Probability-Polarized Optimal Transport for Unsupervised Domain Adaptation

Yan Wang[1], Chuan-Xian Ren[1]*, Yi-Ming Zhai[1], You-Wei Luo[1], Hong Yan[2]

[1]School of Mathematics, Sun Yat-Sen University, China
[2]Department of Electrical Engineering, City University of Hong Kong, Hong Kong
{wangy2277, zhaiym3, luoyw28}@mail2.sysu.edu.cn, rchuan@mail.sysu.edu.cn, h.yan@cityu.edu.hk

Abstract

Optimal transport (OT) is an important methodology to measure distribution discrepancy, which has achieved promising performance in artificial intelligence applications, e.g., unsupervised domain adaptation. However, from the view of transportation, there are still limitations: 1) the local discriminative structures for downstream tasks, e.g., cluster structure for classification, cannot be explicitly admitted by the learned OT plan; 2) the entropy regularization induces a dense OT plan with increasing uncertainty. To tackle these issues, we propose a novel Probability-Polarized OT (PPOT) framework, which can characterize the structure of OT plan explicitly. Specifically, the probability polarization mechanism is proposed to guide the optimization direction of OT plan, which generates a clear margin between similar and dissimilar transport pairs and reduces the uncertainty. Further, a dynamic mechanism for margin is developed by incorporating task-related information into the polarization, which directly captures the intra/inter class correspondence for knowledge transportation. A mathematical understanding for PPOT is provided from the view of gradient, which ensures interpretability. Extensive experiments on several datasets validate the effectiveness and empirical efficiency of PPOT.

Introduction

With the increasing diversity of data sources, domain shift between different data distributions has become an essential problem in machine learning and deep learning (Long et al. 2019). Unsupervised domain adaptation (UDA) addresses this issue by adapting a model trained on a labeled source domain to an unlabeled target domain, i.e., reducing the gap in representations and improving generalization to the target domain (Pan and Yang 2009).

Various UDA approaches have been proposed to reduce domain shift. The main idea is to minimize the domain discrepancy and learn domain-invariant features, which can be broadly classified as distance metric-based domain adaptation (Zellinger et al. 2017; Ren, Luo, and Dai 2023) and adversarial learning-based domain adaptation (Long et al. 2018; Chen et al. 2022). Distance metric-based methods aim to learn a shared feature representation by minimizing the domain discrepancy, e.g., Maximum Mean Discrepancy (MMD) (Tzeng et al. 2014), Central Moment Discrepancy (Zellinger et al. 2017), and H-divergency (Saito et al. 2018). These prior works indicate the importance and effectiveness of reducing domain discrepancy. Adversarial methods use a discriminator to distinguish source and target domain features, which is trained against the feature extractor to generate more transferable representations.

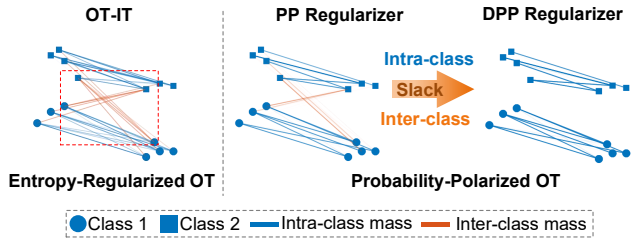


Figure 1: Illustration of the optimal transport plan obtained by entropy-regularized OT (OT-IT), and our probability-polarized OT with PP regularizer and DPP regularizer. Darker lines represent larger transport masses.

Optimal Transport (OT) has also been widely used in UDA since it has solid theoretical supports (Redko, Habrard, and Sebban 2017). Some OT-based methods (Courty, Flamary, and Tuia 2014; Courty et al. 2016) are built upon the squared Euclidean distance cost and explicitly learns the minimal cost for transporting the source distribution into the target. However, it cannot adequately capture the underlying structure in distributions under some complicated tasks. To further enhance the alignment between intra-class samples, some distance-regularized OT methods exploits task-related information to reweight the ground cost matrix (Xu et al. 2020; Luo and Ren 2021; Liu, Zhou, and Sun 2023).

Coupling-regularized OT methods directly impose different regularization terms on the probability coupling, which can introduce some properties to the OT plan. Specifically, the entropy-regularized OT (Cuturi 2013) (OT-IT) offers a smooth and dense OT plan, which enables source samples to transport mass to more target samples. However, high density may blur the margin between intra/inter-class transport pairs and increase the uncertainty, which makes the OT plan challenging to achieve a class-wise transportation, as shown

*Corresponding Author

in Figure 1. There are also some methods introduce regularization based on the label information, e.g., group sparsity (Courty, Flamary, and Tuia 2014) (OT-GL), the cross-domain cluster structures are not sufficiently characterized since only source information is considered. These class regularization terms aim to preserve the local similarity within the source domain, but the overall structural consistency between the source and target domain is neglected. Note that these limitations will induce incorrect transportation plan between the inter-class sample pairs. Thus, it is necessary to explore a novel mechanism that can explicitly characterize inter/intra-class correspondence via OT plan.

To tackle the limitations above, we propose a novel Probability-Polarized OT (PPOT) framework for UDA. Inspired by pioneering work (Chen et al. 2021), which applies polarization regularization to distance metric, we introduce the polarization mechanism to the OT plan and propose the Probability Polarization (PP) regularizer for OT. Mathematically, PPOT imposes a margin constraint on the probability couplings by introducing thresholds for the positive transport and negative transport, i.e., transportations between intra-class pairs and inter-class pairs. As a large probability value suggests the corresponding source and target samples are similar, i.e., a larger likelihood of the sample pair belonging to the same class, PPOT encourages the transportation between similar samples. Such a polarization reduces the uncertainty of probability coupling values and ensures an explicit structure for OT plan. Further, to improve the quality of polarization, Dynamic Probability Polarization (DPP) mechanism is proposed by setting dynamically adjusted margin thresholds, and a mathematical understanding is provided from the view of gradient. As shown in Figure 1, DPP exploits class information via source labels and target pseudo-labels to slack polarization thresholds. Therefore, the OT assignments of the intra-class and inter-class couplings are expected to correctly maximize and minimize, respectively. For optimization, we propose an efficient GCG method for solving PPOT, which is empirically validated to be efficient. The main contributions of this paper are summarized as follows.

- A novel OT framework called Probability-Polarized OT is proposed for learning favorable OT plan. With the PP regularizer, PPOT discriminates probability couplings by generating a clear margin between similar and dissimilar transport pairs, then the learned OT plan will admit explicit structure with small uncertainty.
- A dynamic regularizer is proposed under the PPOT framework. It slacks the margin thresholds in the polarization process, which ensures the OT plan can explicitly characterize the intra/inter-class correspondence. The mathematical relationship of polarization direction and objective gradient is derived, which ensures the interpretability for PPOT.
- A computation-friendly GCG algorithm is proposed for PPOT and extensive experiments are conducted. The numerical results verify the effectiveness and efficiency of PPOT in empirical scenarios.

