MOST: Multi-Source Domain Adaptation via Optimal Transport for Student-Teacher Learning

Tuan Nguyen[1] **Trung Le**[1] **He Zhao**[1] **Quan Hung Tran**[2] **Truyen Nguyen**[3] **Dinh Phung**[1,4]

[1]Department of Data Science and AI, Monash University, Australia
[2]Adobe Research, San Jose, CA, USA
[3]University of Akron, USA
[4]VinAI Research, Vietnam

Abstract

Multi-source domain adaptation (DA) is more challenging than conventional DA because the knowledge is transferred from several source domains to a target domain. To this end, we propose in this paper a novel model for multi-source DA using the theory of optimal transport and imitation learning. More specifically, our approach consists of two cooperative agents: a teacher classifier and a student classifier. The teacher classifier is a combined expert that leverages knowledge of domain experts that can be theoretically guaranteed to handle perfectly source examples, while the student classifier acting on the target domain tries to imitate the teacher classifier acting on the source domains. Our rigorous theory developed based on optimal transport makes this cross-domain imitation possible and also helps to mitigate not only the data shift but also the label shift, which are inherently thorny issues in DA research. We conduct comprehensive experiments on real-world datasets to demonstrate the merit of our approach and its optimal transport based imitation learning viewpoint. Experimental results show that our proposed method achieves state-of-the-art performance on benchmark datasets for multi-source domain adaptation including Digits-five, Office-Caltech10, and Office-31 to the best of our knowledge.

1 INTRODUCTION

Recent advances in deep learning have succeeded in undertaking visual learning tasks under the support of massive annotated data [Krizhevsky et al., 2012, Ren et al., 2015, Shelhamer et al., 2017]. However, directly transferring knowledge of such a learned model to a novel domain can undesirably degrade its performance due to the existence of *domain shift* [Quionero-Candela et al., 2009]. To address this issue, a diverse range of approaches in domain adaptation (DA) has been proposed from shallow domain adaptation [Gong et al., 2013, Courty et al., 2017a,b] to deep domain adaptation [Ganin and Lempitsky, 2015, Long et al., 2015, Shu et al., 2018, Damodaran et al., 2018, Nguyen et al., 2019, 2020]. While the conventional DA aims to transfer knowledge from a labeled source domain to an unlabeled target domain, in many real-world contexts, labeled data are collected from multiple domains, for example, images taken under different conditions (e.g., weather, poses, lighting conditions, distinct backgrounds, and etc) [Zhao et al., 2018]. In this paper, we address a challenging but more practical transfer learning problem named multi-source domain adaptation (MSDA) in which we need to transfer knowledge from multiple distinct domains to a single unlabeled target domain.

Imitation learning method has been known as *learning from demonstration*. Specifically, there are two fundamental agents: an expert teacher and a student. The former agent knows how to do its job perfectly, whilst the latter learns a policy to mimic the teacher's behavior. This learning paradigm has been applied in reinforcement learning and sequence prediction [Abbeel and Ng, 2004, Ho and Ermon, 2016].

Inspired by the principle of imitation learning, we propose in this paper a novel model for MSDA, named ***M**ulti-Source Domain Adaptation via **O**ptimal Transport for **S**tudent-**T**eacher Learning* (MOST). When applying the teacher-student mechanism in the context of MSDA, we seek solutions for two naturally raised questions: i) how is the teacher determined? and ii) what are the principle and mechanisms to enable the student to mimic its teacher? We address these two questions by developing a rigorous and intuitive theory based on the literature of optimal transport (OT) [Villani, 2008, Santambrogio, 2015, Peyré et al., 2019]. Our approach (see Sections 4 and 5) postulates that the teacher is a combination of domain experts learned perfectly under the support of labeled source samples, and the student aims to predict

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

unlabeled target samples via imitating the prediction of the teacher. We summarize our contributions in this work as follows:

- We propose a rigorous OT-based theory to leverage imitation learning into domain adaptation. Our general paradigm can further apply to many learning problems including reinforcement learning.
- Under imitation learning's perspective, we propose a novel model for MSDA, which utilizes two cooperative agents: teacher and student. The implementation of MOST is also available online[1].
- Comprehensive experiments are conducted on benchmark datasets for multi-source domain adaptation including Digits-five, Office-Caltech10, and Office-31. The experimental results show that our MOST achieves state-of-the-art performance on those benchmark datasets to the best of our knowledge.

2 RELATED WORK

2.1 UNSUPERVISED DOMAIN ADAPTATION

A variety of unsupervised domain adaptation (UDA) approaches have been successfully applied to generalize a model learned from a labeled source domain to an unlabeled novel target domain. Several existing methods based on discrepancy-based alignment to minimize a different discrepancy metric to close the gap between the source and target domains [Long et al., 2015, Tzeng et al., 2014, Sun and Saenko, 2016, Yan et al., 2017, Lee et al., 2019]. Another branch of UDA methods have leveraged adversarial learning wherein generative adversarial networks [Goodfellow et al., 2014] were employed to align the source and target domains on either feature-level [Ganin and Lempitsky, 2015, Tzeng et al., 2017, Long et al., 2018] or pixel-level [Ghifary et al., 2016, Bousmalis et al., 2017, Sankaranarayanan et al., 2018, Xu et al., 2020b]. On the category-level, some approaches utilized dual classifier [Saito et al., 2018, Lee et al., 2019], or domain prototype [Xie et al., 2018, Pan et al., 2019, Xu et al., 2020a] to investigate the category relations across domains.

2.2 MULTI-SOURCE DOMAIN ADAPTATION

The aforementioned UDA methods mainly consider single-source domain adaptation, which is less practical than multi-source domain adaptation. The fundamental study in Crammer et al. [2007], Mansour et al. [2009], Ben-David et al. [2010] has shed light upon the wide applications of MSDA, such as in Duan et al. [2012], Xu et al. [2018]. Based on the above works, Hoffman et al. [2018] gave strong theoretical guarantees for cross-entropy and other similar losses, which is a normalized solution for the MSDA problems. Recently, Zhao et al. [2018] deployed domain adversarial networks to align the target domain to source domains. Xu et al. [2018] proposed a new model to deal with the *category shift*, which is the case where sources may not completely share their categories. Peng et al. [2019] introduced a model that aligns moments of source and target feature distributions in latent space. Multi-source distilling model was proposed in Zhao et al. [2020] to fine-tune the generator and classifier separately and utilized the domain weight to aggregate target prediction. Finally, the work in Wang et al. [2020] deployed a graph convolutional network to conduct domain alignment on the category-level.

2.3 OPTIMAL TRANSPORT

Optimal Transport (OT) has raised interest in various fields including domain adaptation. Many works in single-source domain adaptation have used OT as a tool to mitigate the domain gap via minimizing the cost of complex distributions [Villani, 2008, Courty et al., 2014, Santambrogio, 2015, Yan et al., 2018, Damodaran et al., 2018, Nguyen et al., 2021, Le et al., 2021]. Recently, Lee et al. [2019] proposed using the sliced Wasserstein distance on the category-level, whereas Xie et al. [2019] proposed SPOT in which the optimal transport plan is approximated by a pushforward of a reference distribution, and cast the optimal transport problem into a minimax problem. The OT-based DA work in Xu et al. [2020c] has leveraged spatial prototypical information and intra-domain structures of image data to reduce the negative transfer caused by target samples near decision boundaries. Notably, Courty et al. [2017b] developed a new framework to connect the theory of optimal transport and domain adaptation [Courty et al., 2017a], which later inspired an OT-based deep DA method (DeepJDOT) [Damodaran et al., 2018] and a learning from multiple data sources method (JCPOT) [Redko et al., 2019].

