WEIGHTED PARTIAL OPTIMAL TRANSPORT FOR MULTI-SOURCE PARTIAL DOMAIN ADAPTATION

Jayadev Naram, Rebecka Jörnsten & Giuseppe Durisi
Chalmers University of Technology
Gothenburg, Sweden
{jayadev,jornsten,durisi}@chalmers.se

Ziming Wang
Ant Group
Shanghai, China
wzm2256@gmail.com

ABSTRACT

We develop a theoretical and algorithmic framework for multi-source partial domain adaptation (MSPDA) by deriving a generalization bound that relates the target loss to weighted empirical source losses and source-specific partial Wasserstein distances. This bound motivates a partial optimal transport algorithm, termed MS-WARMPOT, that shares a common feature extractor across domains, addressing multi-source heterogeneity. MS-WARMPOT learns source-target sample weights that suppress outlier classes and prevent negative transfer, thereby unifying multi-source domain adaptation (MSDA) and partial domain adaptation within a single framework. Experiments on standard MSDA and MSPDA benchmarks demonstrate competitive performance against the state-of-the-art methods.

1 INTRODUCTION

In many real-world applications, labeled data are collected from multiple related domains—for example, due to differences in acquisition conditions. However, since labeling is often expensive and time-consuming, samples from the target domain are typically unlabeled. Multi-source domain adaptation (MSDA) seeks to leverage the available, diverse, labeled data sources to improve generalization on the target domain, by aligning their distributions. Most MSDA methods assume that all source domains and the target domain share the same label space. In practice, however, some source classes may be absent from the target, which can lead to negative transfer (Cao et al., 2018). Partial domain adaptation (PDA) addresses the label-mismatch problem by assuming that the target label space is an unknown subset of the source label space. Although PDA algorithms handle label mismatch effectively in the single-source setting, when multiple heterogeneous sources are present, these methods have limited efficacy. Indeed, different sources may vary widely in instance distribution and class composition, and treating them as a single aggregated source obscures these differences.

This combination of challenges, i.e., label mismatch and multi-source heterogeneity, gives rise to the multi-source partial domain adaptation (MSPDA) problem. Available MSPDA methods address this problem by performing partial feature selection and partial alignment (Fu et al., 2021), or reweighting source instances based on their relevance to the target (Zhang et al., 2024; Kim et al., 2025). However, a theoretical treatment of this problem is lacking.

Building on recent theoretical advances on the use of optimal transport to perform domain adaptation in various settings (Courty et al., 2017; Redko et al., 2017; Shen et al., 2018; Naram et al., 2025), we develop a theoretically grounded framework for MSPDA. Specifically, we first derive a bound on the empirical target loss in terms of weighted empirical source losses and partial Wasserstein distance between each individual source empirical distribution and the target empirical distribution. We then use the probably approximately correct (PAC) Bayes framework (McAllester, 1999; Catoni, 2007; Alquier, 2024) to translate this empirical-loss bound into a generalization bound. A key feature of our bound is that the weights are obtained naturally from the partial optimal transport problem solved for each source-target pair.

Motivated by our theoretical results, we introduce a partial optimal transport algorithm, termed multi-source weighted and regularized minimizer via partial optimal transport (MS-WARMPOT), that aligns each source domain to the target domain while sharing a common feature extractor across

all domains. The shared representation enables knowledge transfer across sources, while the source-specific alignment selectively emphasizes target-relevant source instances and suppresses outliers. When the source and target label spaces coincide, MS-WARMPOT becomes a MSDA algorithm, while when a single source domain is present, MS-WARMPOT reduces to the PDA algorithm, named WARMPOT, recently proposed by Naram et al. (2025). MSDA and PDA specific algorithms therefore emerge as special cases of a single coherent framework.

Our main contributions are summarized as follows:

- We derive a new PAC-Bayes generalization bound for MSPDA, in which the target loss is expressed in terms of weighted source losses and source-specific partial Wasserstein distances.
- Motivated by this bound, we propose a partial optimal transport algorithm that aligns each source-target pair separately while sharing a common feature extractor.
- The resulting theoretical and algorithmic framework unifies MSDA and PDA under a single MSPDA formulation.
- Extensive experiments on standard MSDA and MSPDA benchmarks validate the effectiveness of the proposed algorithm against the state-of-the-art solutions.

2 Literature Review

Domain adaptation seeks to transfer knowledge from a labeled source domain to an unlabeled target domain with differing distribution. Classical methods promote domain-invariant features through adversarial learning (Ganin et al., 2016; Tzeng et al., 2017). Theoretical analyses based on $\mathcal{H}$-divergence (Ben-David et al., 2010) provided early insights, but rely on VC-dimension bounds that poorly capture generalization in many practically relevant settings (Nagarajan & Kolter, 2019; Zhang et al., 2021). More recent formulations using optimal transport (Courty et al., 2017; Redko et al., 2017) quantify distributional shift using Wasserstein distances, but involve population measures that are translated to empirical measures via an additional concentration step, and overlook feature-level adaptation central to modern deep domain-adaptation algorithms.

PDA relaxes the identical label-space assumption by allowing the target label space to be a subset of the source label space. Accordingly, PDA algorithms aim to identify and down-weight outlier source, *i.e.*, classes that are absent in the target domain during training, thereby preventing negative transfer. Methods such as PADA (Cao et al., 2018) and ETN (Cao et al., 2019) assign class-level weights based on the predictions on unlabeled target samples. MPOT (Nguyen et al., 2022) performs partial feature alignment of source and target distributions using partial Wasserstein distance, but it relies on uniformly weighted empirical source loss. PWAN (Wang et al., 2025) combines class-level weights with partial Wasserstein distance for partial feature alignment. JUMBOT (Fatras et al., 2021) uses uniformly weighted source loss and employs unbalanced optimal transport for partial feature alignment. More recently, Naram et al. (2025) derive generalization bounds on population target loss, motivating an algorithm, named WARMPOT, in which the weights are obtained directly from a partial optimal transport problem.

Single-source adversarial training has been extended to the multi-source setting. Specifically, Zhao et al. (2018) derive a generalization bound for MSDA based on the $\mathcal{H}$-divergence and, motivated by this bound, propose an adversarial training algorithm employing K domain classifiers to learn invariant representations. Similarly, Peng et al. (2019) derive a generalization bound in terms of pairwise moment distances and propose minimizing these discrepancies across domains to align feature distributions. Subsequent works have explored optimal transport-based generalization bounds for MSDA. Redko et al. (2017) derive a generalization bound on the population target loss in terms of a weighted sum of Wasserstein distances between empirical source and target distributions. Building on this result, Montesuma & Mboula (2021) propose an algorithm based on the Wasserstein barycenter (Agueh & Carlier, 2011) of the empirical source distributions. Finally, Turrisi et al. (2022) derive a generalization bound in terms of Wasserstein distance between weighted source distributions and the target distribution. However, in these generalization bounds, the Wasserstein distance is first derived at the population level and then related to empirical measures via an additional concentration step. In contrast, our theoretical result is established directly at the level of empirical measures (see Section 3.3).

Beyond single-source PDA and MSDA formulations, recent works have developed methods explicitly designed for the MSPDA setting. PFSA (Fu et al., 2021) suppresses source features that are irrelevant to the target domain and performs feature alignment by minimizing class-level and domain-level moment discrepancies. Zhang et al. (2024) propose aligning each pair of source and target domains while reweighting source instances according to their relevance to the target. In the regression setting, GAUL (Kim et al., 2025) aligns weighted source domains to target domain, incorporating both domain-level and instance-level weighting schemes. However, these methods lack a theoretical motivation.