Related Work

Unsupervised domain adaptation. UDA methods attempt to reduce the domain discrepancy and then improve the model's performance on the target domain. These methods can be generally divided into two categories, i.e., distance metric-based and adversarial learning-based methods. Deep CORAL (Sun and Saenko 2016) aligns the second-order statistics between the source and target domains. JAN (Long et al. 2017a) introduces the class-specific MMD to mitigate domain discrepancy at the class level. BuresNet (Ren, Luo, and Dai 2023) proposes a conditional Bures metric to align class-conditioned distributions across domains. Adversarial method DANN (Ganin and Lempitsky 2015) aims to learn domain-invariant representations by confusing a domain discriminator. Extended from DANN, CDANs (Long et al. 2018) exploits the discriminative label information to condition the adversarial adaptation models, which enables the alignment of multimodal distributions. BCDM (Li et al. 2021a) achieves feature alignment by maximizing the determinacy between the bi-classifiers in an adversarial manner. DALN (Chen et al. 2022) takes the task-specific classifier as a discriminator to achieve discriminator-free.

Optimal Transport. OT is also a significant line for dealing with the UDA problem. Various OT-based UDA methods can be roughly categorized into distance-regularized OT and coupling-regularized OT. The distance-regularized OT eliminates negative transfer by reweighting the ground cost between samples from different domains. JDOT (Courty et al. 2017) and DeepJDOT (Damodaran et al. 2018) incorporate the discrepancy between labels into the cost function. ETD (Li et al. 2020) reweights the distance matrix by an attention mechanism, which encodes the relationship between samples from different domains. MOT (Luo and Ren 2023) reweights the distance matrix by a label information-based mask operation, which can mitigate the negative transport between inter-class sample pairs. The coupling-regularized OT brings some properties of the probability coupling by imposing different regularization terms. Group sparsity (Courty, Flamary, and Tuia 2014) and Laplacian regularization (Flamary et al. 2014) terms both leverage source labels to regularize the transport. Group sparsity encourages the target point to receive masses from source points with the same labels. Laplacian regularization aims to preserve the structure of the source domain. GOTDA-O (Long et al. 2022) adopts three regularization terms based on the source label information, which promotes learning more discriminative representations. InfoOT (Chuang, Jegelka, and AlvarezMelis 2023) preserves the structure of OT plan by maximizing mutual information, and then produces a better domain alignment with high coherence.

Probability-Polarized Optimal Transport

Following the assumptions of UDA, we define a labeled source domain $\mathcal{D}^s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n_s}$ and an unlabeled target domain $\mathcal{D}^t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$, where sample $\mathbf{x}^{s/t} \in \mathbb{R}^d$ and $y_i^s \in \{1, 2, \ldots, K\}$ denotes the ground-truth label of $\mathbf{x}_i^s$.

Preliminaries. The goal of OT is to map the probability masses from the source distribution to the target with the

least amount of transport cost. Let $\mathcal{X}$ and $\mathcal{Z}$ be metric spaces with marginal distributions μ and ν, respectively. Considering $X \sim \mu$ and $Z \sim \nu$, the Monge problem seeks a map $\mathcal{T} : \mathcal{X} \to \mathcal{Z}$ that pushes the probability mass from the source distribution μ to the target ν (Villani et al. 2009).

To make the Monge formulation more feasible, Kantorovich (Kantorovich 2006) proposes a convex relaxation of the Monge problem, which seeks a general coupling $\boldsymbol{\gamma} \in \Pi(\mu, \nu)$ by minimizing the total transport cost. Note that $\Pi(\mu, \nu)$ is the collection of all possible probability couplings on $\mathcal{X} \times \mathcal{Z}$ with marginals μ and ν.

When μ and ν are only accessible through discrete samples $X = (\mathbf{x}_1, \dots, \mathbf{x}_{n_s})$ and $Z = (\mathbf{x}_1, \dots, \mathbf{x}_{n_t})$, we denote the set of probability couplings as

$$\mathcal{B}(\mu, \nu) = \{\boldsymbol{\gamma} \in (\mathbb{R}^+)^{n_s \times n_t} | \boldsymbol{\gamma} \mathbf{1}_{n_t} = \mu, \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{1}_{n_s} = \nu\},$$

where $\mathbf{1}_{n_{s/t}}$ is a $n_{s/t}$-dimensional vector with one. Then, the Kantorovich problem in the discrete case is defined as

$$\mathcal{D}(\boldsymbol{\gamma}) = \min_{\boldsymbol{\gamma} \in \mathcal{B}(\mu, \nu)} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F, \tag{1}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product and $\mathbf{C} \in \mathbb{R}^{n_s \times n_t}$ is the cost matrix. Specifically, $\mathbf{C}(i, j) = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ represents the cost to move a mass unit form $\mathbf{x}_i^s$ to $\mathbf{x}_j^t$, which the squared Euclidean distance is the classical choice.

To further relax the constraint of probability coupling $\boldsymbol{\gamma}$ and speed up the computation of Eq. (1), Cuturi (Cuturi 2013) introduces an entropy-regularization, i.e.,

$$\Omega_{\mathrm{ent}}(\boldsymbol{\gamma}) = \sum\nolimits_{i=1}^{n_s} \sum\nolimits_{j=1}^{n_t} \boldsymbol{\gamma}_{ij} \ln(\boldsymbol{\gamma}_{ij})$$

Then, the entropy-regularized Kantorovich problem (OT-IT) can be formulated as

$$\min_{\boldsymbol{\gamma} \in \mathcal{B}(\mu, \nu)} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \alpha \Omega_{\mathrm{ent}}(\boldsymbol{\gamma}), \tag{2}$$

where α is a parameter of sparsity penalty. With this entropy regularization, it is expected to obtain an optimal transport plan of Eq. (2) with lower sparsity. When $\alpha \to \infty$, each element of the optimal transport plan converges toward $\frac{1}{n_s n_t}$.

PPOT: Mathematical Formulation

In this section, we explore the mathematical formulation of PPOT, which provides a more discriminative transport by probability polarization.

Recall that the entropy-regularization promotes a smooth and dense probability coupling for the widely used OT-IT, which omits the label information and usually induces mismatch. Differently, our PPOT not only expect source samples to distribute their probability masses toward more target points, but also attempt to seek a discriminative transport with large intra-mass transport masses.

