MOT: Masked Optimal Transport for Partial Domain Adaptation

You-Wei Luo Chuan-Xian Ren*
School of Mathematics, Sun Yat-Sen University, China
luoyw28@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

Abstract

As an important methodology to measure distribution discrepancy, optimal transport (OT) has been successfully applied to learn generalizable visual models under changing environments. However, there are still limitations, including strict prior assumption and implicit alignment, for current OT modeling in challenging real-world scenarios like partial domain adaptation, where the learned transport plan may be biased and negative transfer is inevitable. Thus, it is necessary to explore a more feasible OT methodology for real-world applications. In this work, we focus on the rigorous OT modeling for conditional distribution matching and label shift correction. A novel masked OT (MOT) methodology on conditional distributions is proposed by defining a mask operation with label information. Further, a relaxed and reweighting formulation is proposed to improve the robustness of OT in extreme scenarios. We prove the theoretical equivalence between conditional OT and MOT, which implies the well-defined MOT serves as a computation-friendly proxy. Extensive experiments validate the effectiveness of theoretical results and proposed model.

1. Introduction

For real-world visual data in unconstrained environments, detecting the shift of data distributions and measuring the statistical discrepancy are essential problems for learning generalizable model [29, 30, 51]. Considerable efforts have been made to build well-defined metrics on distributions, where the optimal transport (OT) based distances have shown significant advantages in sample-wise correlation characterization and geometric interpretability [6,7,9,11,12,24,27]. Under the guarantees of appealing theoretical properties, OT-based modules [8, 10, 11, 16, 20, 24] are extensively explored to learn the transferrable models. These advanced models have also been successfully applied to the computer vision and pattern recognition tasks in changing environments, e.g., object classification [31, 39], semantic segmentation [52] and medical images [42].

A typical scenario for learning under distribution shift is known as domain adaptation (DA), where the model trained on the labeled source domain P is expected to be transferrable to unlabeled target domain with different data distribution Q. In DA, source domain and target domain share the same label space $\mathcal{Y}$, then inspired by distribution adaptation theory [1, 48], the OT-based models usually focus on learning invariant properties, e.g., marginal invariant representations [8, 11, 28], adversarial invariant representations [16], conditional invariant representations [19, 24, 38] and joint invariant representations [10, 20].

However, in real-world scenarios, the label space usually changes with the shifting distributions, which induces the partial DA (PDA) problems [2, 3, 16, 21, 25]. Recent theoretical results [38, 49] also imply that the shift on label distributions [47] is non-negligible in application, and the label distribution correction is necessary to achieve a sufficiently small joint risk across domains. Therefore, if label distributions P_Y and Q_Y are different, a correction (e.g., reweighted and relaxation) on source P_Y is usually necessary [22, 47]. Since there are strict constraints on the marginal distributions, classical OT models will learn incorrect sample correlation and further induce the negative transfer problem. Typically, to address these issues, there are two types of variants for OT. 1) Reweighted OT [16, 32, 33] introduces a reweighted operation, which aims to detect the outlier classes (i.e., unseen classes in target domain) and decrease their masses in P_Y. Then the OT assignment, which is constrained by reweighted source distribution, will only transport the information that is shared by both domains. However, the effectiveness of these models directly depends on the reweighted function w, where models could be error-prone when weight w is inaccurate. 2) Unbalanced OT (UOT) [5, 11] introduces relaxation to the strict marginal constraints by penalizing the assignments that don't meet the constraints. Such a relaxation allows the assignments to focus on the correct correlation with low transport cost and ignore the incorrect transportation to outlier classes with higher cost. But, since the relaxation is

*Corresponding Author.

applied to original distributions, the penalty will be unaffordable and the relaxation will fail when cross-domain distributions are significantly different.

Besides, as a sufficient condition to achieve successful DA, the conditional shift correction has received increasing attention recently [14, 19, 23, 38]. Several pioneering works [19, 24, 32] on the OT modeling between conditional distributions have shown great potential in mitigating negative transfer and reducing generalization risk. But, there are still several limitations, i.e., strict assumption on kernel Gaussian prior [24], implicit conditional alignment via reweighted source [32] and implicit proxy with small intra-class discrepancy assumption [19]. Recently, OT with mask is proposed to integrating label information into OT. Zhang et al. [46] first define the mask on transport plan. Gu et al. [15] further establish the mask theory for Gromov-Wasserstein. However, theoretical understanding on the relation between mask mechanism and conditional OT is unexplored. Thus, it is necessary to explore a general and explicit formulation of OT for sufficient conditional distribution matching, and develop computation-friendly algorithms for application.

Generally, to alleviate the difficulties in current OT methodology for label distribution correction and conditional distribution matching, known as generalized label shift (GLS) correction, we are interested in two major problems: 1) rigorous modeling and explicit algorithm for OT between conditional distributions; 2) unbiased and relaxed transport plan learning for extreme generalization scenario, e.g., PDA. In this paper, we propose a novel masked OT (MOT) methodology by introducing a relaxed and reweighted OT formulation and the conditional mask mechanism. Further, we derive an efficient fixed-point method for learning OT assignment, and propose an OT-based invariant risk model to deal with the extreme PDA problem. Our contributions can be summarized as follows.

- The theoretical connection between MOT and conditional OT is proved for commonly used OT formulations. The main result ensures the proposed mask mechanism and MOT model are sufficient to characterize the label-conditioned sample correspondence and measure the conditional discrepancy explicitly.
- A relaxation and reweighting based OT formulation is proposed, which overcomes the sensitivity of current OT modelings to weight estimation and intrinsic label discrepancy. Intuitive analysis is presented to show the advantages of MOT methodology.
- With theoretical guarantees, a computation-friendly empirical estimation with explicit fixed-point algorithm is proposed as proxy for conditional OT. Then an equivalent risk model is proposed to learn minimized risk on both source and transported source domains. The application to PDA is extensively evaluated, which validates the effectiveness of theoretical results and proposed model.

2. Methodology: Masked OT

Notation. Let X and Y be the covariate and label variable, which take their values from $\mathcal{X}$ and $\mathcal{Y}$. P and Q denote the distributions of source and target domains, where lowercase letters p and q denote the probability density functions (PDFs) and subscripts represent the corresponding random variables, e.g., P_Y implies the label distribution on source. Given a positive convex function ϕ with $\phi(1) = 0$, the ϕ-divergence is defined as $D_\phi(P\|Q) \triangleq \mathbb{E}_Q[\phi(\frac{\mathrm{d}P}{\mathrm{d}Q})]$. Especially, when $\phi(t) = t\ln t$, $D_\phi(P\|Q)$ boils down to the well-known KL divergence. For a learning model $f : \mathcal{X} \to \mathcal{Y}$, given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, we denote the risk of f on P_{XY} as $\varepsilon_P(f) = \mathbb{E}_{P_{XY}}[\ell(f(X), Y)]$.

2.1. Preliminary

GLS Correction and PDA. The label distributions (e.g., class proportions) in real-world scenarios usually induces the increasing discrepancy between marginal distributions P_Y and Q_Y [22, 47]. Such a phenomenon is formally known as label/target shift [22, 47]. Note PDA can be taken as an extreme scenario of label shift, where the proportions of outlier classes will shift to 0 in target distribution Q, i.e., $\mathrm{supp}(q_Y) \subset \mathrm{supp}(p_Y)$. Thus, many PDA methods are developed under the shift correction framework, e.g., reweighted moment matching [34,45], reweighted OT [16, 33], relaxation OT [11, 28] and reweighted adversarial adaptation [2, 3]. One the other hands, to achieve better generalization, models are usually required to extract invariant discriminant information and intrinsic pattern across domains, which can be theoretically characterized as the matching between conditional distributions $P_{Z|Y}$ and $Q_{Z|Y}$ [14, 18]. These models aim to develop discrepancy-based objectives for learning conditional invariant representation Z, e.g., conditional adversarial loss [23], conditional Bures metric [24], conditional moment matching [40, 50] and local structure transfer [25, 41].

Recently, by merging label shift and conditional shift, GLS is extensively studied to explore the transferability and discriminability of DA models [19, 35, 38, 47]. Theoretically, shifting distributions lead to the biased risk estimation (i.e., $\varepsilon_P(f) \neq \varepsilon_Q(f)$) and cluster misalignment (i.e., $P_{Z|Y} \neq Q_{Z|Y}$), which explicitly degrades generalization performance. Typically, GLS can be mitigated by introducing the reconstructed source P^w_{ZY} parameterized by importance weight w and conditional invariant representation Z. Ideally, the model trained on weighted source P^w_{ZY} is sufficient for small generalization error and successful transfer [38], which implies GLS-based model can be effective in dealing with complex dataset shift scenarios, e.g., PDA.

UOT. Recently, OT-based models have been successfully applied to transfer learning problems [8, 10, 11, 20, 28, 33].

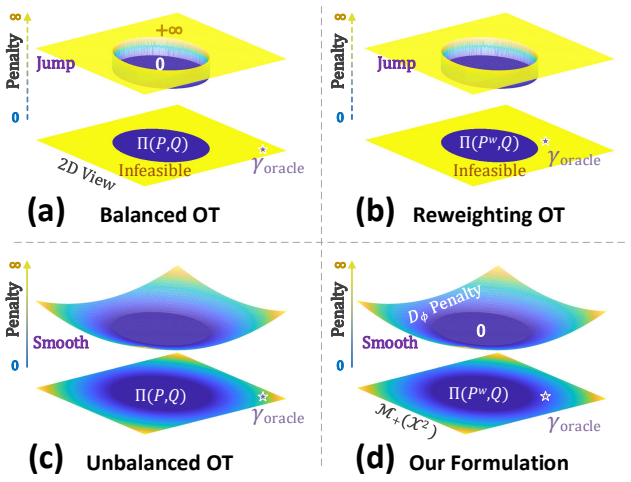


Figure 1. Illustration of different OT formulations. **(a)-(b)**: For OTs with constraints on marginal distributions, though weight w makes ideal solution γ_{oracle} closer to the feasible region Π, γ_{oracle} is still unachievable ($+\infty$ penalty). **(c)**: UOT with relaxation replaces the constraints with smooth penalty, but γ_{oracle} is still far away from the center, which leads to large penalty. **(d)**: Our formulation Eq. (3) with relaxation D_ϕ and reweighted P^w ensures that γ_{oracle} belongs to the feasible region with smaller penalty.