3 BACKGROUND

3.1 OPTIMAL TRANSPORT

Consider two distributions $\mathbb{P}$ and $\mathbb{Q}$ which operate on the domain $\Omega \subseteq \mathbb{R}^d$, let $d(\mathbf{x}, \mathbf{y})$ be a non-negative and continuous cost function or metric. In the modern mathematical language, the very first notion of optimal transport (i.e., Monge problem) [Villani, 2008, Santambrogio, 2015] aims to find the minimum total cost to transport mass from $\mathbb{Q}$ to $\mathbb{P}$ as

$$\mathcal{M}_d(\mathbb{Q}, \mathbb{P}) := \min_{T:T_{\#}\mathbb{Q}=\mathbb{P}} \mathbb{E}_{\mathbf{x}\sim\mathbb{Q}}\left[d(\mathbf{x}, T(\mathbf{x}))\right],$$

where $T_{\#}\mathbb{Q}$ is the push-forward distribution of $\mathbb{Q}$ via the transport map T. A relaxation of the Monge problem (MP), a.k.a the Kantorovich problem (KP), is defined as

$$\mathcal{K}_d(\mathbb{Q}, \mathbb{P}) := \min_{\gamma\in\Gamma(\mathbb{Q},\mathbb{P})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[d(\mathbf{x}, \mathbf{y})\right], \quad (1)$$

where γ is a coupling admitting $\mathbb{Q}, \mathbb{P}$ as marginals.

[1] https://github.com/tuanrpt/MOST

Under some mild conditions as stated in Theorems 1.32 and 1.33 in Santambrogio [2015], KP is identical to MP and for convenience we denote both $\mathcal{M}_d$ and $\mathcal{K}_d$ collectively by $\mathcal{W}_d$ as $\mathcal{W}_d(\mathbb{Q},\mathbb{P}) = \mathcal{K}_d(\mathbb{Q},\mathbb{P}) = \mathcal{M}_d(\mathbb{Q},\mathbb{P})$.

In addition, under some mild conditions as stated in Theorem 5.10 in Villani [2008], we can replace the primal form by its corresponding dual form

$$\mathcal{W}_d(\mathbb{Q},\mathbb{P}) = \max_{\phi\in\mathcal{L}_1(\Omega,\mathbb{P})} \left\{\mathbb{E}_{\mathbb{Q}}\left[\phi^c(\mathbf{x})\right] + \mathbb{E}_{\mathbb{P}}\left[\phi(\mathbf{y})\right]\right\}, \quad (2)$$

where $\mathcal{L}_1(\Omega,\mathbb{P}) := \left\{\psi : \int_\Omega |\psi(\mathbf{y})|\,\mathrm{d}\mathbb{P}(\mathbf{y}) < \infty\right\}$ and ϕ^c is the c-transform of function ϕ defined as $\phi^c(\mathbf{x}) := \min_{\mathbf{y}}\left\{d(\mathbf{x},\mathbf{y}) - \phi(\mathbf{y})\right\}$.

Clustering view of optimal transport. This view of optimal transport has been utilized to study a rich class of hierarchical and multilevel clustering problems [Ho et al., 2019, 2017]. We now present the clustering view of optimal transport which assists us to interpret our method developed in the sequel. Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete distributions defined as

$$\mathbb{P} := \tfrac{1}{m}\textstyle\sum_{i=1}^m \delta_{\mathbf{u}_i} \text{ and } \mathbb{Q} := \tfrac{1}{n}\sum_{j=1}^n \delta_{\mathbf{v}_j},$$

where $\delta_{\mathbf{x}}$ indicates a Dirac measure centered at $\mathbf{x}$. Without loss of generality, we can assume that $n \leq m$ and consider the Wasserstein distance $\mathcal{W}_d(\mathbb{P},\mathbb{Q})$ w.r.t. a metric d. The following theorem characterizes the clustering view of OT.

Theorem 1. *Consider the following optimization problem:* $\min_{\mathbf{v}_{1:n}} \mathcal{W}_d(\mathbb{P},\mathbb{Q})$. *Let* $\mathbf{v}^*_{1:n}$ *and* $\mathbb{Q}^* := \frac{1}{n}\sum_{j=1}^n \delta_{\mathbf{v}^*_j}$ *be its optimal solution and* T^* *be the optimal transport map as*

$$T^* = \mathit{argmin}_{T:T_\#\mathbb{P}=\mathbb{Q}^*} \sum_{i=1}^m d(\mathbf{u}_i, T(\mathbf{u}_i)).$$

Furthermore, let $\mathbf{c}^*_{1:n}$ *and* σ^* *denote the optimal solution of the following clustering problem:*

$$\min_{\mathbf{c}_{1:n},\sigma\in\Pi(m,n)} \sum_{i=1}^m d\left(\mathbf{u}_i, \mathbf{v}_{\sigma(i)}\right),$$

where $\Pi(m,n)$ *is the set of surjective maps from* $\{1,...,m\}$ *to* $\{1,...,n\}$. *We then have* $\mathbf{c}^*_{1:n} = \mathbf{v}^*_{1:n}$ *and* $T^*(\mathbf{u}_i) = \mathbf{v}^*_{\sigma^*(i)}$.

The above theorem states that if we learn the atoms of $\mathbb{Q}$ to minimize $\mathcal{W}_d(\mathbb{P},\mathbb{Q})$ w.r.t. the metric d, the optimal atoms of $\mathbb{Q}$ become the centroids of the clusters formed by the atoms of $\mathbb{P}$ or the atoms of $\mathbb{Q}$ are moving to find the groups of atoms of $\mathbb{P}$ with the aim to minimize the distortion w.r.t. the metric d (see our *supplementary material* for more details).

3.2 ENTROPIC REGULARIZED DUALITY

To enable the application of optimal transport in machine learning and deep learning, Genevay et al. developed an entropic regularized dual form in Genevay et al. [2016]. First, they proposed to add an entropic regularization term to the primal form in (1)

$$\mathcal{W}^\epsilon_d(\mathbb{Q},\mathbb{P}) := \min_{\gamma\in\Gamma(\mathbb{Q},\mathbb{P})} \Big\{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[d(\mathbf{x},\mathbf{y})\right] + \epsilon D_{KL}\left(\gamma\|\mathbb{Q}\otimes\mathbb{P}\right)\Big\}, \quad (3)$$

where ϵ is the regularization rate, $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler (KL) divergence, and $\mathbb{Q}\otimes\mathbb{P}$ represents the specific coupling in which $\mathbb{Q}$ and $\mathbb{P}$ are independent. Note that when $\epsilon\to 0$, $\mathcal{W}^\epsilon_d(\mathbb{Q},\mathbb{P})$ approaches $\mathcal{W}_d(\mathbb{Q},\mathbb{P})$ and the optimal transport plan γ^*_ϵ of (3) also weakly converges to the optimal transport plan γ^* of (1). In practice, we set ϵ to be a small positive number, hence γ^*_ϵ is very close to γ^*.

Second, using the Fenchel-Rockafellar theorem, they obtained the following dual form w.r.t. the potential ϕ

$$\mathcal{W}^\epsilon_d(\mathbb{Q},\mathbb{P}) = \max_\phi \left\{\int \phi^c_\epsilon(\mathbf{x})\,\mathrm{d}\mathbb{Q}(\mathbf{x}) + \int \phi(\mathbf{y})\,\mathrm{d}\mathbb{P}(\mathbf{y})\right\} = \max_\phi \left\{\mathbb{E}_{\mathbb{Q}}\left[\phi^c_\epsilon(\mathbf{x})\right] + \mathbb{E}_{\mathbb{P}}\left[\phi(\mathbf{y})\right]\right\}, \quad (4)$$

where $\phi^c_\epsilon(\mathbf{x}) := -\epsilon\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left\{\frac{-d(\mathbf{x},\mathbf{y})+\phi(\mathbf{y})}{\epsilon}\right\}\right]\right)$.