In OT, a significant probability value indicates that the corresponding source and target samples are similar. Then, The cross-domain sample pair is more likely to belong to the same class. Inspired by this property, our PPOT exploits polarization to maximize or minimize probability coupling values, i.e., the masses received by target samples can be larger if large or smaller if small. Thus, PPOT can enhance the discriminability of the transport plan to a certain degree and achieve clearer pair-wise matching. The formulation of PPOT can be defined as

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma} \in \mathcal{B}} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \alpha \Omega_{\mathrm{ent}}(\boldsymbol{\gamma}) + \beta \Omega_{\mathrm{p}}(\boldsymbol{\gamma}), \tag{3}$$

where α and $\beta > 0$, and $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ is our probability-polarized regularization. To be specific, this term is defined as

$$\Omega_{\mathrm{p}}(\boldsymbol{\gamma}) = \sum\nolimits_{i,j} \|\min\{(\boldsymbol{\gamma}_{ij} - \delta_{y_i^s}^-) \odot (\boldsymbol{\gamma}_{ij} - \delta_{y_i^s}^+), 0\}\|_1, \tag{4}$$

where $\delta_{y_i^s}^-$ and $\delta_{y_i^s}^+$ are the inter-class and intra-class probability thresholds for the k-th class, i.e., $y_i^s = k, k \in \{1, 2, \cdots, K\}$, respectively. $\odot$ is element-wise multiplication. This quadratic (smooth) function $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ optimizes the probability coupling values in the $(\delta_k^-, \delta_k^+) \subseteq (0, 1)$ and encourages to distribute in $(0, \delta_k^-)$ or $(\delta_k^+, 1)$ when minimizing the formulation of PPOT in Eq. (3).

Generally, the probability thresholds δ_k^- and δ_k^+ are free of parameter tuning and can be directly initialized as

$$\delta_k^- \approx \varepsilon / [n_s \cdot (n_t - n_t^k)], \quad \delta_k^+ \approx 1 / (n_s \cdot n_t^k),$$

where n_t^k denote the number of k-th class target samples and ε could be arbitrarily small. The explicit estimation forms above can be mathematically justified by the essential property of ideal transport. Thus, PPOT ensures the OT plan is dominated by the larger intra-class transport masses and leads to more distinct cross-domain sample correspondence.

PPOT: Dynamic Probability Polarization

Motivation for DPP. This polarization strategy is only based on the source information, which lacks target discriminative information. Thus, it may wrongly maximize or minimize some values of the probability coupling for boundary sample pairs. To alleviate the impact of wrong polarization, we instantiate the polarization term in PPOT by proposing Dynamic Probability Polarization (DPP). By exploring the relationship between the gradient of $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ and thresholds $\delta_k^{+/-}$, DPP exploits both source and target class information to dynamically adjust the direction of polarization.

Insights from gradient. According to the gradient descent method, we explore the optimization of the polarization term $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ in Eq. (4). Without loss of generality, we consider the i-th source sample with $y_i^s = k$. As previously described, $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ enforces the i-th row values of $\boldsymbol{\gamma}$ to concentrate in $(0, \delta_k^-)$ or $(\delta_k^+, 1)$. The values in (δ_k^-, δ_k^+) will be polarized. Then, the gradient of $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ w.r.t. $\boldsymbol{\gamma}_{ij}$ is

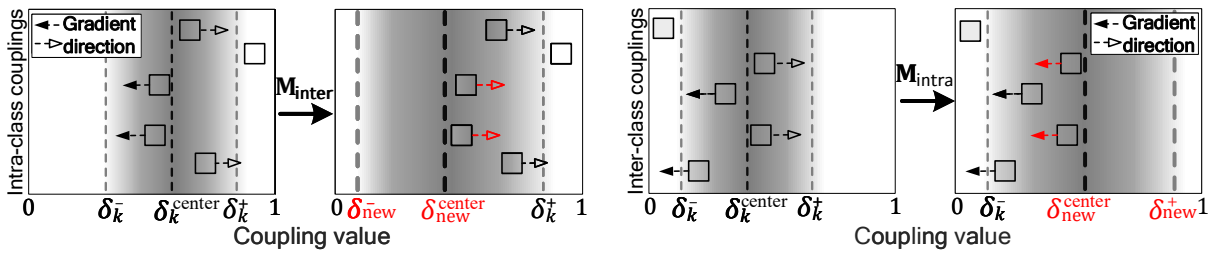
$$g(\boldsymbol{\gamma}_{ij}) = \frac{\partial \Omega_{\mathrm{p}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_{ij}} = \delta_k^- + \delta_k^+ - 2\boldsymbol{\gamma}_{ij}. \tag{5}$$

We set the threshold center $\delta_k^{\mathrm{center}}$ as $\delta_k^{\mathrm{center}} = \frac{1}{2}(\delta_k^- + \delta_k^+)$. Then, the derivation in Eq. (5) can be re-written as

$$g(\boldsymbol{\gamma}_{ij}) = \frac{\partial \Omega_{\mathrm{p}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_{ij}} = 2(\delta_k^{\mathrm{center}} - \boldsymbol{\gamma}_{ij}). \tag{6}$$

With Eq. (6), it is easy to predict the sign of $g(\boldsymbol{\gamma}_{ij})$, i.e.,

$$g(\boldsymbol{\gamma}_{ij}) \begin{cases} > 0, & \text{if } \boldsymbol{\gamma}_{ij} \in (\delta_k^-, \delta_k^{\mathrm{center}}), \\ < 0, & \text{if } \boldsymbol{\gamma}_{ij} \in (\delta_k^{\mathrm{center}}, \delta_k^+), \\ = 0, & \text{otherwise.} \end{cases} \tag{7}$$



(a) Decrease inter-class probability thresholds (b) Increase intra-class probability thresholds

Figure 2: Illustration of two steps of the DPP on probability coupling values. Darker color represents higher values of coupling.

The second-order derivative of $\Omega_{\text{p}}(\boldsymbol{\gamma})$ w.r.t γ_{ij} is $\partial^2\Omega_{\text{p}}(\boldsymbol{\gamma})/\partial^2\gamma_{ij} = -2$. Thus, it is obvious that the polarization term can get the maximum value if $\gamma_{ij} = \delta^k_{\text{center}}$. When minimizing the polarization term, based on the principle of gradient descent, we further find the relationship between polarization direction and gradient: (1) If $\gamma_{ij} > \delta_k^{\text{center}}$, $-g(\gamma_{ij})$ will be larger than 0, and then γ_{ij} will be polarized to the larger; (2) If $\gamma_{ij} < \delta_k^{\text{center}}$, $-g(\gamma_{ij})$ will be smaller than 0, and then γ_{ij} will be polarized to the smaller.

A dynamic mechanism. Inspired by the relationship above, DPP is proposed to correct the wrong polarization in the training process. To be specific, we build the DPP via two polarization slack masks based on source labels and target pseudo-labels, i.e., inter-class polarization slack mask $\mathbf{M}_{\text{inter}}$ and intra-class polarization slack mask $\mathbf{M}_{\text{intra}}$.