In classical OT, by introducing the entropy regularization $H(\gamma) = \mathbb{E}_\gamma[-\ln(\mathrm{d}\gamma)]$, the Sinkhorn divergence [9] between distributions P and Q can be formulated as

$$S^\lambda(P,Q,c) = \min_{\gamma\in\Pi(P,Q)} \int c\,\mathrm{d}\gamma - \lambda H(\gamma), \tag{1}$$

where $\Pi(P,Q)$ is the set of probabilistic couplings over P and Q, λ is parameter of sparsity penalty, c is the cost function for transport, e.g., c is defined on $\mathcal{X}^2$ for marginal OT [8, 20] and $(\mathcal{X}\times\mathcal{Y})^2$ for joint OT [10, 44].

Further, to deal with more general scenarios, e.g., unequal total mass and label shift, UOT [6, 7, 11] is proposed as a relaxation of classical OT Eq. (1). Specifically, by replacing the strict marginal constraint $\gamma\in\Pi(P,Q)$ with relaxed penalty terms on γ, UOT can be formulated as

$$\begin{aligned} S^{\lambda,\beta}(P,Q,c) = \min_{\gamma\in\mathcal{M}_+} &\int c\,\mathrm{d}\gamma - \lambda H(\gamma) \\ &+ \beta\left[D_\phi(\gamma_P\|P) + D_\phi(\gamma_Q\|Q)\right], \end{aligned} \tag{2}$$

where γ_P and γ_Q are the marginals of γ, β is the parameter for marginal penalty, $\mathcal{M}_+$ is the space of distributions, e.g., $\mathcal{M}_+(\mathcal{X}^2)$ over space $\mathcal{X}^2$ for marginal UOT [6, 11].

2.2. New Formulation and Analysis

Motivation. Though the reweighted strategy [16, 32, 38, 47] and UOT methodology [11] have been explored for knowledge transfer, there are still potential weaknesses that we are interested in: 1) though the marginal alignment between P^w and Q mitigates label shift, e.g., MMD [47] and OT [16, 32, 33] on weighted source, it ignores the conditional shift which usually degrades the discriminability; 2) the performance of adaptation with P^w significantly depends on the precision of estimation w as Fig. 1 (b), which may degrade the robustness of shift correction model; 3) the marginal penalty with P in UOT [11] is sensitive to the degree of label shift as Fig. 1 (c), especially for the extreme PDA scenario with large label discrepancy.

To address the challenges above, we first introduce a novel OT with reweighted and relaxation, which reduces the uncertainty of estimation w and alleviates the large penalty on original distribution P simultaneously. Then we propose the *conditional UOT* formulation, which characterizes label-wise correlations and mitigates conditional shift.

Relaxed and Reweighted Formulation. To overcome the bias of risk estimation ε_P with label proportion P_Y on original source domain, the reweighted strategies for modifying distribution are widely applied [19, 35, 38, 47]. To unify the different formulations, we first present a rigorous definition of w-reweighted source as follows.

Definition 1 *Given a reweighted function $w:\mathcal{Y}\to\mathbb{R}_+$ such that $w\cdot p_Y$ is also a PDF on $\mathcal{Y}$.*
(a) The w-reweighted source P^w is defined as $p^w_Y = w\cdot p_Y$, $p^w_{X|Y} = p_{X|Y}$ and $p^w_X = \int_{\mathcal{Y}} p^w_{Y=y}\cdot p_{X|Y=y}\mathrm{d}y$.
(b) The optimal weight w^* satisfies $P^{w^*}_Y = Q_Y$.

An important property of reweighted is that if $P_{X|Y} = Q_{X|Y}$, then dataset shift can be addressed on the w-reweighted source, i.e., $P^{w^*}_{XY} = P_{X|Y}P^{w^*}_Y = Q_{X|Y}Q_Y = Q_{XY}$ and $\varepsilon_{P^{w^*}} = \varepsilon_Q$. Based on the reweighted source, we will next show that the biased sample-wise transports in current OT can be mitigated by the relaxation and reweighted formulation.

As discussed before, though UOT can deal with the label shift scenarios by relaxing the constraints on marginal distribution, the OT model is still required to learn coupling γ with marginal γ_P close to P. It implies such a model may fail in some extreme scenarios where the ideal marginal are actually away from the original source domain P, e.g., γ should not assign values to the outlier classes in P. Therefore, it is necessary to reformulate the original UOT by modifying the marginal penalty on P as P^w:

$$\begin{aligned} S^{\lambda,\beta}(P^w,Q,c) = \min_{\gamma\in\mathcal{M}_+(\mathcal{X}^2)} &\int c\,\mathrm{d}\gamma - \lambda H(\gamma) \\ &+ \beta\left[D_\phi(\gamma_{P^w}\|P^w) + D_\phi(\gamma_Q\|Q)\right]. \end{aligned} \tag{3}$$

An intuitive illustration of Eq. (3) and related OT formulations are shown in Fig. 1, where the ideal γ_{oracle} such that its marginals are Q. For original (i.e., balanced) OT Eq. (1) in Fig 1 (a)-(b), the feasible regions are probabilistic

coupling set Π, which implies ideal solutions are always unachievable unless w is the optimal w^*. Though the marginal constraints are relaxed, and feasible regions are extended to $\mathcal{M}_+$ in UOT Eq. (2), γ_{oracle} may be still hard to learn since the penalty may be large, i.e., Fig 1 (c). Specifically, since the center with 0 penalty value is $\Pi(P, Q)$, the large discrepancy between P and Q makes γ_{oracle} far away from the low penalty region. For our formulation Eq. (3) in Fig 1 (d), the model inherits the advantages of UOT and reweighted OT. The reweighted source P^w further ensures that γ_{oracle} is feasible, and the penalty value is smaller than UOT. Besides, the relaxation on P^w reduces the risk of estimation w, then the neighbours of Π are also feasible. Thus, the ideal solution is still achievable even if $w \neq w^*$, which overcomes the limitation of reweighted OT. In conclusion, new formulation Eq. (3) ensures the existence of unbiased γ_{oracle} for target domain Q and the feasibility of learning.

Conditional OT. Though the reweighted and relaxation formulation Eq. (3) ensures the possibility of learning unbiased γ for target domain Q, the conditional shift correction for intrinsic structure transfer is still not guaranteed. Therefore, to further achieve the conditional invariant property, we now define conditional transport mechanism to learn label-wise correlations and mitigate the negative transports between inter-class sample pairs.

Definition 2 (Conditional OT) *For any non-negative coefficient function $\alpha(\cdot)$ on $\mathcal{Y}$ such that $\text{supp}(\alpha) = \text{supp}(p_Y) \cap \text{supp}(q_Y)$ and $\sum_{y\in\mathcal{Y}} \alpha(y) = 1$, the conditional UOT between P^w and Q is defined as*

$$\text{OT}^{\alpha}_{\text{cond}}(P^w, Q, c) = \sum\nolimits_{y\in\mathcal{Y}} \alpha(y)\text{OT}(P^w_{X|y}, Q_{X|y}, c). \quad (4)$$

Note that the general formulation Eq. (4) is suitable for different OT-based models, e.g., Kantorovich OT, Sinkhorn OT S^{λ} and UOT $S^{\lambda,\beta}$. An intuitive explanation for conditional OT is that it can be taken as the sliced OT between conditional distributions $P_{X|y}$ and $Q_{X|y}$ over all label conditions $y \in \mathcal{Y}$. The coefficient α ensures that the conditional transports are only carried out on shared classes, i.e., $\text{supp}(\alpha) = \text{supp}(p_Y) \cap \text{supp}(q_Y)$. In conclusion, such a formulation ensures that the OT problems for all shared classes are considered, i.e., $\alpha(y) > 0$ if y is shared classes.

For empirical application, the conditional OT in Eq. (4) ensures that the conditional transport model is generally feasible for extreme scenarios, e.g., outlier classes in P-DA and unseen classes in open-set scenario. However, since there are multiple (even 'infinite' for regression scenario with continuous Y) minimization problems (i.e., $\text{OT}(P^w_{X|y}, Q_{X|y}, c)$) in Eq. (4), this formulation is not generally applicable and has no closed-form solution. Besides, the coefficient α is a hyperparameter, which is usually hard to estimate. In the next, we will develop an equivalent proxy for conditional OT, i.e., MOT, which provides an interpretable solution for α via the essential of data.

2.3. Theoretical Proxy and Estimation

In this section, we focus on the computation-friendly proxy of conditional OT. Considering the empirical scenario with finite samples, where labeled source data $\mathcal{D}^s$=$\{\mathbf{x}^s_i, y^s_i\}^n_{i=1}$ and unlabeled target data $\mathcal{D}^t$=$\{\mathbf{x}^t_i\}^m_{i=1}$ are given, we first define the mask matrix for discrete OT and propose the MOT methodology. Then we prove the theoretical connection between *MOT* and *conditional OT* Eq. (4), which ensures that MOT can be a generally applicable proxy with explicit fixed point solution.

Usually, P and Q are denoted as empirical distributions, e.g., $P_X = \frac{1}{n}\sum_i \delta_{\mathbf{x}^s_i}$. Then the discrete formulation of Eq. (3) can be written as

$$S^{\lambda,\beta}(P^w, Q, \mathbf{C}) = \min_{\mathbf{\Gamma}\in\mathcal{M}_+(\mathbb{R}^{n\times m})} \langle \mathbf{\Gamma}, \mathbf{C}\rangle_F + \lambda \langle \mathbf{\Gamma}, \ln \mathbf{\Gamma}\rangle_F + \beta \left[D_\phi(\mathbf{\Gamma}_{P^w} \| P^w) + D_\phi(\mathbf{\Gamma}_Q \| Q)\right], \quad (5)$$

where $\mathbf{\Gamma} \in \mathbb{R}^{n\times m}$ is known as the transport plan matrix, $\mathbf{C}_{ij} = \|\mathbf{x}^s_i - \mathbf{x}^t_j\|^2_2$ is the pair-wise cost, and $\langle \mathbf{\Gamma}, \mathbf{C}\rangle_F = \sum_{ij} \mathbf{\Gamma}_{ij}\mathbf{C}_{ij}$ is the Frobenius inner product. For conditional OT, the empirical conditional distributions are defined as $P_{X|Y=l} = \frac{1}{n_l}\sum_i \mathbb{I}_{[y^s_i=l]}\ \delta_{\mathbf{x}^s_i}$ ($Q_{X|Y}$ is similar), where n_l is the size of the l-th class source data. Then the dimensions of plan and cost matrices in each OT problem of Eq. (4) are $n_l \times m_l$. Therefore, it is clear that the conditional OT induces multiple OT problems between the cross-domain clusters, and the plan in each sub-problem characterizes the class-level dependency between source and target domains.