3 Theoretical Results

In Section 3.1, we formalize the problem setting and establish the notation used throughout the paper. In Section 3.2, we derive a bound on the empirical target loss, which is subsequently employed in Section 3.3 to obtain a generalization bound.

3.1 Problem Setup and Proposed Approach

For each $k \in \{1, \dots, K\}$, let $\mathcal{Z}^{(k)} = \mathcal{X} \times \mathcal{Y}^{(k)}$ denote the kth source domain, where $\mathcal{X} \subseteq \mathbb{R}^d$ represents the input space endowed with the sigma-algebra Σ_X, and $\mathcal{Y}^{(k)} \subseteq \mathbb{R}$ is the corresponding label space equipped with the sigma-algebra $\Sigma_{Y^{(k)}}$. We assume a joint probability distribution $P_{Z^{(k)}}$ defined on $(\mathcal{Z}^{(k)}, \Sigma_X \otimes \Sigma_{Y^{(k)}})$, which we refer to as the kth source distribution. Analogously, the target domain is given by $\tilde{\mathcal{Z}} = \mathcal{X} \times \tilde{\mathcal{Y}}$, where $\tilde{\mathcal{Y}} \subseteq \mathbb{R}$ is equipped with the sigma-algebra $\Sigma_{\tilde{Y}}$. The target data are governed by a joint probability distribution $Q_{\tilde{Z}}$ on $(\tilde{\mathcal{Z}}, \Sigma_X \otimes \Sigma_{\tilde{Y}})$, which we call the target distribution.[1]

Let $\mathcal{Y} = \bigcup_{k=1}^{K} \mathcal{Y}^{(k)}$. In what follows, we study both MSDA and MSPDA problems, defined below.

- MSDA: The label spaces of all domains are identical, *i.e.*,
$$\mathcal{Y}^{(1)} = \mathcal{Y}^{(2)} = \cdots = \mathcal{Y}^{(K)} = \tilde{\mathcal{Y}} = \mathcal{Y}. \tag{1}$$
- MSPDA (Fu et al., 2021): No two domains have identical label space, but the union of the source label spaces subsume the target label space, *i.e.*,
$$\mathcal{Y}^{(1)} \neq \mathcal{Y}^{(2)} \neq \cdots \neq \mathcal{Y}^{(K)} \neq \tilde{\mathcal{Y}}, \quad \tilde{\mathcal{Y}} \subseteq \bigcup_{k=1}^{K} \mathcal{Y}^{(k)} = \mathcal{Y}. \tag{2}$$

Let a hypothesis be a measurable mapping $w : \mathcal{X} \to \mathcal{Y}$ represented as the composition $w = g \circ f$, where f denotes a feature extractor and g a classifier acting on the extracted features. Throughout this work, we consider bounded loss functions $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ for the sake of simplicity. Our goal is to identify a hypothesis w within a suitably chosen hypothesis class $\mathcal{W}$ (specified later) that minimizes the population target loss
$$L_{Q_{\tilde{Z}}}(w) = \mathbb{E}_{(X,Y)\sim Q_{\tilde{Z}}}[\ell(w(X), Y)]. \tag{3}$$

In the multi-source adaptation scenario considered in this work, we are given K labeled source datasets. Specifically, for each source domain $k \in \{1, \dots, K\}$, let $\boldsymbol{z}^{(k)} = (z_1^{(k)}, \dots, z_{n_k}^{(k)}) \in \left(\mathcal{Z}^{(k)}\right)^{n_k}$, with $z_i^{(k)} = (x_i^{(k)}, y_i^{(k)})$, denote a collection of labeled samples drawn independently from the source distribution $P_{Z^{(k)}}$. In addition, we assume that the learner has access to an unlabeled target sample $\boldsymbol{t} = (\tilde{x}_1, \dots, \tilde{x}_{n_t})$ drawn independently from Q_X, which is the marginal distribution over $\mathcal{X}$ induced by $Q_{\tilde{Z}}$. We denote by $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \dots, \tilde{z}_{n_t}) \in \tilde{\mathcal{Z}}^{n_t}$, with $\tilde{z}_j = (\tilde{x}_j, \tilde{y}_j)$, the corresponding (unobserved) labeled target sample. Because these target labels are unavailable to the learner, the empirical target loss
$$L_{\tilde{\boldsymbol{z}}}(w) = \frac{1}{n_t} \sum_{j=1}^{n_t} \ell(w(\tilde{x}_j), \tilde{y}_j) \tag{4}$$

[1] Unless otherwise stated, sigma-algebras are omitted when clear from context.

cannot be directly computed. To address this issue, in Section 3.2 we derive an upper bound on the empirical target loss. The resulting bound comprises K partial Wasserstein distance terms, one for each source-target pair, as well as weighted empirical source losses, where the weighting scheme depends on the optimal coupling obtained from the optimization problem defining the partial Wasserstein distance terms. The formal definition of this distance is presented below.

Definition 3.1 (Figalli, 2010, Eq. (2.1), Caffarelli & McCann, 2010, Eq. (1.8))**.** The partial Wasserstein distance with parameter α between two measures[2] P_X and $Q_{\tilde{X}}$ on $(\mathcal{X}, \Sigma_X)$ is defined as

$$\mathbb{PW}_\alpha(P_X, Q_{\tilde{X}}) = \inf_{\Pi \in \Gamma_\alpha(P_X, Q_{\tilde{X}})} \int c(x, \tilde{x}) \, \mathrm{d}\Pi(x, \tilde{x}), \tag{5}$$

where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the so-called cost function (typically a metric) and $\Gamma_\alpha(P_X, Q_{\tilde{X}})$ is the set of all nonnegative measures Π on $\mathcal{X} \times \mathcal{X}$ for which $\Pi(\mathcal{X} \times \mathcal{X}) = \alpha$ and for which, for all measurable sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{X}$, we have that $\Pi(\mathcal{A} \times \mathcal{X}) \leq P_X(\mathcal{A})$ and $\Pi(\mathcal{X} \times \mathcal{B}) \leq Q_{\tilde{X}}(\mathcal{B})$.

In the discrete setting, where the probability measures of interest are supported on m and n points respectively, it is convenient to represent P_X and $Q_{\tilde{X}}$ as vectors of dimensions m and n. The transport plan (or coupling measure) can then be expressed as a nonnegative matrix $\Pi \in \mathbb{R}^{m \times n}$ with entries Π_{ij}. Under this formulation, the expression in (5) can be rewritten as

$$\mathbb{PW}_\alpha(P_X, Q_{\tilde{X}}) = \min_{\Pi \in \Gamma_\alpha(P_X, Q_{\tilde{X}})} \sum_{i=1}^{m} \sum_{j=1}^{n} c(x_i, \tilde{x}_j) \Pi_{ij} \tag{6}$$

where $\Gamma_\alpha(P_X, Q_{\tilde{X}})$ denotes the set of admissible nonnegative matrices satisfying[3] $\Pi \mathbf{1}_n \leq P_X$, $\Pi^T \mathbf{1}_m \leq Q_{\tilde{X}}$ and $\mathbf{1}_m^T \Pi \mathbf{1}_n = \alpha$.