4 THEORETICAL DEVELOPMENTS

4.1 PRELIMINARIES

We first examine a general supervised learning setting. Consider a hypothesis h in a hypothesis class $\mathcal{H}$ and a labeling function f (i.e., $f(\cdot)\in\mathcal{Y}_\triangle$ and $h(\cdot)\in\mathcal{Y}_\triangle$ where $\mathcal{Y}_\triangle := \left\{\boldsymbol{\pi}\in\mathbb{R}^M : \|\boldsymbol{\pi}\|_1 = 1 \text{ and } \boldsymbol{\pi}\geq\mathbf{0}\right\}$ with the number of classes M). Let $d_{\mathcal{Y}}$ be a metric or divergence over $\mathcal{Y}_\triangle$. We further define the general loss of the hypothesis h w.r.t. the data distribution $\mathbb{P}$ and the labeling function f as:

$$\mathcal{L}(h,f,\mathbb{P}) := \int d_{\mathcal{Y}}\left(h(\mathbf{x}), f(\mathbf{x})\right)\mathrm{d}\mathbb{P}(\mathbf{x}).$$

It is worth noting that by defining the metric or divergence $d_{\mathcal{Y}}$ as $d_{\mathcal{Y}}\left(h(\mathbf{x}), f(\mathbf{x})\right) := \sum_{i=1}^M f_i(\mathbf{x})\, D_{KL}\left(\mathbf{1}_i\|h(\mathbf{x})\right)$, where $\mathbf{1}_i$ is an one-hot vector, we can recover the cross-entropy loss widely used in deep learning.

Next we consider a domain adaptation setting [Ganin and Lempitsky, 2015, Courty et al., 2017a] in which we have a source space $\mathcal{X}^S$ endowed with a distribution $\mathbb{P}^S$ and a target space $\mathcal{X}^T$ endowed with a distribution $\mathbb{P}^T$. Given two pairs $\mathbf{z}_1 = \left(\mathbf{x}_1, y_1^\triangle\right)\in\mathcal{X}^S\times\mathcal{Y}_\triangle$ and $\mathbf{z}_2 = \left(\mathbf{x}_2, y_2^\triangle\right)\in\mathcal{X}^T\times\mathcal{Y}_\triangle$, we define the cost (distance) function between them as:

$$d(\mathbf{z}_1,\mathbf{z}_2) := \lambda d_{\mathcal{X}}(\mathbf{x}_1,\mathbf{x}_2) + d_{\mathcal{Y}}\left(y_1^\triangle, y_2^\triangle\right), \quad (5)$$

where $d_{\mathcal{X}}$ is a metric over $\mathcal{X}^S\times\mathcal{X}^T$ and $\lambda > 0$.

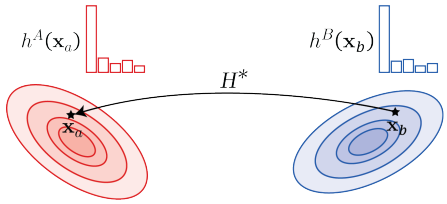

Figure 1: Imitation view explanation. $h^B(\mathbf{x}_b)$ for $\mathbf{x}_b \sim \mathbb{P}^B$ tries to imitate $h^A(\mathbf{x}_a)$ with $\mathbf{x}_a = H^*(\mathbf{x}_b) \sim \mathbb{P}^A$.

4.2 OPTIMAL TRANSPORT BASED IMITATION LEARNING

In what follows, we present the OT based imitation learning which lays foundation for our proposed MOST. Consider two data domains $\mathcal{X}^A$ and $\mathcal{X}^B$ with two data distributions $\mathbb{P}^A$ and $\mathbb{P}^B$ respectively, and assume that $h^A : \mathcal{X}^A \to \mathcal{Y}_\triangle$ is a well-qualified labeling function (classifier) that gives accurate prediction for data instances on $\mathcal{X}^A$ sampled from $\mathbb{P}^A$. We wish to learn a labeling function (classifier) h^B to predict accurately data instances sampled from $\mathbb{P}^B$ by imitating what is done by h^A on $\left(\mathcal{X}^A, \mathbb{P}^A\right)$. Based on the data distribution $\mathbb{P}^A$ and labeling function h^A, we define a distribution $\mathbb{P}_{A,h^A}$ over $\mathcal{X}^A \times \mathcal{Y}_\triangle$ including sample pair $\left(\mathbf{x}, h^A(\mathbf{x})\right)$ by first sampling $\mathbf{x} \sim \mathbb{P}^A$ and then computing $h^A(\mathbf{x})$. Similarly, we can define another distribution $\mathbb{P}_{B,h^B}$ over $\mathcal{X}^B \times \mathcal{Y}_\triangle$ using the data distribution $\mathbb{P}^B$ and the labeling function h^B. To allow h^B to imitate the behavior of h^A, we propose to inspect the Wasserstein distance (WS) between $\mathbb{P}_{A,h^A}$ and $\mathbb{P}_{B,h^B}$ w.r.t. the cost (metric) function d defined in (5). The following proposition is crucial for us to derive the fundamental mechanism of OT-based imitation learning.

Proposition 2. *The WS distance of interest* $\mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{B,h^B}\right)$ *can be expressed as:*

$$\min_{L:L_\#\mathbb{P}^A=\mathbb{P}^B} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^A} \left[\lambda d_\mathcal{X}\left(\mathbf{x}, L(\mathbf{x})\right) + d_\mathcal{Y}\left(h^A(\mathbf{x}), h^B\left(L(\mathbf{x})\right)\right)\right] = \min_{H:H_\#\mathbb{P}^B=\mathbb{P}^A} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^B} \left[\lambda d_\mathcal{X}\left(\mathbf{x}, H(\mathbf{x})\right) + d_\mathcal{Y}\left(h^B(\mathbf{x}), h^A\left(H(\mathbf{x})\right)\right)\right].$$

As indicated by Proposition 2, the optimal transport $H^* : H^*_\#\mathbb{P}^B = \mathbb{P}^A$ is the optimal mover that moves $\mathbb{P}^B$ to $\mathbb{P}^A$ so as to minimize the difference in the predictions of h^B for $\mathbf{x} \sim \mathbb{P}^B$ and h^A for $H^*(\mathbf{x}) \sim \mathbb{P}^A$. In other words, given $\mathbf{x} \sim \mathbb{P}^B$, the optimal transport H^* finds its closest counterpart in the space of $\mathcal{X}^A$ (i.e., $H^*(\mathbf{x})$) so that h^B can conveniently imitate the prediction of h^A on $H^*(\mathbf{x})$) for predicting $\mathbf{x}$ (see Figure 1).

To further elaborate on the proposed OT-based imitation learning, we assume that f^A is the ground-truth labeling function for the domain $\left(\mathcal{X}^A, \mathbb{P}^A\right)$ and theoretically prove that if we minimize the Wasserstein distance $\mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{A,f^A}\right)$, we can obtain the optimal solution $h^A_* = f^A$ and can upper-bound this Wasserstein distance by the general loss of h^A over $\left(\mathcal{X}^A, \mathbb{P}^A\right)$ (the statement (iii) in Theorem 3).

Theorem 3. *The following statements hold*

i) Given $\mathcal{X}^A = \mathcal{X}^B = \mathcal{X}$, $\mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{B,h^B}\right) = 0$ *if only if* $\mathbb{P}^A = \mathbb{P}^B$ *and* $h^A = h^B$.

ii) Consider the problem: $\min_{h_A} \mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{A,f^A}\right)$, *the optimal solution is* $h^A_* = f^A$ *obtained with the optimal mover* $L^* : L^*_\#\mathbb{P}^A = \mathbb{P}^A$ *to be the identity map.*

iii) $\mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{A,f^A}\right) \leq \mathcal{L}\left(h^A, f^A, \mathbb{P}^A\right)$.

iv) $\mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{B,h^B}\right) \geq \lambda \mathcal{W}_{d_\mathcal{X}}\left(\mathbb{P}^A, \mathbb{P}^B\right)$.

5 OUR PROPOSED METHOD

5.1 PROBLEM FORMULATION

In multi-source domain adaptation, we have K multiple source domains with the collected data and labels, and single target domain with the collected data only. We wish to transfer a model learned on the labeled source domains to the unlabeled target domain. Let us denote the collected data and labels for the source domains by $\mathcal{D}^S_k = \left\{\left(\boldsymbol{s}\mathbf{x}^k_i, y^k_i\right)\right\}_{i=1}^{N^S_k}$ with k is the index of a source domain, label $y^k_i \in \{1, 2, ..., M\}$ and collected data without labels for the target domain $\mathcal{D}^T = \{\boldsymbol{t}\mathbf{x}_i\}_{i=1}^{N^T}$.