According to the labels $\mathbf{Y}^s \in \mathbb{R}^{n_s}$ of the source samples $\mathbf{X}^s \in \mathbb{R}^{n_s \times d}$, the inter-class threshold matrix can be constructed as $\boldsymbol{\Delta}^- = [\delta^-_{y^s_1}, \ldots, \delta^-_{y^s_{n_s}}]^{\text{T}} \mathbf{1}^{\text{T}}_{n_t} \in \mathbb{R}^{n_s \times n_t}$. Similarly for intra-class threshold matrix $\boldsymbol{\Delta}^+ \in \mathbb{R}^{n_s \times n_t}$. Besides, denote $\hat{\mathbf{Y}}^t \in \mathbb{R}^{n_t}$ as prediction-based target pseudo-labels.

Decrease inter-class probability thresholds. If samples $\mathbf{x}^s_i$ and $\mathbf{x}^t_j$ have the same label, i.e., $y^s_i = \hat{y}^t_j$, where $\hat{y}^t_j$ is the pseudo-label of $\mathbf{x}^t_j$ and $y^s_i = k$. The corresponding probability mass γ_{ij} is expected to be larger than the intra-class threshold δ^+_k. However, some probability coupling values are wrongly polarized to the direction of δ^-_k since they are smaller than the threshold center δ^k_{center}, as shown in the left of Figure 2(a). To deal with this case, we provide a slack inter-class threshold matrix $\boldsymbol{\Delta}^-_{\text{slack}}$ based on $\mathbf{M}_{\text{inter}}$ to decrease inter-class probability thresholds,

$$\boldsymbol{\Delta}^-_{\text{slack}} = \mathbf{M}_{\text{inter}} \odot \boldsymbol{\Delta}^- = \begin{cases} 1/K \cdot \delta^-_{y^s_i}, & y^s_i = \hat{y}^t_j, \\ 1 \cdot \delta^-_{y^s_i}, & \text{otherwise.} \end{cases} \tag{8}$$

The construction of $\mathbf{M}_{\text{inter}}$ is also illustrated in the left of Figure 3. With the relaxation by $\mathbf{M}_{\text{inter}}$, as shown in the right of Figure 2(a), the probability thresholds δ^-_k will be relaxed to a smaller δ^-_{new}. Then, the probability coupling values of intra-class sample pairs originally wrongly polarized will be larger than $\delta^{\text{new}}_{\text{center}}$ and achieve correct maximization.

Increase intra-class probability thresholds. If samples $\mathbf{x}^s_i$ and $\mathbf{x}^t_j$ have different labels, i.e., $y^s_i \neq \hat{y}^t_j$, the corresponding probability mass γ_{ij} is expected to be optimized in the direction of inter-class threshold δ^-_k. To avoid wrongly maximizing the probability coupling values for samples with different labels, as shown in the left of Figure 2(b), we provide a slack inter-class threshold matrix $\boldsymbol{\Delta}^+_{\text{slack}}$ based on $\mathbf{M}_{\text{intra}}$ to increase intra-class probability thresholds,

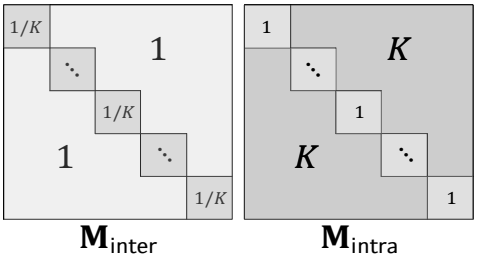


Figure 3: Illustration of inter-class and intra-class polarization slack matrices $\mathbf{M}_{\text{inter}}$ and $\mathbf{M}_{\text{intra}}$. Note the diagonal blocks represent cross-domain intra-class samples.

$$\boldsymbol{\Delta}^+_{\text{slack}} = \mathbf{M}_{\text{intra}} \odot \boldsymbol{\Delta}^+ = \begin{cases} K \cdot \delta^+_{y^s_i}, & y^s_i \neq \hat{y}^t_j, \\ 1 \cdot \delta^+_{y^s_i}, & \text{otherwise.} \end{cases} \tag{9}$$

The construction of $\mathbf{M}_{\text{intra}}$ is illustrated in the right of Figure 3. With the relaxation by $\mathbf{M}_{\text{intra}}$, as shown in the right of Figure 2(b), the probability thresholds δ^+_k will be relaxed to a larger δ^-_{new}. Then, the probability coupling values of inter-class sample pairs originally wrongly polarized will be smaller than $\delta^{\text{new}}_{\text{center}}$ and achieve correct minimization.

The mentioned slack strategy is the core of DPP. Overall, the formulation of DPP $\Omega_{\text{dp}}(\boldsymbol{\gamma})$ can be written as

$$\sum_{i,j} \|\min\{(\gamma_{ij} - \mathbf{M}^{\text{inter}}_{ij}\delta^-_{y^s_i}) \odot (\gamma_{ij} - \mathbf{M}^{\text{intra}}_{ij}\delta^+_{y^s_i}), 0\}\|_1, \tag{10}$$

where $\mathbf{M}^{\text{inter}}_{ij}$ and $\mathbf{M}^{\text{intra}}_{ij}$ are the elements of the intra-class stack masks $\mathbf{M}_{\text{inter}}$ and $\mathbf{M}_{\text{intra}}$, respectively. With the help of class information, DPP exploits slack masks to dynamically adjust the margins of polarization, and improve PPOT robustness with more correct polarization direction.

Optimization Algorithm for PPOT

In this section, we propose a numerical algorithm for solving the PPOT problem and the formulation can be written as

$$\min_{\boldsymbol{\gamma} \in \mathcal{B}} \mathcal{L}(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \alpha\Omega_{\text{ent}}(\boldsymbol{\gamma}) + \beta\Omega_{\text{p}}(\boldsymbol{\gamma}), \tag{11}$$

where $\alpha, \beta \geq 0$, $\Omega_{\mathrm{ent}}(\boldsymbol{\gamma})$ is the entropy regularizer, and $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ is the proposed probability polarization term which can be replaced by $\Omega_{\mathrm{dp}}(\boldsymbol{\gamma})$ in Eq. (10).

The objective of PPOT in Eq. (11) can be optimized by Generalized Conditional Gradient (GCG), which is a generalization of the conditional gradient algorithm (Bredies, Lorenz, and Maass 2009). GCG can address the general case of constrained minimization problem as follows,

$$\min_{\boldsymbol{\gamma} \in \mathcal{H}} \; g_1(\boldsymbol{\gamma}) + g_2(\boldsymbol{\gamma}),$$

where $g_1(\cdot)$ is defined as a differentiable function, and $g_2(\cdot)$ is a convex function. Besides, the constraint set $\mathcal{H}$ denotes any convex and compact subset of $\mathbb{R}^n$.