3.2 Bounds on the Empirical Target Loss

Building on the theoretical results of Redko et al. (2017, Theorem 4) and Naram et al. (2025, Theorem 3.3), we derive in Theorem 3.2 an upper bound on the empirical target loss, in which each of the K source domains is aligned with the target domain through a partial Wasserstein distance computed between their respective joint empirical distributions of features and labels (for the sources) and features and predicted labels (for the target). This joint-distribution formulation is particularly advantageous in scenarios involving labeling distribution shift, where the conditional distribution of labels given inputs differs across domains.

In broad terms, the bound comprises four components: (i) K weighted empirical source losses evaluated on labeled samples from each source domain, (ii) K partial Wasserstein distance terms measuring the discrepancy between each source and the target domain, (iii) a total variation term that captures the discrepancy introduced when replacing a weighted empirical target loss with the actual empirical loss (4), and (iv) a noncomputable term characterizing the intrinsic difficulty of the domain adaptation problem.

Theorem 3.2. *Assume that the loss function ℓ is a metric on $\mathcal{Y}$ and ζ-Lipschitz in each argument. Consider the set $\mathcal{W}$ of hypotheses $w = g \circ f$ for which g is γ-Lipschitz with respect to the Euclidean distance. Let $P^f_{\boldsymbol{z}^{(k)}} = \frac{1}{n_k} \sum_{i=1}^{n_k} \delta_{f(x_i^{(k)}), y_i^{(k)}}$, and $Q^w_{\boldsymbol{t}} = \frac{1}{n_t} \sum_{j=1}^{n_t} \delta_{f(\tilde{x}_j), w(\tilde{x}_j)}$ be the empirical joint feature-label distribution for the kth source and the empirical joint feature-estimated label distribution for the target, respectively, corresponding to the hypothesis $w = g \circ f$. Then, for all $w \in \mathcal{W}$, and all $\alpha_k, \beta_k \in (0, 1], k = 1, \ldots, K$,*

$$L_{\tilde{\boldsymbol{z}}}(w) \leq \sum_{k=1}^{K} \frac{1}{K \alpha_k} \left\{ \sum_{i=1}^{n_k} p_i^{(k)} \ell(w(x_i^{(k)}), y_i^{(k)}) + \mathbb{PW}_{\alpha_k} \left(\tfrac{1}{\beta_k} P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}} \right) \right\} + \frac{1}{2} \sum_{j=1}^{n_t} \left| \frac{1}{n_t} - \sum_{k=1}^{K} \frac{q_j^{(k)}}{K \alpha_k} \right| + L_f, \tag{7}$$

[2] We do not require that the two measures are probability measures. In particular, in our setup we will have $P_X(\mathcal{X}) \geq 1$.

[3] All vector inequalities are to be understood component-wise.

where the underlying cost function in the definition of $\mathbb{PW}_{\alpha_k}$ *is*

$$c((x,y),(\tilde{x},\tilde{y})) = \zeta\gamma\|f(x) - f(\tilde{x})\| + \ell(y,\tilde{y}), \tag{8}$$

the weights $\{p_i^{(k)}\}$ *and* $\{q_j^{(k)}\}$ *are given by*

$$p_i^{(k)} = (\Pi^{(k)^\star}\mathbf{1}_{n_t})_i, \quad i = 1,\dots,n_k \tag{9}$$

$$q_j^{(k)} = \left((\Pi^{(k)^\star})^T\mathbf{1}_{n_k}\right)_j, \quad j = 1,\dots,n_t \tag{10}$$

with $\Pi^{(k)^\star}$ *being the optimal coupling matrix in the definition of* $\mathbb{PW}_{\alpha_k}\left(\frac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right)$*, and*

$$L_f = \min_{g'\in\mathcal{G}}\left\{\sum_{k=1}^{K}\sum_{j=1}^{n_t}\frac{q_j^{(k)}}{K\alpha_k}\ell(g'(f(\tilde{x}_j)),\tilde{y}_j)\right\} + \Xi \tag{11}$$

with Ξ *given in* (28) *(see Appendix D) and* $\mathcal{G}$ *denoting the set of classifiers* g' *associated to hypotheses in* $\mathcal{W}$*.*

The proof of Theorem 3.2 is provided in Appendix D. The degree of partial domain alignment in the bound is governed by the parameters $\alpha_k, \beta_k \in (0,1], k = 1,\dots,K$. Intuitively, β_k controls the proportion of source instances to be aligned with target samples, while α_k specifies the proportion of target instances considered in the transport plan. Consequently, β_k enables partial domain adaptation by mitigating the influence of irrelevant source samples, and α_k helps excluding potential outliers in the target set. Noncomputable terms similar to L_f appear in most theoretical analyses of domain adaptation (Ben-David et al., 2006; Courty et al., 2017; Redko et al., 2017; Turrisi et al., 2022; Naram et al., 2025).

A key aspect of the proposed approach is that all domains share a common feature extractor f, which enables the transfer of knowledge across sources and promotes the learning of domain-invariant representations. At the same time, the bound considers a separate partial optimal transport problem $\mathbb{PW}_{\alpha_k}\left(\frac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right)$ for each source-target pair, through which the empirical joint distributions of the kth source and of the target are partially aligned. This formulation yields alignment weights $p_i^{(k)}$ and $q_j^{(k)}$ derived from the optimal coupling between the joint source and target distributions. Specifically, $p_i^{(k)}$ denotes the transported mass assigned to the source pair $(f(x_i^{(k)}), y_i^{(k)})$, while $q_j^{(k)}$ denotes the transported mass assigned to the target pair $(f(\tilde{x}_j), w(\tilde{x}_j))$. Larger weights are assigned to pairs that exhibit stronger cross-domain similarity, while smaller weights are assigned to dissimilar or weakly aligned pairs. Performing alignment separately for each source-target pair thus allows for fine-grained adaptation. The shared feature representation couples the learning across all domains, facilitating knowledge transfer from multiple sources.

When only a single source domain is available (*i.e.*, $K = 1$), the bound naturally reduces to the partial domain adaptation result of Naram et al. (2025, Theorem 3.2).

3.3 PAC-Bayes Generalization Bounds

In Section 3.2, we established an upper bound on the empirical target loss for a fixed hypothesis. To obtain a generalization bound, we adopt the PAC-Bayes framework (McAllester, 1999; Catoni, 2007; Alquier, 2024), which provides probabilistic generalization guarantees over the choice of the training source and target samples.

In Theorem 3.3, we derive an upper bound on the expected population target loss as a function of the chosen posterior distribution over hypotheses. The bound is expressed as the expectation, under the posterior distribution, of the upper bound in Theorem 3.2, together with a KL term that captures the generalization gap. The proof involves combining Lemma 3.4 by Naram et al. (2025) with the empirical target loss bound from Section 3.2.

Theorem 3.3. *Suppose that the assumptions of Theorem 3.2 hold, and denote the decomposition of the hypothesis* W *into feature extractor and classifier as* $W = G\circ F$*. Let* Q_W *be a prior distribution over* $\mathcal{W}$ *and* $P_{W|\boldsymbol{Z}^{(1)},\dots,\boldsymbol{Z}^{(K)},\boldsymbol{T}}$ *be a posterior distribution over* $\mathcal{W}$ *given the labeled source samples*[4]

[4]We denote by $P_{Z^{(k)}}^{\otimes n_k}$ the n_k-fold product of $P_{Z^{(k)}}$. Similarly, $Q_{\tilde{Z}}^{\otimes n_t}$ stands for the n_t-fold product of $Q_{\tilde{Z}}$.