For the sake of simplification, we denote the common space for source domains by $\mathcal{X}^S$. Note that if source domains have different input spaces, we can resize either input images or use appropriate transformations to map them to a common space. We further equip source domains with data distribution $\mathbb{P}^S_{1:K}$ whose density functions are $p^S_{1:K}(\mathbf{x})$. Let us denote the ground-truth labeling functions for source domains by $f^S_{1:K}(\cdot) \in \mathcal{Y}_\Delta$, implying that $p^S_k(y \mid \mathbf{x}) = f^S_k(\mathbf{x}, y)$ (i.e., $f^S_k(\mathbf{x}, y)$ represents the y-th value of $f^S_k(\mathbf{x})$). Therefore, the joint distribution to generate data instance $\mathbf{x}$ and categorical label $y \in \{1, ..., M\}$ is $p^S_k(\mathbf{x}, y) = p^S_k(\mathbf{x}) f^S_k(\mathbf{x}, y)$.

Regarding the target domain, we define its data space as $\mathcal{X}^T$, data distribution and density function as $\mathbb{P}^T$ and $p^T(\mathbf{x})$, respectively. We further define the ground-truth labeling function for the target domain by f^T which subsequently implies $p^T(y \mid \mathbf{x}) = f^T(\mathbf{x}, y)$ for a categorical label $y \in \{1, ..., M\}$.

Given a discrete distribution $\boldsymbol{\pi}$ over $\{1, ..., K\}$, we define $\mathbb{P}^S_{\boldsymbol{\pi}} := \sum_{k=1}^K \pi_k \mathbb{P}^S_k$ which is a mixture of $\mathbb{P}^S_{1:K}$. For a data instance $\mathbf{x} \sim \mathbb{P}^S_{\boldsymbol{\pi}}$ (i.e., we sample a hidden index $t \sim Cat(\boldsymbol{\pi})$ (i.e., the categorical distribution) and then sample $\mathbf{x} \sim \mathbb{P}^S_t$), we further define f^S as a labeling function such that $f^S(\mathbf{x})$ is identical to $f^S_t(\mathbf{x})$. By this definition, f^S can be viewed as the ground-truth labeling function over the mixture distribution $\mathbb{P}^S_{\boldsymbol{\pi}}$. Finally, the mixing proportion

$\boldsymbol{\pi}$ can be the uniform distribution $[\frac{1}{K}, ..., \frac{1}{K}]$ or proportional to the number of training examples in the source domains (i.e., $N^S_{1:K}$). It is worth noting that the mixing proportion $\boldsymbol{\pi}$ influences the proportion of samples from the individual data sources in the mini-batches. We conduct an ablation study to compare two aforementioned options for $\boldsymbol{\pi}$ and observe that they are comparable in terms of the predictive performances (see the *supplementary material*).

5.2 MULTI-SOURCE EXPERT TEACHER

Using the labeled source training sets $\mathcal{D}^S_{1:K}$, we can train qualified domain expert classifiers $h^S_{1:K}$ (i.e., $h^S_k(\mathbf{x}) \in \mathcal{Y}_\Delta$ represents the prediction probability of h^S_k for a data instance $\mathbf{x}$ in the k^{th} source domain) with good generalization capacity (e.g., $\mathcal{L}\left(h^S_k, f^S_k, \mathbb{P}^S_k\right) \leq \epsilon$ for some small $\epsilon > 0$). The next arising question is how to combine these domain experts to achieve a a multi-source expert teacher h^S that can work well on $\mathbb{P}^S_{\boldsymbol{\pi}}$ (i.e., $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) \leq \epsilon$). To this end, we leverage the weighted ensembling strategy in Mansour et al. [2009], Hoffman et al. [2018] to achieve

$$h^S(\mathbf{x}, y) = \sum_{k=1}^{K} \frac{\pi_k p^S_k(\mathbf{x}, y)}{\sum_{j=1}^{K} \pi_j p^S_j(\mathbf{x}, y)} h^S_k(\mathbf{x}, y), \quad (6)$$

where $y \in \{1, 2, ..., M\}$, and $h^S_k(\mathbf{x}, y)$ and $h^S(\mathbf{x}, y)$ specify the y-th values of $h^S_k(\mathbf{x})$ and $h^S(\mathbf{x})$ respectively.

The following theorem shows that the multi-source expert teacher h^S can work well on the mixture joint distribution $\mathbb{P}^S_{\boldsymbol{\pi}}$. More importantly, it works better than the worst domain expert on its source domain. Hence, if each domain expert is an ϵ-qualified classifier (i.e., $\mathcal{L}\left(h^S_k, f^S_k, \mathbb{P}^S_k\right) \leq \epsilon$), the multi-source expert teacher h^S is also an ϵ-qualified classifier (i.e., $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) \leq \epsilon$).

Theorem 4. *If $d_{\mathcal{Y}}$ can be decomposed as $d_{\mathcal{Y}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \sum_{i=1}^{M} \beta_i \ell(\alpha_i)$ where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{Y}_\Delta$ and ℓ is a convex function, the following statements hold true:*

i) $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) \leq \max_{1 \leq k \leq K} \mathcal{L}\left(h^S_k, f^S_k, \mathbb{P}^S_k\right)$.

ii) If each domain expert is an ϵ-qualified classifier (i.e., $\mathcal{L}\left(h^S_k, f^S_k, \mathbb{P}^S_k\right) \leq \epsilon$), the multi-source expert teacher h^S is also an ϵ-qualified classifier (i.e., $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) \leq \epsilon$).

In what follows, we present how to train the multi-source expert teacher h^S. Our workaround to train h^S comes from the following theoretical observation. Assume that we have K distributions $\mathbb{R}_{1:K}$ with density functions $r_{1:K}(\mathbf{z})$. We form a joint distribution $\mathcal{D}$ of a data instance $\mathbf{z}$ and label $t \in \{1, ..., K\}$ by sampling an index $t \sim \text{Cat}(\boldsymbol{\pi})$, sampling $\mathbf{x} \sim \mathbb{R}_t$, and collecting $(\mathbf{z}, t)$ as a sample from $\mathcal{D}$. With this setting, we have the following corollary.

Corollary 5. *If we train a source domain discriminator $\mathcal{C}$ to classify samples from the joint distribution $\mathcal{D}$ using the cross-entropy loss (i.e., $CE(\cdot, \cdot)$), the optimal source domain discriminator $\mathcal{C}^*$ defined as*

$$\mathcal{C}^* = \textit{argmin}_{\mathcal{C}} \mathbb{E}_{(\mathbf{z},t)\sim\mathcal{D}}\left[CE\left(\mathcal{C}(\mathbf{z}), t\right)\right]$$

satisfies $\mathcal{C}^*(\mathbf{z}) = \left[\frac{\pi_i r_i(\mathbf{z})}{\sum_j \pi_j r_j(\mathbf{z})}\right]_{i=1}^{K}$.

Corollary 5 suggests us a way to compute the weights of the domain experts in (6) in which for a given y, the distributions $p^S_{1:K}(\mathbf{x}, y)$ play roles of $r_{1:K}(\mathbf{z})$ where $\mathbf{z} = (\mathbf{x}, y)$. More specifically, for each $m \in \{1, ..., M\}$, we sample $t \sim Cat(\boldsymbol{\pi})$, then sample $(\mathbf{x}, y = m)$ from $p^S_t(\mathbf{x}, y = m)$, and train a source domain discriminator $\mathcal{C}_m(\mathbf{x}, y = m)$ (i.e., only consider $(\mathbf{x}, y)$ in which $\mathbf{x}$ has label $y = m$) to distinguish the source domain t of $(\mathbf{x}, y = m)$. We finally use $\mathcal{C}_m(\mathbf{x}, y = m)$ to estimate the weights of the domain experts. In addition, to conveniently train the source domain discriminators $\mathcal{C}_m$, we share their parameters, hence having an unique $\mathcal{C}$ that receives a pair $(\mathbf{x}, y)$ and predicts its source domain t. Therefore, in practice, we obtain the expert teacher in (6) as $h^S(\mathbf{x}, y) = \sum_{k=1}^{K} \mathcal{C}(\mathbf{x}, y, k) h^S_k(\mathbf{x}, y)$.