To overcome slice-wise computation problem of conditional OT, *the key idea is incorporating the multiple sub-problems into single OT problem with mask mechanism.*

Definition 3 (Masked OT) *Given labels $\{y^s_i\}^n_{i=1}$ and $\{y^t_i\}^m_{i=1}$, the mask matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ is defined as*

$$\mathbf{M}_{ij} \triangleq \begin{cases} 1, & \text{if } y^s_i = y^t_j, \\ \infty, & \text{if } y^s_i \neq y^t_j. \end{cases} \quad (6)$$

Then the masked cost is defined as $\tilde{\mathbf{C}} = \mathbf{C} \odot \mathbf{M}$, and the masked OT is formulated as

$$\text{OT}_{\text{mask}}(P^w, Q, \tilde{\mathbf{C}}) = \text{OT}(P^w, Q, \tilde{\mathbf{C}}). \quad (7)$$

With the label-conditioned mask $\mathbf{M}$, the inter-class distances in the modified cost $\tilde{\mathbf{C}}$ will be enlarged to infinity. Intuitively, such a mask on cost ensures that the transport plan $\mathbf{\Gamma}$ will only assign values to the intra-class sample pair. For example, the masked UOT is formulated as

$$S^{\lambda,\beta}_{\text{mask}}(P^w, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma}\in\mathcal{M}_+(\mathbb{R}^{n\times m})} \langle \mathbf{\Gamma}, \tilde{\mathbf{C}}\rangle_F + \lambda \langle \mathbf{\Gamma}, \ln \mathbf{\Gamma}\rangle_F + \beta \left[D_\phi(\mathbf{\Gamma}_{P^w} \| P^w) + D_\phi(\mathbf{\Gamma}_Q \| Q)\right]. \quad (8)$$

Now we begin to present the main theoretical results, which connect conditional OT with the computation-friendly OT_{mask}. Proofs are provided in appendix.

Theorem 1 (Proxy) *Assume* $\mathrm{supp}(q_Y) \subseteq \mathrm{supp}(p_Y)$*, the following identities hold.*
(a) ***Kantorovich OT****:*

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \mathrm{OT}^{q_Y}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}).$$

(b) ***Sinkhorn OT****:*

$$S^{\lambda}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + \lambda H(Q_Y) = S^{\lambda, q_Y}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}).$$

(c) ***UOT****: there exists non-negative* $\alpha(\cdot)$ *on* $\mathcal{Y}$ *such that* $\mathrm{supp}(\alpha) = \mathrm{supp}(q_Y)$*,* $\sum_{y\in\mathcal{Y}} \alpha(y) = 1$ *and*

$$S^{\lambda,\beta}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + C_0(\alpha, Q_Y) = S^{\lambda,\beta,\alpha}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}),$$

where C_0 *is a constant depending only on* α *and* Q_Y*.*

Thm. 1 shows that MOT serves as a well-defined approximation for conditional OT. Specifically, for Kantorovich OT, MOT $\mathrm{OT}_{\mathrm{mask}}$ exactly equals to conditional OT $\mathrm{OT}^{\alpha}_{\mathrm{cond}}$; for Sinkhorn OT and UOT, MOT is equivalent (up to a constant) to conditional OT $\mathrm{OT}^{\alpha}_{\mathrm{cond}}$, and minimizing the computation-friendly MOT is sufficient to ensure a small $\mathrm{OT}^{\alpha}_{\mathrm{cond}}$. We also present an intuitive illustration for the connection between $\mathrm{OT}^{\alpha}_{\mathrm{cond}}$ and $\mathrm{OT}_{\mathrm{mask}}$ in Fig. 2. In conclusion, the main results imply that MOT can also achieve conditional transport for class-level knowledge transfer, while overcoming the weaknesses of slice-wise computation and hyperparameter α in $\mathrm{OT}^{\alpha}_{\mathrm{cond}}$.

3. Algorithm and Application

Based on the theoretical guarantee of MOT, we now present the numerical algorithm for solving Eq. (8) explicitly. Then a conditional invariant model with risk minimization on the reweighted source domain and transported source domain is proposed for PDA problem.

3.1. Algorithm: A Fixed Point Method

In this section, we focus on the numerical solution for the masked UOT problem in Eq. (8). Note that such a solution is also applicable for the Sinkhorn OT, since $S^{\lambda,\beta}_{\mathrm{mask}}$ will boil down to $S^{\lambda}_{\mathrm{mask}}$ when $\beta = \infty$ (i.e., strict constraints),

As studied by Chizat et al. [6, Thm. 3.11], if $D_\phi = \mathrm{KL}$, the original UOT problem Eq. (5) admits a fixed point solution with explicit computation. Specifically, defining the kernel $\mathcal{K} = \exp(-\mathbf{C}/\lambda) \in \mathbb{R}^{n\times m}$, the fixed point iterations for non-negative scaling vectors is formulated as

$$\mathbf{s}_1^{(i)} = \left(\frac{p_X^w}{\mathcal{K}\mathbf{s}_2^{(i-1)}}\right)^{\frac{\beta}{\beta+\lambda}}, \quad \mathbf{s}_2^{(i)} = \left(\frac{q_X}{\mathcal{K}^\top \mathbf{s}_1^{(i)}}\right)^{\frac{\beta}{\beta+\lambda}}. \tag{9}$$

Then the sequence $\mathrm{diag}(\mathbf{s}_1^{(i)})\mathcal{K}\mathrm{diag}(\mathbf{s}_2^{(i)})$ will converge to the solution of Eq. (5) (i.e., optimal transport plan $\mathbf{\Gamma}^*$).

Similarly, the masked UOT Eq. (8) can also be solved by the fixed point method. The major difference is that the kernel $\tilde{\mathcal{K}}$ induced by label-conditioned masked cost $\tilde{\mathbf{C}}$ is derived as

$$\tilde{\mathcal{K}}_{ij} = \begin{cases} 0, & \text{if } \tilde{\mathbf{C}}_{ij} = \infty, \\ \exp(-\mathbf{C}_{ij}/\lambda), & \text{else.} \end{cases} \tag{10}$$

Then the transport plan for masked UOT can be deduced by replacing the kernel $\mathcal{K}$ in scaling iteration Eq. (9) with the masked kernel $\tilde{\mathcal{K}}$. Note that since $\mathbf{\Gamma}^* \approx \mathrm{diag}(\mathbf{s}_1^{(i)})\tilde{\mathcal{K}}\mathrm{diag}(\mathbf{s}_2^{(i)})$, then the transport mass $\mathbf{\Gamma}^*_{ij}$ will be 0 if $\tilde{\mathcal{K}}_{ij} = 0$. Intuitively, the mask mechanism ensures that the assignments between the inter-class samples will be significantly reduced. We summarize the fixed point method for masked UOT Eq. (8) in Alg. 1, where the masked Sinkhorn OT ($\beta = \infty$) can be solved by setting $\beta/(\beta+\lambda) = 1$ in step 4. Note that since $\tilde{\mathcal{K}}$ is block diagonal (up to permutations on samples), the iterations in step 4 can be effectively implemented by iterating the block matrices parallelly.

Algorithm 1 Fixed Point Method for Eq. (8)

Input: observations $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^n$ of P and $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^m$ of Q, maximum iteration $I_{\max}$, importance weight w, penalty parameter λ, β;
Output: transport plan $\tilde{\mathbf{\Gamma}}$;
1: Initialize $\mathbf{s}_2^{(0)} = [1, 1, \ldots, 1]^\top/m \in \mathbb{R}^m$;
2: Compute p_X^w as Def. 1 and $\tilde{\mathcal{K}}$ as Eq. (10);
3: **for** $i = 1, 2, \ldots, I_{\max}$ **do**
4: $\quad \mathbf{s}_1^{(i)} \leftarrow \left(\frac{p_X^w}{\tilde{\mathcal{K}}\mathbf{s}_2^{(i-1)}}\right)^{\frac{\beta}{\beta+\lambda}}, \quad \mathbf{s}_2^{(i)} \leftarrow \left(\frac{q_X}{\tilde{\mathcal{K}}^\top \mathbf{s}_1^{(i)}}\right)^{\frac{\beta}{\beta+\lambda}}$;
5: **end for**
6: $\tilde{\mathbf{\Gamma}} \leftarrow \mathrm{diag}(\mathbf{s}_1^{(I_{\max})})\tilde{\mathcal{K}}\mathrm{diag}(\mathbf{s}_2^{(I_{\max})})$.

3.2. Application: PDA

In this section, we focus on the OT-based modeling for extreme PDA scenario, and propose MOT-based model to learn the cross-domain invariant knowledge.

Motivated by the discussion of successful DA in Sec. 2.1, we are interested in two problems: 1) learning conditional invariant representation Z via OT; 2) learning transferrable model with unbiased risk estimation. With these goals in mind, we decompose the learning model f as $f_c \circ f_r$, where $f_r : X \mapsto Z$ is representation learner and $f_c : Z \mapsto Y$ is task learner. These compositions are parameterized by neural networks, and the overall framework can be summarized as following two alternative processes (shown as Fig. 2).

Transport Assignment Learning. In this process, we aim to learn an ideal transport plan $\mathbf{\Gamma}$ for characterizing the cross-domain sample correspondence. Firstly, we estimate weight w with BBSE [22] algorithm, whose convergence property is theoretically ensured. Then the label-

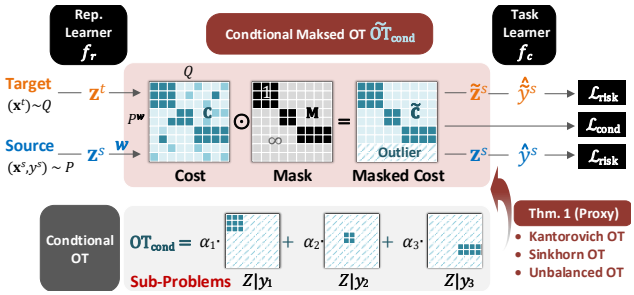


Figure 2. Illustration of the MOT-based model for PDA. **Stage Red**: transport assignment learning for obtaining ideal transport plan $\tilde{\mathbf{\Gamma}}$ with well-defined proxy $\mathrm{OT}_{\mathrm{mask}}$; **Stage Black**: conditional alignment and risk minimization for learning conditional invariant representation Z and transferrable model $f_c \circ f_r$.

conditioned masked UOT Eq. (8) is employed, i.e.,

$$\mathcal{L}_{\mathrm{cond}}(f_r, \mathbf{\Gamma}) = \langle \mathbf{\Gamma}, \tilde{\mathbf{C}} + \lambda \ln \mathbf{\Gamma} \rangle_F + \beta \left[D_\phi(\mathbf{\Gamma}_{P_Z^w} \| P_Z^w) + D_\phi(\mathbf{\Gamma}_{Q_Z} \| Q_Z) \right], \quad (11)$$

where the cost matrix $\mathbf{C}_{ij} = \|\mathbf{z}_i^s - \mathbf{z}_j^t\|_2^2$. Note that this transport assignment learning is implemented in representation space $\mathcal{Z}$ with learner f_r, which ensures the model can dynamically measure and bridge the conditional discrepancy during training procedure. Since there are no labels for target domain, we use the pseudo labels $\hat{y} = f(\mathbf{x})$ for estimating the masked cost $\tilde{\mathbf{C}}$. In this stage, the transport plan will be learned with fixed representation learner f_r, i.e., compute $\tilde{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \mathcal{L}_{\mathrm{cond}}$ according to Alg. 1.

