DOMAIN-INVARIANT REPRESENTATION LEARNING WITH GLOBAL AND LOCAL CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we give an analysis of the existing representation learning framework of unsupervised domain adaptation and show that the learned feature representations of the source domain samples are with discriminability, compressibility, and transferability. However, the learned feature representations of the target domain samples are only with compressibility and transferability. To address this challenge, we propose a novel framework and show from the information theory view that this framework can effectively improve the discriminability of the target domain sample representation. We also propose a method, namely domain-invariant representation learning with global and local consistency (RLGLC), under this framework. In particular, to maintain the global consistency, RLGLC proposes a new metric called asymmetrically-relaxed Wasserstein of Wasserstein distance (AR-WWD), AR-WWD can not only extract the transferability and compressibility of the feature representation of two domains, but also correlates well with human perception. To impose the local consistency structures, we propose a regularized contrastive loss, which can not only keep as much as possible predictive information contained in the feature representation of the target domain, but also alleviates the problem that semantically similar instances are undesirable pushed apart in training processing. Finally, we verify the effectiveness of RLGLC from both theoretical analyses on Bayes error rate and experimental validation on several benchmarks.

1 INTRODUCTION

Despite achieving impressive successes, the standard machine learning models require that the training dataset and the test dataset are drawn from the same distribution. However, in practical applications, deviations in the data collection process and a limited amount of labeled training samples can cause the problem of feature covariate shift. If models trained on the labeled source domain are applied to an unlabeled target domain with a different distribution, the generalization is not guaranteed. In this paper, we expect to learn a model that can handle the distribution gap between two domains, which is the problem that unsupervised domain adaptation (UDA) focuses on.

Adversarial-based representation learning methods for UDA have gained remarkable performance from the theoretical findings to algorithms (Ganin et al., 2016). These methods mainly focus on exactly matching the distributions between two domains. Informally, these approaches minimize the source domain classification risk and the distance between the two domain distributions in the latent space. While current works mainly explore how to measure the distance between two distributions and put forward many original and effective distribution discrepancy metrics, e.g., the Wasserstein distance (Shen et al., 2018; Arjovsky et al., 2017), KL-divergence (Ganin et al., 2016; Zhang et al., 2019a; 2020), and so on, in this paper, we propose to understand these methods from the information theory perspective and design a new learning framework to make the learned feature representations to be with discriminability, compressibility, and transferability (see subsection 4.1), which has also been proven to help reduce the upper bound of the Bayesian error rate.

Specifically, minimizing the distance between two distributions can make the learned feature representations of both two domain samples be with transferability and compressibility. Minimizing source domain classification risk can make the learned feature representations of source domain samples be with discriminability. However, we do not optimize a certain item to explicitly imporve the discriminability of the target domain sample feature representations. Therefore, for the proposed

learning framework, the purpose is to improve the discriminativeness of the target domain sample representations. So, we propose to maximize the MI between the target domain and a generated self-supervised signal. We give theoretical analysis to show that this framework is effective in improving the discriminability for targe domain sample representations.

Based on the proposed framework, we derive our proposed method: domain-invariant representation learning with global and local consistency (RLGLC). RLGLC consists of two parts, including the global consistent module (GCM) and the local consistent module (LCM). For GCM, we proposed a new metric called asymmetrically-relaxed Wasserstein of Wasserstein distance (AR-WWD). Different from the Wasserstein distance, AR-WWD constrains the distribution of source domain contained in the distribution of target domain by an inequality constraint. Then, the ground metric in AR-WWD is selected as the WD. The Wasserstein ground metric can capture the knowledge of an image at the pixel level and is known to correlate well with human perception in computing the similarity between images (Engquist & Yang, 2018; Puthawala et al., 2019). For LCM, we propose a novel loss called regularized contrastive loss (RCL). Different from contrastive loss, RCL, to a certain extent, alleviates the problem that some semantically similar instances are undesirable pushed apart through a regularization item.

To verify the effectiveness of our proposed RLGLC, based on the Bayes error rate, we give the generalization classification error for the learned feature representation and show that both modules can reduce classification error. Experiments on several standard benchmarks also show that the proposed RLGLC is effective. The major contributions of this paper are three-fold: 1) Based on information theory, we demonstrate that the existing learning framework can not guarantee that the learned sample representations of target domain to be with discriminability. 2) We propose a new learning framework and conduct the theoretical analysis to show that maximizing a certain component in this framework can ensure that the learned feature representation of the target domain sample is discriminative. 3) Based on the proposed framework, we derive a new method which consists of a new distribution metric and a regularized contrastive loss. Also, we provide theoretical analysis on the Bayes error rate to show that the proposed method is effective.

2 Related works

Unsupervised domain adaptation aims to transfer knowledge learned from a labeled source domain to a related unlabeled target domain (Kumar et al., 2020; Dhouib et al., 2020; Balaji et al., 2020; Cui et al., 2020b; Combes et al., 2020; Cui et al., 2020a; Hu et al., 2020; Kang et al., 2020; Tang et al., 2020). Remarkable advances have been achieved in UDA, especially these representation learning based methods. The main idea behind these methods is to align the distributions of the source domain and target domain. Therefore, many works are proposed to design an effective metric to measure the differences between distributions. Maximum mean discrepancy (Gretton et al., 2012; Tzeng et al., 2014) is a nonparametric metric that measures the divergence of two distributions in the reproducing kernel Hilbert space. Deep correlation alignment (Sun et al., 2016) aligns two distributions by minimizing the difference in the second-order statistics of the two distributions. Domain Adversarial Neural Network (Ganin et al., 2016) and S-disc (Kuroki et al., 2019) align distributions by minimizing KL-divergence. Wasserstein distance guided representation learning (Shen et al., 2018) introduces the Wasserstein distance for domain adaptation to make the training process stable. Sliced Wasserstein discrepancy (Lee et al., 2019) proposes to utilize the sliced Wasserstein distance to accelerate the training process. Margin disparity discrepancy (Zhang et al., 2019a) aligns distributions based on the scoring function and margin loss. Domain-Symmetric Networks (Zhang et al., 2020) propose a multi-class scoring disagreement divergence. An asymmetrically-relaxed distribution alignment is proposed in (Wu et al., 2019). Reliable weighted optimal transport (Xu et al., 2020) proposes a novel shrinking subspace reliability and weighted optimal transport strategy for UDA. In (Li et al., 2020), an enhanced transport distance for UDA is proposed. Different from these methods, this paper starts from information theory and aims to make the learned feature representations to be with discriminability, compressibility, and transferability.

