

# Discriminative Radial Domain Adaptation

Zenan Huang, Jun Wen, *Member, IEEE*, Siheng Chen, *Member, IEEE*, Linchao Zhu, *Member, IEEE*,  
and Nenggan Zheng, *Senior Member, IEEE*

**Abstract**—Domain adaptation methods reduce domain shift typically by learning domain-invariant features. Most existing methods are built on distribution matching, *e.g.*, adversarial domain adaptation, which tends to corrupt feature discriminability. In this paper, we propose Discriminative Radial Domain Adaptation (DRDA) which bridges source and target domains via a shared radial structure. It's motivated by the observation that as the model is trained to be progressively discriminative, features of different categories expand outwards in different directions, forming a radial structure. We show that transferring such an inherently discriminative structure would enable to enhance feature transferability and discriminability simultaneously. Specifically, we represent each domain with a global anchor and each category a local anchor to form a radial structure and reduce domain shift via structure matching. It consists of two parts, namely isometric transformation to align the structure globally and local refinement to match each category. To enhance the discriminability of the structure, we further encourage samples to cluster close to the corresponding local anchors based on optimal-transport assignment. Extensively experimenting on multiple benchmarks, our method is shown to consistently outperforms state-of-the-art approaches on varied tasks, including the typical unsupervised domain adaptation, multi-source domain adaptation, domain-agnostic learning, and domain generalization.

**Index Terms**—Domain Adaptation, Transfer Learning, Radial Structure Matching

## I. INTRODUCTION

Machine learning methods generally assume that training and test data come from the same data distribution. However, such an assumption may not hold in practice, since a model trained on one distribution or one domain may need to be applied to data from another distribution or domain. Typically, such distribution shifts or domain shifts would cause significant performance drop [1], [2]. To address this issue, domain adaptation methods are proposed, which aim to generalize the learned knowledge from source domain to target domains.

Domain adaptation methods reduce domain shift typically by learning domain-invariant representations [2]. Previously,

Zenan Huang is with the Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, Zhejiang 310007, China and also with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310007, China. (e-mail: lccurious@zju.edu.cn).

Jun Wen is with the Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA. (e-mail: jun\_wen@hms.harvard.edu).

Siheng Chen is with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, 200240, China, and also with Shanghai AI Laboratory, Shanghai, 200240, China. (e-mail: sihengc@sjtu.edu.cn).

Linchao Zhu is with College of Computer Science and Technology, Zhejiang University, Hangzhou, China. (e-mail: zhulinchao@zju.edu.cn).

Nenggan Zheng is with the Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, Zhejiang, 310007, China, also with Collaborative Innovation Center for Artificial Intelligence by MOE and Zhejiang Provincial Government (ZJU) and Zhejiang Lab, Hangzhou, Zhejiang, 311121, China (e-mail: zng@cs.zju.edu.cn).

Nenggan Zheng is the corresponding author.

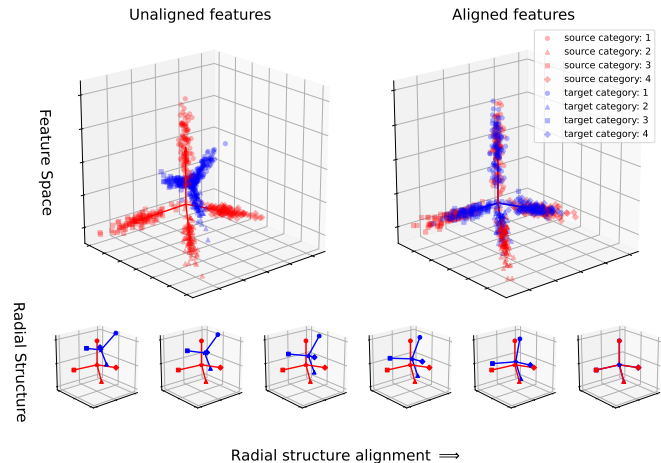


Fig. 1. Illustration of the proposed method which represents each domain using a radial structure and reduce domain shift via structure matching (source: red; target: blue; best viewed in color).

shallow features from both the source and target domains are mapped into a shared subspace [3]. With the success of deep learning, domain-invariant features are learned using deep neural networks [4], upon which various domain discrepancy measures are proposed, *e.g.*, Maximum Mean Discrepancy (MMD) [5], second-order correlations [6], and moments [7]. Recently, adversarial domain adaptation methods [8], [9] have achieved excellent performances and became the most popular approach by training an additional discriminator network to distinguish the features from different domains [10]–[13]. Domain-invariant features are expected to be learned by training the feature extractor to produce features that are indistinguishable by the discriminator.

Though the prevalent adversarial domain adaptation has shown success in many areas, there are still two limitations. Firstly, the *minmax* game of adversarial learning is notoriously known to be difficult to optimize, requiring lots of training tricks [14]. When the domain gap is large or the data distribution is complicated with multi modes, these models tend to collapse with false feature alignment [15], [16], especially when trained from scratch. Secondly, adversarial training is shown to damage the learned feature discriminability [13], [17]. While shown to be alleviated by either balancing the feature singular values [13] or lifting feature norms [17], such a discriminability corruption still persists because of the conflicts between transferability and discriminability which tends to be biased by source labels and with weaker transferability. One more promising approach is to learn a discriminative structure that is inherently transferable across domains.

In this paper, we propose Discriminative Radial Domain

Adaptation, that gets rid of the typical adversarial learning and bridges the source and target domains via a shared discriminative radial structure. It's motivated by the observation that the features initially all cluster together and as the model is trained to be more discriminative, features of different categories expand outwards in different directions to be more separated in the feature space, forming a radial structure, as also observed in [18], [19]. We bridge the source and target domains by aligning the radial structures. Specifically, we first build a radial structure for each domain that consists of a global domain anchor, which is the centroid of the domain data, and a set of local category anchors. Brute-force matching tends to twist the radial structure and damage its discriminability. To alleviate this, we decompose the structure matching into two components, namely *global isometric transformation* and *local refinement* as shown in Fig.1. *Global isometric transformation* aims to align the global shape of the two radial structures by bringing close the domain anchors and rotating the overall radial structure using a Siftfel layer [20]. To achieve fine-grained alignment of each category, *local refinement* further matches the angles and norms of local category anchors across domains.

To enhance the discriminability of the radial-like feature distribution, we encourage local features to cluster close to the corresponding local anchors. This is achieved by first assigning each local feature to the optimal local anchors via optimal transport, which prevents false assignments, and then minimizing an optimal-transport distance. Meanwhile, we enforce a prediction consistency between the radial structure and classifier to prevent conditional shift of the classifier. We observe such a consistency also promotes the radial structure to be more discriminative.

The main contributions of this work can be summarized as follows:

- We propose a novel domain adaptation approach, called Discriminative Radial Domain Adaptation, that gets rid of the typical adversarial training and reduces domain shift by matching radial structures that are inherently discriminative.
- We propose to decompose the alignment of radial structure into global isometric transformation and local anchor refinement to prevent damage to the discriminability of the radial structure.
- We enhance the discriminability of the radial structure by minimizing an optimal-transport distance that optimally assigns each feature to the corresponding local anchors to combat false alignment. Further, we perform a prediction consistency between the radial structure and classifier to alleviate conditional shift.
- Extensively experimenting on several benchmark datasets, our method outperforms the state-of-the-art approaches not only on the typical single-to-single unsupervised domain adaptation but also on multi-source domain adaptation, domain-agnostic adaptation and domain generalization.

## II. RELATED WORKS

In this part, we first review domain adaptation methods and then introduce discriminative structure learning.

### A. Domain Adaptation

To alleviate the domain shift, typical solutions include minimizing domain discrepancy and learning domain invariant features. In the context of domain discrepancy minimization, approaches can be classified according to their discrepancy metrics and their ways of extracting features. Discrepancy metrics include the Proxy  $\mathcal{A}$ -distance [21], the Kullback-Leibler (KL) divergence [22], the Maximum Mean Discrepancy [5], [23], other higher order statistical moments based distance measures [24], and Optimal Transport distance [25], [26]. Many types of feature extraction have also been considered for domain alignment, including handcrafted features [5], shallow features at the pixel level [27], and bottleneck features of deep neural networks [10], [28]. Along with their efficiency in reducing marginal domain discrepancies, these methods were found potentially hinder the learning of feature discriminant information [29]. Therefore, recent advances have focused on the discrepancy in conditional distribution by using labels or soft labels. Such approaches include conditional variants of MMD, Joint Distribution Optimal Transport (JDOT) [30], Moving Semantic Transfer Network (MSTN) [31], Robust Spherical Domain Adaptation (RSDA) [32], Category-Level Adversarial Network (CLAN) [33], Enhanced Transport Distance [34], Discriminative Manifold Propagation [35], and Conditional Kernel Bures (CKB) metric [29]. These improvements resulting from conditional alignment are evident; however, the changes in the class prior distribution and the noise of estimated target labels also pose risks of misalignment. In order to solve these issues, we propose to simultaneously learn the structure of the source and target distributions and align the two domains based on this structure. That is inspired by factorized optimal transport [36], which highlights the benefits of using low-dimensional structures to align data. In our framework, domain adaptation is carried out by aligning these radial structures learned from each domain, without relying on sample-level distribution.

Another approach mainly aims at learning domain invariant features so that the target can share the classifier trained from the labeled source. An effective method to guarantee features transferability is to train the generator to produce indistinguishable features, which can deceive the domain discriminator as a whole, *i.e.* Domain Adversarial Neural Networks (DANN) [9] and Adversarial Discriminative Domain Adaptation (ADDA) [8]. In addition, a number of studies have been published that examine ways to improve training strategies in order to create better transferable features from pixel-level features [37]–[39] or high-level features [40]. A further effort on producing transferable features is conditional adversarial training discriminator on features and class prediction jointly [11], [12]. Later, more works focus on disentangling original features into domain invariant and domain-specific parts [41], [42]. Nevertheless, these models seek to intensify feature transferability at the expense of feature

discriminability. In contrast, DRDA applies domain adaptation based on established discriminative radial structures, so the discriminability of features can be well maintained.