Conditional Alignment and Risk Minimization. In this process, we aim to estimate unbiased risk and cross-domain conditional discrepancy based on the learned ideal assignment $\tilde{\mathbf{\Gamma}}$. Then the model f will be optimized according to the estimated objectives.

For conditional alignment, the masked UOT loss $\mathcal{L}_{\mathrm{cond}}$ Eq. (11) with fixed $\tilde{\mathbf{\Gamma}}$ is sufficient to achieve class-level transport. Theoretically, it can also mitigate the conditional distribution shift as shown in Thm. 1. Besides, since $\mathcal{L}_{\mathrm{cond}}$ is built on reweighted source domain, the risk of negative transfer can be significantly reduced.

For risk minimization, apart from the risk estimation on reweighted source P^w, we propose to learn risk on transported source $\tilde{P}^w$ based on the learned ideal assignment $\tilde{\mathbf{\Gamma}}$. Such transport establishes the explicit connection between empirical risk and target data. Specifically, given $\tilde{\mathbf{\Gamma}}$, the source data can be represented by the target samples via the following *barycenter map problem*:

$$\tilde{\mathbf{z}}_i^s = \arg\min_{\mathbf{z}} \sum_{j=1}^m \tilde{\mathbf{\Gamma}}_{ij} \|\mathbf{z} - \mathbf{z}_j^t\|_2^2. \quad (12)$$

As shown by Courty et al. [8], an analytic solution for Eq. (12), called barycenter mapping ψ, can be written as

$$\tilde{\mathbf{z}}_i^s = \psi_{\tilde{\mathbf{\Gamma}}_{i,:}}(\mathbf{Z}^t) = (< \tilde{\mathbf{\Gamma}}_{i,:}, \mathbf{1}_m >)^{-1} \sum_{j=1}^m \tilde{\mathbf{\Gamma}}_{ij} \mathbf{z}_j^t. \quad (13)$$

Then we denote observations of transported source $\tilde{P}^w$ as $\{(\tilde{\mathbf{z}}_i^s, y_i^s)\}_{i=1}^n$, and propose to learn empirical risk on $\tilde{P}^w$ as

$$\mathbb{E}_{\tilde{P}^w}[\ell(f_c(\tilde{\mathbf{z}}^s), y^s)] = \mathbb{E}_Q[\ell(f_c \circ \psi \circ f_r(\mathbf{x}^t), y^s)]. \quad (14)$$

An intuitive explanation for barycenter mapping ψ is that it considers the minimized cost for reconstructing source sample $\mathbf{z}^s$ in the representation space of target domain, i.e., the range space $\mathrm{Im}(\mathbf{Z}^t)$. Such a reconstruction ensures that the target representation space can be supervisedly optimized via its transport correspondence with source data, i.e., $\psi \circ f_r(\mathbf{x}^t)$ in transported source risk Eq. (14). With P^w and $\tilde{P}^w$, we can formulate the unbiased empirical risk estimations for task learning as

$$\mathcal{L}_{\mathrm{risk}}(f_r, f_c) = \mathbb{E}_{P^w}[\ell(f(\mathbf{x}^s), y^s)] + \mathbb{E}_{\tilde{P}^w}[\ell(f_c(\tilde{\mathbf{z}}^s), y^s)].$$

Finally, the optimization objective in conditional alignment and risk minimization stage can be summarized as

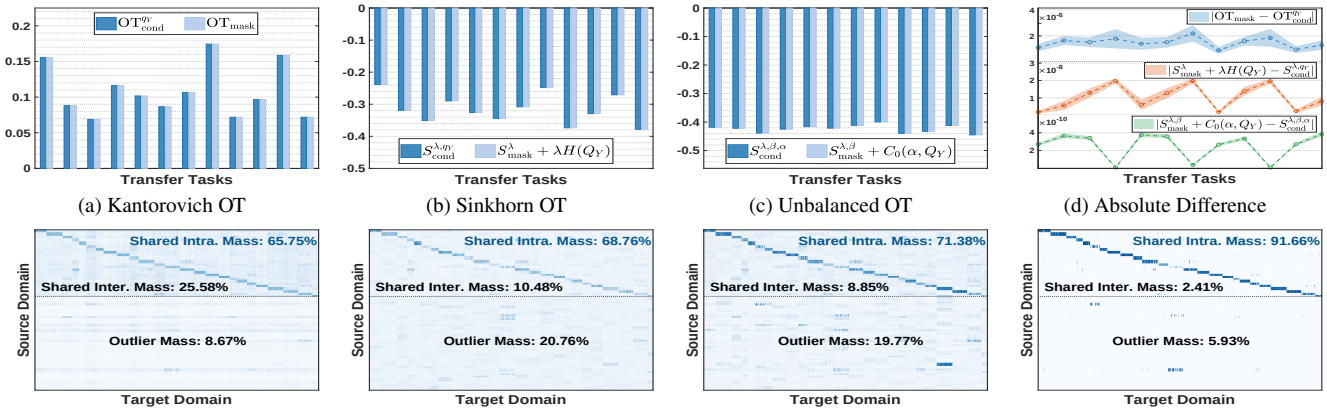
$$\min_{f_r, f_c} \mathcal{L}(f_r, f_c) = \mathcal{L}_{\mathrm{risk}}(f_r, f_c) + \eta \mathcal{L}_{\mathrm{cond}}(f_r, \tilde{\mathbf{\Gamma}}). \quad (15)$$

4. Experiments

In this section, we validate MOT methodology and evaluate the proposed PDA model by conducting experiments on standard PDA datasets, i.e., Office-Home [39], VisDA-2017 [31], Office-31 [36] and ImageCLEF [4]. Details of datasets and implementations are provided in appendix.

Proxy Analysis. To validate the proxy property of masked OT in Thm. 1, we compare values of $\mathrm{OT}_{\mathrm{cond}}^{\alpha}$ and $\mathrm{OT}_{\mathrm{mask}}$ on Office-Home. Results in Fig. 3a-3d imply that the absolute difference between $\mathrm{OT}_{\mathrm{cond}}^{\alpha}$ and $\mathrm{OT}_{\mathrm{mask}}$ (with constant in Thm. 1) is almost zero and is negligible compared to the scales of OT values. These results demonstrate that the theoretical results are valid, and $\mathrm{OT}_{\mathrm{cond}}^{\alpha}$ and proxy $\mathrm{OT}_{\mathrm{mask}}$ are indeed equivalent in application.

OT Formulations. To compare the existing OT formulations with proposed modelings, we visualize the plan values (darker color is higher in value) and report the masses assigned to outlier, (shared) inter-class and intra-class on Office-Home. Results in Fig. 3e-3h demonstrate that reweighted OT actually reduces assignments to outlier classes, but estimated w may be less accurate for shared classes and inter-class mass is large. Though UOT learns better assignments for shared classes, there is large mass for outlier classes since marginal penalty may be unaffordable for extreme scenario. In Fig. 3g, our reweighted and relaxation formulation Eq. (3) is superior and the intra-class mass is larger. Further, MOT ensures a significant block diagonal structure for transport plan in Fig. 3h, where the inter-class mass and outlier mass are significantly reduced by penalty on cost matrix. These results validate that MOT methodology can learn better plan for task knowledge transfer.



(a) Kantorovich OT (b) Sinkhorn OT (c) Unbalanced OT (d) Absolute Difference

(e) Reweighted OT (f) Unbalanced OT (g) Formulation Eq. (3) (h) MOT Eq. (8)

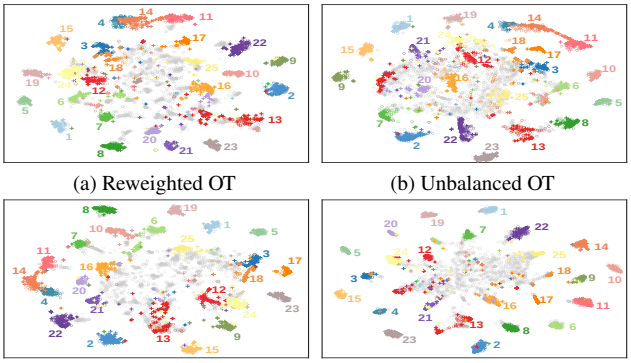
Figure 3. (a)-(d): Numerical validation of the proxy theorem on all transfer tasks of Office-Home, where $\mathrm{OT}^{\alpha}_{\mathrm{cond}}$ and $\mathrm{OT}_{\mathrm{mask}}$ are actually equivalent in empirical scenarios. (e)-(h): Comparison of different OT formulations, where proposed methods Eq. (3) and Eq. (8) are better.

Modules			Office	VisDA	Office	Image	Mean
OT	$\mathbb{E}_{\tilde{P}^w}[\ell]$	$\mathcal{L}_{\mathrm{cond}}$	Home	2017	31	CLEF	
ROT	✓	✓	73.9	73.1	93.7	88.2	82.2
UOT	✓	✓	69.1	67.6	92.5	87.1	79.1
Eq. (3)	✓	✓	74.2	76.8	94.3	89.2	83.6
MOT	✓		76.6	84.9	95.4	90.4	86.8
MOT		✓	74.6	89.3	97.5	91.9	88.3
MOT	✓	✓	**80.6**	**92.4**	**98.4**	**93.6**	**91.3**

Table 1. Ablation study (ResNet-50).

Ablation Study. To evaluate the effectiveness of different modules, we present results of ablation experiments in Tab. 1. We first consider different OT formulations for model proposed in Sec. 3.2. The 1st-3rd rows imply that proposed formulation Eq. (3) achieve higher accuracies than reweighted OT (ROT) and UOT. Further, the 6th row implies mask mechanism can improve accuracies significantly with label-conditioned knowledge transfer. These results validate the superiority of new OT formulations. Besides, we consider the transported source risk $\mathbb{E}_{\tilde{P}^w}[\ell]$ in Eq. (14) and conditional alignment $\mathcal{L}_{\mathrm{cond}}$ in Eq. (15). Comparing the results in 4th-6th, we observe that both $\mathbb{E}_{\tilde{P}^w}[\ell]$ and $\mathcal{L}_{\mathrm{cond}}$ can improve model performance, and the full MOT model is significantly better. These results demonstrate that proposed invariant learning model with transported risk minimization is effective in dealing with PDA problem.