5.3 TARGET-DOMAIN IMITATING STUDENT

Inspired by the statement (ii) in Theorem 3, recall that f^T is the ground-truth labeling function and h^T is the classifier on the target domain, we propose to learn h^T on this domain to further minimize with the aim to obtain $h^T = f^T$:

$$\min_{h^T} \mathcal{W}_d\left(\mathbb{P}_{T,h^T}, \mathbb{P}_{T,f^T}\right).$$

To proceed our theory, we assume that $d_{\mathcal{Y}}$ is a metric over $\mathcal{Y}_\Delta$, which together with the metric $d_{\mathcal{X}}$ forms the metric d (cf. (5)), implying that $\mathcal{W}_d(\mathbb{P}_{.,.}, \mathbb{P}_{.,.})$ is a proper metric. We can thus bound the quantity of interest $\mathcal{W}_d\left(\mathbb{P}_{T,h^T}, \mathbb{P}_{T,f^T}\right)$:

$$\begin{aligned}\mathcal{W}_d\left(\mathbb{P}_{T,h^T}, \mathbb{P}_{T,f^T}\right) &\leq \mathcal{W}_d\left(\mathbb{P}_{T,h^T}, \mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}\right) \\ &+ \mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}, \mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}\right) + \mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}, \mathbb{P}_{T,f^T}\right) \overset{(1)}{\leq} \\ \mathcal{W}_d\left(\mathbb{P}_{T,h^T}, \mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}\right) &+ \mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}, \mathbb{P}_{T,f^T}\right), \end{aligned} \quad (7)$$

where $\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}$, a joint distribution over $\mathcal{X}^S \times \mathcal{Y}_\Delta$, consists of pairs $(\mathbf{x}, y_\Delta)$ in which $\mathbf{x} \sim \mathbb{P}^S_{\boldsymbol{\pi}}$ and $y_\Delta = f^S(\mathbf{x})$, h^S is a classifier on the mixture of source domains (i.e., $\mathbb{P}^S_{\boldsymbol{\pi}}$), and the definition of $\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}$ is similar to $\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}$ by changing the role of f^S to h^S. Note that we achieve the inequality (1) because $\mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}, \mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}\right)$ is upper-bounded by $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)$ (thanks to the statement (iii) in Theorem 3).

Moreover, $\mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}, \mathbb{P}_{T,f^T}\right)$ is a constant. Hence, to minimize the upper-bound in (7), we seek a classifier h^S working well on the mixture of source domains with a sufficiently small $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)$, while encouraging h^T to imitate h^S by minimizing $\mathcal{W}_d\left(\mathbb{P}_{T,h^T}, \mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}\right)$. To this end, we employ

the multi-source expert teacher h^S as in Section 5.2, which can operate well on $\mathbb{P}^S_{\boldsymbol{\pi}}$ as long as we can train good domain experts $h^S_{1:K}$, hence leading to the following optimization problem:

$$\min_{h^T}\left\{\mathcal{W}_d\left(\mathbb{P}_{T,h^T},\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}\right)+\mathcal{L}\left(h^S,f^S,\mathbb{P}^S_{\boldsymbol{\pi}}\right)\right\}. \quad (8)$$

The optimization problem in (8) is in line with the context of imitation learning for which the teacher classifier h^S has been trained effectively on the mixture source domain (i.e., $\mathbb{P}^S_{\boldsymbol{\pi}}$) and the student classifier h^T tries to imitate the teacher on the target domain. Specifically, Proposition 2 implies finding the optimal transport map H^*: $H^*_{\#}\mathbb{P}^T=\mathbb{P}^S_{\boldsymbol{\pi}}$ so that for any $\mathbf{x}\sim\mathbb{P}^T$, $h^T(\mathbf{x})$ should mimic the prediction of the expert teacher h^S over $H^*(\mathbf{x})\sim\mathbb{P}^S_{\boldsymbol{\pi}}$. This observation forms the foundation of our proposed MOST.

Proposition 2 further illustrates that among the transport maps H transporting $\mathbb{P}^T$ to $\mathbb{P}^S_{\boldsymbol{\pi}}$, we need to seek the map incurring the minimal label shift and enabling the student h^T easiest to imitate its teacher h^S. Inspired by the statement (iv) in Theorem 3 where $\mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S},\mathbb{P}_{T,h^T}\right)$ is lower-bounded by $\lambda\mathcal{W}_{d_{\mathcal{X}}}\left(\mathbb{P}^S_{\boldsymbol{\pi}},\mathbb{P}^T\right)$ (the discrepancy gap between the mixture of source distributions and the target one), to reduce the data shift, we propose to map both $\left(\mathcal{X}^S,\mathbb{P}^S_{\boldsymbol{\pi}}\right)$ and $\left(\mathcal{X}^T,\mathbb{P}^T\right)$ to a common joint space via two generators G^S and G^T and solve the following optimization problem:

$$\min_{h^T,G^T}\left\{\mathcal{L}\left(h^S\circ G^S,f^S,\mathbb{P}^S_{\pi}\right)+\mathcal{W}_d\left(\mathbb{Q}_{T,h^T},\mathbb{Q}^{\pi}_{S,h^S}\right)\right\}, \quad (9)$$

where $\mathbb{Q}_{T,h^T}$ is similar to $\mathbb{P}_{T,h^T}$ but on the joint space and consists of the pairs $\left(G^T(\mathbf{x}),h^T\left(G^T(\mathbf{x})\right)\right)$ for $\mathbf{x}\sim\mathbb{P}^T$ and $\mathbb{Q}^{\boldsymbol{\pi}}_{S,h^S}$ is similar to $\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}$ but on the joint space and consists of the pairs $\left(G^S(\mathbf{x}),h^S\left(G^S(\mathbf{x})\right)\right)$ for $\mathbf{x}\sim\mathbb{P}^S_{\boldsymbol{\pi}}$. Note that both h^S and $h^S_{1:K}$ now act on $G^S(\cdot)$.

Theorem 6. *Let $h^S_*\circ G^S_*$ be the optimal teacher and h^T_*,G^T_* be the optimal solutions of the optimization problem in (9). Assume that G^T,h^T are in the families having infinite capacity (i.e., those can approximate any continuous function up to any level of precision, e.g., neural nets), we have*[2]

$$\min_{h^T,G^T}\mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T},\mathbb{P}^{G^T}_{T,f^{G^T}_T}\right)\leq\mathcal{L}\left(h^S_*\circ G^S_*,f^S,\mathbb{P}^S_{\pi}\right)+\mathcal{W}_d\left(\mathbb{P}^{G^S_*}_{S,f^S_*},\mathbb{P}^{G^T_*}_{T,f^T_*}\right), \quad (10)$$

where $f^S_*:=f^{G^S_*}_S$ and $f^T_*:=f^{G^T_*}_T$.

In Theorem 6, $\mathbb{P}^{G^T}_{T,h^T}$ is the distribution consisting of samples of pairs $\left(G^T(\mathbf{x}),h^T\left(G^T(\mathbf{x})\right)\right)$ where $\mathbf{x}\sim\mathbb{P}^T$and same definition for other similar distributions. Theorem 6 demonstrates that our MOST with the support of the generators and the joint space can mitigate data and label shifts as $\mathcal{W}_d\left(\mathbb{P}^{G^S_*}_{S,f^S_*},\mathbb{P}^{G^T_*}_{T,f^T_*}\right)$ is the natural shift between two ground-truth labeling functions f^S and f^T in the joint space.

[2]We define f^G as the induced labeling function over the joint space such that f^G predicts $G(\mathbf{x})$ as same as f predicts $\mathbf{x}$.