### B. Discriminative Structure Learning

Discriminative learning is aimed at pushing dissimilar features away from each other and enclosing the similar ones to be compact. Many efforts have been made to minimize intra-class feature distances and maximize inter-class feature distances, such as contrastive loss [18] and center loss [19], originally proposed in face recognition tasks. Inspired by softmax objective, L-Softmax (large margin softmax) [43] is introduced as another extension of enhancing discriminability by lifting angular separability between learned features. The principle of discriminative learning also enhances the performance of domain adaptation tasks. Follow the discriminative clustering, entropy minimization is introduced into domain adaptation [12], [44] to encourage classifier to produce ideal one-hot predictions are promising method. More recent studies in adversarial-based domain adaptation have shown that the discriminability of target features can be damaged by adversarial feature alignment. Based on the observations of close relation between the singular values of learned features and discriminative power, [13] correct the degeneration of discriminability by adding the penalties corresponding to these singular values. Also, [17] identified the connection between norm values and discriminatory power, and then lift the norm values for target features in order to increase the discriminatory power.

As well as features being discriminative during the learning process, the feature distribution is likely to perform in a particular low-dimensional format. The whole domain can therefore be well sketched by several clusters rather than using entire samples, allowing for a more robust approach to domain adaptation. In line with this idea, MSTN [31] is proposed to align the centroids of each category across domains, which reduces the noise influence of false pseudo labels compared to direct matching distributions. Prototypical networks [45] is proposed to learn prototypes of each category region and reduce conditional domain discrepancy by learning similar prototypes across domains. And [46] recognizes the importance of structure, connects the statistical property to geometric structures of data, and integrates feature selection and structure preservation into a unified optimization process. Moreover, [47] considered unsupervised domain adaptation as a clustering problem with missing labels using the structure preserve framework. Compared to these methods, our approach direct models feature distribution with a radial structure which maintains the intrinsic structure of the data while increasing the feature discriminability.

## III. DISCRIMINATIVE RADIAL DOMAIN ADAPTATION

In this section, we first introduce the construction of the radial structure. Then, we describe a proposed structure alignment strategy which decouple alignment into two independent components, namely global isometric transformation and local anchor refinement.

### A. Notations and Overview

In an unsupervised domain adaptation task, we are given labeled source domain  $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$  of  $n_s$  labeled examples and unlabeled target domain  $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$  of  $n_t$  unlabeled examples. Our model mainly contains a shared backbone  $G(\cdot)$  with parameter  $\theta$ , a shared classifier  $F(\cdot)$  with parameter  $\varphi$ , and a Stiefel layer  $S(\cdot)$  whose parameters  $\Delta$  are defined on Stiefel manifold  $\mathbf{V}_k(\mathbb{R}^d) = \{\Delta \in \mathbb{R}^{d \times k} | \Delta^\top \Delta = \mathbf{I}_k\}$ . Let  $\mathbf{z}_i^s = G(\mathbf{x}_i^s)$  and  $\hat{\mathbf{y}}_i^s = F(\mathbf{z}_i^s)$  be the feature representation and the estimated label of the  $i$ -th sample in the source domain, respectively.

With insights that linear classification output probability  $p_{ik} \propto \exp(W_k^\top \mathbf{z}_i + b) = \exp(\|W_k\| \|\mathbf{z}_i\| \cos(W_k, \mathbf{z}_i) + b)$  supposing importance of feature direction and norm in discrimination, we suggest a radial expansion-like structure for modeling features. Therefore, our framework is aiming to learn and align radial structures  $\mathcal{G}^s = \{\mathbf{a}^s, \mathcal{N}^s\}$  and  $\mathcal{G}^t = \{\mathbf{a}^t, \mathcal{N}^t\}$  from source and target domains, each structure containing a **global anchor**  $\mathbf{a}^{s/t}$  and a set  $\mathcal{N}^{s/t} = \{\mathbf{a}_i^{s/t}\}_{i=1}^k$  of  $k$  **local anchors**  $\mathbf{a}_i^{s/t} \in \mathbb{R}^d$ . From an intuitive viewpoint, a radial structure in latent space can be understood as a structure with a group of arrows that point from a global anchor to local anchors. Thus, for emphasizing the radial characteristic of structures, we also use the egocentric representation version  $\mathcal{V}^{s/t} := \{\mathbf{v}_i^{s/t} = (\mathbf{a}_i^{s/t} - \mathbf{a}^{s/t}) | \mathbf{a}^{s/t} \in \mathcal{N}^{s/t}\}$  of radial structure when comparing the shape differences of the structures. Finally, domain shifts and class prior differences are then manifested in terms of the isometric transformation and shape differences between two radial structures. We align the  $\mathcal{G}^s$  and  $\mathcal{G}^t$  by reducing isometric transformation to match each other globally in latent space and then refine them into the same shape. Where the isometric transformation and shape refinement are applied in a non-interfering manner for avoiding negative alignment. The DRDA approach can be viewed as an alternative optimization strategy that iteratively updates the radial structures  $\mathcal{G}^s, \mathcal{G}^t$  to be more representative and aligns radial structures in order to obtain more accurate label predictions in the target domain.

### B. Discriminative Radial Structure

Extraction of radial structure  $\mathcal{G}^{s/t}$  includes aggregating **global anchor**  $\mathbf{a}^{s/t}$  and a collection of **local anchors**  $\mathcal{N}^{s/t}$ . We represent the global and local anchors using vectorial embeddings, and iteratively update the anchors and model parameters.

1) *Global anchors*: For each domain, we define the global anchor as the centroid of overall features extracted by the shared feature extractor  $G_\theta(\cdot)$ . Formally, the global anchors  $\mathbf{a}^s, \mathbf{a}^t$  in the source and target domains are:

$$\mathbf{a}^s = \frac{1}{n_s} \sum_{i=1}^{n_s} G_\theta(\mathbf{x}_i^s), \quad \mathbf{a}^t = \frac{1}{n_t} \sum_{j=1}^{n_t} G_\theta(\mathbf{x}_j^t). \quad (1)$$

As an indicator of the mean position of features, global anchor is ideal reference point for contrasting two feature vector under the context of linear classification. They are also the reference points for comparing the radial structures. And displacement

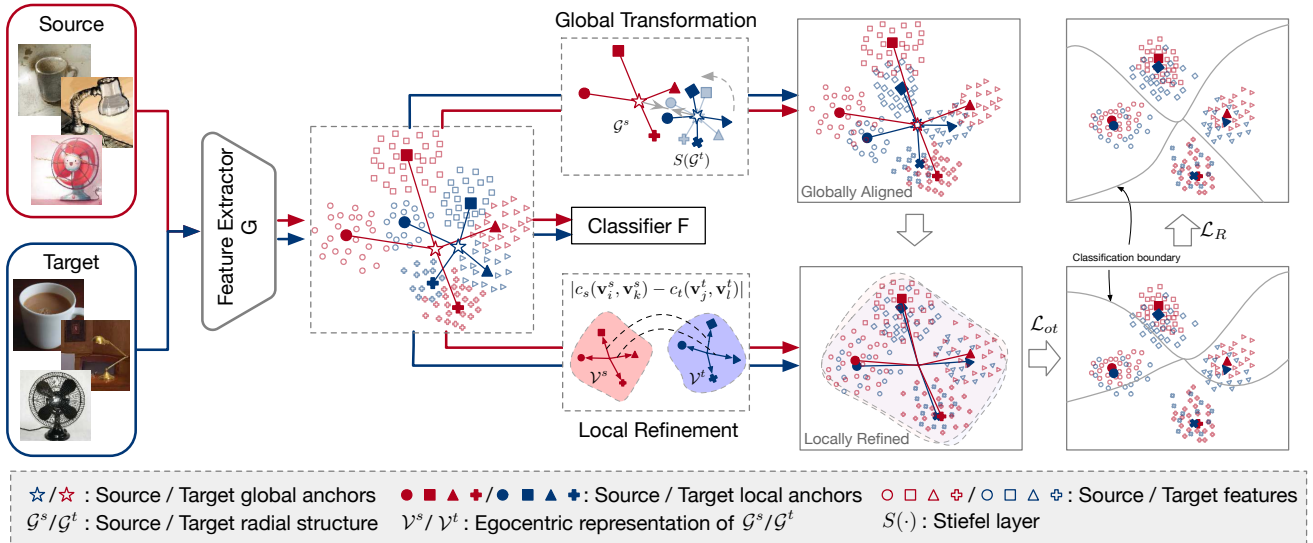


Fig. 2. Architecture of Discriminative Radial Domain Adaptation (DRDA). It bridges source and target domains by matching the radial structure which consists of a global domain anchor and a set of local category anchors. DRDA aligns the radial structures across domains via global isometric transformation and local anchor refinement (best viewed in color).

between global anchors naturally represent the mean feature shift  $\mathbb{E}[\mathbf{z}^s] - \mathbb{E}[\mathbf{z}^t]$ . We then use the distance between  $\mathbf{a}^s$  and  $\mathbf{a}^t$  as the global translation distance between two domains.

2) *Local anchors*: Source and the target radial structures contains  $k_s$  and  $k_t$  local anchors, respectively. Local anchors are located within high-density regions in each domain, each region containing a set of semantically related features.

