**	69.2	81.6	**93.5**	61.9	89.3	81.4	**93.5**	**81.6**	84.5	49.9	80.4
DeepJDOT [7]	85.4	73.4	77.3	87.3	84.1	64.7	91.5	79.3	91.9	44.4	88.5	61.8	77.4
RWOT	**95.1**	80.3	**83.7**	**90.0**	92.4	68.0	**92.5**	**82.2**	87.9	78.4	**90.4**	**68.2**	**84.0**

Table 3. Classification accuracy (%) on Digits dataset for unsupervised domain adaptation (LeNet)

Method	S→M	M→U	U→M	Avg
LeNet [19]	68.3±0.3	65.3±0.5	66.2±0.2	66.6
DANN [1]	85.5±0.4	84.9±0.6	86.3±0.3	85.6
ADDA [37]	89.2±0.4	85.4±0.4	96.5±0.4	90.4
DeepCORAL [36]	88.3±0.2	84.1±0.3	93.6±0.2	88.7
DeepJDOT [7]	96.1±0.3	96.3±0.5	96.7±0.2	96.4
RWOT-M	97.2±0.2	97.5±0.3	96.8±0.4	97.1
RWOT-D	97.9±0.2	98.0±0.1	97.3±0.3	97.7
RWOT-C	98.5±0.1	**98.5**±0.2	**97.5**±0.2	98.1
RWOT	**98.8**±0.1	**98.5**±0.2	**97.5**±0.2	**98.3**

in 65 object classes in office and home settings, with four dissimilar domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**) and Real-World (**Rw**).

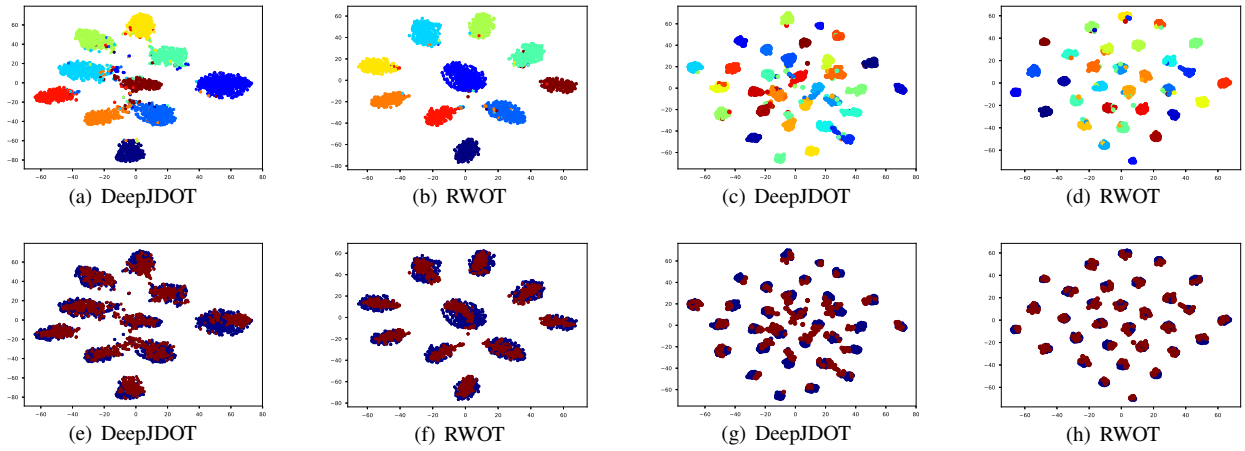
VisDA-2017 [29] is a large-scale computer vision dataset with two domains: **Synthetic**, renderings of 3D models from different angles and different lighting conditions; **Real**, real-world images. It has 280K images in 12 classes. This scale brings challenges to domain adaptation.

We compare the proposed **RWOT** model with state-of-the-art domain adaptation methods: (1) **ResNet-50** [14]. (2) Domain Adversarial Neural Network (**DANN**) [1] matches different domains by making them indistinguishable for a domain discriminator. (3) Adversarial Discriminative Domain Adaptation (**ADDA**) [37] designs a robust two-stage unsupervised domain adaptation model based on adversarial learning objectives. (4) Deep Correlation Alignment (**DeepCORAL**) [36] applies correlational matrix for marginal alignment in deep domain adaptation. (5) Conditional Domain Adversarial Network (**CDAN**) [22] designs a conditional alignment network based on adversarial learning. (6) Deep Joint Distribution Optimal Transport (**DeepJDOT**) [7] adapts optimal transport strategy in deep domain adaptation. (7) Transferrable Prototypical Network (**TPN**) [27] explores the prototypical information for discriminative feature alignment. (8) Sliced Wasserstein Discrepancy (**SWD**) [20] utilizes the Wasserstein Distance in domain alignment.