Feature Visualization. To analyze the quality of learned representations, we use t-SNE [26] to visualize the 2-D features of different OT formulations on Office-Home. For reweighted OT and UOT, though the negative impacts of outlier classes are reduced as Fig. 4a-4b, the discriminability for shared classes are not sufficiently learned, and some samples are transferred to incorrect classes with marginal adaptation. For formulation Eq. (3), the representations are more compact, and fewer samples are overlapping with outlier classes. Further, the conditional alignment in MOT ensures the class-wise alignment for cross domain samples,



(a) Reweighted OT (b) Unbalanced OT

(c) Formulation Eq. (3) (d) MOT Eq. (8)

Figure 4. t-SNE [26] visualization of learned representations, where shared classes have different color and outlier classes are gray. 'o': source samples; '+': target samples; 'number': labels.

then the intra-class compactness and inter-class separability are significantly improved as Fig. 4d. These results demonstrate that the proposed OT formulations and invariant risk learning model indeed learn correct transport relations and discriminative representations.

Different PDA Scenarios. We evaluate OT-based models under different PDA scenarios by varying the number of shared classes on Office-Home. The results in Fig 5a-5b imply that all models achieve high accuracy when there are less shared classes, since there are relatively more samples/knowledge on the source domain. With the increase of shared classes, the label discrepancy is smaller, but the target classification task is more challenging since there are more target samples. Thus, we can observe that the accuracies increase at interval $[20, 30]$ and then decrease with the increasing target samples. Generally, MOT is superior to other OT modelings, which demonstrate the effectiveness of proposed methodology in different learning scenarios.

Comparison with SOTA. To evaluate the model performance, we compare MOT with several SOTA PDA methods, and present results in Tab. 2-3. 1) For adversarial adapta-

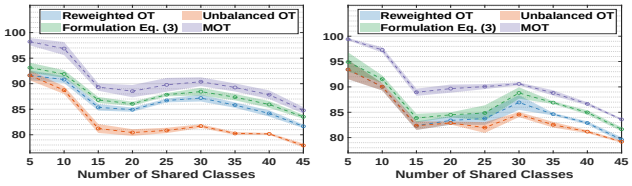
Methods	Office-Home													VisDA-2017
	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Mean	S→R
Source [17]	46.3	67.5	75.9	59.1	59.9	62.7	58.2	41.8	74.9	67.4	48.2	74.2	61.4	45.3
DANN [13]	43.8	67.9	77.5	63.7	59.0	67.6	56.8	37.1	76.4	69.2	44.3	77.5	61.7	51.0
PADA [2]	52.0	67.0	78.7	52.2	53.8	59.0	52.6	43.2	78.8	73.7	56.6	77.1	62.1	53.5
ETN [3]	59.2	77.0	79.5	62.9	65.7	75.0	68.3	55.4	84.4	75.7	57.7	84.5	70.5	-
SAFN [43]	58.9	76.3	81.4	70.4	73.0	77.8	72.4	55.3	80.4	75.8	60.4	79.9	71.8	67.7
DRCN [21]	51.6	75.8	82.0	62.9	65.1	72.9	67.4	50.0	81.0	76.4	57.7	79.3	68.5	58.2
DMP [25]	59.0	81.2	86.3	68.1	72.8	78.8	71.2	57.6	84.9	77.3	61.5	82.9	73.5	72.7
JUMBOT [11]	62.7	77.5	84.4	76.0	73.3	80.5	74.7	60.8	85.1	80.2	66.5	83.9	75.5	84.0†
AR [16]	**67.4**	85.3	90.0	77.3	70.6	85.2	79.0	**64.8**	89.5	80.4	66.2	86.4	78.3	88.7
m-POT [28]	64.6	80.6	87.2	76.4	77.6	83.6	77.1	63.7	87.6	81.4	**68.5**	87.4	78.0	87.0†
MOT	63.1	**86.1**	**92.3**	**78.7**	**85.4**	**89.6**	**79.8**	62.3	**89.7**	**83.8**	67.0	**89.6**	**80.6**	**92.4**

Table 2. Classification accuracies (%) on Office-Home and VisDA-2017 datasets (ResNet-50). † Results reported by Salvador et al. [37].

Office-31	A→W	D→W	W→D	A→D	D→A	W→A	Mean
Source [17]	75.6	96.3	98.1	83.4	83.9	85.0	87.1
DANN [13]	73.6	96.3	98.7	81.5	82.8	86.1	86.5
PADA [2]	86.5	99.3	**100**	82.2	92.7	95.4	92.7
SAFN [43]	87.5	96.6	99.4	89.8	92.6	92.7	93.1
DRCN [21]	90.8	**100**	**100**	94.3	95.2	94.8	95.9
DMP [25]	96.6	**100**	**100**	96.4	95.1	95.4	97.2
AR [16]	93.5	**100**	99.7	96.8	95.5	96.0	96.9
MOT	**99.3**	**100**	**100**	**98.7**	**96.1**	**96.4**	**98.4**

ImageCLEF	I→P	P→I	I→C	C→I	C→P	P→C	Mean
Source [17]	78.3	86.9	91.0	84.3	72.5	91.5	84.1
DANN [13]	78.1	86.3	91.3	84.0	72.1	90.3	83.7
PADA [2]	81.7	92.1	94.6	89.8	77.7	94.1	88.3
SAFN [43]	79.5	90.7	93.0	90.3	77.8	94.0	87.5
DMP [25]	82.4	94.5	96.7	94.3	78.7	96.4	90.5
MOT	**87.7**	**95.0**	**98.0**	**95.0**	**87.0**	**98.7**	**93.6**

Table 3. Classification accuracies (%) on Office-31 and ImageCLEF datasets (ResNet-50).



(a) Office-Home P→R (b) Office-Home R→P

Figure 5. Accuracy curves and 95% confidence intervals of OT-based models by varying the number of shared classes.

tion, PADA improves DANN by introducing the reweighted source P^w to adversarial training. The better performance validates the effectiveness of label distribution correction for PDA. Compared with these adversarial models with marginal alignment, MOT learns conditional invariant representations with masked mechanism. 2) For metric-based adaptation models SAFN and DMP, though label-conditioned information is incorporated, the explicit conditional distribution alignment is not ensured. 3) Compared with other OT models (JUMBOT, AR and m-POT), MOT overcomes the limitations of existing OT modelings as discussed before. Specifically, AR is reweighted-based model, which applies P^w to the dual OT problem, but the performance may be sensitive to the precision of w. JUMBOT and m-POT are respectively unbalanced transport model and partial transport model, which can be both regarded as relaxation-based OT without strict marginal constraints, but the relaxation may be less effective when label shift is severe, e.g., PDA. Fortunately, our formulation Eq. (3) integrating the advantages of relaxation and reweighted, and mask mechanism further ensures the sufficiency of conditional shift correction. Therefore, MOT achieve significant improvements (about 1%∼4%) on four PDA datasets, which validates the effectiveness of proposed methodology.

5. Conclusion

In this paper, we systematically study the limitations in current OT modelings for real-world learning scenarios, and then develop novel OT methodology with theoretical guarantees. For theoretical aspect, a novel formulation is proposed with reweighted and relaxation operations, and a new OT variant called MOT is proposed by exploring the mask mechanism; the equivalent relation between MOT and conditional OT is proved, which implies the computation-friendly MOT can also characterize the conditional information in transportation. For methodology, we propose an invariant learning model with MOT, whose effectiveness is validated by extensively numerical experiments and analysis. An interesting future direction is studying theory and algorithm for masked partial OT and MOT with soft labels.

Acknowledgement

This work is supported in part by the National Key R&D Program of China under Grant 2021YFA1003001; in part by the National Natural Science Foundation of China under Grant 61976229; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023B1515020004; and in part by the Open Research Projects of Zhejiang Lab under Grant 2021KH0AB08.

References

[1] 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(1-2):151–175, 2010. 1

[2] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *ECCV*, pages 135–150, 2018. 1, 2, 8

[3] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *CVPR*, pages 2985–2994, 2019. 1, 2, 8

[4] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211, 2014. 6

[5] Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. In *NeurIPS*, volume 34, pages 23270–23282, 2021. 1

[6] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. 1, 3, 5

[7] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018. 1, 3

[8] Nicolas Courty, Rmi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE TPAMI*, 39(9):1853–1865, 2017. 1, 2, 3, 6

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, volume 26, 2013. 1, 3

[10] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, pages 447–463, 2018. 1, 2, 3

[11] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, pages 3186–3197. PMLR, 2021. 1, 2, 3, 8

[12] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTAT*, pages 2131–2141, 2020. 1

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 8

[14] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *ICML*, pages 2839–2848, 2016. 2

[15] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. In *NeurIPS*, 2022. 2

[16] Xiang Gu, Xi Yu, Jian Sun, Zongben Xu, et al. Adversarial reweighting for partial domain adaptation. In *NeurIPS*, volume 34, pages 14860–14872, 2021. 1, 2, 3, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 8

[18] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *ICML*, pages 4816–4827, 2020. 2

[19] Matthieu Kirchmeyer, Alain Rakotomamonjy, Emmanuel de Bezenac, and Patrick Gallinari. Mapping conditional distributions for domain adaptation under generalized target shift. In *ICLR*, 2022. 1, 2, 3

[20] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *CVPR*, pages 13936–13944, 2020. 1, 2, 3

[21] Shuang Li, Chi Harold Liu, Qiuxia Lin, Qi Wen, Limin Su, Gao Huang, and Zhengming Ding. Deep residual correction network for partial domain adaptation. *IEEE TPAMI*, 43(7):2329–2344, 2021. 1, 8

[22] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, pages 3122–3130. PMLR, 2018. 1, 2, 5

[23] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, volume 31, 2018. 2

[24] You-Wei Luo and Chuan-Xian Ren. Conditional Bures metric for domain adaptation. In *CVPR*, pages 13989–13998, 2021. 1, 2

[25] You-Wei Luo, Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. Unsupervised domain adaptation via discriminative manifold propagation. *IEEE TPAMI*, 44(3):1653–1669, 2022. 1, 2, 8

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11):2579–2605, 2008. 7

[27] Khai Nguyen, Dang Nguyen, Quoc Nguyen, Tung Pham, Hung Bui, Dinh Phung, Trung Le, and Nhat Ho. On transportation of mini-batches: A hierarchical approach. In *ICML*, page 34, 2022. 1

[28] Khai Nguyen, Dang Nguyen, Tung Pham, Nhat Ho, et al. Improving mini-batch optimal transport via partial transportation. In *ICML*, pages 16656–16690, 2022. 1, 2, 8