In PPOT, the polarization regularizer terms focus on the transport coupling $\boldsymbol{\gamma}_{ij}$ between the margin thresholds. This means we can initialize $\boldsymbol{\gamma}^0$ in this margin, then $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ and $\Omega_{\mathrm{dp}}(\boldsymbol{\gamma})$ are always optimized until the $\boldsymbol{\gamma}_{ij}$ approaches the inter-class or intra-class threshold. Besides, the formulation of entropy regularizer is differentiable and convex. Therefore, the GCG algorithm can be applied to optimize PPOT. The formulation of PPOT in Eq. (11) can be set as

$$g_1(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \beta \Omega_{\mathrm{p}}(\boldsymbol{\gamma}), \quad g_2(\boldsymbol{\gamma}) = \alpha \Omega_{\mathrm{ent}}(\boldsymbol{\gamma}).$$

Numerical optimization process of PPOT (Eq. (11)) is provided in Algorithm 1. Specifically, step 4 boils down to

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma} \in \mathcal{B}} \; \langle \boldsymbol{\gamma}, \mathbf{C} + \beta \nabla \Omega_{\mathrm{p}}(\boldsymbol{\gamma}) \rangle_F + \alpha \Omega_{\mathrm{ent}}(\boldsymbol{\gamma}). \quad (12)$$

Note that Eq. (12) can also be efficiently solved by Sinkhorn algorithm. In the training process, we utilize $\delta^- = \min_k \; \delta_k^-$ and $\delta^+ = \max_k \; \delta_k^+$ to unify intra-class and inter-class probability thresholds in the $\Omega_{\mathrm{p}}(\boldsymbol{\gamma})$ to simplify the PPOT.

Modeling for UDA

In this section, we focus on the PPOT-based modeling for UDA and learning a well-transferable model f for the target domain. DPP is employed as the regularizer term in the framework of PPOT. The learning model f can be decomposed as $f_c \circ f_r$, where $f_r : X \mapsto Z$ is the feature extractor and $f_c : Z \mapsto Y$ is the classifier.

According to Ben-David's (BenDavid et al. 2006) transfer theory, it is necessary to minimize the expected error on the source domain and reduce the domain discrepancy. Thus, a supervised learning task based on cross-entropy loss l_{ce} is conducted on the source domain, i.e.,

$$\mathcal{L}_{\mathrm{s}}(f_c, f_r) = \frac{1}{n_s} \sum\nolimits_{i=1}^{n_s} l_{\mathrm{ce}}(f_c(\mathbf{z}_i^s), y_i^s), \quad (13)$$

Further, PPOT framework is applied for distribution alignment. With DPP, PPOT makes that the polarized plan can explicitly assigns larger weights for intra-class sample alignment.

$$\mathcal{L}_{\mathrm{ot}}(f_r, \boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \alpha \langle \boldsymbol{\gamma}, \ln \boldsymbol{\gamma} \rangle_F$$
$$+ \beta \| \min\{(\boldsymbol{\gamma} - \boldsymbol{\Delta}_{\mathrm{slack}}^-) \odot (\boldsymbol{\gamma} - \boldsymbol{\Delta}_{\mathrm{slack}}^+), 0\} \|_1.$$

When the feature extractor f_r is fixed, the OT plan can be learned via $\boldsymbol{\gamma}^\star = \arg\min_{\boldsymbol{\gamma}} \mathcal{L}_{\mathrm{ot}}$. Such transportation establishes an explicit connection between the source and target samples. Specifically, given $\boldsymbol{\gamma}^\star$, the transported source samples $\tilde{\mathbf{z}}_i^s$ can be represented by the target via the *barycenter mapping* ψ (Courty et al. 2016):

$$\tilde{\mathbf{z}}_i^s = \psi_{\boldsymbol{\gamma}^\star_{i,:}}(\mathbf{Z}^t) = (\langle \boldsymbol{\gamma}^\star{}_{i,:}, \mathbf{1}_{n_t} \rangle)^{-1} \sum\nolimits_{j=1}^{n_t} \boldsymbol{\gamma}_{ij}^\star \mathbf{z}_j^t.$$

An intuitive explanation for ψ is that it considers the minimal cost for finding the image of source samples in the representation space of the target domain. Then, the target representation space can be further optimized via the transported source samples $\{\tilde{\mathbf{z}}_i^s, y_i^s\}_{i=1}^{n_s}$ in a supervised way,

$$\mathcal{L}_{\mathrm{t}}(f_c, f_r) = \frac{1}{n_s} \sum\nolimits_{i=1}^{n_s} l_{\mathrm{ce}}(f_c(\tilde{\mathbf{z}}_i^s, y_i^s)). \quad (14)$$

Combining the learning objectives above, the training principle of PPOT can be formulated as

$$\min_{f_c, f_r, \boldsymbol{\gamma}} \; \mathcal{L}_{\mathrm{s}}(f_c, f_r) + \lambda_1 \mathcal{L}_{\mathrm{ot}}(f_r, \boldsymbol{\gamma}) + \lambda_2 \mathcal{L}_{\mathrm{t}}(f_c, f_r), \quad (15)$$

where $\lambda_1, \lambda_2 > 0$. The model reduces domain discrepancy by minimizing the OT distance $\mathcal{L}_{\mathrm{ot}}$, and learns a discriminant classifier by minimizing the source risk $\mathcal{L}_{\mathrm{s}}$. Besides, the risk on transported source samples $\mathcal{L}_{\mathrm{t}}$ benefits knowledge learning on the target space. To obtain reliable pseudo-labels, we first pre-train the model with the source risk $\mathcal{L}_{\mathrm{s}}$.

Algorithm 1: PPOT

Input: Maximum iteration I_{max}, parameters α and β;
Output: OT plan $\boldsymbol{\gamma}^\star$;

1: Initialize $\boldsymbol{\gamma}^0 \leftarrow \frac{1}{n_s n_t} \mathbf{1}_{n_s} \mathbf{1}_{n_t}^{\mathrm{T}}$, $\varepsilon^0 \leftarrow 1$, $i \leftarrow 0$.
2: **while** $i \leq I_{max}$ and $\varepsilon^i \geq 1e^{-9}$ **do**
3: Compute $\nabla \Omega_{\mathrm{p}}(\boldsymbol{\gamma}^i)$ and $\mathbf{C} + \beta \nabla \Omega_{\mathrm{p}}(\boldsymbol{\gamma}^i)$;
4: Obtain $\boldsymbol{\gamma}^*$ of Eq. (12) via Sinkhorn algorithm;
5: Search the optimal step α^i via
$\alpha^i = \arg\min_{0 \leq \alpha \leq 1} f(\boldsymbol{\gamma}^i + \alpha \Delta\boldsymbol{\gamma}) + g(\boldsymbol{\gamma}^i + \alpha \Delta\boldsymbol{\gamma})$,
where $\Delta\boldsymbol{\gamma} = \boldsymbol{\gamma}^* - \boldsymbol{\gamma}^i$;
6: Update $\boldsymbol{\gamma}^{i+1} = \boldsymbol{\gamma}^i + \alpha^i \Delta\boldsymbol{\gamma}$;
7: Obtain OT distances $\mathcal{L}(\boldsymbol{\gamma}^{i+1})$ and $\mathcal{L}(\boldsymbol{\gamma}^i)$ via Eq. (11);
8: $\varepsilon^{i+1} = |\mathcal{L}(\boldsymbol{\gamma}^{i+1}) - \mathcal{L}(\boldsymbol{\gamma}^i)|$
9: $i \leftarrow i + 1$.
10: **end while**