On par with the domain adaptation algorithms findings, there are rich advances in the domain adaptation theoretical findings. In (Mansour et al., 2009; Ben-David et al., 2010), a rigorous learning bound is proposed for UDA. Then, a series of theories have been proposed to extend this theory (Mohri & Medina, 2012; Germain et al., 2013; Cortes et al., 2015). Based on reproducing kernel

Hilbert space, Redko et al. (2017) proposes to bound the target error by the WD. Then, Shen et al. (2018) proposes to use the Kantorovich-Rubinstein dual formulation to obtain a generalization bound. In (Wu et al., 2019), a good target domain performance is demonstrated theoretically under the setting of relaxed alignment. In (Zhang et al., 2019a), based on Rademacher complexity, a margin-aware generalization bound is provided to bridge the gaps between the theories and algorithms. Then, Zhang et al. (2020) extends the generalization bound provided in (Zhang et al., 2019a) and can better explain the effectiveness of UDA related to multiple classes. In (Zhao et al., 2019), both upper and lower bounds for the target and joint errors are provided. In (Dhouib et al., 2020), based on large margin separation, a new theoretical analysis is provided to uniform the margin, adversarial learning, and domain adaptation. In (Kumar et al., 2020), the self-train is proved to be effective for larger distribution shifts. Different from these theoretical findings, this paper is motivated by the information theory and Bayes error rate.

3 PRELIMINARIES

3.1 PROBLEM DEFINITION AND NOTATIONS

This paper focuses on the classification task of unsupervised domain adaptation. Let $\mathcal{X}$ denote the input feature space and $\eta : X \to Y$ be the domain-invariant ground truth labeling function, where $X \in \mathcal{X}$, and Y is the label. Let P_s be the input distribution over X for the source domain and P_t be the input distribution over X for the target domain. Let $\mathcal{Z}$ be a latent space and $\Phi : X \to Z$ be a class of feature extractors, where $Z \in \mathcal{Z}$. For a domain $u \in \{s, t\}$, $P_u^\phi(Z) = P_u\left(\phi^{-1}(Z)\right)$ represents the induced probability distribution over $\mathcal{Z}$. For a given $Z \in \mathcal{Z}$, let $\phi_u(X|Z)$ be the induced conditional distribution over $\mathcal{X}$ that satisfies $\int P_u^\phi(Z)\,\phi_u(X|Z)dZ = P_u(X)$, where $\phi \in \Phi$. Denote $\Psi : Z \to Y$ as a class of prediction functions. Then, the learned classifier can be represented as $\psi(\phi(X))$, where $\psi \in \Psi$. The goal is to learn a classifier that can minimize the following expected target risk:

$$R_{P_t}(X) = \int P_t(X)\,|\eta(X) - \psi(\phi(X))|\,dX. \tag{1}$$

where $R_{P_t}(X)$ denotes the expected target risk. Also, we denote a sample in the support set of $P_u(X)$ as X_u and a sample in the support set of $P_u^\phi(Z)$ as Z_u.

3.2 REPRESENTATION LEARNING FOR UNSUPERVISED DOMAIN ADAPTATION

The framework of representation learning based methods is divided into three parts including a metric to measure the distance of two distributions in the latent space $\mathcal{Z}$, a loss function to measure the source risk, and a regularization term. The whole objective function is formulated as

$$\min_{\phi,\psi} D\left(P_s^\phi(Z), P_t^\phi(Z)\right) + L_{cl}\left(\psi(\phi(X_s)), \eta(X_s)\right) + \Delta \tag{2}$$

where $D\left(P_s^\phi(Z), P_t^\phi(Z)\right)$ is a metric to measure the difference between two distributions in the latent space, L_{cl} is the classification loss function, and Δ is the regularization term. In this paper, we consider the Wasserstein distance as the baseline metric. Specifically, for $P_s^\phi(Z), P_t^\phi(Z)$, the corresponding support sets are denoted as Σ_s, Σ_t, the p-th Wasserstein distance is defined as

$$W_p\left(P_s^\phi(Z), P_t^\phi(Z)\right) = \left(\inf_{\mu(Z_s,Z_t)\in\Pi(Z_s,Z_t)} \int c(Z_s, Z_t)^p d\mu\right)^{\frac{1}{p}}, \tag{3}$$

where $Z_s \in \Sigma_s, Z_t \in \Sigma_t$, $c(Z_s, Z_t)$ is the distance of two patterns, and $\Pi(Z_s, Z_t)$ is the set of all joint distributions $\mu(Z_s, Z_t)$ that satisfies $P_s^\phi = \int_{Z_t} u(Z_s, Z_t)dZ_t, P_t^\phi = \int_{Z_s} u(Z_s, Z_t)dZ_s$.

4 METHODOLOGY

In this section, we give an analysis to the existing RL-based domain adaptation method from the perspective of information theory to motivate the proposed representation learning framework and derive the proposed method called domain-invariant representation learning with global and local consistency (RLGLC).

4.1 INFORMATION-THEORETICAL BASED ANALYSIS

From the perspective of the information bottleneck principle (Tishby et al., 2000), a good feature representation not only contains as little task-independent information as possible, but also contains as much task-related information as possible. From the perspective of the domain adaptation (Chen et al., 2019), a good feature representation is not only discriminable for the downstream tasks, but also is transferable between different domains. To bridge the gap between these two concepts, the represetations learned in domain adaptation should satisfy three properties: discriminability, compressibility, and transferability.

Definition 4.1. *Discriminability: A representation Z_u of X_u is discriminant for the label Y if and only if $I\left(X_u; Y\left|Z_u\right.\right) = I\left(Z_u; Y\left|X_u\right.\right) = 0$.*

Definition 4.2. *Compressibility: A representation Z_u of X_u is minimal for the label Y if and only if the task-irrelevant information equals to zero: $I\left(X_u; Z_u\left|Y\right.\right) = 0$.*

Definition 4.3. *Transferability: A representation Z_u of X_u is transferable if and only if the domain-specific information equals to zero: $I\left(Z_u; Z_s\left|X_t\right.\right) = I\left(X_t; Z_t\left|X_s\right.\right) = 0$.*

From **Definition 4.1**, we can obtain that $I\left(X_u; Z_u\right) = I\left(X_u; Y\right) = I\left(Z_u; Y\right)$ (see Appendix Proposition A.1), which means that a discriminant representation Z can predict Y at least as accurately as the original data X. Base on the chain rule of mutual information, we can obtain $I\left(X_u; Z_u\right) = I\left(X_u; Z_u\left|Y\right.\right) + I\left(Z_u; Y\right)$, we can see that $I\left(X_u; Z_u\left|Y\right.\right)$ represents the information in Z_u that is not predictive of Y and $I\left(Z_u; Y\right)$ represents the predictive information. Therefore, from **Definition 4.2**, we can obtain that a minimal representation contains less superfluous information.