$\boldsymbol{Z}^{(k)} \sim P_{Z^{(k)}}^{\otimes n_k}$ for $k = 1, \ldots, K$ and the unlabeled target samples $\boldsymbol{T}$. Here, $\boldsymbol{T}$ is the projection on $\mathcal{X}^{n_t}$ of $\tilde{\boldsymbol{Z}} \sim Q_{\tilde{Z}}^{\otimes n_t}$. Then, for every choice of $\lambda > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $(\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(K)}, \boldsymbol{T})$,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}^{(1)},\ldots,\boldsymbol{Z}^{(K)},\boldsymbol{T}}}[L_{Q_{\tilde{Z}}}(W)] \leq B + \mathbb{E}_{P_{W|\boldsymbol{Z}^{(1)},\ldots,\boldsymbol{Z}^{(K)},\boldsymbol{T}}}\Bigg[\sum_{k=1}^{K} \frac{1}{K\alpha_k}\Bigg\{\sum_{i=1}^{n_k} p_i^{(k)} \ell(W(X_i^{(k)}), Y_i^{(k)}) + \mathbb{PW}_{\alpha_k}\left(\tfrac{1}{\beta_k} P^F_{\boldsymbol{Z}^{(k)}}, Q^W_{\boldsymbol{T}}\right)\Bigg\} + \frac{1}{2}\sum_{j=1}^{n_t}\left|\frac{1}{n_t} - \sum_{k=1}^{K}\frac{q_j^{(k)}}{K\alpha_k}\right| + L_F\Bigg], \quad (12)$$

where $(X_i^{(k)}, Y_i^{(k)}) = Z_i^{(k)}$ are the entries of $\boldsymbol{Z}^{(k)}$, and we use the shorthand

$$B = \frac{\lambda}{8n_t} + \frac{D_{\mathrm{KL}}(P_{W|\boldsymbol{Z}^{(1)},\ldots,\boldsymbol{Z}^{(K)},\boldsymbol{T}} \,||\, Q_W) + \log\frac{1}{\delta}}{\lambda}. \quad (13)$$

The proof techniques employed to derive our main results offer several key advantages over that of Redko et al. (2017, Theorem 4). First, the relationship between the target-domain loss and the source-domain loss in Theorem 3.2 is established directly at the level of empirical measures. As a result, the partial Wasserstein distances appearing in our bound (12) are entirely empirical, eliminating the need for additional concentration steps that arise in Redko et al. (2017) when transitioning from population to empirical quantities. Second, the bounds are expressed explicitly in terms of the learned feature representations rather than fixed input distributions. This formulation naturally supports the feature-based alignment strategy using partial optimal transport and provides a theoretical justification for its effectiveness.

4 Algorithm: Extending WARMPOT

Motivated by the bound on the empirical target loss in Theorem 3.2, we propose MS-WARMPOT. Specifically, focusing on the first two computable terms of the upper bound on the empirical target loss given in (7), we consider the following optimization problem:

$$\min_{w} \sum_{k=1}^{K}\left\{\sum_{i=1}^{n_k} p_i^{(k)} \ell(w(x_i^{(k)}), y_i^{(k)}) + \mathbb{PW}_{\alpha_k}\left(\tfrac{1}{\beta_k} P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right)\right\}. \quad (14)$$

Here, the cost function in the definition of the partial Wasserstein distance is $c\big((x, y), (\tilde{x}, \tilde{y})\big) = \eta_1 \|f(x) - f(\tilde{x})\| + \eta_2 \ell(y, \tilde{y})$, and the weights $p_i^{(k)}$ are defined in (9). The parameters η_1 and η_2 determine the influence of the inter-feature and inter-label distances, respectively, acting as regularization terms, while α_k and β_k control the alignment between the kth source and the target distribution.

When $K = 1$, the optimization problem (14) reduces exactly to the WARMPOT objective (Naram et al., 2025) for partial domain adaptation. For $K > 1$, rather than aggregating all sources into a single mixture distribution, as would be done by applying WARMPOT verbatim, we propose to align each source-target pair separately while sharing a common feature extractor across all domains, thereby preserving source-specific structure while enabling cross-domain knowledge transfer.

5 Experiments

We now experimentally evaluate our proposed algorithm, MS-WARMPOT, for multi-source adaptation tasks. We detail our experimental setup in Section 5.1. In Section 5.2, we discuss the implementation aspects of our algorithm. Then, we compare the performance of our algorithm against that of existing MSDA and MSPDA methods in Sections 5.3 and 5.4, respectively. Additional details on the experiments are provided in Appendix A.

Table 1: MSDA test accuracy on the Office-Home dataset. The baseline results are taken from Li et al. (2023) and Ng et al. (2025). WARMPOT† combines all source domains into one and performs domain adaptation.

Algorithm	Art	Clipart	Product	RealWorld	Average
ResNet-50 (He et al., 2016)	64.5	52.3	77.6	80.7	68.8
DANN (Ganin et al., 2016)	64.2	58.0	76.4	78.8	69.3
DAN (Long et al., 2015)	68.2	57.9	78.4	81.9	71.6
DCTN (Xu et al., 2018)	66.9	61.8	79.2	77.7	71.4
MFSAN (Zhu et al., 2019)	72.1	62.0	80.3	81.8	74.1
WADN (Shui et al., 2021)	75.2	61.0	83.5	84.4	76.1
iMSDA (Kong et al., 2022)	75.4	61.4	83.5	84.4	76.2
SIG (Li et al., 2023)	76.4	**63.9**	**85.4**	**85.8**	**77.8**
GAMA (Ng et al., 2025)	**76.6**	62.6	84.9	84.9	77.3
WARMPOT† (Naram et al., 2025)	74.6 (0.2)	60.7 (0.9)	83.7 (0.6)	85.8 (0.1)	76.2 (0.2)
MS-WARMPOT (ours)	75.6 (0.3)	63.6 (0.9)	84.5 (0.4)	85.7 (0.5)	77.3 (0.2)

5.1 SETUP

We evaluate the proposed method on multiple benchmark datasets commonly used in MSDA and MSPDA.

For MSDA, we consider the Office-Home dataset (Venkateswara et al., 2017), which contains approximately $15\,500$ images across 65 object categories drawn from four distinct domains: art, clipart, product, and real-world.

For MSPDA, we construct tasks from Office-Home by selecting 43 classes for each source and target domain. In addition, we consider two benchmark datasets used in prior work (Fu et al., 2021), namely Digit-Five (Peng et al., 2019) and Office-31 (Saenko et al., 2010). Digit-Five comprises five digit recognition datasets—MNIST-M, MNIST, USPS, SVHN, and Synthetic Digits—each containing ten classes (digits 0–9). In the MSPDA setting, five classes are selected for each source domain and seven for the target domain. Office-31 includes 4652 images spanning 31 office-related categories across three domains (amazon, webcam, and DSLR); 21 classes are selected for each source domain and target domain. The detailed class splits for all datasets are provided in Appendix B.

For both MSDA and MSPDA experiments, one domain is designated as the target, while the remaining domains serve as sources. The experiments are repeated for 6 random seeds, and we report the average and the standard deviation.