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2009. 1

[30] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 1

[31] 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. 1, 6

[32] Alain Rakotomamonjy, Rémi Flamary, Gilles Gasso, M El Alaya, Maxime Berar, and Nicolas Courty. Optimal transport for conditional domain matching and label shift. *Machine Learning*, 111(5):1651–1670, 2022. 1, 2, 3

[33] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *AISTAT*, pages 849–858. PMLR, 2019. 1, 2, 3

[34] Chuan-Xian Ren, Pengfei Ge, Peiyi Yang, and Shuicheng Yan. Learning target-domain-specific classifier for partial domain adaptation. *IEEE TNNLS*, 32(5):1989–2001, 2021. 2

[35] Chuan-Xian Ren, Xiao-Lin Xu, and Hong Yan. Generalized conditional domain adaptation: A causal perspective with low-rank translators. *IEEE TCYB*, 50(2):821–834, 2018. 2, 3

[36] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 6

[37] Tiago Salvador, Kilian Fatras, Ioannis Mitliagkas, and Adam Oberman. A reproducible and realistic evaluation of partial domain adaptation methods. *arXiv preprint arXiv:2210.01210*, 2022. 8

[38] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *NeurIPS*, volume 33, pages 19276–19289, 2020. 1, 2, 3

[39] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 1, 6

[40] Pengfei Wei, Yiping Ke, Xinghua Qu, and Tze-Yun Leong. Subdomain adaptation with manifolds discrepancy alignment. *IEEE TCYB*, 52(11):11698–11708, 2022. 2

[41] Haifeng Xia, Taotao Jing, and Zhengming Ding. Maximum structural generation discrepancy for unsupervised domain adaptation. *IEEE TPAMI*, 2022. 2

[42] Geng-Xin Xu, Chen Liu, Jun Liu, Zhongxiang Ding, Feng Shi, Man Guo, Wei Zhao, Xiaoming Li, Ying Wei, Yaozong Gao, Chuan Xian Ren, and Dinggang Shen. Cross-site severity assessment of covid-19 from ct images via domain adaptation. *IEEE Transactions on Medical Imaging*, 41(1):88–102, 2021. 1

[43] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, pages 1426–1435, 2019. 8

[44] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *CVPR*, pages 4394–4403, 2020. 3

[45] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, pages 2272–2281, 2017. 2

[46] Jiying Zhang, Xi Xiao, Long-Kai Huang, Yu Rong, and Yatao Bian. Fine-tuning graph neural networks via graph topology induced optimal transport. In *IJCAI*, 2022. 2

[47] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *ICML*, pages 819–827. PMLR, 2013. 1, 2, 3

[48] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, pages 7404–7413. PMLR, 2019. 1

[49] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *ICML*, pages 7523–7532, 2019. 1

[50] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE TNNLS*, 32(4):1713–1722, 2021. 2

[51] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 1

[52] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018. 1

MOT: Masked Optimal Transport for Partial Domain Adaptation (Supplementary Material)

You-Wei Luo Chuan-Xian Ren*
School of Mathematics, Sun Yat-Sen University, China
luoyw28@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

Abstract

This supplementary material contains the proofs of theoretical results, implementation details for numerical experiments and illustrations of experiment datasets.

S.1. Proof of Theorem 1

Theorem 1 (Proxy) *Assume* $\mathrm{supp}(q_Y) \subseteq \mathrm{supp}(p_Y)$*, the following identities hold.*
(a) ***Kantorovich OT****:*

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) \;=\; \mathrm{OT}^{q_Y}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}).$$

(b) ***Sinkhorn OT****:*

$$S^{\lambda}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + \lambda H(Q_Y) \;=\; S^{\lambda, q_Y}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}).$$

(c) ***UOT****: there exists non-negative* $\alpha(\cdot)$ *on* $\mathcal{Y}$ *such that* $\mathrm{supp}(\alpha) = \mathrm{supp}(q_Y)$*,* $\sum_{y\in\mathcal{Y}} \alpha(y) = 1$ *and*

$$S^{\lambda,\beta}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + C_0(\alpha, Q_Y) \;=\; S^{\lambda,\beta,\alpha}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}),$$

where C_0 *is a constant depending only on* α *and* Q_Y*.*

Proof *For convenience, we first introduce some notations for proof. For finite sample setting, denote* $|\mathcal{Y}| = k$ *as class number,* n_l/m_l *as the sample size of* l*-th source/target class. Since* $\mathrm{supp}(q_Y) \subseteq \mathrm{supp}(p_Y)$*, we denote* $\mathrm{supp}(p_Y) = \mathcal{Y} = \{1, 2, \ldots, k\}$*,* $\mathrm{supp}(p_Y) = \{1, 2, \ldots, k_0\}$ *and* $n_0 = \sum_{l=1}^{k_0} n_l$*, where* $k_0 \leq k$ *is the number of shared classes and* n_0 *the sample size of shared classes on source domain. Without loss of generality, we denote the data matrix with cluster data as* $\mathbf{X}^s = [\mathbf{X}^s_1, \mathbf{X}^s_2, \ldots, \mathbf{X}^s_k] \in \mathbb{R}^{d\times n}$ *and* $\mathbf{X}^t = [\mathbf{X}^t_1, \mathbf{X}^t_2, \ldots, \mathbf{X}^t_{k_0}] \in \mathbb{R}^{d\times m}$*, where* d *is data dimension,* $\mathbf{X}^s_l \in \mathbb{R}^{d\times n_l}$ *and* $\mathbf{X}^t_l \in \mathbb{R}^{d\times m_l}$ *are the data matrix of* l*-th source class and* l*-th target class, respectively. Generally, for a matrix* $\mathbf{A}$*, let the uppercase letters* $\mathbf{A}_{ij}$ *denote the blocks of* $\mathbf{A}$ *and lowercase letters* a_{ij} *the entries of* $\mathbf{A}$*. Note that for the reweighted source we have*

$$p^{w^*}_X = \sum_{y\in\mathcal{Y}} p^{w^*}_y p_{X|y} = \sum_{y\in\mathcal{Y}} q_y p_{X|y} = \sum_{l=1}^{k_0} q_{Y=l} p_{X|l}, \tag{S.1}$$

which implies the proportions of outlier classes are 0 in reweighted distribution. Then a submatrix $\tilde{\mathbf{C}}^{\mathrm{sub}} \in \mathbb{R}^{n_0\times m}$ *of* $\tilde{\mathbf{C}}$*, which considers the cost between samples of shared classes, is defined as the first* n_0 *rows of* $\tilde{\mathbf{C}}$*. Now we begin to prove the main results.*

(1) Kantorvich OT.

*Corresponding Author.

Recall the masked Kantorvich OT is formulated as

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma} \in \Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F .$$

Let the source distribution of shared classes be $r_X^{w^*} \in \mathbb{R}^{n_0}$, which consists of the first n_0 elements of p^{w^*}. Since the values of outlier classes' samples are 0 in $p_X^{w^*} \in \mathbb{R}^n$ as Eq. (S.1), there transport plan for outlier classes will be 0, i.e., $\gamma_{ij} = 0$ if $i > n_0$. Then the original problem boils down to the transportation between shared classes:

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma} \in \Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F = \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}^{\mathrm{sub}}, \tilde{\mathbf{C}}^{\mathrm{sub}} \right\rangle_F .$$

On the other hands, note that the transport plan between inter-class sample pair will be 0 since, i.e., $\gamma_{ij}^{\mathrm{sub}} = 0$ if $y_i^s \neq y_j^t$, since otherwise the overall transport cost will be infinity and the problem will not be well-defined. It implies the plan $\mathbf{\Gamma}^{\mathrm{sub}}$ under masked cost admits a block diagonal structure, then we have

$$\begin{aligned}
\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) &= \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}^{\mathrm{sub}}, \tilde{\mathbf{C}}^{\mathrm{sub}} \right\rangle_F \\
&= \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \left\langle \begin{bmatrix} \mathbf{\Gamma}_{11}^{\mathrm{sub}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Gamma}_{k_0 k_0}^{\mathrm{sub}} \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{C}}_{11}^{\mathrm{sub}} & \cdots & \infty \\ \vdots & \ddots & \vdots \\ \infty & \cdots & \tilde{\mathbf{C}}_{k_0 k_0}^{\mathrm{sub}} \end{bmatrix} \right\rangle_F \\
&= \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \sum_{l=1}^{k_0} \left\langle \mathbf{\Gamma}_{ll}^{\mathrm{sub}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
&= \sum_{l=1}^{k_0} \min_{\mathbf{\Gamma}_{ll}^{\mathrm{sub}} \in \Pi(q_{Y=l} R_{X|l}^{w^*}, q_{Y=l} Q_{X|l})} \left\langle \mathbf{\Gamma}_{ll}^{\mathrm{sub}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
&= \sum_{l=1}^{k_0} \min_{\frac{\mathbf{\Gamma}_{ll}^{\mathrm{sub}}}{q_{Y=l}} \in \Pi(R_{X|l}^{w^*}, Q_{X|l})} q_{Y=l} \left\langle \frac{\mathbf{\Gamma}_{ll}^{\mathrm{sub}}}{q_{Y=l}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
\left(\bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}} \triangleq \frac{\mathbf{\Gamma}_{ll}^{\mathrm{sub}}}{q_{Y=l}} \right) \quad &= \sum_{l=1}^{k_0} q_{Y=l} \min_{\bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}} \in \Pi(R_{X|l}^{w^*}, Q_{X|l})} \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
&= \sum_{l=1}^{k_0} q_{Y=l} \mathrm{OT}(R_{X|l}^{w^*}, Q_{X|l}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}}) \\
&= \sum_{l=1}^{k_0} q_{Y=l} \mathrm{OT}(P_{X|l}^{w^*}, Q_{X|l}, \mathbf{C}_{ll}^{\mathrm{sub}}) \qquad \text{(S.2)} \\
&= \mathrm{OT}_{\mathrm{cond}}^{q_Y}(P^{w^*}, Q, \mathbf{C}),
\end{aligned}$$

where $\bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}}$ and $\tilde{\mathbf{C}}_{ll}^{\mathrm{sub}}$ are $n_l \times m_l$ blocks of l-th class, Eq. (S.2) holds since $R_{X|l}^{w^*} = P_{X|l}^{w^*}$ for shared classes and $\tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} = \mathbf{C}_{ll}^{\mathrm{sub}}$ for intra-class sample pairs.

(2) Sinkhorn OT.