5.4 TRAINING PROCESS OF MOST

5.4.1 Training Multi-Source Expert Teacher

To work out the multi-source expert teacher h^S, we simultaneously train domain experts $h^S_{1:K}$ on the labeled training sets $\mathcal{D}^S_{1:K}$ and the source domain discriminator $\mathcal{C}$ to offer the weights of the domain experts. Basically, we minimize: $\sum_{k=1}^K\mathcal{L}^{de}_k+\mathcal{L}^{\mathcal{C}}$, where we define

$$\mathcal{L}^{de}_k=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^s_k}\left[CE\left(h^S_k\left(G^S(\mathbf{x})\right),y\right)\right],$$
$$\mathcal{L}^{\mathcal{C}}=\mathbb{E}_{(\mathbf{x},y,t)\sim\mathcal{D}}\left[CE\left(\mathcal{C}(\mathbf{x},y),t\right)\right]$$

with $\mathcal{D}$ is formed by sampling $t\sim Cat(\boldsymbol{\pi})$ and $(\mathbf{x},y)\sim\mathcal{D}^S_t$ and $CE(\cdot,\cdot)$ is the cross-entropy loss.

5.4.2 Training Target-Domain Imitating Student

We use the *entropic regularized dual form* in (4) to solve the optimization problem of interest in (9) by minimizing $\mathcal{W}^{\epsilon}_d\left(\mathbb{Q}_{T,h^T},\mathbb{Q}^{\boldsymbol{\pi}}_{S,h^S}\right)$, hence arriving at the following optimization problem:

$$\min_{h^T,G^T}\mathcal{L}^{WS}=\min_{h^T,G^T}\max_{\phi}\left\{\mathbb{E}_{\mathbb{P}^T}\left[-\epsilon\log\left(\mathbb{E}_{\mathbb{P}^S_{\boldsymbol{\pi}}}\left[\exp\left\{\frac{1}{\epsilon}\gamma\left(\mathbf{x}^S,\mathbf{x}^T\right)\right\}\right]\right)\right]+\mathbb{E}_{\mathbb{P}^S_{\boldsymbol{\pi}}}\left[\phi\left(G^S\left(\mathbf{x}^S\right)\right)\right]\right\}$$

, where

$$\gamma\left(\mathbf{x}^S,\mathbf{x}^T\right)=\phi\left(G^S\left(\mathbf{x}^S\right)\right)-d\left(G^S\left(\mathbf{x}^S\right),G^T\left(\mathbf{x}^T\right)\right)$$

, ϕ is a neural net named Kantorovich potential network and

$$d\left(G^S\left(\mathbf{x}^S\right),G^T\left(\mathbf{x}^T\right)\right)=d_{\mathcal{Y}}\Big(h^T\left(G^T\left(\mathbf{x}^T\right)\right),h^S\left(G^S\left(\mathbf{x}^S\right)\right)\Big)+\lambda\left\|G^T\left(\mathbf{x}^T\right)-G^S\left(\mathbf{x}^S\right)\right\|$$

, while $\mathbf{x}^T\sim\mathbb{P}^T,\mathbf{x}^S\sim\mathbb{P}^S_{\boldsymbol{\pi}}$.

Clustering view explanation of the WS distance term. More specifically, according to the cluster view of optimal transport $\mathcal{W}^{\epsilon}_d\left(\mathbb{Q}_{T,h^T},\mathbb{Q}^{\boldsymbol{\pi}}_{S,h^S}\right)$, at the optimal solution, each $G^T\left(\mathbf{x}^T\right)$ finds a cluster of $G^S\left(\mathbf{x}^S\right)$ (s) to minimize the distortion w.r.t. the metric $d\left(G^T\left(\mathbf{x}^T\right),G^S\left(\mathbf{x}^S\right)\right)$ defined as

$$d_{\mathcal{Y}}\left(h^T\left(G^T\left(\mathbf{x}^T\right)\right),h^S\left(G^S\left(\mathbf{x}^S\right)\right)\right)+\lambda\left\|G^T\left(\mathbf{x}^T\right)-G^S\left(\mathbf{x}^S\right)\right\|$$

, which further implies that $G^T\left(\mathbf{x}^T\right)$ should move closely to a cluster of $G^S\left(\mathbf{x}^S\right)$(s) with the same predicted label

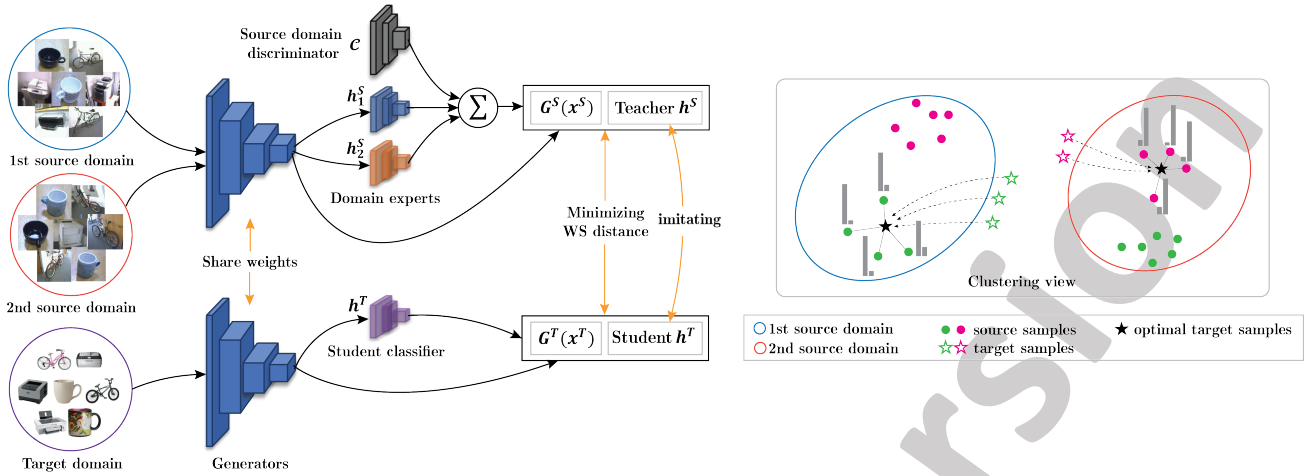

Figure 2: Left: The overall structure of our proposed method for multi-source domain adaptation. MOST consists of two cooperative agents: an expert teacher h^S, a weighted combination of domain experts and a student h^T that tries to imitate the prediction of the teacher via the OT-based imitation learning. Right: Clustering view explanation of the WS distance term.

regarding h^S so as to imitate the prediction of h^S (i.e., $\min d_{\mathcal{Y}}\left(h^T\left(G^T\left(\mathbf{x}^T\right)\right), h^S\left(G^S\left(\mathbf{x}^S\right)\right)\right)$). This certainly helps to mitigate the label shift problem (see Figure 2).

The teacher h^S also offers pseudo labels on source and target examples for the student h^T to imitate, hence we minimize:

$$\mathcal{L}^{pl} = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^S_{\boldsymbol{\pi}},\mathbb{P}^T}\left[CE\left(h^S\left(G^S\left(\mathbf{x}\right)\right), h^T\left(G^T\left(\mathbf{x}\right)\right)\right)\right].$$

Virtual adversarial training (VAT) [Miyato et al., 2019] in conjunction with minimizing entropy of prediction [Grandvalet and Bengio, 2005] with the aim to ensuring the clustering assumption [Chapelle and Zien, 2005] has been applied successfully to UDA [Shu et al., 2018, Kumar et al., 2018]. Inspired by this success, we propose to minimize:

$$\mathcal{L}^{clus} = \mathcal{L}^{ent} + \mathcal{L}^{vat},$$

where

$$\mathcal{L}^{ent} = \mathbb{E}_{\mathbb{P}^T}\left[\mathbb{H}\left(h^T\left(G^T\left(\mathbf{x}\right)\right)\right)\right]$$

, $\mathbb{H}$ is the entropy and

$$\mathcal{L}^{vat} = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^T}$$
$$\left[\max_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|<\theta} D_{KL}\left(h^T\left(G^T\left(\mathbf{x}\right)\right), h^T\left(G^T\left(\mathbf{x}'\right)\right)\right)\right]$$

with which D_{KL} represents a Kullback-Leibler divergence and θ is very small positive number.