5.2 IMPLEMENTATION OF MS-WARMPOT

We consider hypotheses of the form $w = g \circ f$, where f is a ResNet (He et al., 2016) feature extractor pretrained on ImageNet (Russakovsky et al., 2015), and g is a fully connected classifier with one hidden layer of dimension 512, *i.e.*, feature $\to 512 \to D$, where D is the number of classes in $\mathcal{Y}$. We use ResNet-50 for all experiments, except for the MSPDA Digit-Five setting, for which the architecture is described in Appendix A.

The objective in (14) is optimized using stochastic gradient descent with learning rate 0.001 for a maximum of 5000 iterations. At each iteration, we compute $\mathbb{PW}_{\alpha_k}\left(\frac{1}{\beta_k} P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right)$ for all K source-target mini-batch pairs, using the PWAN method of Wang et al. (2025). In addition, we incorporate three auxiliary loss terms as in Wang et al. (2025): (i) entropy regularization (Grandvalet & Bengio, 2004), (ii) complement objective regularization (Chen et al., 2019), and (iii) label smoothing (Szegedy et al., 2016). Additional experimental details are provided in Appendix A.[5]

[5]Open source Python implementation of MS-WARMPOT: `https://github.com/JayD2106/WARMPOT`.

Table 2: MSPDA test accuracy on the Office-31 dataset. PFSA† (reproduced) indicates the results we obtained by evaluating PFSA on the class splits of Office-31 dataset detailed in Appendix B.

Algorithm	Amazon	DSLR	Webcam	Avg
PFSA† (reproduced)	56.1 (2.9)	86.3 (2.0)	**84.1** (2.5)	75.5 (2.3)
MS-WARMPOT (ours)	**61.9** (3.8)	**89.1** (3.7)	81.5 (1.3)	**77.5** (2.1)

Table 3: MSPDA test accuracy on the Digit-Five dataset. PFSA† (reproduced) indicates the results we obtained by evaluating PFSA on the class splits of Digit-Five dataset detailed in Appendix B.

Algorithm	MNIST	MNIST-M	SVHN	Synthetic	USPS	Avg
PFSA† (reproduced)	91.4 (3.1)	59.9 (3.9)	50.1 (8.3)	74.4 (5.7)	95.3 (1.0)	74.2 (3.2)
MS-WARMPOT (ours)	**91.7** (1.6)	**64.1** (1.5)	**51.5** (3.8)	**74.6** (1.1)	**96.4** (0.3)	**75.7** (0.9)

5.3 Results for the MSDA Setting

In Table 1, we report the MSDA classification accuracies on the Office-Home dataset. Each column in the table corresponds to a different target domain, while the remaining domains serve as sources. MS-WARMPOT consistently achieves competitive performance across all domains, with an average accuracy of 77.3%, comparable to recent state-of-the-art methods such as SIG (Li et al., 2023) and GAMA (Ng et al., 2025). These empirical results are consistent with the theoretical insights provided by Theorem 3.2, which shows that aligning each source-target pair via the partial Wasserstein distance yields a small target loss. Additionally, we aggregate all source domains into a single mixture and perform domain adaptation using WARMPOT (Naram et al., 2025). This approach performs comparably to recent MSDA methods such as WADN (Shui et al., 2021) and iMSDA (Kong et al., 2022), highlighting the effectiveness of optimal transport for domain adaptation. As expected, however, it underperforms MS-WARMPOT, underscoring the advantage of aligning each source-target pair individually.

5.4 Results for the MSPDA Setting

In Table 2, we report MSPDA classification accuracies on the Office-31 dataset. MS-WARMPOT performs competitively across all domains relative to the state-of-the-art MSPDA method PFSA (Fu et al., 2021). Since the class splits used in Fu et al. (2021) are not publicly available, a direct comparison with their reported results is not possible. Instead, we evaluated PFSA on the class splits of the Office-31 dataset detailed in Appendix B. MS-WARMPOT consistently matches or outperforms these reproduced PFSA results, demonstrating the practical effectiveness of our method in addition to its theoretical motivation. Results from a similar experiment on the Digit-Five dataset (see Table 3) show comparable trends.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

REFERENCES

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Found. Trends Mach. Learn.*, 17 (2):174–303, Jan. 2024.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2006.

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

Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and Monge-Ampere obstacle problems. *Ann. Math.*, 171(2):673–730, Mar. 2010.

Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018.

Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, June 2019.

Olivier Catoni. *PAC-Bayesian Supervised Classification: the Thermodynamics of Statistical Learning*. IMS Lecture Notes Monogr. Ser., 56, Beachwood, OH, USA, 2007.

Hao Yun Chen, Pei Hsin Wang, Chun Hao Liu, Shih Chieh Chang, Jia Yu Pan, Yu Ting Chen, Wei Wei, and Da Cheng Juan. Complement objective training. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *Proc. Int. Conf. Mach. Learning (ICML)*, Virtual Conference, July 2021.

Alessio Figalli. The optimal partial transport problem. *Arch. Ration. Mech. Anal.*, 195(2):533–560, Jan. 2010.

Yangye Fu, Ming Zhang, Xing Xu, Zuo Cao, Chao Ma, Yanli Ji, Kai Zuo, and Huimin Lu. Partial feature selection and alignment for multi-source domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Virtual Conference, June 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res. (JMLR)*, 17(1):2096–2030, Jan. 2016.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2004.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, June 2016.

Jihyun Kim, Hansam Cho, Minjung Lee, and Seoung Bum Kim. Multi-source partial domain adaptation with gaussian-based dual-level weighting for ppg-based heart rate estimation. *Knowledge-Based Systems*, 309:112769, 2025.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *Proc. Int. Conf. Mach. Learning (ICML)*, Baltimore, MA, USA, July 2022.

Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. Subspace identification for multi-source domain adaptation. *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proc. Int. Conf. Mach. Learning (ICML)*, Lille, France, July 2015.

David A. McAllester. PAC-Bayesian model averaging. In *Proc. Conf. Comput. Learn. Theory (COLT)*, Santa Cruz, CA, USA, July 1999.

Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. Wasserstein barycenter for multi-source domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Virtual Conference, June 2021.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

Jayadev Naram, Fredrik Hellström, Ziming Wang, Rebecka Jörnsten, and Giuseppe Durisi. Theoretical performance guarantees for partial domain adaptation via partial optimal transport. In *Proc. Int. Conf. Mach. Learning (ICML)*, Vancouver, Canada, July 2025.

Ignavier Ng, Yan Li, Zijian Li, Yujia Zheng, Guangyi Chen, and Kun Zhang. A general representation-based approach to multi-source domain adaptation. In *Proc. Int. Conf. Mach. Learning (ICML)*, Vancouver, Canada, July 2025.

Khai Nguyen, Dang Nguyen, Tung Pham, Nhat Ho, et al. Improving mini-batch optimal transport via partial transportation. In *Proc. Int. Conf. Mach. Learning (ICML)*, Baltimore, MA, USA, July 2022.

Yuki Ohnishi and Jean Honorio. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Virtual Conference, Apr. 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, June 2019.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Proc. Mach. Learn. Knowl. Discov. Databases (ECML PKDD)*, Skopje, Macedonia, Sep. 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)*, 115(3):211–252, Apr. 2015.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Crete, Greece, Sep. 2010.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, New Orleans, LA, USA, Apr. 2018.

Changjian Shui, Zijian Li, Jiaqi Li, Christian Gagné, Charles X Ling, and Boyu Wang. Aggregating from multiple target-shifted sources. In *Proc. Int. Conf. Mach. Learning (ICML)*, Virtual Conference, July 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, June 2016.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012. Lecture slides.