Recall the masked Sinkhorn OT is formulated as

$$S_{\mathrm{mask}}^{\lambda}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma} \in \Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}, \ln \mathbf{\Gamma} \right\rangle_F .$$

Similarly, we have

$$
\begin{aligned}
& S^{\lambda}_{\text{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) \\
=& \min_{\mathbf{\Gamma}\in\Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}, \ln \mathbf{\Gamma} \right\rangle_F \\
=& \min_{\mathbf{\Gamma}^{\text{sub}}\in\Pi(R_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}} \right\rangle_F \\
=& \min_{\mathbf{\Gamma}^{\text{sub}}\in\Pi(R_X^{w^*}, Q_X)} \left\langle \begin{bmatrix} \mathbf{\Gamma}^{\text{sub}}_{11} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Gamma}^{\text{sub}}_{k_0 k_0} \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{C}}^{\text{sub}}_{11} & \cdots & \infty \\ \vdots & \ddots & \vdots \\ \infty & \cdots & \tilde{\mathbf{C}}^{\text{sub}}_{k_0 k_0} \end{bmatrix} + \lambda \ln \begin{bmatrix} \mathbf{\Gamma}^{\text{sub}}_{11} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Gamma}^{\text{sub}}_{k_0 k_0} \end{bmatrix} \right\rangle_F \\
=& \min_{\mathbf{\Gamma}^{\text{sub}}\in\Pi(R_X^{w^*}, Q_X)} \sum_{l=1}^{k_0} \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll} \right\rangle_F \\
=& \sum_{l=1}^{k_0} \min_{\mathbf{\Gamma}^{\text{sub}}_{ll}\in\Pi(q_{Y=l} R^{w^*}_{X|l}, q_{Y=l} Q_{X|l})} \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll} \right\rangle_F \\
=& \sum_{l=1}^{k_0} \min_{\frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} q_{Y=l} \left[\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll} \right\rangle_F \right] \\
=& \sum_{l=1}^{k_0} \min_{\frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} q_{Y=l} \left[\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \ln \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}} \right\rangle_F + \lambda \left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, (\ln q_{Y=l}) \mathbf{1}_{n_l \times m_l} \right\rangle_F \right] \\
=& \sum_{l=1}^{k_0} q_{Y=l} \left[\min_{\bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \ln \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \ln q_{Y=l} \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \mathbf{1}_{n_l \times m_l} \right\rangle_F \right] \\
=& \sum_{l=1}^{k_0} q_{Y=l} \left[\min_{\bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \ln \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \ln q_{Y=l} \right] \\
=& \sum_{l=1}^{k_0} q_{Y=l} S^{\lambda}(R^{w^*}_{X|l}, Q_{X|l}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}) + \lambda \sum_{l=1}^{k_0} q_{Y=l} \ln q_{Y=l} \\
=& \sum_{l=1}^{k_0} q_{Y=l} S^{\lambda}(P^{w^*}_{X|l}, Q_{X|l}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}) - \lambda H(Q_Y) \\
=& \, \mathrm{S}^{\lambda, q_Y}_{\text{cond}}(P^{w^*}, Q, \mathbf{C}) - \lambda H(Q_Y).
\end{aligned}
$$

Therefore, we have $\mathrm{S}^{\lambda, q_Y}_{\text{cond}}(P^{w^*}, Q, \mathbf{C}) = S^{\lambda}_{\text{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + \lambda H(Q_Y)$

(3) Unbalanced OT.

Recall the masked unbalanced OT is formulated as

$$
S^{\lambda,\beta}_{\text{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma}\in\mathcal{M}_+(\mathbb{R}^{n\times m})} \langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \rangle_F + \lambda \left\langle \mathbf{\Gamma}, \ln \mathbf{\Gamma} \right\rangle_F + \beta \left[D_\phi(\mathbf{\Gamma}_{P_X^{w^*}} \| P_X^{w^*}) + D_\phi(\mathbf{\Gamma}_{Q_X} \| Q_X) \right],
$$

where D_ϕ is KL divergence. The major difference between unbalanced OT and other OTs with marginal constraints is that the $\mathbf{\Gamma}$ is only required to be a distribution over $\mathbb{R}^{n\times m}$, i.e., $\mathbf{\Gamma} \in \mathcal{M}_+(\mathbb{R}^{n\times m})$ will satisfy that $\gamma_{ij} >= 0$ and $\sum_{ij} \gamma_{ij} = 1$. Since $\mathbf{\Gamma}$ is no longer a coupling of $(P^{w^*}, Q, \tilde{\mathbf{C}})$, it is necessary to consider whether the 0 transport plans for outlier classes still hold.

Note the KL penalty $D_\phi(\mathbf{\Gamma}_{P_X^{w^*}} \| P_X^{w^*})$ implies that $\mathbf{\Gamma}_{P_X^{w^*}}$ should be absolutely continuous with respect to $P_X^{w^*}$, since otherwise the penalty value will be infinity and the problem is not well-defined. Therefore, for the i-th source sample, if it belongs to outlier classes, the corresponding values in $P_X^{w^*}$ are 0 (i.e., $[P_X^{w^*}]_i = 0$), and the transport plan will also be 0 (i.e., $[\mathbf{\Gamma}_{P_X^{w^*}}]_i = \sum_j \gamma_{ij} = 0 \implies \gamma_{ij} = 0$). Therefore, the original problem can also be written as the transportation between shared classes, i.e.,

$$S^{\lambda,\beta}_{\text{mask}}(P^{w^*},Q,\tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma}^{\text{sub}}\in\mathcal{M}_+(\mathbb{R}^{n_0\times m})} \langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}}\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}}\right\rangle_F + \beta\left[D_\phi(\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}\|R^{w^*}_X) + D_\phi(\mathbf{\Gamma}^{\text{sub}}_{Q_X}\|Q_X)\right].$$

Let $\mathbf{\Gamma}^{\text{sub}^*}$ be the optimal solution for the objective above. Similarly, $\mathbf{\Gamma}^{\text{sub}^*}$ is also block-diagonal since the non-zero plan values for inter-class sample pairs $(\mathbf{x}^s_i,\mathbf{x}^t_j)$ will induce infinite transport cost with $\tilde{c}^{\text{sub}}_{ij}$. Then we consider the following coefficient

$$\alpha(l) = \sum_{ij}[\mathbf{\Gamma}^{\text{sub}^*}_{ll}]_{ij},$$

which represents the values assigned to the transportation between l-th source class and l-th target class. It is clear that $\alpha(\cdot)$ is non-negative on $\mathcal{Y}$ and satisfies that $\text{supp}(\alpha) = \text{supp}(q_Y)$. For simplicity, we denote the blocks of $\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}$ and $\mathbf{\Gamma}^{\text{sub}}_{Q_X}$ as

$$\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X} = \begin{bmatrix}\mathbf{o}^s_1\\ \vdots \\ \mathbf{o}^s_{k_0}\end{bmatrix}\in\mathbb{R}^{n_0}, \qquad \mathbf{\Gamma}^{\text{sub}}_{Q_X} = \begin{bmatrix}\mathbf{o}^t_1\\ \vdots \\ \mathbf{o}^t_{k_0}\end{bmatrix}\in\mathbb{R}^{m}$$

where $\mathbf{o}^s_l\in\mathbb{R}^{n_l}$ and $\mathbf{o}^t_l\in\mathbb{R}^{m_l}$. Then we have

$$\begin{aligned}
&S^{\lambda,\beta}_{\text{mask}}(P^{w^*},Q,\tilde{\mathbf{C}})\\
&= \min_{\mathbf{\Gamma}^{\text{sub}}\in\mathcal{M}_+(\mathbb{R}^{n_0\times m})} \left\langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}}\right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}}\right\rangle_F + \beta\left[D_\phi(\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}\|R^{w^*}_X) + D_\phi(\mathbf{\Gamma}^{\text{sub}}_{Q_X}\|Q_X)\right]\\
&= \min_{\mathbf{\Gamma}^{\text{sub}}\in\mathcal{M}_+(\mathbb{R}^{n_0\times m})} \left\langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}}\right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}}\right\rangle_F + \beta\left[\left\langle \mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}, \ln\frac{\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}}{R^{w^*}_X}\right\rangle_F + \left\langle \mathbf{\Gamma}^{\text{sub}}_{Q_X}, \ln\frac{\mathbf{\Gamma}^{\text{sub}}_{Q_X}}{Q_X}\right\rangle_F\right]\\
&= \sum_{l=1}^{k_0}\min_{\mathbf{\Gamma}^{\text{sub}}_{ll}\in\alpha(l)\mathcal{M}_+(\mathbb{R}^{n_l\times m_l})} \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}\right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll}\right\rangle_F + \beta\left[\left\langle \mathbf{o}^s_l, \ln\frac{\mathbf{o}^s_l}{r^{w^*}_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \left\langle \mathbf{o}^t_l, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\right]
\end{aligned} \tag{S.3}$$

$$\begin{aligned}
&= \sum_{l=1}^{k_0}\min_{\frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{\alpha(l)}\in\mathcal{M}_+(\mathbb{R}^{n_l\times m_l})} \alpha(l)\Bigg[\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{\alpha(l)}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}\right\rangle_F + \lambda\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{\alpha(l)}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll}\right\rangle_F\\
&\qquad\qquad + \beta\left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{q_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \beta\left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\Bigg]\\
&= \sum_{l=1}^{k_0}\min_{\bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\in\mathcal{M}_+(\mathbb{R}^{n_l\times m_l})} \alpha(l)\Bigg[\left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}\right\rangle_F + \lambda\left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \ln \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\right\rangle_F + \lambda\ln\alpha(l)\\
&\qquad\qquad + \beta\left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{q_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \beta\left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\Bigg],
\end{aligned} \tag{S.4}$$

where Eq. (S.3) holds since $\mathbf{\Gamma}^{\text{sub}^*}$ is block diagonal, which implies the minimization problem can be divided into k_0 sub-problems and the mass assigned to l-th class is $\alpha(l)$. Note that for the KL terms, we have

$$\begin{aligned}
&\left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{q_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\\
&= \left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{\alpha(l)R^{w^*}_{X|l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\alpha(l)\mathbf{1}_{n_l}}{q_{Y=l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{\alpha(l)Q_{X|l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\alpha(l)\mathbf{1}_{m_l}}{q_{Y=l}}\right\rangle_F.
\end{aligned} \tag{S.5}$$

Denote $\bar{\mathbf{o}}^s_l = \frac{\mathbf{o}^s_l}{\alpha(l)}\in$ and $\bar{\mathbf{o}}^t_l = \frac{\mathbf{o}^t_l}{\alpha(l)}$, then $\sum_i[\bar{\mathbf{o}}^s_l]_i = \sum_i[\bar{\mathbf{o}}^t_l]_i = 1$ since the mass assigned to l-th class is $\alpha(l)$. Then Eq. (S.5)

can be further written as

$$
\begin{aligned}
&\left\langle \frac{\mathbf{o}_l^s}{\alpha(l)}, \ln \frac{\mathbf{o}_l^s}{q_{Y=l} R_{X|l}^{w^*}} \right\rangle_F + \left\langle \frac{\mathbf{o}_l^t}{\alpha(l)}, \ln \frac{\mathbf{o}_l^t}{q_{Y=l} Q_{X|l}} \right\rangle_F \\
&= \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}} \left\langle \bar{\mathbf{o}}_l^s, \mathbf{1}_{n_l} \right\rangle_F + \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}} \left\langle \bar{\mathbf{o}}_l^t, \mathbf{1}_{m_l} \right\rangle_F \\
&= \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}} + \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}}
\end{aligned} \tag{S.6}
$$