5.4.3 Simultaneous Training of Student and Teacher

We have two scenarios to train our teacher and student paradigm: (i) sequential training and (ii) simultaneous training of teacher and student. As suggested by the ablation study (see Section 6.4.1), we follow the strategy of simultaneous training of teacher and student in which we minimize:

$$\mathcal{L} = \sum_{k=1}^{K} \mathcal{L}_k^{de} + \mathcal{L}^{\mathcal{C}} + \alpha\mathcal{L}^{WS} + \beta\mathcal{L}^{pl} + \gamma\mathcal{L}^{clus}, \quad (11)$$

where $\alpha, \beta, \gamma > 0$ are trade-off parameters.

We note that the loss $\mathcal{L}^{WS}$ has the form of maximizing over ϕ which is parameterized by a neural net. In training MOST, we update ϕ several times for each mini-batch of data. Due to the effect of the envelope theorem, the term $\mathcal{L}^{WS}$ (hence the total loss $\mathcal{L}$) smoothly decreases (see Figure 3). Finally, we present the overview of our approach in Figure 2.

6 EXPERIMENTS

6.1 MODEL EVALUATION

We evaluate our proposed MOST on several commonly-used benchmark domain adaptation datasets.

Digits-five [Peng et al., 2019] consists of five-digit datasets: MNIST (**mt**), MNIST-M (**mm**), USPS (**up**), SVHN (**sv**), Synthetic Digits (**sy**). There are 10 classes corresponding to digits ranging from 0 to 9 in each domain.

Office-Caltech10 [Gong et al., 2012] is categorized in four different domains: Amazon (**A**), Caltech (**C**), DSLR (**D**), and Webcam (**W**) with 10 common classes and 2533 images in total.

Office-31 [Saenko et al., 2010] contains 4,110 images with 31 classes, and is categorized into three domains: Amazon (**A**), DSLR (**D**), and Webcam (**W**).

The details of the data preparation and preprocessing of all datasets are described in our *supplementary material*. Similar to Wang et al. [2020], we compare our MOST with the MSDA standards: (1) *Single Best*: the best classification accuracy on the test set among single-source transfer results; (2) *Source Combine*: the evaluation on single-source domain adaptation whereas the single-source is combined by all source data; (3) *Multi-Source*: results on the adaptation from multiple source domains to the target domain.

Table 1: Classification results with mean and standard deviation on Digits-five.

Standards	Methods	→**mm**	→**mt**	→**up**	→**sv**	→**sy**	Avg
Single Best	Source-only	59.2±0.6	97.2±0.6	84.7±0.8	77.7±0.8	85.2±0.6	80.8
	DAN [Long et al., 2015]	63.8±0.7	96.3±0.5	94.2±0.9	62.5±0.7	85.4±0.8	80.4
	CORAL [Sun and Saenko, 2016]	62.5±0.7	97.2±0.8	93.5±0.8	64.4±0.7	82.8±0.7	80.1
	DANN [Ganin et al., 2016]	71.3±0.6	97.6±0.8	92.3±0.9	63.5±0.8	85.4±0.8	82.0
	ADDA [Tzeng et al., 2017]	71.6±0.5	97.9±0.8	92.8±0.7	75.5±0.5	86.5±0.6	84.8
Source Combine	Source-only	63.4±0.7	90.5±0.8	88.7±0.9	63.5±0.9	82.4±0.6	77.7
	DAN [Long et al., 2015]	67.9±0.8	97.5±0.6	93.5±0.8	67.8±0.6	86.9±0.5	82.7
	DANN [Ganin et al., 2016]	70.8±0.8	97.9±0.7	93.5±0.8	68.5±0.5	87.4±0.9	83.6
	JAN [Long et al., 2017]	65.9±0.7	97.2±0.7	95.4±0.8	75.3±0.7	86.6±0.6	84.1
	ADDA [Tzeng et al., 2017]	72.3±0.7	97.9±0.6	93.1±0.8	75.0±0.8	86.7±0.6	85.0
	MCD [Saito et al., 2018]	72.5±0.7	96.2±0.8	95.3±0.7	78.9±0.8	87.5±0.7	86.1
Multi-Source	MDAN [Zhao et al., 2018]	69.5±0.3	98.0±0.9	92.4±0.7	69.2±0.6	87.4±0.5	83.3
	DCTN [Xu et al., 2018]	70.5±1.2	96.2±0.8	92.8±0.3	77.6±0.4	86.8±0.8	84.8
	M^{3}SDA [Peng et al., 2019]	72.8±1.1	98.4±0.7	96.1±0.8	81.3±0.9	89.6±0.6	87.7
	MDDA [Zhao et al., 2020]	78.6±0.6	98.8±0.4	93.9±0.5	79.3±0.8	89.7±0.7	88.1
	LtC-MSDA [Wang et al., 2020]	85.6±0.8	99.0±0.4	98.3±0.4	83.2±0.6	93.0±0.5	91.8
	MOST (ours)	**91.5**±1.7	**99.6**±0.0	**98.4**± 0.0	**90.9**±0.6	**96.4**±2.7	**95.4**

Table 2: Classification accuracy (%) on Office-Caltech10 using pretrained ResNet-101.

Standards	Methods	→**W**	→**D**	→**C**	→**A**	Avg
Source Combine	Source-only	99.0	98.3	87.8	86.1	92.8
	DAN [Long et al., 2015]	99.3	98.2	89.7	94.8	95.5
Multi-Source	Source-only	99.1	98.2	85.4	88.7	92.9
	DAN [Long et al., 2015]	99.5	99.1	89.2	91.6	94.8
	DCTN [Xu et al., 2018]	99.4	99.0	90.2	92.7	95.3
	JAN [Long et al., 2017]	99.4	99.4	91.2	91.8	95.5
	MEDA [Wang et al., 2018]	99.3	99.2	91.4	92.9	95.7
	MCD [Saito et al., 2018]	99.5	99.1	91.5	92.1	95.6
	M^{3}SDA [Peng et al., 2019]	99.5	99.2	92.2	94.5	96.4
	MOST (ours)	**100**	**100**	**96.0**	**96.4**	**98.1**

Table 3: Classification accuracy (%) on Office-31 using pretrained AlexNet.

Standards	Methods	→**D**	→**W**	→**A**	Avg
Single Best	Source-only	99.0	95.3	50.2	81.5
	RevGrad [Ganin and Lempitsky, 2015]	99.2	96.4	53.4	83.0
	DAN [Long et al., 2015]	99.0	96.0	54.0	83.0
	RTN [Long et al., 2016]	99.6	96.8	51.0	82.5
	ADDA [Tzeng et al., 2017]	99.4	95.3	54.6	83.1
Source Combine	Source-only	97.1	92.0	51.6	80.2
	DAN [Long et al., 2015]	98.8	96.2	54.9	83.3
	RTN [Long et al., 2016]	99.2	95.8	53.4	82.8
	JAN [Long et al., 2017]	99.4	95.9	54.6	83.3
	ADDA [Tzeng et al., 2017]	99.2	96.0	55.9	83.7
	MCD [Saito et al., 2018]	99.5	96.2	54.4	83.4
Multi-Source	MDAN [Zhao et al., 2018]	99.2	95.4	55.2	83.3
	DCTN [Xu et al., 2018]	99.6	96.9	54.9	83.8
	M^{3}SDA [Peng et al., 2019]	99.4	96.2	55.4	83.7
	MDDA [Zhao et al., 2020]	99.2	97.1	56.2	84.2
	LtC-MSDA [Wang et al., 2020]	99.6	97.2	56.9	84.6
	MOST (ours)	**100**	**98.7**	**60.6**	**86.4**

Table 4: Results (%) on different training strategies.