Rosanna Turrisi, Rémi Flamary, Alain Rakotomamonjy, and Massimiliano Pontil. Multi-source domain adaptation via weighted joint distributions optimal transport. In *Proc. Conf. Uncertainty Artif. Intell. (UAI)*, Eindhoven, Netherlands, Aug. 2022.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, Hawaii, USA, July 2017.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, Hawaii, USA, July 2017.

Zi-Ming Wang, Nan Xue, Ling Lei, Rebecka Jörnsten, and Gui-Song Xia. Partial distribution matching via partial wasserstein adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 47(9):7944–7959, 2025.

Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, June 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, Feb. 2021.

Guowei Zhang, Xianguang Kong, Qibin Wang, Jingli Du, Kun Xu, Jinrui Wang, and Hongbo Ma. Multi-source partial domain adaptation method based on pseudo-balanced target domain for fault diagnosis. *Knowledge-Based Systems*, 284:111255, 2024.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, Canada, Dec. 2018.

Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Honolulu, Hawaii, USA, Jan. 2019.

A Additional Details on the Experiments

We set $\alpha_k = 1$ and $\beta_k = \bar{\beta}_k \times s^v$ for all $k = 1, \ldots, K$, where v denotes the iteration index and $s = \exp(\log(u)/5000)$ is an annealing factor. The parameter u is fixed at 0.03162 across all experiments. For mini-batch construction, we use a balanced sampler as in Damodaran et al. (2018), where each mini-batch contains one randomly selected sample per class. The label smoothing parameter is set to 0.1 across all experiments.

The PWAN model is implemented as a fully connected network with two hidden layers (feature $\to 512 \to 512 \to 1$) using LeakyReLU activations. Following Wang et al. (2025), we train PWAN with a gradient penalty regularizer (Gulrajani et al., 2017), weighted by 1000. The PWAN model is trained using RMSProp (Tieleman & Hinton, 2012) for 15 iterations per update unless otherwise specified. We set $\eta_1 = 1$ for all partial Wasserstein computations.

The loss function of MS-WARMPOT comprises four components: (i) K weighted empirical source losses, one for each source domain; (ii) K partial Wasserstein distances between each source and the target domain; (iii) K complement objective losses (Chen et al., 2019) on the source domains; and (iv) a single entropy loss (Grandvalet & Bengio, 2004) on the target domain. The partial Wasserstein terms are weighted by λ_{dom}, and the complement objective losses by λ_{cot}, where each weight is shared across the K source domains. The entropy loss is weighted separately by λ_{ent}. Let λ_{leaky} denote the LeakyReLU parameter. The hyperparameters for each experiment are specified below.

- **Office-Home MSDA** (Table 1): The batch size is 65. We set $\eta_2 = 0.723$, $\bar{\beta}_1 = 0.614$, $\bar{\beta}_2 = 0.712$, $\bar{\beta}_3 = 0.9$, $\lambda_{\text{dom}} = 0.064$, $\lambda_{\text{ent}} = 0.197$, $\lambda_{\text{cot}} = 0.636$, $\lambda_{\text{leaky}} = 0.368$. The learning rate for PWAN model is set to 0.0005.
- **Office-31 MSPDA** (Table 2): The batch size is 42. We set $\eta_2 = 0.01$, $\bar{\beta}_1 = 0.702$, $\bar{\beta}_2 = 0.217$, $\lambda_{\text{dom}} = 0.0611$, $\lambda_{\text{ent}} = 0.415$, $\lambda_{\text{cot}} = 0.897$, $\lambda_{\text{leaky}} = 0.2$. The PWAN model, with hidden dimension 256, was trained for 5 iterations. The learning rate for PWAN model is set to 0.0001.
- **Digit-Five MSPDA** (Table 3): The batch size is 128. We set $\eta_2 = 0.145$, $\bar{\beta}_1 = 0.752$, $\bar{\beta}_2 = 0.603$, $\bar{\beta}_3 = 0.587$, $\bar{\beta}_4 = 0.989$, $\lambda_{\text{dom}} = 0.271$, $\lambda_{\text{ent}} = 0.348$, $\lambda_{\text{cot}} = 0.774$, $\lambda_{\text{leaky}} = 0.594$. The learning rate for PWAN model is set to 0.0001. Following Fu et al. (2021), we use a feature extractor consisting three convolutional layers ($3 \to 64 \to 64 \to 128$, kernel size 5) with batch normalization and ReLU activations, followed by two fully connected layers ($8192 \to 3072 \to 2048$). A final linear layer ($2048 \to 10$) is used for classification.
- **Office-Home MSPDA** (Table 4, Appendix C): The batch size is 65. We set $\eta_2 = 0.411$, $\bar{\beta}_1 = 0.317$, $\bar{\beta}_2 = 0.482$, $\bar{\beta}_3 = 0.963$, $\lambda_{\text{dom}} = 0.026$, $\lambda_{\text{ent}} = 0.378$, $\lambda_{\text{cot}} = 0.387$, $\lambda_{\text{leaky}} = 0.756$. The learning rate for PWAN model is set to 0.0001.

For PFSA† (reproduced) (see Tables 2 and 3), we set all hyperparameters as suggested in Fu et al. (2021).

B MSPDA Dataset Class Splits

For each MSPDA dataset, we order alphabetically the class labels, and we index them by $0, \ldots, D-1$, where D is the number of classes in $\mathcal{Y}$. In the MSPDA experiments, we consider the following class splits.

- **Digit-Five:** The dataset consists of 5 domains, each containing 10 classes. For each experiment, one domain is designated as the target, while the remaining domains serve as sources. We select 5 classes from each source domain and 7 classes from the target domain to construct the MSPDA task. The specific class splits for the source and target domains are given below:
  $\mathcal{Y}^{(1)} = \{0, 1, 2, 3, 4\}$, $\mathcal{Y}^{(2)} = \{2, 3, 4, 5, 6\}$, $\mathcal{Y}^{(3)} = \{4, 5, 6, 7, 8\}$,
  $\mathcal{Y}^{(4)} = \{6, 7, 8, 9, 0\}$, $\tilde{\mathcal{Y}} = \{2, 3, 4, 5, 6, 7, 8\}$.
  Based on the class splits, we construct the following MSPDA tasks:

Table 4: MSPDA test accuracy on the Office-Home dataset

Algorithm	Art	Clipart	Product	RealWorld	Average
MS-WARMPOT (ours)	63.4 (1.0)	49.5 (1.8)	77.0 (1.5)	80.2 (0.5)	67.5 (0.7)

1. target $=$ MNIST-M, source 1 $=$ MNIST, source 2 $=$ USPS, source 3 $=$ SVHN, source 4 $=$ Synthetic.
2. target $=$ MNIST, source 1 $=$ USPS, source 2 $=$ SVHN, source 3 $=$ Synthetic, source 4 $=$ MNIST-M.
3. target $=$ USPS, source 1 $=$ SVHN, source 2 $=$ Synthetic, source 3 $=$ MNIST-M, source 4 $=$ MNIST.
4. target $=$ SVHN, source 1 $=$ Synthetic, source 2 $=$ MNIST-M, source 3 $=$ MNIST, source 4 $=$ USPS.
5. target $=$ Synthetic, source 1 $=$ MNIST-M, source 2 $=$ MNIST, source 3 $=$ USPS, source 4 $=$ SVHN.