For the general UDA task, it is straightforward to set  $k_s, k_t$  equal to the number of categories to be classified. Then, the local anchor is equivalent to the centroid of the features with same category. In labeled source data, such local anchors can be obtained directly from labels, while in the target domain local anchors can be obtained from pseudo-labels:

$$\mathbf{a}_k^s = \frac{1}{M_k} \sum_{i=1}^{n_s} \mathbf{z}_i \mathbb{I}[y_i = k], \quad \mathbf{a}_k^t = \frac{1}{M_k} \sum_{i=1}^{n_t} \mathbf{z}_i \mathbb{I}[\hat{y}_i = k], \quad (2)$$

where  $\mathbb{I}[\cdot]$  is indicator function,  $M_k = \sum \mathbb{I}[y_i = k]$  is a normalize constant.

### C. Radial Structure Alignment

We disentangle the structure alignment into two parts, namely global isometric transformation and local anchor refinement, to prevent its corruption to the discriminability of the radial structure. It's shown that when  $\mathcal{D}_s$  and  $\mathcal{D}_t$  are not aligned, the learned features would be arbitrarily rotated, translated or permuted [48]. By disentangling the alignment process into two independent processes, it is hopefully to best prevent false feature alignment.

1) *Isometric transformation*: We first globally align the shape of the radial structures to reduce the isometric transformation  $\tilde{T}(\cdot)$  between source  $\mathcal{G}^s$  and target  $\mathcal{G}^t$ . It is equivalent to minimize the isometric transformation  $\tilde{T}(\cdot)$  between the source and target, which is defined as:

$$\tilde{T} := \arg \min_T \|\mathcal{G}^s - T(\mathcal{G}^t)\|.$$

We seek to optimize the backbone  $G_\theta(\cdot)$ , thereby making  $\tilde{T}(\cdot)$  to be an identical transformation  $I(\cdot) : I(\mathcal{G}) = \mathcal{G}$ .

We disentangle the objective into translation and rotation parts in order to optimize feature extractor to achieve isometric alignment. The translation reduction is performed by minimizing the distance between global anchors among domains as follows:

$$\mathcal{L}_{\text{global}} = d(\mathbf{a}^s, \mathbf{a}^t) = \|\mathbf{a}^s - \mathbf{a}^t\|_F, \quad (3)$$

where the global distance measures the common distribution difference between source and target. By applying such global distance minimization, we force two global anchors to align and, consequently, we shift two feature distributions so that they share the same centroid. In addition, it is intuitively possible to make the entire radial structures (as the structures shown in Fig.1) on which the feature points lie roughly coincide.

The rotation reduction is accomplished by adding a Stiefel layer  $S(\cdot)$  to rotate the target features. According to a strict definition of the Stiefel manifold  $\mathbf{V}_k(\mathbb{R}^d) = \{\Delta \in \mathbb{R}^{d \times k} | \Delta^\top \Delta = \mathbf{I}_k\}$ , the Stiefel layer would perform rotation transform without causing any side effects to the features. In the backbone networks, we embed the Stiefel layer for target-specific use. In addition, for emphasizing the radial characteristic, we shift the radial structures  $\mathcal{G}^s, \mathcal{G}^t$  with respect to the global anchors, so that they share the same center and are referred to as egocentric version  $\mathcal{V}^s, \mathcal{V}^t$ , respectively. Thanks to the benefits of manifold optimization methods [49], it is easy to optimize the Stiefel layer as following objective:

$$\Delta^* = \arg \min_{\mathbf{V}_k(\mathbb{R}^d)} d(\mathcal{V}^s, S(\mathcal{V}^t)), \quad (4)$$

where the parameter  $\Delta$  is optimized with respect to a shape difference metric  $d(\cdot, \cdot)$  (induced from local alignment) between the radial structures  $\mathcal{G}^s, \mathcal{G}^t$  of the target and the source. It is noteworthy that the backbone is shared by the source and

target, but  $\Delta^*$  is embed in backbone network for target-specific use.

Optimization on Eq.(3) and Eq.(4) enables a coarse alignment among domains, and increases the reliability of target pseudo labels given by the classifier, as discriminative radial structures achieve more and more overlapping as shown in Fig.2.

2) *Local refinement*: Global alignment is intended to eliminate isometric discrepancies between the source and the target, whereas local alignment involves refining the two structures to be identical in shape. In order to avoid the occurrence of a contradiction between global and local alignments, the fine-grained structure difference measured here need to be independent of global alignment. In light of this, we apply Gromov-Wasserstein (GW) [50] distance to compare the shapes of two radial structures. According to the definition, GW distance is solely based on intra-space measurements, it has many desirable properties, especially in terms of invariances. And it terms out the invariant of translations, permutations, and rotations when Euclidean distance is used for intra-space measurement. Accordingly, whenever two structures  $\mathcal{G}^s$  and  $\mathcal{G}^t$  are shifted with different offsets or rotations, GW distance only determines the shape difference between them. For emphasizing the radial chararistic differences of the source and target structures we use the egocentric representation  $\mathcal{V}^s, \mathcal{V}^t$  instead of standard form  $\mathcal{G}^s, \mathcal{G}^t$ . Accordingly, the GW distance is defined as follows:

$$GW_2^2(c_s, c_t, \mu, \nu) = \min_{\pi \in \Pi(\mathcal{V}^s, \mathcal{V}^t)} J(c_s, c_t, \pi), \quad (5)$$

where

$$J(c_s, c_t, \pi) = \sum_{i,j,k,l} |c_s(\mathbf{v}_i^s, \mathbf{v}_k^s) - c_t(\mathbf{v}_j^t, \mathbf{v}_l^t)|^2 \pi_{i,j} \pi_{k,l},$$

where,  $\mu = \sum_{i=1}^{k_s} \delta_{\mathbf{a}_i^s}$ ,  $\nu = \sum_{i=1}^{k_t} \delta_{\mathbf{a}_i^t}$ , are the measure of anchors.  $\pi$  is the transport plan,  $\Pi(\cdot, \cdot)$  represent set of total transport permutation combination.  $c_s$  and  $c_t$  are specific intra-distance metrics defined on the radial structures of the source and the target, respectively. To incorporate the discriminative information, we recall classifier formulation  $p(y = k | \mathbf{z}_i) \propto \exp(\|W_k\| \|\mathbf{z}_i\| \cos(W_k, \mathbf{z}_i) + b)$  suggests angular and norm value is critical for vectors discrimination, we combine the both information in intra-distance function for  $c_s, c_t$ , and define them with same formulation:

$$c(\mathbf{v}_i, \mathbf{v}_j) = [1 - \frac{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}] + \lambda_{\text{dist}} \frac{1}{2} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2, \quad (6)$$

with  $\lambda_{\text{dist}}$  weight parameter to tradeoff angular difference loss between the structures. The first term calculates the cosine distances between the corresponding pairs of displacement vectors as angular difference. The second term captures the length difference by calculating the  $\ell_2$  distances between the corresponding pairs of displacement vectors. Furthermore, the transport plan  $\pi$  can be fixed due to one-to-one correspondences of discriminative vectors  $\mathcal{V}^s$  and  $\mathcal{V}^t$  from source and target are known, i.e. we force  $\pi_{i,j} = 0$  when  $i \neq j$  which gives:

$$GW(\mathcal{V}^s, \mathcal{V}^t) = \sum_{ij} |c(\mathbf{v}_i^s, \mathbf{v}_j^s) - c(\mathbf{v}_i^t, \mathbf{v}_j^t)|^2,$$

where the GW distance with a fixed transport plan implies that a certain property of GW are lost, that is rotational invariance. However, a loss of rotational difference yields a metric of shape difference that is useful for optimizing the Stiefel layer. This completed the distance metric  $d(\cdot, \cdot)$  in Eq.(4). Which has the advantage of providing a more efficient formula for the loss of local alignment and we defined it as  $\phi(\mathcal{V}^s, \mathcal{V}^t)$ . Where the  $\phi(\mathcal{V}^s, \mathcal{V}^t)$  can be expressed by the expectation of pairs of elements difference across two domains:

$$\phi(\mathcal{V}^s, \mathcal{V}^t) = \mathbb{E}_{(\mathbf{v}_k^s, \mathbf{v}_k^t) \in (\mathcal{V}^s, \mathcal{V}^t)} [c(\mathbf{v}_k^s, \mathbf{v}_k^t)]. \quad (7)$$

The simplified objective Eq.(7) gradually forces corresponding vectors being the same length and pointing to the same direction, thus ensuring the local structure alignment.

With synchronously minimizing the discrepancy among domains based on the radial structures by isometric transformation and structure refinement, data distributions of the source and target will move towards to each other and finally present the identical radial structure. In this way, the posterior probability expectations of each category in the source and the target domains can also be consistent.

#### D. Radial Structure Enhancement

We further improve the learning of radial structures from the following two aspects; 1) First one is structure faithfulness requirement, which encourages samples to enclose their corresponding local anchors. 2) Second one is semantic meaningfulness requirement, extracted radial structure should be informative for the semantic information of data distribution, i.e., the consensus between geometrical assignment labels and classifier labels.