4.2. Implementation Details

Standard protocols for unsupervised domain adaptation have been followed. For Office-31 and VisDA dataset, ResNet-50 [14] is used as the backbone network and the models are fine-tuned from ResNet-50 pretrained on ImageNet. For experiments on digits dataset, we use the sample LeNet architecture following previous work [4]. We perform five random experiments and record the averaged accuracy for all transfer tasks.

For RWOT, we adopt a Gaussian kernel $K(\boldsymbol{a}, \boldsymbol{b}) = \exp(-||G_f(\boldsymbol{a}) - G_f(\boldsymbol{b})||/\sigma)$ with the bandwidth σ set to the median pair-wise distances on the training data. Following existing works [21, 13], we consider a family of m Gaussian kernels $\{k_u\}_{u=1}^{m}$ by varying bandwidth σ_u between $2^{-8}\sigma$ and $2^{8}\sigma$ with a multiplicative step-size of $2^{\frac{1}{2}}$. The multi-kernel strategy can measure prototypical distances and capture useful transferable information. We use all labeled



(a) DeepJDOT (b) RWOT (c) DeepJDOT (d) RWOT

(e) DeepJDOT (f) RWOT (g) DeepJDOT (h) RWOT

Figure 3. The t-SNE visualization of **SVHN→MNIST** and **A→D** tasks. Figure (a-d) represents category information (Each color denotes a class). Figure (e-h) represents domain information (Blue: Source domain; Red: Target domain).

source examples and unlabeled examples for training.

We optimize the network by the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and batch-size 128, namely total 256 images from the source and target domains. The learning rate and progressive training strategies are the same as [7]. Note that λ, as the hyper-parameter in centroid clustering, is set to 0.001 and the constraint margin ν is fixed as 50 throughout experiments. The constraint temperature parameter τ is fixed as 0.5. As for trade-off hyper-parameters $\alpha \in [10^{-3}, 1]$ and $\beta \in [10^{-2}, 10]$, we select $\alpha = 0.01$ and $\beta = 0.1$ for all transfer tasks. And we choose $m_b = 4$ for class center computation. In our experiments, we compare the average accuracy of each method on five random experiments.

4.3. Result and Discussion

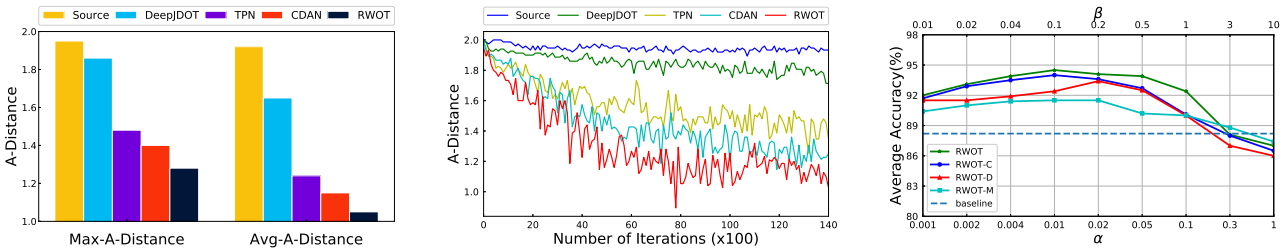
The unsupervised adaptation results on six *Office-31* and two *ImageNet-Caltech* transfer tasks are reported in Table 1. For a fair comparison, the results for most comparison methods are from their original papers. We can observe that RWOT significantly outperforms all previous methods on most tasks. It is worth noting that our proposal improves the classification accuracy substantially on hard transfer tasks, e.g. **A→D** and **D→A**, and achieves comparable classification performance on easy transfer tasks, e.g. **D→W** and **W→D**, where source and target are similar. Compared with DeepJDOT, the proposed framework exploits spatial prototypical information and aligns discriminative representation of each class centers of both domains, achieving intra-class compactness and inter-class separability.

The results on *VisDA-2017* are shown in Table 2. Due to the large domain gap between the source and target domains, we observe that the comparison methods obtain poor performance in some classes. The RWOT approach achieves performance boost in total, indicating that our architecture of reliable weight optimal transport via SSR is able to transfer more dissimilar categories. Considering the large size of *VisDA-2017* dataset, our proposal gains significant improvement.