- **Office-31:** The dataset consists of 3 domains, each containing 31 classes. From each domain, we select 21 classes to construct the MSPDA task. The specific class splits for the source and target domains are given below:
  $\mathcal{Y}^{(1)} = \{0, \ldots, 20\},\ \mathcal{Y}^{(2)} = \{10, \ldots, 30\},\ \tilde{\mathcal{Y}} = \{5, \ldots, 25\}$.
  Based on the class splits, we construct the following MSPDA tasks:
  1. target $=$ Amazon, source 1 $=$ Webcam, source 2 $=$ DSLR.
  2. target $=$ DSLR, source 1 $=$ Webcam, source 2 $=$ Amazon.
  3. target $=$ Webcam, source 1 $=$ DSLR, source 2 $=$ Amazon.
- **Office-Home:** The dataset consists of 4 domains, each containing 65 classes. From each domain, we select 43 classes to construct the MSPDA task. The specific class splits for the source and target domains are given below:
  $\mathcal{Y}^{(1)} = \{0, \ldots, 42\},\ \mathcal{Y}^{(2)} = \{11, \ldots, 53\},\ \mathcal{Y}^{(3)} = \{22, \ldots, 64\},$
  $\tilde{\mathcal{Y}} = \{33, \ldots, 64, 0, \ldots, 10\}$.
  Based on the class splits, we construct the following MSPDA tasks:
  1. target $=$ Art, source 1 $=$ Clipart, source 2 $=$ Product, source 3 $=$ RealWorld.
  2. target $=$ Clipart, source 1 $=$ Product, source 2 $=$ RealWorld, source 3 $=$ Art.
  3. target $=$ Product, source 1 $=$ RealWorld, source 2 $=$ Art, source 3 $=$ Clipart.
  4. target $=$ RealWorld, source 1 $=$ Art, source 2 $=$ Clipart, source 3 $=$ Product.

C Additional Numerical Results

The MSPDA performance of MS-WARMPOT on the Office-Home dataset is reported in Table 4. For this dataset, no benchmark results are available in the literature.

D Proof of Theorem 3.2

The proof follows along the same lines as the proof of (Naram et al., 2025, Theorem 3.3). We start by noting that there may be duplicate entries in $\{(f(x_i^{(k)}), y_i^{(k)})\}_{i=1}^{n_k}, k = 1, \ldots, K$ and $\{(f(\tilde{x}_j), w(\tilde{x}_j))\}_{j=1}^{n_t}$. Hence, strictly speaking, $P^f_{\boldsymbol{z}^{(k)}}, k = 1, \ldots, K$ and $Q^w_{\boldsymbol{t}}$ are probability vectors whose dimensions are given by the number of distinct features, and multiplicities need to be accounted for. However, in our proof, this yields the same result as if we treat the duplicate values as separate features with identical cost values. Hence, for simplicity but without loss of generality, we assume that the entries in both $\{(f(x_i^{(k)}), y_i^{(k)})\}_{i=1}^{n_k}, k = 1, \ldots, K$ and $\{(f(\tilde{x}_j), w(\tilde{x}_j))\}_{j=1}^{n_t}$ are distinct. This allows us to view $P^f_{\boldsymbol{z}^{(k)}}$ and $Q^w_{\boldsymbol{t}}$ as probability vectors of dimensions n_k and n_t respectively, where all entries of each vector are equal, *i.e.*, $P^f_{\boldsymbol{z}^{(k)}} = [1/n_k, \ldots, 1/n_k]^T$ and $Q^w_{\boldsymbol{t}} = [1/n_t, \ldots, 1/n_t]^T$.

For $k = 1, \dots, K$, consider the $n_k \times n_t$ cost matrix $C^{(k)}$ with entries $C_{ij}^{(k)} = \zeta\gamma\|f(x_i^{(k)}) - f(\tilde{x}_j)\| + \ell(y_i^{(k)}, w(\tilde{x}_j))$. We consider the partial Wasserstein distance between $P^f_{\boldsymbol{z}^{(k)}}$ and $Q^w_{\boldsymbol{t}}$, which is given by (see the definition in (6))

$$\mathbb{PW}_{\alpha_k}\left(\tfrac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right) = \min_{\Pi^{(k)} \in \Gamma_{\alpha_k}\left(\frac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right)} \sum_{i=1}^{n_k}\sum_{j=1}^{n_t} C_{ij}^{(k)}\Pi_{ij}^{(k)}, \tag{15}$$

where $\Gamma_{\alpha_k}\left(\frac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right) = \left\{\Pi^{(k)} \in \mathbb{R}^{n_k \times n_t} : \Pi^{(k)}\mathbf{1}_{n_t} \le \frac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, (\Pi^{(k)})^T\mathbf{1}_{n_k} \le Q^w_{\boldsymbol{t}}, \mathbf{1}_{n_s}^T\Pi^{(k)}\mathbf{1}_{n_t} = \alpha_k\right\}$.

Let $\boldsymbol{q} = [q_1^{(1)}, \dots, q_{n_t}^{(1)}, \dots, q_1^{(K)}, \dots, q_{n_t}^{(K)}]^T \in \mathbb{R}^{Kn_t}$, where the weights $q_j^{(k)}$ are defined in (10). Also, let

$$Q_{\tilde{\boldsymbol{z}}} = \sum_{j=1}^{n_t}\frac{1}{n_t}\delta_{\tilde{z}_j}, \qquad Q^{\boldsymbol{q}}_{\tilde{\boldsymbol{z}}} = \sum_{k=1}^{K}\sum_{j=1}^{n_t}\frac{q_j^{(k)}}{K\alpha_k}\delta_{\tilde{z}_j}. \tag{16}$$

Then, given the feature map f, for every fixed hypothesis $w' \in \mathcal{W}$ that can be decomposed as $w' = g' \circ f$, we have

$$L_{\tilde{\boldsymbol{z}}}(w) \le \mathrm{TV}(Q_{\tilde{\boldsymbol{z}}}, Q^{\boldsymbol{q}}_{\tilde{\boldsymbol{z}}}) + \sum_{k=1}^{K}\sum_{j=1}^{n_t}\frac{q_j^{(k)}}{K\alpha_k}\ell(w(\tilde{x}_j), \tilde{y}_j) \tag{17}$$

$$\le \mathrm{TV}(Q_{\tilde{\boldsymbol{z}}}, Q^{\boldsymbol{q}}_{\tilde{\boldsymbol{z}}}) + \sum_{k=1}^{K}\sum_{j=1}^{n_t}\frac{q_j^{(k)}}{K\alpha_k}\left(\ell(w(\tilde{x}_j), w'(\tilde{x}_j)) + \ell(w'(\tilde{x}_j), \tilde{y}_j)\right) \tag{18}$$

$$= \mathrm{TV}(Q_{\tilde{\boldsymbol{z}}}, Q^{\boldsymbol{q}}_{\tilde{\boldsymbol{z}}}) + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{p_i^{(k)}}{K\alpha_k}\ell(w'(x_i^{(k)}), y_i^{(k)}) + \sum_{k=1}^{K}\sum_{j=1}^{n_t}\frac{q_j^{(k)}}{K\alpha_k}\ell(w'(\tilde{x}_j), \tilde{y}_j)$$
$$+ \sum_{k=1}^{K}\frac{1}{K\alpha_k}\left(\sum_{j=1}^{n_t}q_j^{(k)}\ell(w(\tilde{x}_j), w'(\tilde{x}_j)) - \sum_{i=1}^{n_k}p_i^{(k)}\ell(w'(x_i^{(k)}), y_i^{(k)})\right). \tag{19}$$

Here, (17) follows from a change of measure (Ohnishi & Honorio, 2021, Lemma 4); in (18) we used that the weights $\{q_j^{(k)}\}$ are nonnegative as well as triangle inequality; to obtain (19) we summed and subtracted the term $\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{p_i^{(k)}}{K\alpha_k}\ell(w'(x_i^{(k)}), y_i^{(k)})$. We now focus on the last two terms of (19).