1) *Enclose features to local anchors*: According to the structure faithfulness requirement, features are expected to be located near desired anchors. Since the distribution of features is unknown, we model the assignment of features to desired anchors by optimal transport plan [51]. In the case where the distances between features and anchors determine the transport cost, the optimal transport plan is the one which has the lowest total cost (also referred as optimal transport distance or Wasserstein distance) for moving features to corresponding anchors. The optimal transport plan can also be viewed as an adaptive distribution model allowing different anchors correspond to different probability densities. Then, to fairly push instances toward the desired local anchors, shared backbone network is learned to minimize optimal transport. Further, for relaxing the objective and stabilizing the end-to-end training we use entropic optimal transport [52] distance defined by:

$$\text{OT}_\theta^\epsilon(\mathcal{X}, \mathcal{N}) = \min_{\pi \in \Pi(\mu, \mu_a)} \sum_{i,j} d(G_\theta(\mathbf{x}_i), \mathbf{a}_j) \pi_{i,j} + \epsilon \text{KL}(\pi \| \mu \otimes \mu_a), \quad (8)$$

where  $\mu = \sum_{i=1}^n \delta_{\mathbf{x}_i}$  the measure of data instances and  $\mu_a = \sum_{j=1}^k \delta_{\mathbf{a}_j}$  the measure of anchors,  $d(\cdot, \cdot)$  is euclidean distance metric,  $\pi$  is the transport plan,  $\Pi(\cdot, \cdot)$  represent set of total transport permutation combination,  $\epsilon \geq 0$  is the regularization coefficient. As a relevant metric for assigning

samples to the best-fitted anchors, optimal transport distance can lead to a more reliable assignment than nearest neighbor search [53]. Therefore, by optimizing  $G_\theta(\cdot)$  in minimizing  $\mathcal{L}_{ot} = \text{OT}_\theta^\epsilon(\mathcal{X}^s, \mathcal{N}^s) + \text{OT}_\theta^\epsilon(\mathcal{X}^t, \mathcal{N}^t)$  in both source and target domain independently, the extracted features in both domains are more compactly arranged around their radial structures, the structure faithfulness requirement can be indirectly achieved.

2) *Consensus regularization*: For semantic meaningfulness requirement, we regard that instances assigned to the same local anchor have the same label, and for each instance, the label assigned by the classifier must match the label assigned by the radial structure. Hence, to train a network meets semantic meaningfulness requirements, a consensus regularization is designed to force the labels assigned by the classifier match the labels assigned by the radial structure,

$$\mathcal{R}_\varphi(\mathbf{Q}, \mathbf{P}) = \text{KL}(\mathbf{Q} \parallel \mathbf{P}) + H(\mathbf{P}), \quad (9)$$

where regularization is performed at classifier parameter  $\varphi$ ,  $\text{KL}(\cdot \parallel \cdot)$  is Kullback-Leibler divergence,  $H(\cdot)$  is entropy that balances the discriminability negative effects in this regularization,  $\mathbf{Q} = \{q_{i,k}\}$  is soft-assignments given by the transport plan  $\pi$ ,  $\mathbf{P} = \{(p_{i,1}, \dots, p_{i,k})\} \in [0, 1]^{K \times N}$  indicates the posteriors given by classifier. Consensus between data distribution structures and classifications can be improved by minimizing terms of regularization  $\mathcal{L}_R = \mathcal{R}_\varphi(\mathbf{Q}^s, \mathbf{P}^s) + \mathcal{R}_\varphi(\mathbf{Q}^t, \mathbf{P}^t)$ .

Intuitively, the objective based on Eq.(8) and Eq.(9) gradually enhances the representative and discriminative of radial structures in each domain through minimizing optimal transport distance from samples to local anchors and consensus regularization between radial structure assignment and classification.

### E. Optimization

The optimization is conducted in two steps, *i.e.*, radial structures extraction and alignment.

a) *Radial structure update*: The ideal implementation of calculating local anchors in Eq.(2) requires iterating over the entire dataset, which is computationally expensive. By employing an appropriate exponential moving average update strategy, we can easily perform end-to-end training:

$$\mathbf{a}_k = \eta \frac{1}{M_k} \sum_{i=1}^B \mathbf{z}_i \mathbb{I}[y_i = k] + (1 - \eta) \mathbf{a}'_k, \quad (10)$$

where  $B$  indicates the batch size and  $M_k = \sum_i^B \mathbb{I}[y_i = k]$  is a normalization constant,  $\mathbf{a}'_k$  indicates the last updated anchors and  $\mathbf{a}_k$  indicates new anchors computed in current iteration.

b) *Network update*: Recall objective of Eq.(8) and Eq.(9), a critical insights on behind successful optimization is similar to Expectation–Maximization (EM) algorithm. To optimize optimal transport distance from samples to local anchors, we fixed local anchors and update  $\theta$  according to Eq.(8), then update local anchors make use of updated  $\theta$  next iteration according to Eq.(2). To optimize the consensus between geometrical assignments and classifier assignments, we fixed  $\mathbf{Q}$  and only update classifier  $\varphi$  according to Eq.(9) with insights that classifier shall make trade off to respect

---

### Algorithm 1: DRDA Training

---

**Data**: Labeled source  $\mathcal{D}^s$ , Unlabeled target  $\mathcal{D}^t$

**Result**:  $\theta, \varphi, \Delta$

Initialization:  $\theta \leftarrow \theta_0, \varphi \leftarrow \varphi_0, \Delta \leftarrow \mathbf{I}$ ;

**while** *Not Converge* **do**

Sample  $\{(\mathcal{X}^s, \mathcal{Y}^s)\}$  and  $\{\mathcal{X}^t\}$  from  $\mathcal{D}^s$  and  $\mathcal{D}^t$ ;

$(\hat{\mathbf{P}}^s, \hat{\mathbf{P}}^t) \leftarrow (f_\varphi(G(\mathcal{X}^s), f_\varphi(S(G(\mathcal{X}^t))))$ );

Update radial structures  $\mathcal{G}^s, \mathcal{G}^t$  according to Eq.(2), Eq.(1);

Calculate source classification loss  $\mathcal{L}_{ce}(\hat{\mathbf{P}}^s, \mathcal{Y}^s)$  ;

Calculate alignment loss  $\mathcal{L}_{\text{global}}, \phi(\mathcal{V}^s, \mathcal{V}^t)$  according to Eq.(3), Eq.(7) ;

Calculate OT distance  $\mathcal{L}_{ot}$  by Eq.(8);

Calculate prediction discrepancy  $\mathcal{L}_R$  between classifier and radial structure in Eq.(9);

// Update parameters according to gradients;

$\Delta \leftarrow \Delta - \nabla_\Delta \lambda_\phi \phi(\mathcal{V}^s, \mathcal{V}^t)$ ;

$\varphi \leftarrow \varphi - \nabla_\varphi (\mathcal{L}_{ce} + \lambda_R \mathcal{L}_R)$ ;

$\theta \leftarrow \theta - \nabla_\theta (\mathcal{L}_{ce} + \lambda_{ot} \mathcal{L}_{ot} + \lambda_T \mathcal{L}_{\text{global}} + \lambda_\phi \phi(\mathcal{V}^s, \mathcal{V}^t))$ ;

**end**

**return**  $\theta, \varphi, \Delta$

---

intrinsic data distribution. Finally, alternative network update approach can be easily implemented by *stop gradient* tricks, then the overall objective respectively:

$$\begin{aligned} \min_{\theta, \varphi, \Delta} \quad & \mathcal{L}_{ce} + \lambda_T \mathcal{L}_{\text{global}} + \lambda_\phi \phi(\mathcal{V}^s, \mathcal{V}^t) \\ & + \lambda_{ot} [\text{OT}_\theta^\epsilon(\mathcal{X}^s, \text{SG}[\mathcal{N}^s]) + \text{OT}_\theta^\epsilon(\mathcal{X}^t, \text{SG}[\mathcal{N}^t])] \\ & + \lambda_R [\mathcal{R}_\varphi(\text{SG}[\mathbf{Q}^s], \mathbf{P}^s) + \mathcal{R}_\varphi(\text{SG}[\mathbf{Q}^t], \mathbf{P}^t)], \end{aligned} \quad (11)$$

where  $\text{SG}[\cdot]$  indicates the stop-gradient operation. This operation prevents parameters from being updated by the gradients. In the light of alternative network update approach, stop-gradient operation is critical for preventing degeneration of the structure during learning. Specifically, in Eq.(11), first term  $\mathcal{L}_{ce}$  is classification error; the second term  $\mathcal{L}_{\text{global}}$  and third term  $\phi(\mathcal{V}^s, \mathcal{V}^t)$  jointly perform isometric transformation and structure refinement for aligning feature distributions of different domains; the rest terms enhance the representativity and discriminability of radial structures. To balance the scale of terms in overall objective, the global loss (*i.e.* global translation distance) is scaled by  $\lambda_T$ , intra-structures difference is scaled by  $\lambda_\phi$ , OT distance is scaled by  $\lambda_{ot}$  and consensus regularization is scaled by  $\lambda_R$ . Notice, based upon differences Eq.(7) in the radial structures, the global rotation transformation distance minimization is implicitly optimized with respect to Stiefel layer parameters.