Finally, by substituting KL terms in Eq. (S.6) into main proof Eq. (S.4), we have

$$
\begin{aligned}
&S_{\text{mask}}^{\lambda,\beta}(P^{w^*}, Q, \tilde{\mathbf{C}}) \\
&= \sum_{l=1}^{k_0} \min_{\bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \in \mathcal{M}_+(\mathbb{R}^{n_l \times m_l})} \alpha(l) \Bigg[\left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \tilde{\mathbf{C}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \ln \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \ln \alpha(l) \\
&\qquad\qquad + \beta \left\langle \frac{\mathbf{o}_l^s}{\alpha(l)}, \ln \frac{\mathbf{o}_l^s}{q_{Y=l} R_{X|l}^{w^*}} \right\rangle_F + \beta \left\langle \frac{\mathbf{o}_l^t}{\alpha(l)}, \ln \frac{\mathbf{o}_l^t}{q_{Y=l} Q_{X|l}} \right\rangle_F \Bigg] \\
&= \sum_{l=1}^{k_0} \min_{\bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \in \mathcal{M}_+(\mathbb{R}^{n_l \times m_l})} \alpha(l) \Bigg[\left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \tilde{\mathbf{C}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \ln \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \ln \alpha(l) \\
&\qquad\qquad + \beta \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \beta \ln \frac{\alpha(l)}{q_{Y=l}} + \beta \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F + \beta \ln \frac{\alpha(l)}{q_{Y=l}} \Bigg] \\
&= \sum_{l=1}^{k_0} \alpha(l) \Bigg[\min_{\bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \in \mathcal{M}_+(\mathbb{R}^{n_l \times m_l})} \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \tilde{\mathbf{C}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \ln \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \right\rangle_F + \beta \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \beta \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F \Bigg] \\
&\qquad\qquad + \sum_{l=1}^{k_0} \left[\lambda \alpha(l) \ln \alpha(l) + 2\beta \ln \alpha(l) \frac{\alpha(l)}{q_{Y=l}} \right] \\
&= \sum_{l=1}^{k_0} \alpha(l) S^{\lambda,\beta}(P_{X|l}^{w^*}, Q_{X|l}, \tilde{\mathbf{C}}_{ll}^{\text{sub}}) - \lambda H(Q_Y) + 2\beta D_\phi(\alpha \| Q_Y) \\
&= S_{\text{cond}}^{\lambda,\beta,\alpha}(P^{w^*}, Q, \mathbf{C}) - \lambda H(Q_Y) + 2\beta D_\phi(\alpha \| Q_Y).
\end{aligned}
$$

Therefore, the non-negative $\alpha(\cdot)$ such that $S_{\text{cond}}^{\lambda,\beta,\alpha}(P^{w^*}, Q, \mathbf{C}) = S_{\text{mask}}^{\lambda,\beta}(P^{w^*}, Q, \tilde{\mathbf{C}}) + C_0(\alpha, Q_Y)$, where $C_0(\alpha, Q_Y) = \lambda H(Q_Y) - 2\beta D_\phi(\alpha \| Q_Y)$

S.2. Experiment Details and Additional Discussions

S.2.1. Implementation Details

The network-based model is implemented in PyTorch [11] platform. For network architectures, f_r consists of ResNet-50 [6] and two Fully-Connected (FC) layers ($\mathbb{R}^{2048} \to \mathbb{R}^{1024} \to \mathbb{R}^{512}$) with batch normalization, where the FC layers are activated by Leaky ReLU ($\alpha = 0.2$) and Tanh, respectively; f_c is a single FC layer ($\mathbb{R}^{512} \to \mathbb{R}^{|\mathcal{Y}|}$) with SoftMax activation. For optimization, we use batch gradient descent with Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), where the learning rate is set as $1e$-3. Entropic parameter λ is empirically set as $1e$-2 in numerical experiments. To ensure the more accurate OT estimation with larger batch-size, we load the pretrained parameter on ImageNet for the ResNet-50 in representation learner f_r, and then froze them during the training. Therefore, the overall model is trained with batch gradient descent on Office-Home, Office-31, ImageCLEF and mini-batch gradient descent (batch size is 5k) on VisDA-2017. The importance weight w is estimated by BBSE algorithm [7] and updated on the fly. In training stage, we first warm up the model on source domain with risk $\mathbb{E}_P[\ell(f(\mathbf{x}^s), y^s)]$ for 20 epochs, and then train the model with full objective. Such a warm up will reduce the uncertainty induced by pseudo labels effectively. The overall training pipeline for full objective is summarized in Alg. S.1. Note that

Algorithm S.1 MOT-based Model for PDA

Input: labeled source data $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^n$ and unlabeled target data $\{(\mathbf{x}_i^t)\}_{i=1}^m$, training epochs $E_{\max}$, conditional alignment parameter η;

Output: representation learner f_r, task learner f_c;

Initialize f_r and f_c as neural networks;

for $i = 1, 2, \ldots, E_{\max}$ **do**

  Forward propagate $\{\mathbf{x}_i^s\}_{i=1}^{n_s}$ and $\{\mathbf{x}_i^t\}_{i=1}^{n_t}$ and obtain $\{(\mathbf{z}_i^s, \hat{y}_i^s)\}_{i=1}^{n_s}$ and $\{(\mathbf{z}_i^t, , \hat{y}_i^t)\}_{i=1}^{n_t}$;

  # Weight Estimation

  Estimate importance weight w on-the-fly with BBSE algorithm [7];

  # Transport Assignment Learning

  Compute reweighted source p_X^w as Def. 1 and maksed kernel $\tilde{\mathcal{K}}$ as Eq. (10) with pseudo target labels $\{\hat{y}_i^t\}_{i=1}^{n_t}$;

  Compute transport plan $\tilde{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \mathcal{L}_{\text{cond}}$ for MOT according to Alg. 1;

  # Conditional Alignment and Risk Minimization

  Compute alignment loss $\mathcal{L}_{\text{cond}}(f_r, \tilde{\mathbf{\Gamma}})$ and risk loss $\mathcal{L}_{\text{risk}}(f_r, f_c)$ with transport plan $\tilde{\mathbf{\Gamma}}$ and barycenter mapping ψ;

  Update learners with overall loss $\mathcal{L}(f_r, f_c) = \mathcal{L}_{\text{risk}}(f_r, f_c) + \eta \mathcal{L}_{\text{cond}}(f_r, \tilde{\mathbf{\Gamma}})$:

$$f_r \leftarrow f_r - \lambda \nabla_{f_r} \mathcal{L}(f_r, f_c), \quad f_c \leftarrow f_c - \lambda \nabla_{f_c} \mathcal{L}(f_r, f_c)$$

end for

studying deep model-based implementation with mini-batch OT [3, 9] is also a meaningful direction. Compared with the shallow networks with larger batch-size, mini-batch OT algorithm ensures larger capacity of deep model.

S.2.2. Dataset Details

• **Office-Home** [14] contains 15k images from 4 domains with 65 classes, i.e., *Art* (**A**), *Clipart* (**C**), *Product* (**P**) and *Real-World* (**R**). In PDA setting, target domain consists of the first 25 classes (alphabetical order).

• **VisDA-2017** [12] contains 152k synthetic images from domain **S** and 55k real images from domain **R**. There are 12 classes, and we form target domain with the first 6 classes.

• **Office-31** [13] contains 4k images and 31 classes from *Amazon* (**A**), *Webcam* (**W**), *Dslr* (**D**). We follow standard protocol [1] to form target domain with 10 classes.

• **ImageCLEF** [2] contains 3 domains with 12 classes, i.e., *Caltech* (**C**), *ImageNet* (**I**), *Pascal* (**P**). We form target domain with the first 6 classes as protocol [8].

S.2.3. About Parameters

There are two major parameters for MOT model, i.e., entropic regularization parameter λ and relaxation parameter β for UOT. For sensitivity of parameters, the model performance is generally robust under different entropic parameter λ, while the larger λ (i.e., closer to original OT) may reduce the convergence speed of Sinkhorn iteration. For relaxation parameter β, its impact is related with the degree of label shift. When label shift is severer, β (i.e., penalty on marginals) should be smaller to reduce negative transfer. In this case, the impact of β will be significant, and vice versa.

S.2.4. About Barycenter Map

Note that the barycenter maps learned with original plan and entropy regularized plan are generally different. Empirically, we observed that intuitive strategy for increasing the sparsity of Sinkhorn plan is effective. For example, truncating the small values in plan $\mathbf{\Gamma}$ with threshold or contribution ratio can improve the proportion of accurate connection pairs in $\mathbf{\Gamma}$. Therefore, learning de-biased map and reducing the density of $\mathbf{\Gamma}$, e.g., low entropy regularization in Eq. (12), could be meaningful problems.

S.2.5. About Partial OT

For relaxation strategy, there is another methodology called partial OT (POT) [4, 5, 10]. For masked version of POT, the empirical modeling can be directly achieved by replacing the UOT-based relaxation model Eq. (8) with partial OT. But, the theoretical understanding of mask mechanism with partial OT needs an in-depth analysis, which could be a meaningful problem. We will provide preliminary discussions on masked partial OT.

References

[1] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *ECCV*, pages 135–150, 2018. 6

[2] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211, 2014. 6

[3] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTAT*, pages 2131–2141, 2020. 6

[4] Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010. 6

[5] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. In *NeurIPS*, 2022. 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[7] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, pages 3122–3130. PMLR, 2018. 5, 6

[8] You-Wei Luo, Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. Unsupervised domain adaptation via discriminative manifold propagation. *IEEE TPAMI*, 44(3):1653–1669, 2022. 6

[9] Khai Nguyen, Dang Nguyen, Quoc Nguyen, Tung Pham, Hung Bui, Dinh Phung, Trung Le, and Nhat Ho. On transportation of mini-batches: A hierarchical approach. In *ICML*, page 34, 2022. 6

[10] Khai Nguyen, Dang Nguyen, Tung Pham, Nhat Ho, et al. Improving mini-batch optimal transport via partial transportation. In *ICML*, pages 16656–16690, 2022. 6

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 5

[12] 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. 6

[13] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 6

[14] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 6