Let $\Pi^{(k)\star}$ be the coupling matrix achieving $\mathbb{PW}_{\alpha_k}\left(\frac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right)$. We have

$$\begin{aligned}
&\sum_{j=1}^{n_t} q_j^{(k)} \ell(w(\tilde{x}_j), w'(\tilde{x}_j)) - \sum_{i=1}^{n_k} p_i^{(k)} \ell(w'(x_i^{(k)}), y_i^{(k)}) \\
&\qquad = \sum_{j=1}^{n_t} \ell(w(\tilde{x}_j), w'(\tilde{x}_j)) \sum_{i=1}^{n_k} \Pi_{ij}^{(k)\star} - \sum_{i=1}^{n_k} \ell(w'(x_i^{(k)}), y_i^{(k)}) \sum_{j=1}^{n_t} \Pi_{ij}^{(k)\star} && (20)\\
&\qquad = \sum_{i=1}^{n_k}\sum_{j=1}^{n_t} \Pi_{ij}^{(k)\star} \Big(\ell(w(\tilde{x}_j), w'(\tilde{x}_j)) - \ell(w'(x_i^{(k)}), y_i^{(k)})\Big) && (21)\\
&\qquad \le \sum_{i=1}^{n_k}\sum_{j=1}^{n_t} \Pi_{ij}^{(k)\star} \Big|\ell(w(\tilde{x}_j), w'(\tilde{x}_j)) - \ell(w'(x_i^{(k)}), y_i^{(k)})\Big| && (22)\\
&\qquad \le \sum_{i=1}^{n_k}\sum_{j=1}^{n_t} \Pi_{ij}^{(k)\star} \Big[\Big|\ell(w(\tilde{x}_j), w'(\tilde{x}_j)) - \ell(w(\tilde{x}_j), w'(x_i^{(k)}))\Big| \\
&\qquad\qquad + \Big|\ell(w(\tilde{x}_j), w'(x_i^{(k)})) - \ell(w'(x_i^{(k)}), y_i^{(k)})\Big| \Big] && (23)\\
&\qquad \le \sum_{i=1}^{n_k}\sum_{j=1}^{n_t} \Pi_{ij}^{(k)\star} \Big[\zeta \Big|w'(\tilde{x}_j) - w'(x_i^{(k)})\Big| + \ell(w(\tilde{x}_j), y_i^{(k)})\Big] && (24)\\
&\qquad \le \sum_{i=1}^{n_k}\sum_{j=1}^{n_t} \Pi_{ij}^{(k)\star} \Big[\zeta\gamma \Big\|f(\tilde{x}_j) - f(x_i^{(k)})\Big\| + \ell(w(\tilde{x}_j), y_i^{(k)})\Big] && (25)\\
&\qquad = \mathbb{PW}_{\alpha_k}\left(\tfrac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right). && (26)
\end{aligned}$$

Here, (20) follows from the definitions of $p_i^{(k)}$ in (9) and $q_j^{(k)}$ in (10); to obtain (23) we added and subtracted the term $\ell(w(\tilde{x}_j), w'(x_i^{(k)}))$ and used the triangle inequality of $|\cdot|$; (24) follows because the loss is ζ-Lipschitz and because of the reverse triangle inequality; (25) follows since g' is γ-Lipschitz with respect to the Euclidean distance.

By substituting (26) into (19) and decomposing w' as $w' = g' \circ f$, we obtain

$$\begin{aligned}
L_{\tilde{\boldsymbol{z}}}(w) \le &\sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{p_i^{(k)}}{K\alpha_k} \ell(g'(f(x_i^{(k)})), y_i^{(k)}) + \sum_{k=1}^{K}\sum_{j=1}^{n_t} \frac{q_j^{(k)}}{K\alpha_k} \ell(g'(f(\tilde{x}_j)), \tilde{y}_j) \\
&+ \sum_{k=1}^{K} \frac{1}{K\alpha_k} \mathbb{PW}_{\alpha_k}\left(\tfrac{1}{\beta_k}P^f_{\boldsymbol{z}^{(k)}}, Q^w_{\boldsymbol{t}}\right) + \mathrm{TV}(Q_{\tilde{\boldsymbol{z}}}, Q^{\boldsymbol{q}}_{\tilde{\boldsymbol{z}}}). \qquad (27)
\end{aligned}$$

Next, we define

$$\begin{aligned}
\Xi = &\min_{g'\in\mathcal{G}} \left\{ \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{p_i^{(k)}}{K\alpha_k} \ell(g'(f(x_i^{(k)})), y_i^{(k)}) + \sum_{k=1}^{K}\sum_{j=1}^{n_t} \frac{q_j^{(k)}}{K\alpha_k} \ell(g'(f(\tilde{x}_j)), \tilde{y}_j) \right\} \\
- &\left(\min_{g'\in\mathcal{G}} \left\{ \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{p_i^{(k)}}{K\alpha_k} \ell(g'(f(x_i^{(k)})), y_i^{(k)}) \right\} + \min_{g'\in\mathcal{G}} \left\{ \sum_{k=1}^{K}\sum_{j=1}^{n_t} \frac{q_j^{(k)}}{K\alpha_k} \ell(g'(f(\tilde{x}_j)), \tilde{y}_j) \right\} \right). \qquad (28)
\end{aligned}$$

We now minimize over g' in the two summations of (27), and note that

$$\begin{aligned}
&\min_{g'\in\mathcal{G}} \left\{ \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{p_i^{(k)}}{K\alpha_k} \ell(g'(f(x_i^{(k)})), y_i^{(k)}) + \sum_{k=1}^{K}\sum_{j=1}^{n_t} \frac{q_j^{(k)}}{K\alpha_k} \ell(g'(f(\tilde{x}_j)), \tilde{y}_j) \right\} \\
&\qquad \le \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{p_i^{(k)}}{K\alpha_k} \ell\Big(g(f(x_i^{(k)})), y_i^{(k)}\Big) + \min_{g'\in\mathcal{G}} \left\{ \sum_{k=1}^{K}\sum_{j=1}^{n_t} \frac{q_j^{(k)}}{K\alpha_k} \ell(g'(f(\tilde{x}_j)), \tilde{y}_j) \right\} + \Xi. \qquad (29)
\end{aligned}$$

We obtain the desired result by recalling that $g(f(x_i^{(k)})) = w(x_i^{(k)})$, by using the definition of L_f in (11), and by noting that

$$\mathrm{TV}(Q_{\tilde{\boldsymbol{z}}}, Q_{\tilde{\boldsymbol{z}}}^{\boldsymbol{q}}) = \frac{1}{2} \sum_{j=1}^{n_t} \left| \frac{1}{n_t} - \sum_{k=1}^{K} \frac{q_j^{(k)}}{K\alpha_k} \right|. \tag{30}$$