## IV. EXPERIMENTS

We compare the proposed method with several state-of-art methods on three types of UDA tasks, including single UDA, Domain-Agnostic UDA and Multi-Source UDA. The experimental results show that our method outperforms the other methods in terms of the average classification accuracy. In addition, we present a series of visualization results and



TABLE I  
ACCURACY (%) ON OFFICE-31 FOR UNSUPERVISED DOMAIN ADAPTATION (RESNET-50)

Method	A $\rightarrow$ W	D $\rightarrow$ W	W $\rightarrow$ D	A $\rightarrow$ D	D $\rightarrow$ A	W $\rightarrow$ A	Average
ResNet-50	68.4 $\pm$ 0.2	96.7 $\pm$ 0.1	99.3 $\pm$ 0.1	68.9 $\pm$ 0.2	62.5 $\pm$ 0.3	60.7 $\pm$ 0.3	76.1
RevGrad [54]	82.0 $\pm$ 0.4	96.9 $\pm$ 0.2	99.1 $\pm$ 0.1	79.7 $\pm$ 0.4	68.2 $\pm$ 0.4	67.4 $\pm$ 0.5	82.2
DAN [10]	80.5 $\pm$ 0.4	97.1 $\pm$ 0.2	99.6 $\pm$ 0.1	78.6 $\pm$ 0.2	63.6 $\pm$ 0.3	62.8 $\pm$ 0.2	80.4
JAN [55]	85.4 $\pm$ 0.3	97.4 $\pm$ 0.2	99.8 $\pm$ 0.2	84.7 $\pm$ 0.3	68.6 $\pm$ 0.3	70.0 $\pm$ 0.4	84.3
MADA [56]	90.0 $\pm$ 0.2	97.4 $\pm$ 0.1	99.6 $\pm$ 0.1	87.8 $\pm$ 0.2	70.3 $\pm$ 0.3	66.4 $\pm$ 0.3	85.2
CDAN+E* [12]	94.1 $\pm$ 0.1	98.6 $\pm$ 0.1	100.0 $\pm$ .0	92.9 $\pm$ 0.2	71.0 $\pm$ 0.3	69.3 $\pm$ 0.3	87.7
ALDA [57]	95.6 $\pm$ 0.5	97.7 $\pm$ 0.5	100.0 $\pm$ .0	94.0 $\pm$ 0.4	72.2 $\pm$ 0.4	72.5 $\pm$ 0.2	88.7
DRDA (w/o Angular)	92.6 $\pm$ 0.5	98.3 $\pm$ 0.2	100.0 $\pm$ .0	91.9 $\pm$ 0.5	71.0 $\pm$ 0.3	70.6 $\pm$ 0.1	87.4
DRDA (w/o Stiefel)	92.3 $\pm$ 0.5	98.7 $\pm$ 0.1	100.0 $\pm$ .0	92.1 $\pm$ 0.5	74.7 $\pm$ 0.2	75.3 $\pm$ 0.2	88.8
DRDA (w/o $\mathcal{R}_\varphi$ )	94.9 $\pm$ 0.3	98.2 $\pm$ 0.1	100.0 $\pm$ .0	93.8 $\pm$ 0.3	74.2 $\pm$ 0.1	75.8 $\pm$ 0.1	89.4
DRDA (w/o $\text{OT}_\theta^\epsilon$ )	94.8 $\pm$ 0.2	98.0 $\pm$ 0.1	100.0 $\pm$ .0	94.0 $\pm$ 0.2	74.8 $\pm$ 0.1	75.4 $\pm$ 0.1	89.5
DRDA	<b>95.8</b> $\pm$ 0.4	<b>98.8</b> $\pm$ 0.4	<b>100.0</b> $\pm$ .0	<b>94.5</b> $\pm$ 0.3	<b>75.6</b> $\pm$ 0.2	<b>76.6</b> $\pm$ 0.4	<b>90.2</b>

TABLE II  
ACCURACY (%) ON OFFICE-HOME FOR UNSUPERVISED DOMAIN ADAPTATION (RESNET-50)

Method	Ar $\rightarrow$ Cl	Ar $\rightarrow$ Pr	Ar $\rightarrow$ Rw	Cl $\rightarrow$ Ar	Cl $\rightarrow$ Pr	Cl $\rightarrow$ Rw	Pr $\rightarrow$ Ar	Pr $\rightarrow$ Cl	Pr $\rightarrow$ Rw	Rw $\rightarrow$ Ar	Rw $\rightarrow$ Cl	Rw $\rightarrow$ Pr	Average
ResNet-50	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
DANN [9]	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 [55]	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
CDAN+E [12]	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
ALDA [57]	53.7	70.1	76.4	60.2	72.6	71.5	56.8	51.9	77.1	70.2	56.3	82.1	66.6
MDD [58]	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
DRDA (w/o Angular)	54.3	70.3	74.8	60.7	69.2	69.8	59.1	52.8	76.4	70.9	58.3	82.0	66.5
DRDA (w/o Stiefel)	57.4	74.5	79.3	64.8	75.6	<b>74.0</b>	62.9	56.2	79.7	72.0	62.9	84.1	70.4
DRDA (w/o $\mathcal{R}_\varphi$ )	57.3	74.3	80.4	64.7	74.3	73.0	64.9	55.8	79.7	74.5	63.0	84.3	70.5
DRDA (w/o $\text{OT}_\theta^\epsilon$ )	57.0	73.9	80.2	64.1	73.8	73.1	64.4	56.1	78.9	73.4	62.8	84.1	70.1
DRDA	<b>58.2</b>	<b>74.2</b>	<b>81.2</b>	<b>65.6</b>	<b>75.1</b>	73.3	<b>65.8</b>	<b>57.1</b>	<b>80.4</b>	<b>75.6</b>	<b>63.2</b>	<b>85.1</b>	<b>71.2</b>

ablation studies to demonstrate the insights of our method and the effectiveness of each component in our model.

### A. Experimental Setup

1) *Office-31*: [59] is a widely used dataset for visual domain adaptation, which consists of 31 categories count up to 4,652 images from three distinct domains: 2,817 Amazon(A) images, 795 Webcam(W) images, and 498 DSLR(D) images. We evaluate methods upon all 6 in pairs of transfer tasks.

2) *Office-Home*: [60] is a better organized dataset and more difficult dataset compared to Office-31, which consists of 65 categories count up to 15,500 images in office and home setting, formed with four extremely dissimilar domains: Artistic images (Ar), Clipart images (Cl), Product images (Pr), and Real-World images (Rw).

3) *Office-Caltech10*: [61] is collected from Office31 and Caltech formed with four domains: A (Amazon), C (Caltech), W (Webcam), and D (DSLR). It consists of 10 object categories, each domain includes 958, 295, 157, and 1,123 images, respectively.

We compared the proposed (DRDA) with state-of-the-art domain adaptation methods: Domain Adversarial Neural Network (DANN) [9], Joint Adaptation Network (JAN) [55], Conditional Domain Adversarial Network with Entropy (CDAN+E) [12], Adversarial-Learned Loss for Domain Adaptation (ALDA) [57] and Margin Disparity Discrepancy (MDD)

[58]. For **multi-source domain adaptation** we compared our model with state-of-the-art domain adaptation methods: Deep Alignment Network (DAN) [11], Domain Adversarial Neural Network (DANN) [9], Manifold Embedded Distribution Alignment (MEDA) [62], Maximum Classifier Discrepancy (MCD) [28] and Moment Matching for Multi-Source Domain Adaptation (M<sup>3</sup>SDA) [24]. For **domain agnostic domain adaptation**, we compared our model with state-of-the-art methods: Self-Ensembling (SE) [38], Maximum Classifier Discrepancy (MCD) [28], Domain Adversarial Neural Network (DANN) [9] and Deep Adversarial Disentangled Autoencoder (DADA) [42].

We follow the standard protocols of unsupervised domain adaptation. We use all labeled source samples and unlabeled target samples and compare the average classification accuracy based on three experiments. The overall architecture consists of a backbone, **ResNet-50**, a bottleneck layer with 256 units and a full-connected layer. The Stiefel layer is a simple full-connected layer whose parameters are manipulated on *Stiefel Manifold* implemented with *geoopt* [49]. And this Stiefel layer is used for processing target features only. We implement our method in **Pytorch**. We finetune from ImageNet pre-trained models as the feature extract backbone. We essentially tune the hyper-parameters in Eq.(11),  $\lambda_T \sim 200$ ,  $\lambda_\phi \sim 0.6$ ,  $\lambda_{ot} \sim 0.0005$ ,  $\lambda_R \sim 1$ , they control the scaling of each loss term in overall objective. Both backbone layers and task-specific layers are trained through back-propagation using Stochastic

Gradient Descent (SGD). The Stiefel layer is optimized using Riemannian SGD [49]. The backbone layers is finetuned based on pre-trained ResNet-50 on ImageNet, while the task-specific layers are trained from scratch whose learning rate is 10 times that of backbone layers.

We use mini-batch stochastic gradient descent as the optimizer and apply momentum of 0.9 and learning rate schedule rule [9] with  $\eta_p = \eta_0(1 + \gamma p)^{-\beta}$ , where  $p$  is within the range of  $[0, 1]$ , and  $\eta_0 = 0.01$ ,  $\gamma = 10$ ,  $\beta = 0.75$ . We conduct the grid hyper-parameter selection base on loss curve fitting to obtain optimal weighted combination of objective in the experiments. To reduce noise influence and stable optimization convergence, we adopt progressive domain transfer weights with factor  $\lambda_p = \frac{2}{1 + \exp(-\alpha p)} - 1$  increasing from 0 to 1 where  $\alpha = 10$  as a default setting. Especially, for Office31 dataset, due to the small number of samples in DSLR and Webcam domain, we add temperature factor  $t = 0.85$  to adjust the convergence speed of a cross-entropy loss.

For the experiments on Multiple source unsupervised domain adaptation using Office-Caltech and OfficeHome, we sample instances uniformly from the combined source domains as source inputs. For the experiments on Domain agnostic unsupervised domain adaptation using Office-Caltech we sample instances uniformly from the combined target domains as target inputs. For the experiments on domain-generalized unsupervised domain adaptation, we train models similar to the setting of one-to-one UDA, and demonstrate the validation accuracies in domains which were not the sources or targets.

### B. Comparison with the State-of-The-Art Methods

1) *Single source to single target UDA*: To testify the effectiveness of the proposed DRDA, we first compare our method with state-of-art single domain UDA tasks. For a fair comparison, we report previous domain adaptation methods whose results are based on ResNet-50 and test in same validation setting.

The results on Office-31 are reported in Table I. In most transfer sub-tasks, DRDA attains the highest classification accuracy and improves the average accuracy over state-of-the-art methods. Our method works particularly well for small-to-large transfer tasks, such as  $\mathbf{D} \rightarrow \mathbf{A}, \mathbf{W} \rightarrow \mathbf{A}$ . Even though the sample size in the source domain is small, the proposed discriminative radial structure is sufficiently representative and discriminative to serve as a guide to domain alignment and more robust to noise.