Experiments and Analysis

Set Up. We evaluate PPOT on three datasets. **Office-31** (Saenko et al. 2010) has 4,652 images from 3 domains with 31 classes, i.e., *Amazon* (A), *Webcam* (W), and *Dslr* (D); **ImageCLEF** (Long et al. 2017b) has 3 domains with 12 classes, i.e., *Caltech* (C), *ImageNet* (I) and *Pascal* (P); **DomainNet** (Peng et al. 2019) is a challenging large-scale dataset includes over 0.6 million images distributed across 345 classes from 6 different domains, i.e., *Clipart* (clp), *Infograph* (inf), *Painting* (pnt), *Quickdraw* (qdr), *Real* (**rel**) and *Sketch* (skt). The implementation details and algorithm are presented in the appendix.

Comparison with SOTA Methods. We employ some state-of-the-art methods for comparison. (1) Adversarial-based methods: DANN, ADDA (Tzeng et al. 2017),

Method	Office-31							ImageCLEF						
	A→W	A→D	W→D	D→W	D→A	W→A	Avg.	I→P	P→I	I→C	C→I	C→P	P→C	Avg.
Source-only	82.3	84.5	99.8	98.0	75.6	74.5	85.8	74.8	83.9	91.5	78.0	65.5	91.3	80.7
DANN	84.5	78.6	99.6	96.8	63.6	62.8	81.6	75.0	86.0	96.2	87.0	74.3	91.5	85.0
CDAN+E	94.1	92.9	**100.0**	98.6	71.0	69.3	87.7	77.7	90.7	97.7	91.3	74.2	94.3	87.7
BCDM	95.4	93.8	**100.0**	98.6	73.1	73.0	89.0	79.5	93.2	96.8	91.3	78.9	95.8	89.3
DALN	95.2	95.4	**100.0**	99.1	76.4	76.5	90.4	80.5	93.8	97.5	92.8	78.3	95.0	89.7
DeepJDOT	88.9	88.2	99.6	98.5	72.1	70.1	86.2	77.5	90.5	95.0	88.3	74.9	94.2	86.7
RWOT	**95.1**	94.5	**100.0**	99.5	77.5	77.9	90.8	81.3	92.9	97.9	92.7	79.1	96.5	90.0
DDW-OT	92.1	90.8	**100.0**	**100.0**	73.9	68.5	87.6	**82.6**	92.8	**98.5**	93.9	79.9	96.7	90.7
PCT	94.6	93.8	99.9	98.7	77.2	76.0	90.0	78.5	93.1	97.0	92.2	75.7	95.4	88.7
DMP	93.0	91.0	**100.0**	99.0	71.4	70.2	87.4	80.7	92.5	97.2	90.5	77.7	96.2	89.1
PGFL	90.7	93.8	**100.0**	99.1	78.1	76.4	89.6	78.7	92.9	96.2	94.2	80.2	96.8	89.8
PPOT	$94.6^{0.3}$	**96.1**$^{0.4}$	**100.0**$^{0.0}$	$98.5^{0.2}$	**86.3**$^{0.5}$	**85.4**$^{0.3}$	**93.5**	$82.0^{0.4}$	**95.0**$^{0.4}$	$98.2^{0.2}$	**96.1**$^{0.2}$	**81.8**$^{0.4}$	**97.8**$^{0.2}$	**91.9**

Source-only	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	19.3	37.5	11.1	52.2	41.0	32.2
inf	30.2	-	31.2	3.6	44.0	27.9	27.4
pnt	39.6	18.7	-	4.9	54.5	36.3	30.8
qdr	7.0	0.9	1.4	-	4.1	8.3	4.3
rel	48.4	22.2	49.4	6.4	-	38.8	33.0
skt	46.9	15.4	37.0	10.9	47.0	-	31.4
Avg.	34.4	15.3	31.3	7.4	40.4	30.5	26.6

DANN	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	15.5	34.8	9.5	50.8	41.4	30.4
inf	31.8	-	30.2	3.8	44.8	25.7	27.3
pnt	39.6	15.1	-	5.5	54.6	35.1	30.0
qdr	11.8	2.0	4.4	-	9.8	8.4	7.3
rel	47.5	17.9	47.0	6.3	-	37.3	31.2
skt	47.9	13.9	34.5	10.4	46.8	-	30.7
Avg.	35.7	12.9	30.2	7.1	41.4	29.6	26.1

BCDM	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	19.9	38.5	15.1	53.2	43.9	34.1
inf	31.9	-	32.7	6.9	44.7	28.5	28.9
pnt	42.5	19.8	-	7.9	54.5	38.5	32.6
qdr	23.0	4.0	9.5	-	16.9	16.2	13.9
rel	51.9	24.9	51.2	8.7	-	40.6	35.5
skt	53.7	20.5	46.0	13.1	53.4	-	37.1
Avg.	40.6	17.8	35.6	10.3	44.3	33.5	30.4

SCDA	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	18.6	39.3	5.1	55.0	44.1	32.4
inf	29.6	-	34.0	1.4	46.3	25.4	27.3
pnt	44.1	19.0	-	2.6	56.2	42.0	32.8
qdr	30.0	4.9	15.0	-	25.4	19.8	19.0
rel	54.0	22.5	51.9	2.3	-	42.5	34.6
skt	55.6	18.5	44.7	6.4	53.2	-	35.7
Avg.	42.6	16.7	37.0	3.6	47.2	34.8	30.3

DCAN	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	18.5	43.6	17.1	60.3	45.8	37.1
inf	39.7	-	38.4	5.9	54.6	28.5	33.4
pnt	48.6	19.7	-	9.9	61.7	41.2	36.2
qdr	33.2	5.6	16.1	-	18.4	16.2	17.9
rel	53.7	18.5	50.5	4.0	-	33.4	32.0
skt	57.6	17.3	47.3	10.1	55.3	-	37.5
Avg.	46.6	15.9	39.2	9.4	50.1	33.0	32.4

PPOT	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	18.7	38.4	15.7	55.8	45.2	34.8
inf	51.5	-	36.3	9.1	54.1	36.9	37.6
pnt	54.1	22.8	-	11.1	59.3	43.9	38.2
qdr	25.1	2.6	6.5	-	12.3	16.1	14.0
rel	58.0	24.6	50.8	18.4	-	44.3	39.2
skt	60.4	20.9	44.8	23.0	58.1	-	41.4
Avg.	49.8	17.9	35.4	15.5	47.7	37.3	**34.0**

Table 1: Accuracies (%) on Office-31, ImageCLEF (ResNet-50), and DomainNet (ResNet-101). For DomainNet, in each subtable, the column-wise domains are selected as the source domain and the row-wise domains are selected as the target domain.