We further compare RWOT with previous approaches on the *Digits* dataset, as reported in Table 3. In contrast to *Office-31* datasets, *Digits* dataset has a much larger domain size. We observe that RWOT overpasses all comparison methods on most transfer tasks and achieves almost state-of-the-art performance. It is remarkable that our proposal promotes the classification accuracy substantially on hard transfer tasks, *e.g.* **SVHN→MNIST**, where the source and target domains are substantially different (in scale, background clutter, blurring, slanting). The significant results suggest that RWOT is robust to a large domain gap and able to learn more transferable representations for unsupervised domain adaptation. Due to the limit of space, the results of *Office-Home* and *ImageNet-Caltech* are shown in the Supplementary Material.

4.4. Analysis

Ablation Study. To tooth apart the separate contributions of the weighted optimal transport strategy and discriminative centroid loss, we compare RWOT with DeepJDOT and three variants of RWOT on *Office-31* and *Digits* datasets: (1) **RWOT-M**, the variant only with reliable intra-domain structure ($\mathcal{L}_s + \beta\mathcal{L}_g$, **Q=M**). (2) **RWOT-D**, the variant only with spatial prototypical information ($\mathcal{L}_s + \beta\mathcal{L}_g$, **Q=D**). (3) **RWOT-C**, the variant without centroid loss ($\mathcal{L}_s + \beta\mathcal{L}_g$). RWOT-C without SSR is essentially DeepJDOT. Since RWOT-C and all the variants of RWOT outperform DeepJDOT in each task of our experiments significantly, and RWOT-C works better on those difficult tasks than RWOT-M and RWOT-D, it demonstrates that SSR weighted optimal transport is vital to match the



(a) Maximum and Average discrepancy (b) Average discrepancy (c) Parameter sensitivity

Figure 4. Analysis of domain discrepancy and model parameter w.r.t. α and β on **A**→**D** task.

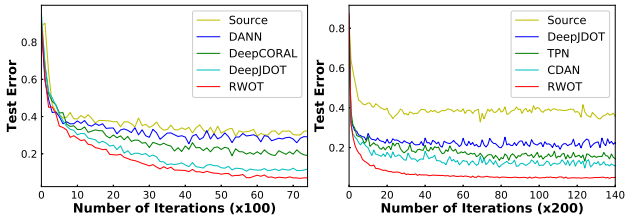


Figure 5. Comparison between RWOT and other state-of-the-art approaches in the convergence performance on the **SVHN**→**MNIST** (left) and the **A**→**D** (right) tasks.

source and target samples associated with the same class. Moreover, RWOT also improves with a large room over all three variants and DeepJDOT, validating the complement of weighted optimal transport strategy and discriminative centroid loss.

Feature visualization. To show the feature transferability, we visualize the t-SNE embeddings [18] of the bottleneck representation by DeepJDOT and RWOT on **SVHN**→**MNIST** and **A**→**D** tasks. Figure 3(a)-3(c) display that the features learned by DeepJDOT for different categories are mixed up. Figure 3(e)-3(g) shows that the domains are not well aligned while even worse, the target samples are aligned to the entire source data with possibly wrong classes, causing negative transfer. Note that, Figure 3(f)-3(h) show that the representations generated by RWOT achieves exactly 31 clusters with clear boundaries on **A**→**D** task. The significant visualization results suggest that our proposal is able to match the complex structures of the source and target domains and maximize the margin between different classes.

Distribution Discrepancy. The domain adaptation theory [3] suggests proxy $\mathcal{A}$-distance as a measure of cross-domain discrepancy. We adopt $d_{\mathcal{A}(i)}(\mathcal{D}_i^s, \mathcal{D}_i^t) = 2(1 - 2\epsilon(h_i))$ to analyze the distance between two domains. Note that, $\epsilon(h_i)$ is the generalization error of a linear classifier h discriminating the source domain $\mathcal{D}_i^s$ and the target domain $\mathcal{D}_i^t$ [22]. Figure 4(a) demonstrates the maximum discrepancy $(\max_i d_{\mathcal{A}(i)})$ and average discrepancy $(\text{avg}_i d_{\mathcal{A}(i)})$ on **A**→**D** task with the features of Source only, DeepJDOT, CDAN, TPN and RWOT. A much smaller discrepancy of either maximum or average has been observed by using RWOT features than compared approach features, which implies an effective reduction in domain gap. Figure 4(b) shows the convergence performance of our approach via $\mathcal{A}$-distance of the training process on **A**→**D**. We observe that RWOT achieves rapid convergence performance and significantly lower $\mathcal{A}$-distance due to it considers the intra-domain structure of the same classes in training.