The results on Office-Home are reported in Table II. The proposed DRDA achieves the best accuracy in all transfer tasks and improves the average accuracy over the state-of-the-art methods by 3%. Compared to Office-31, Office-Home is more challenging because it has more categories and greater discrepancies between domains. As the task becomes increasingly challenging, our approach outperforms our competitors by a greater margin. As explained in the hypothesis, radial-like structures are beneficial for sketching and preserving discriminative structures, and this structure is well suited for domain alignment. Therefore, in the case of more categories and greater discrepancies between domains,

the radial-like structure shows greater superior performance in domain alignment than the other methods. In the context of domain alignment tasks, these results illustrate the importance of discrimination preservation and low-dimensional structures (i.e. the proposed radial-like structure).

TABLE III  
ACCURACY (%) ON OFFICE-CALTECH FOR MULTI-SOURCE  
UNSUPERVISED DOMAIN ADAPTATION (RESNET-50)

Method	A,C,D→W	A,C,W→D	A,D,W→C	C,D,W→A	Avg
ResNet-50	97.1	99.2	89.4	94.7	95.6
DANN [9]	96.5	99.1	89.2	94.7	94.8
MEDA [62]	99.3	99.2	91.4	92.9	95.7
MCD [28]	99.5	99.1	91.5	92.1	95.6
M <sup>3</sup> SDA- $\beta$ [24]	99.5	99.2	92.2	94.5	96.4
DRDA (w/o Angular)	100.0	100.0	95.7	96.5	98.1
DRDA (w/o Stiefel)	100.0	100.0	95.7	96.8	98.1
DRDA (w/o $\mathcal{R}_\varphi$ )	100.0	100.0	95.8	96.5	98.0
DRDA (w/o OT $_\theta^c$ )	100.0	100.0	96.0	96.8	98.2
DRDA (ours)	<b>100.0</b>	<b>100.0</b>	<b>96.4</b>	<b>96.9</b>	<b>98.3</b>

2) *Multi-source to single target UDA: multi-source unsupervised domain adaptation* [24], which transfers knowledge from multiple source domains to one unlabeled target domain. Compared to one-to-one unsupervised domain adaptation, this task is much more difficult as the source domain is a mixture of multiple domains. In this task, we merge the multiple source domain into a single one. The results on the Office-Caltech10 dataset are reported in Table III. According to the observation, the proposed DRDA surpasses state-of-the-art methods even those developed for such tasks specifically. Recall the conception of radial-like structure, the key idea is sketching the discriminative structure of data distributions. In this respect, the more domains the algorithm uses as sources, the greater the generalization power the source structure has. Therefore, the proposed DRDA is naturally fit for UDA tasks involving multiple source domains.

TABLE IV  
ACCURACY (%) ON OFFICE-CALTECH FOR DOMAIN-AGNOSTIC  
UNSUPERVISED DOMAIN ADAPTATION (RESNET-50)

Method	A→C,D,W	C→A,D,W	D→A,C,W	W→A,C,D	Avg
ResNet-50	90.5±0.3	94.3±0.2	88.7±0.4	82.5±0.3	89
SE [38]	90.3±0.4	94.7±0.4	88.5±0.3	85.5±0.4	89.7
MCD [28]	91.7±0.4	95.3±0.3	89.5±0.2	84.3±0.2	90.2
DANN [9]	91.5±0.4	94.3±0.4	90.5±0.3	86.3±0.3	90.6
DADA [42]	92.0±0.4	95.1±0.3	91.3±0.4	93.1±0.3	92.9
DRDA (w/o Angular)	97.6±0.4	97.8±0.5	96.2±0.4	96.7±0.1	97.2
DRDA (w/o Stiefel)	97.7±0.3	97.1±0.7	96.8±0.1	97.0±0.2	97.2
DRDA (w/o $\mathcal{R}_\varphi$ )	97.6±0.5	97.6±0.3	96.6±0.5	96.7±0.2	97.1
DRDA (w/o OT $_\theta^c$ )	97.8±0.5	<b>98.0±0.2</b>	<b>97.1±0.5</b>	96.8±0.3	<b>97.5</b>
DRDA (ours)	<b>98.1±0.2</b>	97.5±0.2	96.6±0.4	<b>96.8±0.2</b>	97.3

3) *Single source to agnostic multi-target UDA*: We also consider another type of unsupervised domain adaptation task: **domain agnostic unsupervised domain adaptation** [42], which transfers knowledge from a labeled source domain to unlabeled data in one of multiple target domains. In this task, we regard the mixture target domain as a single one. The Table IV shows that our model gets a 97.3% average accuracy and improves the other methods by 4.4% in the domain agnostic unsupervised domain adaptation task. It appears that the radial-like structure is consistently effective at representing



TABLE V  
ACCURACY (%) ON OFFICE-HOME FOR DOMAIN GENERALIZE UNSUPERVISED DOMAIN ADAPTATION (RESNET-50)

Train	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
	Pr	Rw	Cl	Rw	Cl	Pr	Rw	Pr	Ar	Rw	Ar	Pr	Cl	Rw	Ar	Rw	Ar	Cl	Pr	Cl	Ar	Pr	Ar	Cl	
ResNet-50	50.0	58.0	34.9	58.0	34.9	50.0	46.2	41.9	37.4	41.9	37.4	41.9	31.2	60.4	38.5	60.4	38.5	31.2	59.9	41.2	53.9	59.9	53.9	41.2	46.1
DANN [9]	60.0	68.7	37.8	70.0	42.3	66.3	63.0	59.1	49.4	64.3	54.8	63.2	37.4	71.9	48.0	68.0	58.6	39.7	75.1	41.2	59.2	73.3	63.1	43.4	57.4
MDD [58]	61.6	68.5	38.1	71.4	40.7	67.0	63.9	60.2	47.1	62.3	53.7	61.9	34.7	70.1	46.7	67.9	54.6	35.9	74.5	40.5	61.5	74.3	60.6	40.0	56.5
DRDA	<b>65.6</b>	<b>74.1</b>	<b>50.6</b>	<b>75.8</b>	<b>50.3</b>	<b>72.1</b>	<b>71.6</b>	<b>68.9</b>	<b>60.1</b>	<b>71.4</b>	<b>61.7</b>	<b>68.9</b>	<b>48.2</b>	<b>76.9</b>	<b>59.7</b>	<b>75.7</b>	<b>64.4</b>	<b>51.2</b>	<b>80.7</b>	<b>52.9</b>	<b>70.8</b>	<b>79.3</b>	<b>72.0</b>	<b>52.2</b>	<b>65.6</b>

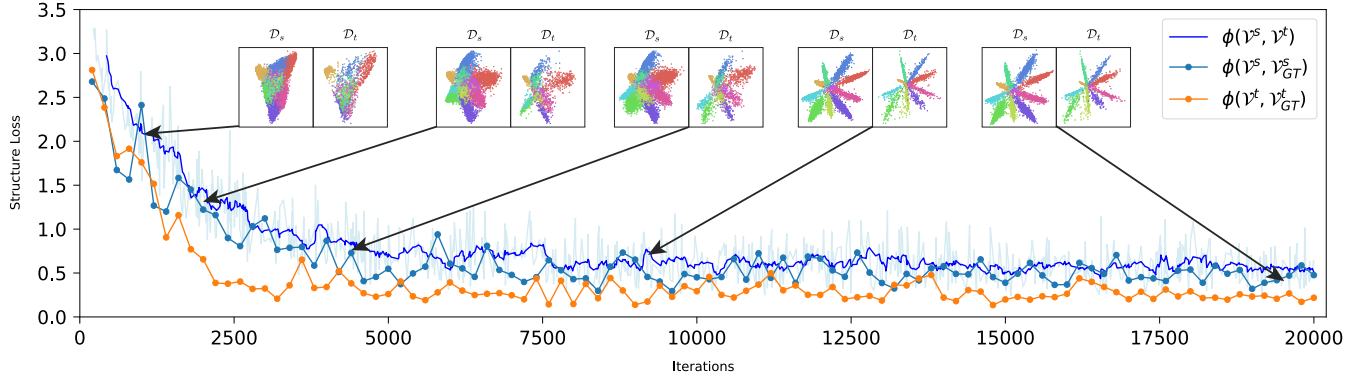


Fig. 3. Training curve and latent features visualization at difference stages, the colors of points indicate instances colors. We can see a clear structure evolution progress while structure loss decreasing. Where  $\phi(\mathcal{V}^s, \mathcal{V}_{GT}^t)$  is computed every validation,  $\mathcal{V}_{GT}^s$  indicates the structure calculated with ground truth labels (resp  $\phi(\mathcal{V}^t, \mathcal{V}_{GT}^t)$ ).  $\phi(\mathcal{V}^s, \mathcal{V}^t)$  is passed through a median filter (original with light blue color) for visualization purpose, best view in color.

discriminative structures regardless of domain heterogeneity. Hence, the domain alignment can be well assured with the help of the radial structure.

4) *Domain generalize UDA*: A further extension to illustrate the utility of DRDA for knowledge abstraction is to extend it to domain-generalized UDA problems, that is, to train the model on one task and to test it on another domain that is different from the source and target domains.

The detailed results on OfficeHome were reported in Table V, where the first row indicates the standard single-to-single UDA task that the models are trained for, and the second row indicates the test domain used only for evaluation. These results indicate that our method is capable of generalizing to domain generalization tasks. The DRDA method performed well in a number of subtasks. In many domains of the test, our method was superior to those classical UDA methods that directly optimize adaptation performance on those domains, even though our method did not incorporate this domain information for training. These results presented here provide evidence for the effectiveness of radial-like structures in discriminative feature modeling. The in-depth explanation is that the alignment using discriminative radial structures forces the network to learn more meaningful features as a result of regularizing its optimization pathway. As a consequence, when the trained model encounters instances from domains that have not previously been encountered, they can also be classified on the basis of their semantic features.