A basic assumption in the multi-view representation learning field (Xu et al., 2013) shows that the shared information between the two views contains all the information related to the task. In this paper, we mainly focus on solving the covariance shift problem of the domain adaptation. We assume that $P_s\left(X\right) \neq P_t\left(X\right)$ and $P\left(Y\left|X_s\right.\right) = P\left(Y\left|X_t\right.\right)$. Therefore, we can safely regard the two domains as two views and have that a representation Z containing all information shared between both domains is able to contain the necessary label information: $I(Z_s; Z_t) = I(Z_s Z_t; Y) = I(Z_s; Y) = I(Z_t; Y)$ (see Appendix Proposition A.2), where $Z_s Z_t$ is the observation of the joint distribution of P_s and P_t. Furthermore, a transferable Z should eliminate the domain-specific details and reduce the sensitivity of the representation to domain-changes. **Definition 4.3** states that the transferability of the learned representation is mainly caused by eliminate the domain-specific information.

Based on the three definitions, we can factorize the mutual information into two components, that is:

$$\begin{aligned} I\left(X_s; Z_s\right) &= I\left(X_s; Z_s\left|Y\right.\right) + I\left(Z_s; Y\right) = I\left(X_s; Z_s\left|X_t\right.\right) + I\left(X_t; Z_s\right) \\ I\left(X_t; Z_t\right) &= I\left(X_t; Z_t\left|Y\right.\right) + I\left(Z_t; Y\right) = I\left(X_t; Z_t\left|X_s\right.\right) + I\left(X_s; Z_t\right) \end{aligned} \tag{4}$$

Because we have that $I\left(X_t; Z_s\right) = I\left(Z_s; Y\right)$ and $I\left(X_s; Z_t\right) = I\left(Z_t; Y\right)$, so we can easily obtain that $I\left(X_s; Z_s\left|Y\right.\right) = I\left(X_t; Z_t\left|Y\right.\right)$ and $I\left(X_t; Z_s\right) = I\left(X_s; Z_t\right)$.

For objective function 2, the first term is to align the distributions of the source and target domains in the latent space $\mathcal{Z}$, the second term is to make the learned feature representation related to the label. To this end, we have:

Theorem 4.1. *Suppose the representations for source domain and target domain are obtained by minimizing the objective function (2). Then, Z_s is with the discriminability, compressibility, and transferability, while Z_t is only with compressibility and transferability.*

The proof can be seen in Appendix Theorem A.1. From **Theorem** 4.1, we can obtain that the discriminability of the sample representation of the target domain is not fully considered. Therefore, we propose a novel representation learning framework to address this problem.

4.2 THE PROPOSED INFORMATION-THEORETICAL BASED FRAMEWORK

First, for each sample in the target domain, a stochastic data augmentation operation is implemented to transform any given sample in the target domain resulting in a new views called target self-supervised signal. We denote the distribution of the target self-supervised signal as $P_{tss}\left(X\right)$, and the induced distribution in the latent space as $P_{tss}^{\phi}\left(Z\right)$. Also, we denote a sample in the support set of $P_{tss}\left(X\right)$

or $P_{tss}^{\phi}(Z)$ as X_{tss} or Z_{tss}. Then, based on the objective 2, the proposed framework can be written as:

$$\min_{\phi,\psi} D\left(P_s^{\phi}, P_t^{\phi}\right) + D\left(P_t^{\phi}, P_{tss}^{\phi}\right) - I\left(Z_t; Z_{tss}\right) + L_{cl}\left(\psi\left(\phi\left(X_s\right)\right), \eta\left(X_s\right)\right) + \Delta \tag{5}$$

where Δ is the regularization term, D is the distribution metric, and I is the mutual information (MI). Compared the objective 5 with the objective 2, there are two additional items including $D\left(P_t^{\phi}, P_{tss}^{\phi}\right)$ and $I\left(Z_t; Z_{tss}\right)$. The motivations behind these two items are as following:

Theorem 4.2. *Maximizing $I\left(Z_t; Z_{tss}\right)$ while minimizing $D\left(P_t^{\phi}, P_{tss}^{\phi}\right)$ can make the learned feature representation of the target domain sample be with discriminability.*

The proof can be seen in Appendix Theorem A.2. From **Theorem** 4.2, we can see that the learned feature representations of the source domain and target domain samples are with discriminability, compressibility, and transferability. Based on the proposed representation learning framework, we propose a novel learning method as follows.

4.3 Domain-invariant representation learning with global and local consistency

4.3.1 Learning with global consistency module

The global consistency module aims to align the source domain distribution P_s^{ϕ} with the target domain distribution P_t^{ϕ} and align the target domain distribution P_t^{ϕ} with the target self-supervised signal P_{tss}^{ϕ} in the learned latent space. This module relaxes the exact aligning constraint to a loose one by requiring the support set of P_s^{ϕ} is contained in that of P_τ. Motivated by (Wu et al., 2019), this can be achieved by the inequality constraint: $\sup_{Z \in \mathcal{Z}} \frac{P_t^{\phi}(Z)}{P_s^{\phi}(Z)} \leq 1 + \beta$. The 'global' is corresponding to the whole distribution, not an instance. To achieve this, we propose a novel metric called asymmetrically-relaxed Wasserstein of Wasserstein distance (AR-WWD).

For AR-WWD, the first 'Wasserstein' refers to the Wasserstein distance (WD) between the probability distributions of two domains on image space. The second 'Wasserstein' refers to using the WD as the ground metric. From (Dukler et al., 2019), we can know that using WD as the ground metric is closely related to human perception for natural images, e.g., being robust to translations and rotations.

Specifically, we consider $Z \in \mathbb{R}^{c \times w \times h}$, where w, h, and c are the width, height, and the number of feature channels. For AR-WWD, we regard the learned representation as a probability distribution over pixels. Therefore, the ground metric is defined as

$$d\left(Z_s, Z_t\right) = W_{q,d_\Omega}\left(Z_s, Z_t\right) = \left(\inf_{\mu(z_s,z_t) \in \Pi(z_s,z_t)} \int d_\Omega(z_s, z_t)^q d\mu\right)^{\frac{1}{q}} \tag{6}$$

where $\Pi\left(z_s, z_t\right)$ denotes the set of all joint distributions $\mu\left(z_s, z_t\right)$ that satisfies $Z_s = \int_{z_t} \mu\left(z_s, z_t\right) dz_t$, $Z_t = \int_{z_s} \mu\left(z_s, z_t\right) dz_s$. Z_s and Z_t are representations of source and target domain samples, respectively. z_s and z_t are the pixels in Z_s and Z_t, respectively. $d_\Omega\left(z_s, z_t\right)$ is defined as the spatial distance between two-pixel locations. We set $p = 1$ and $q = 2$. Then, AR-WWD is defined as

$$W_{1,W_{2,d_\Omega}}\left(P_s^{\phi}, P_t^{\phi}\right) = \inf_{\mu(Z_s,Z_t) \in \Pi(Z_s,Z_t)} \int W_{2,d_\Omega}\left(Z_s, Z_t\right) d\mu \tag{7}$$

where $\Pi\left(Z_s, Z_t\right)$ denotes the set of all joint distributions $\mu\left(Z_s, Z_t\right)$ that satisfies $P_s^{\phi} = \int_{Z_t} u\left(Z_s, Z_t\right) dZ_t, \left(1+\beta\right) P_t^{\phi} \geq \int_{Z_s} u\left(Z_s, Z_t\right) dZ_s$.