Methods	→**mm**	→**mt**	→**up**	→**sv**	→**sy**	Avg
Two-phase training	89.7	**99.6**	98.2	**92.0**	97.7	95.4
Simultaneous training	**93.4**	**99.6**	**98.4**	90.9	**97.8**	**96.0**

6.2 ARCHITECTURE/HYPERPARAMETERS

We follow the training paradigms in Saito et al. [2018], Peng et al. [2019] where G^S are shared weights with G^T. All the experiments on *Office-Caltech10* and *Office-31* are based on pre-trained ResNet-101 [He et al., 2016] and AlexNet [Krizhevsky et al., 2012], respectively. The network architecture and hyperparameter settings are presented in the *supplementary material*.

6.3 RESULTS

We first compare MOST with recent state-of-the-art works on *Digits-five* whose results are reported in Table 1. Our MOST surpasses all transfer tasks, with a sizable margin especially in the following adaptation tasks: "→**mm**", "→**sv**", and "→**sy**". Overall, our proposed method achieves a high average accuracy of 96.0%, which is a 4.2% increase compared to LtC-MSDA [Wang et al., 2020].

The experimental results on *Office-Caltech10* are shown in Table 2. Compared to the baselines, our MOST obtains impressive scores on all the settings: 100% on the adaptation tasks from corresponding source domains to **W** and **D**, and significant improvements on "→**C**", and "→**A**" tasks. As a result, our proposed method experiences a rise of 1.7% on average compared to the runner-up method M^{3}SDA [Peng et al., 2019].

Finally, we report the performance on *Office-31* and compare results in Table 3. MOST continues to perform the best with 1.8% improvement on average over the second. Additionally, on the challenging task "→**A**", MOST significantly surpasses the state-of-the-art method by 3.7%.

6.4 ABLATION STUDY

6.4.1 Training Strategy

We consider two training strategies for MOST, which are two-phase training and simultaneous training. In the former, we train a perfect teacher and then train a student to imitate it, while in the latter, we train all in once with the loss in (11). Table 4 shows that simultaneous training is more effective with an improvement of 0.6% on the average accuracy. We hence stick to this strategy for our main experiments.

6.4.2 Effect of Losses

We investigate the effectiveness of the component losses in (11) w.r.t. the source domain (a.k.a. *source only* setting), i.e., $\mathcal{L}_k^{de} + \mathcal{L}^{\mathcal{C}}$, and w.r.t. the target domain to perform domain adaptation, i.e., $\mathcal{L}^{pl}$, $\mathcal{L}^{WS}$, $\mathcal{L}^{ent}$ and $\mathcal{L}^{vat}$. The average results reported in Table 5 show that by only incorporating

Table 5: Average accuracies (%) on Digits-five and Office-Caltech10 datasets with different settings.

$\mathcal{L}_k^{de} + \mathcal{L}^{\mathcal{C}}$	$\mathcal{L}^{pl}$	$\mathcal{L}^{WS}$	$\mathcal{L}^{ent}$	$\mathcal{L}^{vat}$	Digits-five	Office-Caltech10
✓					82.6	95.8
✓	✓				89.9	96.7
✓	✓	✓			94.2	96.9
✓	✓	✓	✓		93.8	97.9
✓	✓	✓		✓	94.8	97.4
✓	✓	✓	✓	✓	**96.0**	**98.1**

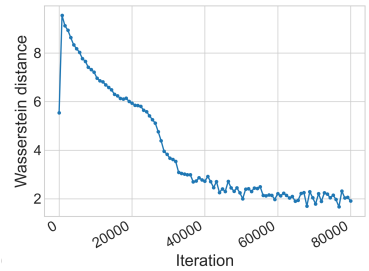

Figure 3: Values of $\mathcal{W}_d^{\epsilon}\left(\mathbb{Q}_{T,h^T}, \mathbb{Q}_{S,h^S}^{\boldsymbol{\pi}}\right)$ during training.

2 losses $\mathcal{L}^{pl} + \mathcal{L}^{WS}$ (fourth row) to align the target samples to source samples, MOST already achieves the state-of-the-art results (94.2% on *Digits-five* and 96.9% on *Office-Caltech10*) compared to the runner-up baselines (91.8% on *Digits-five* and 96.4% on *Office-Caltech10*). While the performance is improved further with the help of clustering assumption (the last row).

6.4.3 Wasserstein Distance

We further observe the values of the WS distance between $\mathbb{Q}_{T,h^T}$ and $\mathbb{Q}_{S,h^S}^{\boldsymbol{\pi}}$ in (3) on "→**mm**" task during training. As shown in Figure 3, the WS values tend to go down, which signals the decline of the data shift and label shift between the two domains.

7 CONCLUSION

In this paper, inspired by the principle of imitation learning and the theory of optimal transport, we propose Multi-Source Domain Adaptation via Optimal Transport for Student-Teacher Learning (MOST). Via rigorous theoretical guarantees, we introduce a model with two fundamental components: a teacher and a student for multi-source domain adaptation to actualize the cross-domain imitation capability. Comprehensive experiments demonstrate that MOST outperforms the state-of-the-art methods on several benchmark domain adaptation datasets.

Acknowledgements

This work was supported by the US Air Force grant FA2386-19-1-4040.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 2010.

K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *AISTATS*, 2005.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases*, 2014.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NIPS*, 2017a.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE TPAMI*, 2017b.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. In *NIPS*. 2007.

B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, 2018.

L. Duan, D. Xu, and S. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, 2012.

Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *NIPS*. 2016.

M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, 2016.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*. 2005.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *NIPS*, 2016.

N. Ho, X. L. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via wasserstein means. In *ICML*, 2017.

N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. In *AISTATS*, 2019.

J. Hoffman, M. Mohri, and N. Zhang. Algorithms and theory for multiple-source adaptation. In *NeurIPS*. Curran Associates, Inc., 2018.

A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, B. Freeman, and G. Wornell. Co-regularized alignment for unsupervised domain adaptation. In *NeurIPS*. 2018.

T. Le, T. Nguyen, N. Ho, H. Bui, and D. Phung. Lamda: Label matching deep domain adaptation. In *ICML*, 2021.

C. Lee, T. Batra, M. H. Baig, and D. Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, 2019.

M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*. 2016.

M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*. 2018.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*. 2009.

T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 2019.

T. Nguyen, T. Le, H. Zhao, H. Q. Tran, T. Nguyen, and D. Phung. Tidot: A teacher imitation learning approach for domain adaptation with optimal transport. In *IJCAI*, 2021.

V. Nguyen, T. Le, T. Le, K. Nguyen, O. De Vel, P. Montague, L. Qu, and D. Phung. Deep domain adaptation for vulnerable code function identification. In *IJCNN*, 2019.

V. Nguyen, T. Le, O. De Vel, P. Montague, J. Grundy, and D. Phung. Dual-component deep domain adaptation: A new approach for cross project software vulnerability detection. In *PAKDD*, 2020.

Y. Pan, T. Yao, Y. Li, Y. Wang, C. Ngo, and T. Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, 2019.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 2019.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

I. Redko, N. Courty, R. Flamary, and D. Tuia. Optimal transport for multi-source domain adaptation under target shift. In *AISTATS*, 2019.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 2015.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

S. Sankaranarayanan, Y. Balaji, Carlos D. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *CVPR*, 2018.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser*, 2015.

E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 2017.

R. Shu, H. H. Bui, H. Narui, and S. Ermon. A DIRT-t approach to unsupervised domain adaptation. In *ICLR*, 2018.

B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

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

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

H. Wang, M. Xu, B. Ni, and W. Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *ECCV*, 2020.

J. Wang, Wenjie Feng, Y. Chen, H. Yu, M. Huang, and Philip S. Yu. Visual domain adaptation with manifold embedded distribution alignment. *ACM international conference on Multimedia*, 2018.

S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, 2018.

Y. Xie, M. Chen, H. Jiang, T. Zhao, and H. Zha. On scalable and efficient computation of large scale optimal transport. In *ICML*, 2019.

M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, 2020a.

M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020b.

R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*, 2018.

R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *CVPR*, 2020c.

H. Yan, Yukang Ding, P. Li, Qilong Wang, Yong Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. *CVPR*, 2017.

Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, 2018.

H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J Gordon. Adversarial multiple source domain adaptation. In *NeurIPS*. 2018.

S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, and K. Keutzer. Multi-source distilling domain adaptation. In *AAAI*, 2020.