### C. Analysis

In this subsection, we present various experiments that illustrate the intuition behind radial-like structures and demonstrate

the effectiveness of our approach.

1) *Low-dimensional radial structures*: To illustrate the convergence of the structure, we built a simplified LetNet similar to [57] while reducing the bottleneck dimension to 2 and training it with the task MNIST→USPS. We visualize the two dimensional features from different stages of the overall training procedure, with different colored points showing different categories. To facilitate visual understanding, the source domain features and the target domain features were plotted side-by-side in separate plots, and instance points were uniformly sampled from entire datasets with stride 10. In Fig.3, we observe that the features coming from the source and target domains gradually evolve the radial-like structures. It is worth noting that, at the beginning of the process, there is a significant structure discrepancy between the source and target. Then, as the training progresses, the structures of two domains become more and more discriminative (i.e. the features in latent space present a more and more clear radial-like manner). Also, the decreasing of both source structure error  $\phi(\mathcal{V}^s, \mathcal{V}_{GT}^s)$  and target structure error  $\phi(\mathcal{V}^t, \mathcal{V}_{GT}^t)$  indicates these radial-like structures become more and more close to ground-truth ones. As radial-like structures become more reliable and discriminative, domain alignment is expected to be more accurate as well.

2) *Global isometric effects*: To evaluate the proposed Stiefel layer, we implement a network without the Stiefel layer for comparison, denoting as DRDA (w/o Stiefel). The detailed performance results presented in Table I,II,III,IV indicate the importance of the Stiefel layer to the final accuracy. Based on the results of the performance drop, we can confirm the hypothesis that features will be rotated globally and that

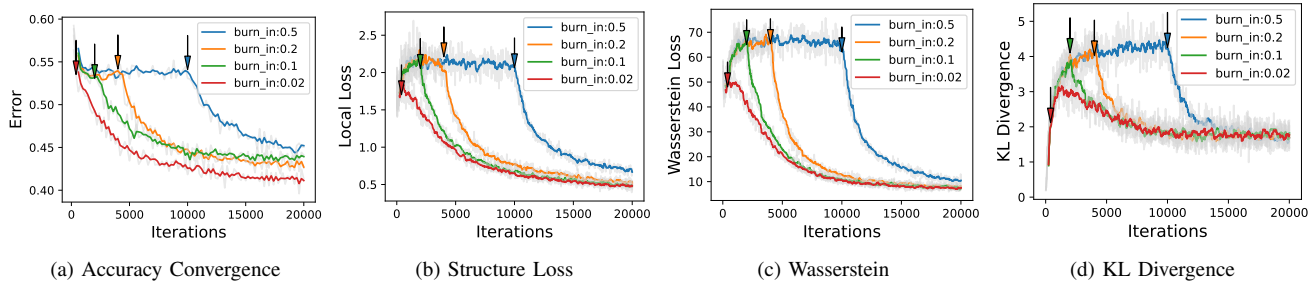


Fig. 4. The component loss along with the train iterations. We use ‘burn\_in’ to indicate the percentage of training progress before adding the structure alignment operations. (a). validation accuracies of Clipart along with training iterations (b).  $\phi(\mathcal{V}^s, \mathcal{V}^t)$  along with training iterations. (c). wasserstein distance between instances and local anchors in the target domain, along with training iterations. (d). The KL divergence between target geometrical label assignments and the classifier label assignments. (best view in color)

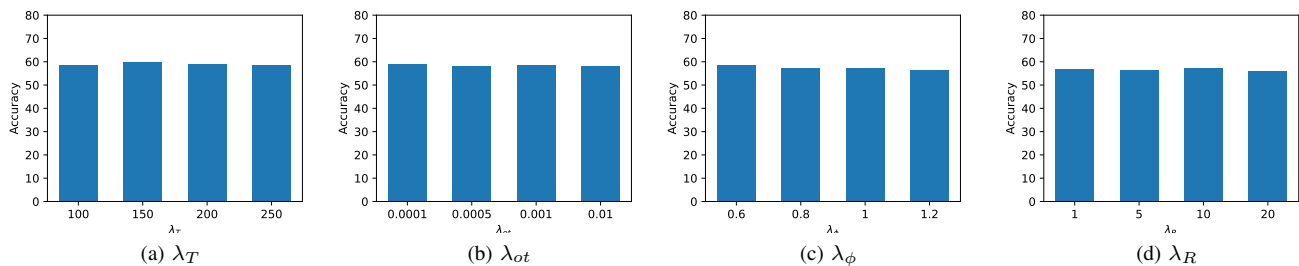


Fig. 5. The effect of different hyperparameters on the performance of OfficeHome dataset. Here we report subtask results on Art→Clipart.

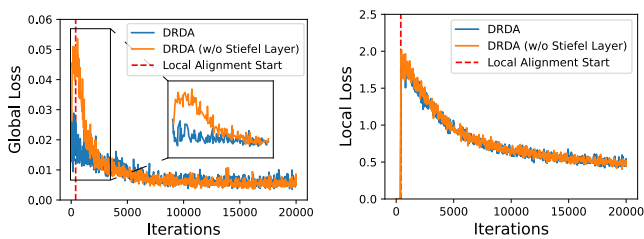


Fig. 6. Global translation distance  $\mathcal{L}_{\text{global}}$  and local structure loss  $\phi(\mathcal{V}^s, \mathcal{V}^t)$ .

the Stiefel layer can reduce the negative effect on domain alignment. Furthermore, we tried to elaborate on the in-depth explanations in Fig.6. As we can see, when we use the Stiefel layer, we can reduce the structure loss to a small range quickly, and the maximum loss values are significantly less than those obtained without the Stiefel layer. This is partly due to the fact that the global rotation difference between the source and target domains is particularly a misleading factor in the calculation of domain discrepancies in the early stages of domain alignment. Furthermore, this misleading factor can have irreversible negative impacts on overall domain alignment. The final accuracy drops for the models without the Stiefel layer also confirmed these irreversible negative impacts. The global rotational component can be easily extracted from the difference between two radial structures when there is a Stiefel layer, which would naturally mitigate such negative impacts in the early stages of domain alignment. Additionally, the results of this study demonstrate the necessity of decoupling global and local transformations when performing alignment and the

Stiefel layer is a suitable choice.

3) *Effects of structures alignment:* To better understand the effects of radial structure local alignment on domain alignment, we perform an ablation study that does not optimize for local structure loss at the beginning of the training period. The detailed results are shown in Fig.4. As we can see accuracy on target increasing during the early learning stage and decreases while training moves on. As we can see obviously when ‘burn\_in:0.5’ the accuracy on target increases during the early learning stage and then gets stuck. Meanwhile, the structure loss, the Wasserstein distance from instances to local anchors, and the KL divergence between geometrical labels and classifier labels were increased. It is clear from these simultaneous losses increasing that the structure of the target domain is crumbling. This is because, with the progress of training, the network gradually learns the common semantic information at the beginning, and then begins to over-fit the data in the source domain. Moreover, this over-fit phenomenon is accompanied by arbitrary distortions to discriminative structures. As shown in Fig.4, stretching the ‘burn\_in’ results in irreversible damage to the final accuracy. By comparing the influences of different ‘burn\_in’ on the rest component losses in Fig.4, we notice that once the structure alignment operations are restored, the corresponding losses drop rapidly. Correspondingly, the test errors on the target domain also decreased rapidly after structure alignment was restored. The results show that our structure alignment can always reconstruct and align discriminative structures, which supports the validity of our model in the domain alignment.

4) *Angular term effects in intra-structure comparison:* When the angular losses are removed from the intra-structure

comparison loss function, denoting as DRDA (w/o Angular), the performance returns to baseline, which indicates that discrepancy based on angular distance between discriminative vectors is very critical. The reason can be two folds. First, the angular loss is more consistent with the formulation of classification. Secondly, in high-dimensional space, the mass of the sphere is primarily concentrated on the shell and the distance between any two point pairs becomes even smaller. Therefore, angular loss confirms that modeling data distribution with radial-like structures that are well suited to angular comparison is an effective strategy.

5) *Effects of optimal transport distance minimization:* To verify the effectiveness of optimal transport distance minimization between instance and local anchors, we conduct the ablation studies by removing this minimization term in training, denoting as DRDA (w/o OT). The detailed results reported in Table I and II illustrate the performance drop when instances are not restricted to being located nearby local anchors. We note that the performance degradation of DRDA (w/o OT) is much smaller than that of DRDA (w/o Angular), indicating that our proposed radial structure based alignment is efficient and robust in domain adaptation even if the structures are not forced to be compact.

6) *Effects of consensus regularization:* We evaluate the proposed classifier regularization terms by implementing the model without regularization loss  $\mathcal{R}(\mathbf{P}, \mathbf{Q})$  denoting as DRDA (w/o  $\mathcal{R}_\phi$ ). The results reported in Table I and II indicate that such a regularization term enhances the performance of domain alignment across all subtasks. It is shown that the consensus regularization between geometrical assignments and classifier assignments can enhance the classification performance of the classifier by encouraging the classifier to admit geometrical assignments.

7) *Parameters sensitivity analysis:* In this section, we conduct sensitivity analysis on the hyper-parameters for our proposed method. The detailed results are shown in Fig.5. Parameters  $\lambda_T$ ,  $\lambda_\phi$ ,  $\lambda_{ot}$  and  $\lambda_R$  are mainly for scaling the loss value. From the observation, we choose  $\lambda_T = 150$  which controls the impacts of global translation between domains. The parameters  $\lambda_{ot}$  balance the radial-like structure compactness and alignment effects, and we find when  $\lambda_{ot}$  is around 0.0001-0.005 the model performance reaches the peaks. The parameters  $\lambda_R$  regularize the agreement between classifier and geometric assignments, when  $\lambda_R = 20.0$  the model performance reaches the peaks. As for structure alignment,  $\lambda_\phi = 3$  provides the best performance. It is also obvious that the performance of the system is quite stable across a wide range of transfer losses when they are ranged in respective orders of magnitude.