To this end, the duality of AR-WWD can be defined as

$$\begin{aligned} &W_{1,W_{2,d_\Omega}}\left(P_s^{\phi}, P_t^{\phi}\right) = \sup_{f \in C(Z)} E_{Z \sim P_s^{\phi}} f\left(Z\right) - \left(1+\beta\right) E_{Z \sim P_t^{\phi}} f\left(Z\right) \\ &s.t. \int_\Omega \left\| \nabla_z \delta_Z f\left(Z\left(z\right)\right) \right\|_{d_\Omega}^2 Z\left(z\right) dz \leq 1 \end{aligned} \tag{8}$$

where $C : Z \to \mathbb{R}$ represents all of the functions that are continuous and bounded everywhere, ∇ is the gradient operator in pixel space Ω, $\beta > 0$ is the fixed, and δ is the L^2 gradient in latent space $\mathcal{Z}$.

For discrete case, we can rewrite objective 8 as

$$\sup_{f\in C(Z)}\left[\frac{1}{N}\sum_{i=1}^{N}f\left(Z_s\right)-\frac{1+\beta}{N}\sum_{i=1}^{N}f\left(Z_t\right)+\frac{\lambda}{N}\sum_{i=1}^{N}\left(\left\|\mathrm{grad}f\left(Z_{st}\right)\right\|_W-1\right)^2\right] \tag{9}$$

where N is the number of samples, Z_{st} is the data sampled from distribution P_{st}^{ϕ}, P_{st}^{ϕ} is the distribution taken to be the uniform on "Euclidean" lines connecting points drawn from P_s^{ϕ} and P_t^{ϕ}, λ is the hyperparameter, and $\left\|\mathrm{grad}f\left(Z_{st}\right)\right\|_W$ is denoted as

$$\left\|\mathrm{grad}f\left(Z_{st}\right)\right\|_W=\sqrt{\sum_{i.j\in E}\omega_{ij}\left(\nabla_{Z_j}f\left(Z_{st}\right)-\nabla_{Z_i}f\left(Z_{st}\right)\right)^2\frac{Z_i/d_i+Z_j/d_j}{2}} \tag{10}$$

where $G=(V,E,\omega)$ is a defined pixel space graph, E is the edge set, $V=\{1,...,n\}$ is the vertex set $\left(e.g.,\, for Z\in\mathbb{R}^{c\times w\times h}, n=c\times w\times h\right)$, ω is a symmetric matrix of weights associated with the edges which define a ground metric of pixels, $N\left(i\right)=\left\{j\in V:(i,j)\in E\right\}$ is the neighborhood of node i, and $d_i=\sum_{j\in N(i)}\omega_{ij}\Big/\sum_{i=1}^{n}\sum_{i'\in N(i)}\omega_{ii}$. For a pixel of Z, we only consider the pixels in the fixed-size window area centered on it as neighbors. Similarly, we can give the formulation of the item $W_{1,W_{2,d_\Omega}}\left(P_t^{\phi},P_{tss}^{\phi}\right)$.

4.3.2 Learning with local consistency module

The local consistency module is to maximize the MI between the target domain P_t^{ϕ} and the target self-supervised signal P_{tss}^{ϕ} obtained by transforming each sample in the target domain into an augmented one. The MI is achieved by contrastive loss (Oord et al., 2018; Chen et al., 2020; Chen & Li, 2020). The local refers that the contrastive loss is implemented in an instance-based manner. The main idea is to encourage the learned feature for positive pairs to be similar while pushing features from the randomly sampled negative pairs apart. To achieve this, we propose a novel loss called regularized contrastive loss (RCL).

In our setting, a sample and its corresponding augmented one are regarded as positive pairs and the remaining samples as negative samples. One challenge is that negative samples include the semantically similar instance, directly minimizing the contrastive loss could lead to some semantically similar instances are undesirably pushed apart.

Intuitively, minimizing contrastive loss makes the similarity of positive pairs to 1 and the similarity of negative pairs to 0. One way to alleviate the mentioned challenge is to constraint the distribution of negative pair similarities to fill the entire $[0,1]$ interval. Because the similarity of semantically similar pairs are is more likely to appear near 1. To achieve this constraint, the proposed RCL learns representations with a regularization item.

Specifically, given a minibatch of N examples from the target domain, we transform each sample to its augmented one. So, we obtain $2N$ samples within a minibatch. We denote this as $Z_{mb}=\{Z_1,...,Z_{2N}\}$ in the latent space. Then the proposed RCL is presented as

$$L_{RCL}\left(P_t^{\phi},P_{tss}^{\phi}\right)=\sum_{i=1}^{2N}-\log\frac{\exp\left(\mathrm{sim}\left(Z_i,\hat{Z}_i\right)\Big/\tau\right)}{\sum_{k=1}^{2N}\mathbb{1}_{k\neq i}\exp\left(\mathrm{sim}\left(Z_i,Z_k\right)/\tau\right)}+\alpha\mathrm{Re}\left(Z_{mb}\right) \tag{11}$$

where α, τ are two temperature hyperparameter, $\hat{Z}_i$ is the augmentation corresponding to Z_i, and $\mathbb{1}_{[ii\neq i]}$ is an indicator function evaluating to 1 if $k\neq i$. We assume $P\left(\mathrm{sim}\left(Z_i,Z_j\right)\right)=\exp\left(\mathrm{sim}\left(Z_i,Z_j\right)\Big/\sqrt{d}\right)\Big/\sum_{k,l=1}^{2N}\exp\left(\mathrm{sim}\left(Z_k,Z_l\right)\Big/\sqrt{d}\right)$ and $Q\left(\mathrm{sim}\left(Z_i,Z_j\right)\right)=1/2N$, where d is the dimension of Z_i after being flattened into a vector. Then, we have $\mathrm{Re}\left(Z_{mb}\right)=KL\left(Q\left\|P\right.\right)\doteq\ln\sum_{i,j=1}^{2N}e^{\mathrm{sim}(Z_i,Z_j)/\sqrt{d}}-\frac{1}{4N^2}\sum_{i,j=1}^{2N}\mathrm{sim}\left(Z_k,Z_l\right)\Big/\sqrt{d}$. Note that minimizing the regularization item $\mathrm{Re}\left(Z_{mb}\right)$ can lead to the distribution of negative pair similarities to fill the entire $[0,1]$ interval.