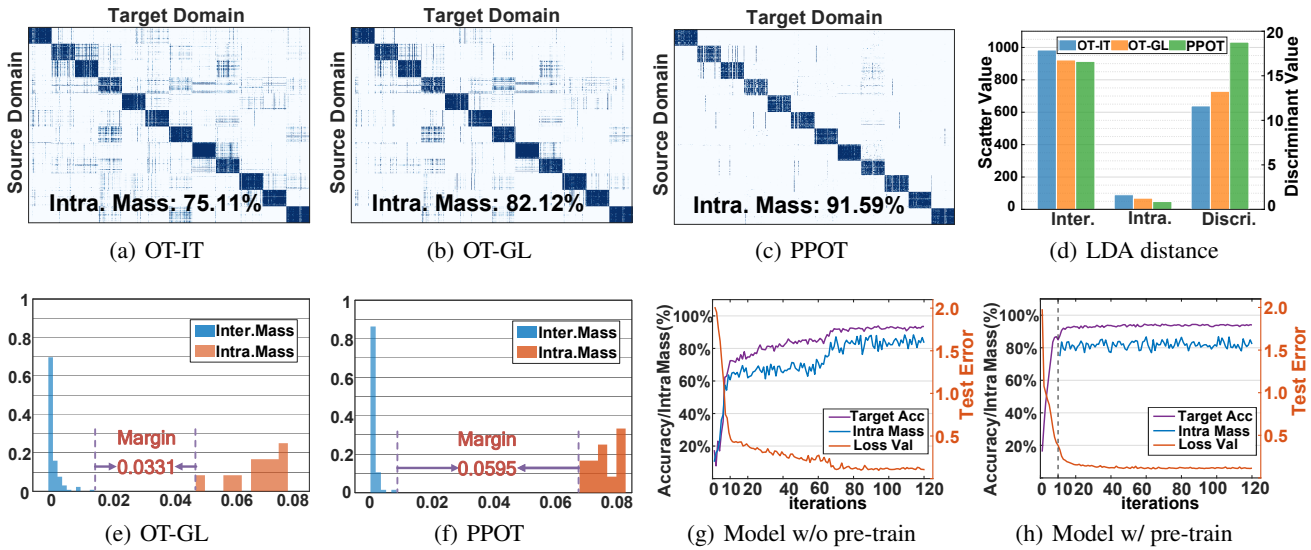
CDAN+E, MDD (Zhang et al. 2019), BCDM, SCDA (Li et al. 2021b), and DALN. (2) Distance-metric and OT-based methods: PCT (Tanwisuth et al. 2021), DMP (Luo et al. 2022), DCAN (Li et al. 2022), PGFL (Du et al. 2023) and DeepJDOT, RWOT, DDW-OT (Wang et al. 2022).

The results on Office-31, ImageCLEF and DomainNet are shown in Table 1. PPOT outperforms other methods and achieves the best average accuracies **93.5**% on Office-31 and **91.9**% on ImageCLEF. Compared with RWOT and DDW-OT, which propose weighted OT to align domains, PPOT directly explores the structure of OT plan via the dynamic polarization mechanism. Thus, PPOT obtains superior results over these methods. Office31 tasks D→A and W→A are common difficulties for all methods. PPOT not only achieves the highest accuracies on the two tasks but also surpasses the second-best results by a large margin. Besides, the average accuracy on ImageCLEF is at least **1.2**% higher than others. On DomainNet, PPOT improves the average accuracy to **34.0**%, and outperforms other methods by at least **1.6**%. The results above validate the superiority of PPOT in dealing with UDA under varying difficulties and data scales.

Visualization of OT Plan. To make an intuitive comparison among OT plans obtained by different methods, we visualize $\boldsymbol{\gamma}^\star$ in Figure 4(a)-(c). The block diagonal structure implies intra-class mass and the remaining parts imply inter-class mass. Higher intra-class mass represents a more certain OT plan. Compared to OT-IT and OT-GL, PPOT has a more clear diagonal structure and obtains the highest intra-class mass with **91.59**%, which validates that PPOT can learn a better OT plan in reducing domain discrepancy.

Margin of OT Plan. To verify that PPOT can polarize the values of the OT plan, we provide the histograms of intra-class and inter-class mass in Figure 4(e)-(f). Compared to OT-GL, PPOT has a much larger margin between the intra-class and inter-class mass. Besides, the transport masses of PPOT concentrate around a larger or smaller value. These results indicate the effectiveness of DPP, which is helpful in learning a classifier with a separable decision boundary.

LDA Distance. To compare the discriminability of OT models, we exploit the LDA criterion to compute the distance of representations across domains. LDA distance, i.e., the value of Discri., is the ratio of the inter-class distance to the intra-class distance, represents the class separability of representations. The result is shown in Figure 4(d), compared to OT-IT and OT-GL, PPOT achieves the highest LDA distance, which quantitatively demonstrates that PPOT can



(a) OT-IT (b) OT-GL (c) PPOT (d) LDA distance

(e) OT-GL (f) PPOT (g) Model w/o pre-train (h) Model w/ pre-train

Figure 4: Method analysis on ImageCLEF task P→I. (a)-(c): Heat-maps of OT plan $\boldsymbol{\gamma}^*$, where darker colors represent larger probability coupling values. (d): LDA distance for OT methods with different regularizers. (e)-(f): Histogram of intra/inter-class masses. (g)-(h): Accuracy, intra class mass sum of $\boldsymbol{\gamma}^*$ and loss w.r.t. different iterations. Best viewed in color.

Method	$\mathcal{L}_{\text{ot}}$	$\mathcal{L}_{\text{t}}$	P→I	I→P	Avg.
OT-IT	✓	✓	91.5	78.5	85.0
OT-GL	✓	✓	93.1	80.5	86.8
PPOT (w/ PP)	✓	✓	94.0	81.0	87.8
PPOT (w/ DPP)	✓		93.5	80.3	86.9
PPOT (w/ DPP)		✓	94.7	81.4	88.1
PPOT (w/ DPP)	✓	✓	**95.0**	**82.0**	**88.8**

Table 2: Accuracies (%) of ablation study on ImageCLEF.

Task	OT-GL	PPOT (w/ PP)	PPOT (w/ DPP)
P→I	2.32s (145it)	0.21s (65it)	0.22s (74it)
I→P	1.56s (123it)	0.16s (63it)	0.25s (59it)

Table 3: Time (s) and loops (it) comparison on ImageCLEF.

learn distinguishable representations.