8) *The evolution of structures discrepancies:* We also demonstrate the differences of the radial structure between domains with increasing iteration numbers. As illustrated in Fig.7, after adding the structure alignment, the discrepancies  $\phi(\mathcal{V}_{GT}^s, \mathcal{V}_{GT}^l)$  of structures derived from ground truth labels appear to consistently decrease, as well as local differences between structures being minimized. It is evident from the results that our method is able to consistently produce positive alignment with the increasing number of training iterations.

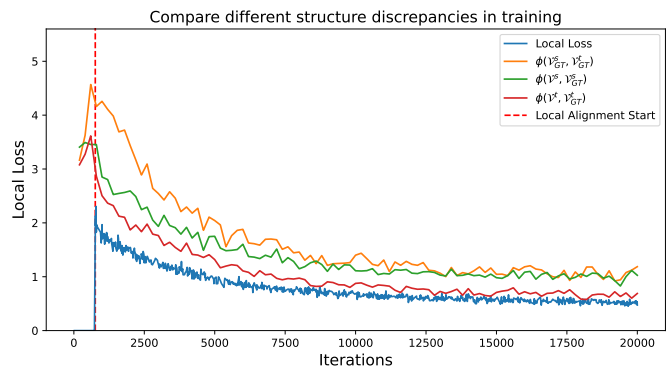


Fig. 7. Radial structures alignment convergence visualization with an increasing number of iterations, where  $\phi(\mathcal{V}^s, \mathcal{V}_{GT}^s)$ ,  $\phi(\mathcal{V}^l, \mathcal{V}_{GT}^l)$ ,  $\phi(\mathcal{V}_{GT}^s, \mathcal{V}_{GT}^l)$  are computed every validation,  $\mathcal{V}_{GT}^s$  indicates the structure calculated with ground truth labels (resp  $\mathcal{V}_{GT}^l$ ).

## V. CONCLUSION

This paper presents a new structure-preserved domain adaptation method, which has two key features: a new discriminative radial structure and a new alignment strategy based on radial structure. The discriminative radial structure preserves both representative and discriminative information in feature distribution. The decoupled global alignment and fine-grained morphological alignment reduce the common domain shifts and conditional domain shifts. Experimental results on several benchmark datasets showed that i) our method consistently outperforms state-of-the-art methods on four types of unsupervised domain adaptation tasks, and ii) our method leads to more superiority when the task is more challenging.

## ACKNOWLEDGMENT

This work is supported by the National Key R&D Program of China(2020YFB1313501), Zhejiang Provincial Natural Science Foundation (LR19F020005), National Natural Science Foundation of China (61972347, T2293723) and the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

## REFERENCES

- [1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
- [2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1-2, pp. 151–175, May 2010.
- [3] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Proceedings of the 30th International Conference on Machine Learning*. PMLR, Feb. 2013, pp. 222–230.
- [4] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1717–1724.
- [5] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.
- [6] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, vol. 9915, pp. 443–450.

- [7] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.
- [8] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.
- [9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.
- [10] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. Lille, France: JMLR.org, Jul. 2015, pp. 97–105.
- [11] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., Dec. 2016, pp. 136–144.
- [12] M. Long, Z. CAO, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.
- [13] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 1081–1090.
- [14] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, "Generalization and equilibrium in generative adversarial nets (gans)," in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 224–232.
- [15] H. Tang and K. Jia, "Discriminative adversarial domain adaptation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5940–5947, Apr. 2020.
- [16] V. K. Kurmi and V. P. Namboodiri, "Looking back at Labels: A Class based Domain Adaptation Technique," in *2019 International Joint Conference on Neural Networks (IJCNN)*. Budapest, Hungary: IEEE, Jul. 2019, pp. 1–8.
- [17] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.
- [18] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.
- [19] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, vol. 9911, pp. 499–515.
- [20] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008.
- [21] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, Jun. 2012, pp. 1627–1634.
- [22] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. Buenos Aires, Argentina: AAAI Press, Jul. 2015, pp. 4119–4125.
- [23] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," Dec. 2014.
- [24] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1406–1415.
- [25] L. V. Kantorovich, "On the translocation of masses," *Journal of Mathematical Sciences*, vol. 133, no. 4, pp. 1381–1382, Mar. 2006.
- [26] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.
- [27] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3722–3731.
- [28] 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*, Jun. 2018, pp. 3723–3732.
- [29] Y.-W. Luo and C.-X. Ren, "Conditional bures metric for domain adaptation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 13 984–13 993.
- [30] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
- [31] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 5423–5432.
- [32] X. Gu, J. Sun, and Z. Xu, "Spherical space domain adaptation with robust pseudo-label loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.
- [33] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.
- [34] M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren, "Enhanced Transport Distance for Unsupervised Domain Adaptation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 13 933–13 941.
- [35] Y.-W. Luo, C.-X. Ren, D.-Q. Dai, and H. Yan, "Unsupervised domain adaptation via discriminative manifold propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1653–1669, Mar. 2022.
- [36] C.-H. Lin, M. Azabou, and E. Dyer, "Making transport more robust and interpretable by moving data through a small number of anchor points," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 6631–6641.
- [37] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 1989–1998.
- [38] Y. Xu, B. Du, L. Zhang, Q. Zhang, G. Wang, and L. Zhang, "Self-ensembling attention networks: Addressing domain shift for semantic segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5581–5588, Jul. 2019.
- [39] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 6502–6509, Apr. 2020.
- [40] H. Liu, M. Long, J. Wang, and M. Jordan, "Transferable adversarial training: A general approach to adapting deep classifiers," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 4013–4022.
- [41] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao, "Learning Disentangled Semantic Representation for Domain Adaptation," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 2060–2066.
- [42] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 5102–5112.
- [43] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. New York, NY, USA: JMLR.org, Jun. 2016, pp. 507–516.
- [44] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, Jun. 2012, pp. 1275–1282.
- [45] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.
- [46] J. Li, J. Zhao, and K. Lu, "Joint feature selection and structure preservation for domain adaptation," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. New York, New York, USA: AAAI Press, Jul. 2016, pp. 1697–1703.



- [47] H. Liu, M. Shao, Z. Ding, and Y. Fu, "Structure-preserved unsupervised domain adaptation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 799–812, Apr. 2019.
- [48] V. Titouan, R. Flamary, N. Courty, R. Tavenard, and L. Chapel, "Sliced gromov-wasserstein," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.
- [49] M. Kochurov, R. Karimov, and S. Kozlukov, "Geoopt: Riemannian optimization in pytorch," Jul. 2020.
- [50] F. Mémoli, "Gromov—wasserstein distances and the metric approach to object matching," *Foundations of Computational Mathematics*, vol. 11, no. 4, pp. 417–487, Aug. 2011.
- [51] C. Villani, *Optimal Transport*, ser. Grundlehren Der Mathematischen Wissenschaften, M. Berger, B. Eckmann, P. de la Harpe, F. Hirzebruch, N. Hitchin, L. Hörmander, A. Kupiainen, G. Lebeau, M. Ratner, D. Serre, Y. G. Sinai, N. J. A. Sloane, A. M. Vershik, and M. Waldschmidt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 338.
- [52] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2292–2300, 2013.
- [53] H. Yang and E. G. Tabak, "Clustering, factor discovery and optimal transport," *Information and Inference: A Journal of the IMA*, vol. 10, no. 4, pp. 1353–1387, 2021.
- [54] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. Lille, France: JMLR.org, Jul. 2015, pp. 1180–1189.
- [55] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. Sydney, NSW, Australia: JMLR.org, Aug. 2017, pp. 2208–2217.
- [56] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-Adversarial Domain Adaptation," vol. 32, no. 1, Apr. 2018.
- [57] M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3521–3528, Apr. 2020.
- [58] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning*, 2019, pp. 7404–7413.
- [59] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, Sep. 2010, pp. 213–226.
- [60] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.
- [61] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2066–2073.
- [62] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 402–410.



**Jun Wen** received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2020. He is currently a Postdoctoral Research Fellow at the Harvard Medical School. His research interests include transfer learning and biomedical informatics.



**Siheng Chen** is a tenure-track associate professor of Shanghai Jiao Tong University and Co-PI at Shanghai AI Laboratory. Dr. Chen received his doctorate from Carnegie Mellon University. Dr. Chen's work on sampling theory of graph data received the 2018 IEEE Signal Processing Society Young Author Best Paper Award. His co-authored paper on structural health monitoring received ASME SHM/NDE 2020 Best Journal Paper Runner-Up Award and another paper on 3D point cloud processing received the Best Student Paper Award at 2018 IEEE Global

Conference on Signal and Information Processing. Dr. Chen contributed to the project of scene-aware interaction, winning MERL President's Award. His research interests include collective intelligence, autonomous driving and graph neural networks.



**Linchao Zhu** (Member, IEEE) received the B.E. degree from Zhejiang University, China, in 2015, and the Ph.D. degree in computer science from the University of Technology Sydney, Australia, in 2019. He is a Research Professor with the College of Computer Science and Technology, Zhejiang University, China. His research interests are video analysis and understanding.



**Nenggan Zheng** received the bachelor's and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2002 and 2009, respectively. He is currently a Full Professor in computer science with the Academy for Advanced Studies, Zhejiang University. His research interests include artificial intelligence, brain-computer interface, and embedded systems.



**Zenan Huang** received B.E. degree in the Computer Science from Zhejiang University of Technology, in 2018. He is currently pursuing the Ph.D degree in the College of Computer Science and Technology, Zhejiang University. His research interests include computer vision, causality, and machine learning.