4.3.3 The proposed objective

Based on the global and local consistency modules, we present the proposed RLGLC. RLGLC first maps the input data into the latent space $\mathcal{Z}$ by a projection function ϕ. Then, classifier ψ is learned

Table 1: Accuracy (%) on Office-31 dataset for unsupervised domain adaptation (ResNet-50)

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50	68.4±0.2	96.7±0.1	99.3±0.1	68.9±.2	62.5±0.3	60.7±0.3	76.1
DAN	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
DANN	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
JAN	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
GTA	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
CDAN	93.1±0.2	98.2±0.2	**100.0±0.0**	89.8±0.3	70.1±0.4	68.0±0.4	86.6
CDAN+E	94.1±0.1	98.6±0.1	**100.0±0.0**	92.9±0.2	71.0±0.3	69.3±0.3	87.7
BSP+DANN	93.0±0.2	98.0±0.2	**100.0±0.0**	90.0±0.4	71.9±0.3	73.0±0.3	87.7
BSP+CDAN	93.3±0.2	98.2±0.2	**100.0±0.0**	93.0±0.2	73.6±0.3	72.6±0.3	88.5
ADDA	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
MCD	88.6±0.2	98.5±0.1	**100.0±0.0**	92.2±0.2	69.5±0.1	69.7±0.3	86.5
MDD	94.5±0.3	98.4±0.1	**100.0±0.0**	93.5±0.2	74.6±0.3	72.2±0.1	88.9
SymmNets	94.2±0.1	98.8±0.0	**100.0±0.0**	93.5±0.3	74.4±0.1	73.4±0.2	89.1
GVB-GD	94.8±0.5	98.7±0.3	**100.0±0.0**	95.0±0.4	73.4±0.3	73.7±0.4	89.3
ETD	92.1± -	**100.0± -**	**100.0± -**	88.0± -	71.0± -	67.8± -	86.2
SRDC	95.7±0.2	99.2±0.1	**100.0±0.0**	95.8±0.2	76.7±0.3	77.1±0.1	90.8
RLGC	95.1±0.3	98.6±0.3	**100.0±0.0**	95.2±0.3	74.1±0.2	76.2±0.2	89.9
RLGC*	95.9±0.2	99.2±0.2	**100.0±0.0**	95.8±0.3	77.1±0.5	76.9±0.2	90.8
RLGLC	**96.6±0.3**	99.7±0.1	**100.0±0.0**	**96.5±0.4**	**77.6±0.3**	**77.7**±0.3	**91.4**

based on the source domain samples in the latent space. The overall objective is formulated as

$$\min_{\phi,\psi} W_{1,W_{2,d_\Omega}}\left(P_s^\phi, P_t^\phi\right) + W_{1,W_{2,d_\Omega}}\left(P_t^\phi, P_{tss}^\phi\right) + L_{RCL}\left(P_t^\phi, P_{tss}^\phi\right) + L_{cl}\left(\psi\left(\phi\left(X_s\right)\right), \eta\left(X_s\right)\right) + \Delta \tag{12}$$

where Y_s is the label of source domain samples, L_{cl} is the cross-entropy loss, and Δ is the regularization term, which is used to punish the parameters of ϕ and ψ.

5 Bayes error rate

In this section, we provide a theoretical analysis of the generalization classification error for the learned feature representation. We utilize the Bayes error rate (Feder & Merhav, 1994) to measure the quality of the learned feature representation. Specifically, let P_e be the Bayes error rate of the learned representation Z_t and $\hat{Y}_t$ as the estimation for ground truth label Y_t from the learned classifier. Then, we have

$$P_e := E_{Z_t \sim P_t^\phi}\left[1 - \max_{y_t \in Y_t} P\left(\hat{Y}_t = y_t \,|Z_t\right)\right] \tag{13}$$

Formally, let $|Y_t|$ be the cardinality of Y_t, and let $\mathrm{Th}\left(x\right) = \min\left\{\max\left\{x, 0\right\}, 1 - 1/|Y_t|\right\}$ be a thresholding function. Then we have:

Theorem 5.1. *For an arbitrary learned feature representation Z_t, we have: $P_e = \mathrm{Th}\left(\bar{P}_e\right)$ with*

$$\bar{P}_e \leq 1 - \exp\left[-H\left(Y\right) + I\left(X_t; Z_t \,|X_s\right) + I\left(X_s; Z_t\right)\right] \tag{14}$$

The proof can be seen in Appendix Theorem A.3. From Theorem 5.1, we can decrease the corresponding Bayes error rate by 1) reducing $I\left(X_t; Z_t \,|X_s\right)$, and $I\left(X_s; Z_t\right)$; 2) increasing the number of training samples N. This result supports our proposed framework is better than the traditional representation learning based method for domain adaptation.

6 Experiments

6.1 Setup

We perform the proposed approach on multiple datasets: Office-Home (Venkateswara et al., 2017), including 4 domains; 2) Office-31 dataset (Saenko et al., 2010), including 3 domains; 3) VisDa-2017 dataset (Peng et al., 2017), including 12 categories; 4) Digits datasets (Ganin et al., 2016), including SVHN (S) (Hull, 2002), USPS (U) (Netzer et al., 2011), and MNIST (M) (Lecun et al., 1998). We compare our proposed RLGLC with state-of-the-art unsupervised domain adaptation

Table 2: Accuracy (%) on Office-Home dataset for unsupervised domain adaptation (ResNet-50)

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
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
DAN	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN	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	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	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
CDAN+E	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
BSP+DANN	51.4	68.3	75.9	56.0	67.8	68.8	57.0	49.6	75.8	70.4	57.1	80.6	64.9
BSP+CDAN	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
MDD	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
SymmNets	48.1	74.3	78.7	64.6	71.8	74.1	64.4	50.0	80.2	74.3	53.1	83.2	68.1
ETD	51.3	71.9	85.7	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
GVB-GD	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
HDAN	56.8	75.2	79.8	65.1	73.9	75.2	66.3	**56.7**	81.8	75.4	59.7	84.7	70.9
SRDC	52.3	76.3	81.0	69.5	76.2	78.0	**68.7**	53.8	81.7	**76.3**	57.1	85.0	71.3
RLGC	55.8	75.4	80.1	68.8	75.2	77.6	64.7	53.5	80.4	71.4	59.4	84.4	70.6
RLGC*	56.9	76.1	81.3	70.0	75.7	78.2	65.4	53.9	82.2	72.1	60.5	85.1	71.5
RLGLC	**57.5**	**76.7**	**81.9**	**71.1**	**76.4**	**78.3**	66.2	54.5	**82.6**	73.4	**61.3**	**85.9**	**72.2**

methods including: ResNet-50(He et al., 2016), ResNet-101(He et al., 2016), DAN(Long et al., 2015), DANN(Ganin et al., 2016), JAN(Long et al., 2017b), GTA(Sankaranarayanan et al., 2018), ADDA(Tzeng et al., 2017), UNIT(Liu et al., 2017), CyCADA(Hoffman et al., 2018), CDAN(Long et al., 2017a), CDAN+E(Long et al., 2017a), BSP+DANN(Chen et al., 2019), BSP+ADDA(Chen et al., 2019), BSP+CDAN(Chen et al., 2019), ADDA(Tzeng et al., 2017), MCD(Saito et al., 2018), MDD(Zhang et al., 2019b), SWD(Lee et al., 2019), CAN(Kang et al., 2019), SymmNets(Zhang et al., 2020), GVB-GD(Cui et al., 2020b), ETD(Li et al., 2020), SRDC(Tang et al., 2020), HDAN(Cui et al., 2020a). We also conduct ablation study on all datasets. The classification accuracy is reported in this paper including dataset average accuracy and specific transfer task accuracy, and we repeat the experiments 5 times and then report the average accuracy of the five results. For a fair comparison, the reported results of most comparison methods come from their original papers. For more implementation details, please refer to Appendix A.1.