Ablation Study. To compare PPOT (w/ DPP) with IT, GL and PPOT (w/ PP), and evaluate the effectiveness of the loss items in Eq. (15), we conduct ablation experiments on ImageCLEF. In Table 2, we can find that PPOT (w/ DPP) achieves the best accuracy on tasks P→I and I→P. PPOT (w/ DPP) also achieves better accuracies than OT-IT and OT-GL. These results prove the superiority of our PPOT framework. The 2nd column shows that $\mathcal{L}_t$ based on the formulation of PPOT (w/ DPP) is helpful in exploring the structure of the target domain and improving the accuracy significantly.

Training Stability. We evaluate the training stability of PPOT on ImageCLEF task P→I. In the training process, we pre-train the model on the source domain to obtain more reliable target pseudo-labels. Thus, we explore the training stability w.r.t w/o pre-train and w/ pre-train. In Figure 4(h), the curves have more fluctuations and the model w/o pre-train takes a longer time to converge. Figure 4(g) shows more stable and smooth curves. The better performance of model w/ pre-train also proves the effectiveness of pre-train strategy.

Time Comparison. We evaluate the efficiency of PPOT by comparing the iteration loops and convergence time for obtaining the OT solution. The convergence criteria are identical across all experiments. In Table 3, we can see that both PPOT (w/ PP) and PPOT (w/ DPP) can significantly reduce the loops of iteration and improve the convergence rate with the GCG algorithm. Despite the additional computation introduced by the dynamic threshold slack, PPOT (w/ DPP) gets comparable speed compared to PPOT (w/ PP). Specially, the complexity of OT-GL is $\mathcal{O}(Kn^2 \log n)$ while PPOT induces smaller $\mathcal{O}(n^2)$ complexity. Compared with the linear convergence rate of smooth (quadratic) function $\Omega_{\text{p}}(\gamma)$ or $\Omega_{\text{dp}}(\gamma)$, the non-smoothness of OT-GL leads to slower convergence.

Conclusion

In this paper, we propose a novel PPOT framework to characterize the structure of the OT plan. Specifically, the probability polarization introduces a margin between similar and dissimilar transport pairs, which can guide the polarization direction of the OT plan. Besides, the dynamic probability polarization adjusts the margin by incorporating class information, which further captures more clearer cross-domain sample correspondence and improve robustness. The mathematical relationship between the polarization direction and gradient guarantees the interpretability of PPOT. Empirical evaluations on UDA demonstrate the effectiveness of PPOT. An interesting future direction is studying algorithm and application of PPOT in multi-source domain adaptation.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62376291, 61976229), in part by Guangdong Basic and Applied Basic Research Foundation (2023B1515020004), in part by Guangdong Province Key Laboratory of Computational Science at Sun Yat-sen University (2020B1212060032), in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA), and in part by the Hong Kong Research Grants Council (Project 11204821).

References

BenDavid, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19.

Bredies, K.; Lorenz, D. A.; and Maass, P. 2009. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42: 173–193.

Chen, L.; Chen, H.; Wei, Z.; Jin, X.; Tan, X.; Jin, Y.; and Chen, E. 2022. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7181–7190.

Chen, S.; Niu, G.; Gong, C.; Li, J.; Yang, J.; and Sugiyama, M. 2021. Large-margin contrastive learning with distance polarization regularizer. In *International Conference on Machine Learning*, 1673–1683.

Chuang, C.; Jegelka, S.; and AlvarezMelis, D. 2023. InfoOT: Information Maximizing Optimal Transport. In *International Conference on Machine Learning*, 6228–6242.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30.

Courty, N.; Flamary, R.; and Tuia, D. 2014. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, 274–289.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2016. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1: 1–40.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26.

Damodaran, B. B.; Kellenberger, B.; Flamary, R.; Tuia, D.; and Courty, N. 2018. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 447–463.

Du, Y.; Zhou, D.; Xie, Y.; Lei, Y.; and Shi, J. 2023. Prototype-Guided Feature Learning for Unsupervised Domain Adaptation. *Pattern Recognition*, 135: 109154.

Flamary, R.; Courty, N.; Rakotomamonjy, A.; and Tuia, D. 2014. Optimal transport with Laplacian regularization. In *Advances in Neural Information Processing Systems*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189.

Kantorovich, L. V. 2006. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4): 1381–1382.

Li, M.; Zhai, Y.; Luo, Y.; Ge, P.; and Ren, C. 2020. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13936–13944.

Li, S.; Lv, F.; Xie, B.; Liu, C. H.; Liang, J.; and Qin, C. 2021a. Bi-classifier determinacy maximization for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8455–8464.

Li, S.; Xie, B.; Lin, Q.; Liu, C. H.; Huang, G.; and Wang, G. 2022. Generalized domain conditioned adaptation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4093–4109.

Li, S.; Xie, M.; Lv, F.; Liu, C. H.; Liang, J.; Qin, C.; and Li, W. 2021b. Semantic concentration for domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9102–9111.

Liu, Y.; Zhou, Z.; and Sun, B. 2023. COT: Unsupervised Domain Adaptation With Clustering and Optimal Transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19998–20007.

Long, M.; Cao, Y.; Cao, Z.; Wang, J.; and Jordan, M. I. 2019. Transferable Representation Learning with Deep Adaptation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 3071–3085.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017a. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, 2208–2217.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017b. Deep transfer learning with joint adaptation networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2208–2217.

Long, T.; Sun, Y.; Gao, J.; Hu, Y.; and Yin, B. 2022. Domain adaptation as optimal transport on Grassmann manifolds. *IEEE Transactions on Neural Networks and Learning Systems*.

Luo, Y.; and Ren, C. 2021. Conditional bures metric for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13989–13998.

Luo, Y.; and Ren, C. 2023. MOT: Masked Optimal Transport for Partial Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3531–3540.

Luo, Y.; Ren, C.; Dai, D.; and Yan, H. 2022. Unsupervised domain adaptation via discriminative manifold propagation.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3): 1653–1669.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1406–1415.

Redko, I.; Habrard, A.; and Sebban, M. 2017. Theoretical analysis of domain adaptation with optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 10*, 737–753.

Ren, C.; Luo, Y.; and Dai, D. 2023. BuresNet: Conditional bures metric for transferable representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4198–4213.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision*, 213–226.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3723–3732.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 443–450.

Tanwisuth, K.; Fan, X.; Zheng, H.; Zhang, S.; Zhang, H.; Chen, B.; and Zhou, M. 2021. A prototype-oriented framework for unsupervised domain adaptation. In *Advances in Neural Information Processing systems*, volume 34, 17194–17208.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Villani, C.; et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Wang, B.; Wang, S.; Zhang, Z.; Zhao, X.; and Fu, Z. 2022. Decomposed-distance weighted optimal transport for unsupervised domain adaptation. *Applied Intelligence*, 52(12): 14070–14084.

Xu, R.; Liu, P.; Wang, L.; Chen, C.; and Wang, J. 2020. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4394–4403.

Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; and Saminger-Platz, S. 2017. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. In *International Conference on Learning Representations*.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, 7404–7413.