Parameter Sensitivity. We evaluate the effects of the parameter α, β which balances the contribution of weighted optimal transport strategy and discriminative centroid loss, respectively. Figure 4(c) shows the variation of average accuracy as α or β on **A**→**D** task. We find that the accuracy increases first and then decreases as α or β increases, which demonstrates a proper trade-off can improve transfer performance. The wide-range significant performances validate the robustness and flexibility of RWOT.

Convergence. To illustrate the convergence of RWOT, we evaluate the test errors of all comparison methods on **SVHN**→**MNIST** and **A**→**D** tasks, as shown in Figure 5. The result reveals that our achieves stable convergence performance and significantly lower test error on the target domain. What is more, the trend of convergence curve suggests that RWOT considers spatial prototypical information and intra-domain structure. Such a phenomenon implies that RWOT can be trained efficiently and stably than previous domain adaptation methods.

5. Conclusions

This paper presented Reliable Weighted Optimal Transport (RWOT) for unsupervised domain adaptation with powerful Shrinking Subspace Reliability (SSR) and discriminative centroid loss. It exploits spatial prototypical information and intra-domain structure to reduce negative transfer brought by the samples near decision boundaries in the target domain. The proposed centroid loss also substantially enhances the performance of the hard-aligned samples in the more difficult transfer tasks. Comprehensive experiments show that our proposal outperforms state-of-the-art results on various domain adaptation datasets.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 1, 2, 6

[2] Sigurd Angenent, Steven Haker, and Allen Tannenbaum. Minimizing flows for the monge–kantorovich problem. *SIAM journal on mathematical analysis*, 35(1):61–97, 2003. 4

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007. 4, 8

[4] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3296–3303, 2019. 2, 3, 5, 6

[5] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014. 1, 3

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

[7] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 467–483. Springer, 2018. 2, 3, 4, 6, 7

[8] Luca Dieci and JD Walsh III. The boundary method for semi-discrete optimal transport partitions and wasserstein distance computation. *Journal of Computational and Applied Mathematics*, 353:318–344, 2019. 4

[9] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2780–2789, 2018. 2

[10] Léo Gautheron, Ievgen Redko, and Carole Lartizien. Feature selection for unsupervised domain adaptation using optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 759–776. Springer, 2018. 2

[11] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268, 2011. 2

[12] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Xu Bing, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, 2014. 2

[13] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012. 3, 4, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

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

[16] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 16(5):550–554, 2002. 5

[17] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 2

[18] Van Der Maaten Laurens and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605):2579–2605, 2008. 8

[19] Y. L. Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. proc ieee. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5, 6

[20] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 6

[21] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 2, 6

[22] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 2, 6, 8

[23] 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*, pages 2208–2217. JMLR. org, 2017. 2

[24] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017. 1

[25] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. *Nips Workshop on Deep Learning & Unsupervised Feature Learning*, 2011. 5

[26] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 1

[27] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2239–2247, 2019. 1, 2, 6

[28] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[29] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018. 6

[30] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. On minimum discrepancy estimation for deep domain adaptation. *arXiv preprint arXiv:1901.00282*, 2019. 1

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, 2010. 5

[33] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015. 1, 3

[34] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *arXiv preprint arXiv:1707.01217*, 2017. 2

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 2

[36] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016. 1, 2, 6

[37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 1, 6

[38] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[39] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. *CoRR*, abs/1706.07522, 2017. 5

[40] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 1

[41] Jindong Wang, Yiqiang Chen, Wenjie Feng, Han Yu, Meiyu Huang, and Qiang Yang. Transfer learning with dynamic distribution adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1):1–25, 2020. 2

[42] Jindong Wang, Yiqiang Chen, Shuji Hao, Wenjie Feng, and Zhiqi Shen. Balanced distribution adaptation for transfer learning. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1129–1134. IEEE, 2017. 2

[43] Jindong Wang, Yiqiang Chen, Han Yu, Meiyu Huang, and Qiang Yang. Easy transfer learning by exploiting intra-domain structures. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1210–1215. IEEE, 2019. 1, 2, 3

[44] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 402–410. ACM, 2018. 1, 2, 4

[45] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing. 5

[46] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017. 2

[47] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, and Mingkui Tan. Semi-supervised optimal transport for heterogeneous domain adaptation. In *Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18*, 2018. 1

[48] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. Transfer learning with dynamic adversarial adaptation network. In *International Conference on Data Mining (ICDM)*, 2019. 2

[49] Xu Zhang, Felix Xinnan Yu, Shih-Fu Chang, and Shengjin Wang. Deep transfer network: Unsupervised domain adaptation. *arXiv preprint arXiv:1503.00591*, 2015. 1, 2

[50] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2019. 2

[51] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 1

[52] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 1