6.2 Results and discussions

The classification results on the Office-31 dataset are reported in Table 1. The average classification accuracy of our proposed RLGLC is the highest among all compared methods. Also, RLGLC gets better results than all compared methods on 5 specific transfer tasks, which is the most numerous. It worth noting that RLGLC gets the best results on two hard specific transfer tasks: A $\rightarrow$ D and D $\rightarrow$ A. The classification results on the Office-Home dataset are shown in Table 2. It is worth noting that the specific transfer tasks for this dataset is quite challenging. As we can see, RLGLC gets the highest average result than the compared methods. As for specific transfer tasks, RLGLC wins 9 of the 12 tasks, which is also the most numerous of all methods. Table 3 records the classification results on the VisDa-2017 dataset, we can observe that RLGLC gains significant improvement, that is the average result of RLGLC is 0.3 higher than CAN, which is with the second rank among the compared methods. For specific transfer tasks, RLGLC wins 7 of the 12 tasks in this dataset, which is also the most numerous. We further compare RLGLC with previous methods on the Digits dataset, the results are shown in Table 4. Compared with the Office-31 dataset, the size of this dataset is much larger. We can know that the classification results of RLGLC exceed all compared approaches in terms of average accuracy. Also, RLGLC achieves almost state-of-the-art performance on most specific transfer tasks. Therefore, we can conclude that our proposed RLGLC is effective and contain compressed label-relevant information.

6.3 Ablation study and parameter sensitivity

The proposed RLGLC mainly composes of two parts including the global consistency module and local consistency module. For the ablation study, a simplified version of RLGLC, which does not contain the local consistency module, is verified and is named RLGC*. Also for RLGC*, a simplified version of RLGC*, which does not contain $W_{1,W_{2,d_\Omega}}\left(P_t^\phi, P_{tss}^\phi\right)$, is verified and is named RLGC. We can see from the four tables that RLGC* gets comparable classification results on both specific transfer tasks and average classification accuracy. This can demonstrate that the global consistency module is effective. Also, compared RLGC* with RLGC, we can observe that the

Table 3: Accuracy (%) on VisDA-2017 dataset for unsupervised domain adaptation (ResNet-101)

Method	plane	bcybl	bus	car	horse	knife	mcyle	person	plant	sktbrd	train	truck	Avg
ResNet-101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CDAN	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.7
BSP+DANN	92.2	72.5	83.8	47.5	87.0	54.0	86.8	72.4	80.6	66.9	84.5	37.1	72.1
BSP+CDAN	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
SWD	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
CAN	97.0	**87.2**	82.5	74.3	**97.8**	96.2	90.8	**80.7**	**96.6**	96.3	87.5	**59.9**	87.2
RLGC	95.7	86.2	84.1	70.3	92.0	93.9	90.4	77.9	93.3	90.3	84.9	54.7	84.5
RLGC*	96.6	86.4	85.0	73.7	94.6	95.6	92.3	78.2	94.7	93.8	88.7	55.6	86.3
RLGLC	**97.3**	86.6	**86.2**	**75.1**	96.5	**96.7**	**92.4**	79.4	96.1	**96.8**	**89.3**	57.9	**87.5**

Table 4: Accuracy (%) on Digits dataset for unsupervised domain adaptation (ResNet-50)

Method	M→U	U→M	S→M	Avg
DANN	90.4	94.7	84.2	89.8
ADDA	89.4	90.1	86.3	88.6
UNIT	96.0	93.6	90.5	93.4
CyCADA	95.6	96.5	90.4	94.2
CDAN	93.9	96.9	88.5	93.1
CDAN+E	95.6	98.0	89.2	94.3
BSP+DANN	94.5	97.7	89.4	93.9
BSP+ADDA	93.3	94.5	91.4	93.1
BSP+CDAN	95.0	98.1	92.1	95.1
ETD	96.4	96.3	**97.9**	96.9
RLGC	96.4	98.2	95.2	96.6
RLGC*	97.6	98.7	96.2	97.5
RLGLC	**97.9**	**99.1**	96.9	**98.0**

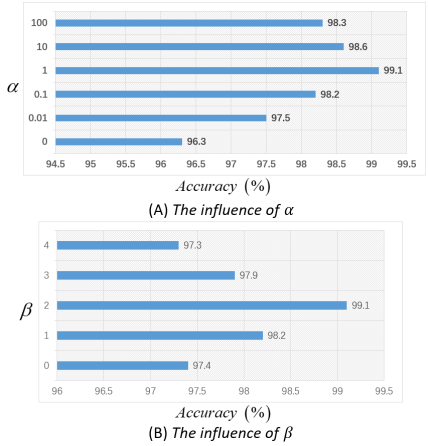

Figure 1: Study on hyper-parameters

average classification accuracy of RLGC* is significantly better than that of RLGC, which prove that the proposed AR-WWD metric is valuable for RLGLC. Compare RLGLC with RLGC*, we can observe that the average classification accuracy of RLGLC is significantly better than that of RLGC*, which demonstrates that the local consistency module is effective.

Based on the specific transfer task U→M, we evaluate the effects of parameter α, which balances the influence of the regularization item in the RCL, and parameter β, which controls the proportion of $\sup_{Z\in\mathcal{Z}} P_t^\phi(Z) \Big/ P_s^\phi(Z)$. For α, we first fix $\beta = 2$, and then select α from the range of $\left\{0, 10^{-2}, 10^{-1}, ..., 10^2\right\}$. The results are shown in Figure 1 (a), we can see that when $\alpha = 1$, we get the best accuracy, this illustrates that the regularization item is effective. Also, we first fix $\alpha = 1$ and select β from the range of $\{0, 1, ..., 4\}$. As we can see from Figure 1 (b), when $\beta = 2$, we obtain the best results, this indicates that constraining the distribution of target domain contained in the distribution of source domain can reduce the label-irrelevant information. This also proves the effectiveness of the proposed information-theoretical based framework. Note that when $\beta = 4$, the accuracy is the lowest, we can conclude that if the relaxed constraint is too loose, then an amount of task-relevant information will also be discarded.

7 Conclusions

In this paper, we first give an analysis of the traditional representation learning based domain adaptation methods based on information theory. We prove that the representation of the target domain data learned by traditional methods is not discriminative. To improve the problem, we propose a new information-theoretical based framework. We provide a theoretical analysis to this new framework to verify the effectiveness. Then, we derive the proposed method called domain-invariant representation learning with global and local consistency (RLGLC). RLGLC consists of two modules including the global consistency module and local consistency module. We also give a theoretical analysis that RLGLC is conducive to minimize the Bayes error rate. Experiment results on four domain adaptation datasets show the effectiveness of the proposed method.

8 ETHICS STATEMENT

All data used in this paper is public. References we cited related to the dataset is also public. This paper do not involve human subjects and do not have potentially harmful insights. Also, this paper do not have discrimination/bias/fairness concerns and do not involve privacy and security issues.

9 REPRODUCIBILITY STATEMENT

The implementation details of our proposed RLGLC are shown in subsection 6.1 and subsection A.1. For theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix. For the code of this paper, we will make it public after the article is finally accepted.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *arXiv preprint arXiv:2010.05862*, 2020.

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

Ting Chen and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1081–1090. PMLR, 2019.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, 2015.

Shuhao Cui, Xuan Jin, Shuhui Wang, Yuan He, and Qingming Huang. Heuristic domain adaptation. *arXiv preprint arXiv:2011.14540*, 2020a.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2020b.

Sofien Dhouib, Ievgen Redko, and Carole Lartizien. Margin-aware adversarial domain adaptation with optimal transport. In *International Conference on Machine Learning*, pp. 2514–2524. PMLR, 2020.

Yonatan Dukler, Wuchen Li, Alex Lin, and Guido Montúfar. Wasserstein of wasserstein loss for learning generative models. In *International Conference on Machine Learning*, pp. 1716–1725. PMLR, 2019.

Björn Engquist and Yunan Yang. Seismic imaging and optimal transport. *arXiv preprint arXiv:1808.04801*, 2018.

Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pp. 738–746. PMLR, 2013.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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*, pp. 770–778, 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1994–2003. PMLR, 2018.

Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4043–4052, 2020.

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

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

Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *arXiv preprint arXiv:2011.00147*, 2020.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4122–4129, 2019.

Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998. doi: 10.1109/5.726791.

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

Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13936–13944, 2020.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 700–708, 2017.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 97–105. JMLR.org, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017a.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. pp. 2208–2217, 2017b.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pp. 124–138. Springer, 2012.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, 01 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Michael A Puthawala, Cory D Hauck, and Stanley J Osher. Diagnosing forward operator error using optimal transport. *Journal of Scientific Computing*, 80(3):1549–1576, 2019.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios (eds.), *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pp. 213–226. Springer, 2010.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732, 2018.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. pp. 8503–8512, 2018.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8725–8735, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. pp. 7167–7176, 2017.

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

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pp. 6872–6881. PMLR, 2019.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4394–4403, 2020.

Yabin Zhang, Bin Deng, Hui Tang, Lei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413. PMLR, 2019a.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I. Jordan. Bridging theory and algorithm for domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7404–7413. PMLR, 2019b.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

A Appendix

A.1 Implementation details

All experiments are implemented with PyTorch and optimized by the Adam optimizer. All hyper-parameters are fixed. For a fair comparison, we use the same network architecture as the compared methods on each dataset. Specifically, the backbone in the Office-31 dataset, the Office-Home dataset, and the Digits dataset is set to ResNet-50, the backbone in the VisDa-2017 dataset is set as ResNet-101. For all datasets, the backbone is first fine-tuned by pretraining on ImageNet. The hyperparameter τ in the RCL is set to 2.5, and the hyperparameter λ in the AR-WWD is set to 10. The f that is named as the critic in AR-WWD is implemented by a convolutional neural network with 3 hidden layers and leaky ReLU activations. For all tables, the 'Avg' represents the average classification accuracy of all specific transfer tasks

A.2 Proofs

Proposition A.1. *Based on **Definition 4.1**, we can obtain:* $I(X_u; Z_u) = I(X_u; Y) = I(Z_u; Y)$.

Proof.

$$\begin{aligned} & I(X_u; Y \,|Z_u) \\ &= I(X_u; Y) - I(X_u; Y; Z_u) \\ &= I(X_u; Y) - I(Y; Z_u) - I(Z_u; Y \,|X_u) \\ &= I(X_u; Y) - I(Y; Z_u) \end{aligned} \tag{15}$$

$$
\begin{aligned}
&I\left(Z_u; Y \,|X_u\right) \\
&= I\left(Z_u; Y\right) - I\left(Z_u; Y; X_u\right) \\
&= I\left(Z_u; Y\right) - I\left(Y; X_u\right) - I\left(Z_u; Y \,|X_u\right) \\
&= I\left(Z_u; Y\right) - I\left(Y; X_u\right)
\end{aligned} \tag{16}
$$

Because $I\left(X_u; Y \,|Z_u\right) = I\left(Z_u; Y \,|X_u\right) = 0$, therefore, we have $I\left(X_u; Z_u\right) = I\left(X_u; Y\right) = I\left(Z_u; Y\right)$. □

Proposition A.2. $I(Z_s; Z_t) = I(Z_s Z_t; Y) = I(Z_s; Y) = I(Z_t; Y)$.

Proof. Since Z_s is a represenntation of X_s, we have $I\left(Y; Z_s \,|X_t X_s\right) = 0$, therefore:

$$
\begin{aligned}
&I\left(Y; X_s \,|Z_s\right) \\
&= I\left(X_s; Y \,|X_t Z_s\right) + I\left(X_s; X_t; Y \,|Z_s\right) \\
&= I\left(X_s; Y \,|X_t\right) - I\left(X_s; Z_s; Y \,|X_t\right) + I\left(X_s; X_t; Y \,|Z_s\right) \\
&= I\left(X_s; Y \,|X_t\right) - I\left(Z_s; Y \,|X_t\right) + I\left(Z_s; Y \,|X_t X_s\right) + I\left(X_s; X_t; Y \,|Z_s\right) \\
&\leq I\left(X_s; Y \,|X_t\right) + I\left(Z_s; Y \,|X_t X_s\right) + I\left(X_s; X_t; Y \,|Z_s\right) \\
&= I\left(X_s; Y \,|X_t\right) + I\left(X_s; X_t; Y \,|Z_s\right) \\
&= I\left(X_s; Y \,|X_t\right) + I\left(X_s; X_t \,|Z_s\right) - I\left(Z_s; Z_t \,|Z_s Y\right) \\
&\leq I\left(X_s; Y \,|X_t\right) + I\left(X_s; X_t \,|Z_s\right)
\end{aligned} \tag{17}
$$

Similarly, we have $I\left(Y; X_t \,|Z_t\right) \leq I\left(X_t; Y \,|X_s\right) + I\left(X_s; X_t \,|Z_t\right)$.

Then, we have $I\left(X_s; Y \,|X_t\right) + I\left(X_s; X_t \,|Z_s\right) = I\left(X_s; X_t \,|Z_t\right)$ and $I\left(X_t; Y \,|X_s\right) + I\left(X_s; X_t \,|Z_t\right) = I\left(X_s; X_t \,|Z_s\right)$. Therefore, $I\left(X_s; X_t \,|Z_s\right) = 0 \Rightarrow I\left(X_s; Y \,|Z_s\right) = 0$.

Then, we have:

$$
\begin{aligned}
&I\left(Y; X_s\right) \\
&= I\left(Y; Z_s \,|X_s X_t\right) + I\left(Y; X_s X_t; Z_s\right) \\
&= I\left(Y; X_s X_t; Z_s\right) \\
&= I\left(Y; X_s X_t\right) - I\left(Y; X_s X_t \,|Z_s\right) \\
&= I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|Z_s\right) - I\left(Y; X_t \,|Z_s X_s\right) \\
&= I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|Z_s\right) - I\left(Y; X_t \,|X_s\right) + I\left(Y; X_t; Z_s \,|X_s\right) \\
&= I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|Z_s\right) - I\left(Y; X_t \,|X_s\right) + I\left(Y; Z_s \,|X_s\right) - I\left(Y; Z_s \,|X_s X_t\right) \\
&= I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|Z_s\right) - I\left(Y; X_t \,|X_s\right) + I\left(Y; Z_s \,|X_s\right) \\
&\geq I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|Z_s\right) - I\left(Y; X_t \,|X_s\right) \\
&> I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|X_t\right) - I\left(X_s; X_t \,|Z_s\right) - I\left(Y; X_t \,|X_s\right)
\end{aligned} \tag{18}
$$

Similarly, we have $I\left(Y; X_t\right) \geq I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|X_t\right) - I\left(X_s; X_t \,|Z_t\right) - I\left(Y; X_t \,|X_s\right)$.

Then, we have:

$$
\begin{aligned}
&I\left(Y; X_t\right) \\
&\geq I\left(Y; X_s X_t\right) - I\left(Y; X_s \,|X_t\right) - I\left(X_t; X_s \,|Z_s\right) - I\left(X_t; Y \,|X_s\right) \\
&= I\left(Y; X_s X_t\right) - I\left(X_t; X_s \,|Z_s\right) \\
&= I\left(Y; X_s X_t\right)
\end{aligned} \tag{19}
$$

Since $I\left(Y; X_t\right) \leq I\left(Y; X_s X_t\right)$ is a consequence of the data processing inequality, we conclude that $I\left(Y; X_t\right) = I\left(Y; X_s X_t\right)$. Similarly, we can have $I\left(Y; X_s\right) = I\left(Y; X_s X_t\right)$. Therefore, we have $I(Z_s; Z_t) = I(Z_s Z_t; Y) = I(Z_s; Y) = I(Z_t; Y)$. □

Theorem A.1. *Suppose the representations for source domain and target domain are obtained by minimizing the objective function (2). Then, Z_s is with the discriminability, compressibility, and transferability, while Z_t is only with compressibility and transferability.*

Proof.

$$
\begin{aligned}
&I\left(X_s; Z_s \,|X_t\right) \\
&= E_{X_s, X_t \sim P(X_s, X_t)} E_{Z_s \sim P(Z_s|X_s)} \left[\log \tfrac{P(Z_s|X_s)}{P(Z_s|X_t)}\right] \\
&= E_{X_s, X_t \sim P(X_s, X_t)} E_{Z_s \sim P(Z_s|X_s)} \left[\log \tfrac{P(Z_s|X_s) P(Z_t|X_t)}{P(Z_s|X_t) P(Z_t|X_t)}\right] \\
&= KL\left(P\left(Z_s \,|X_s\right) \| P\left(Z_t \,|X_t\right)\right) - KL\left(P\left(Z_t \,|X_s\right) \| P\left(Z_s \,|X_t\right)\right) \\
&\leq KL\left(P\left(Z_s \,|X_s\right) \| P\left(Z_t \,|X_t\right)\right)
\end{aligned} \tag{20}
$$

Similarly, we can obtain: $I\left(X_t; Z_t \,|X_s\right) \leq KL\left(P\left(Z_t \,|X_t\right) \| P\left(Z_s \,|X_s\right)\right)$. Because that minimizing $KL\left(P\left(Z_t \,|X_t\right) \| P\left(Z_s \,|X_s\right)\right)$ and $KL\left(P\left(Z_s \,|X_s\right) \| P\left(Z_t \,|X_t\right)\right)$ equal to minimize the Wasserstein distance between the distribution $P\left(Z_s \,|X_s\right)$ and the distribution $P\left(Z_t \,|X_t\right)$. Also, we can know that $I\left(X_s; Z_s \,|Y\right) = I\left(X_t; Z_t \,|Y\right)$. So, we can obtain that minimizing the first term in the objective 2 equals to make the learned sample feature representations of both domain to be with compressibility, and transferability. Minimizing the second term in the objective 2 can result that the learned sample feature representations of source domain to be with discriminability. But there is no obvious term to constrain the learned sample feature representations of target domain to be with discriminability. □

Theorem A.2. *Maximizing $I\left(Z_t; Z_{tss}\right)$ while minimizing $D\left(P_t^{\phi}, P_{tss}^{\phi}\right)$ can make the learned feature representation of the target domain sample be with discriminability.*

Proof. compared the objective 5 with the objective 2, we can see that there are two additional terms in objective 5. We can directly see that maximizing the $I\left(Z_t; Z_{tss}\right)$ is to make the learned sample feature representations of target domain to be with discriminability and minimizing the $D\left(P_t^{\phi}, P_{tss}^{\phi}\right)$ can be regard as to make the learned sample feature representations of target domain to be with compressibility and transferability. □

Theorem A.3. *For an arbitrary learned feature representation Z_t, then we have: $P_e = \mathrm{Th}\left(\bar{P}_e\right)$ with*

$$\bar{P}_e \leq 1 - \exp\left[-H\left(Y\right) + I\left(X_t; Z_t \,|X_s\right) + I\left(X_s; Z_t\right)\right] \tag{21}$$

Proof. From the Feder & Merhav (1994), we can obtain that

$$-\log\left(1 - P_e\right) \leq H\left(Y \,|Z_t\right) \tag{22}$$

Because we have

$$\begin{aligned} &H\left(Y \,|Z_t\right) = H\left(Y\right) - I\left(Z_t; Y\right) \\ &I\left(Z_t; Y\right) = I\left(X_t; Z_t\right) \\ &I\left(X_t; Z_t\right) = I\left(X_t; Z_t \,|Y\right) + I\left(Z_t; Y\right) = I\left(X_t; Z_t \,|X_s\right) + I\left(X_s; Z_t\right) \end{aligned} \tag{23}$$

Therefore, we can obtain

$$\bar{P}_e \leq 1 - \exp\left[-H\left(Y\right) + I\left(X_t; Z_t \,|X_s\right) + I\left(X_s; Z_t\right)\right] \tag{24}$$

Next, by definition of the Bayes error rate, we know $0 \leq P_e \leq 1 - \frac{1}{|T|}$. □