MOST: Multi-Source Domain Adaptation via Optimal Transport for Student-Teacher Learning (Supplementary Material)

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

This is the supplementary material for "MOST: Multi-Source Domain Adaptation via Optimal Transport for Student-Teacher Learning". Our supplementary material is organized into four main sections. In the first and second sections, we provide complete proofs for our theory. The third section is dedicated to the clustering view of optimal transport with a link to our imitation learning view, whilst the detail of network architecture, implementation specification, and further ablation studies are presented in the fourth section.

1 THEORETICAL DEVELOPMENTS

In what follows, we present all proofs of our theory. Given a pair of 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. We now investigate the Wasserstein distance between $\mathbb{P}_{A,h^A}$ and $\mathbb{P}_{B,h^B}$ w.r.t the cost (metric) function $d = \lambda d_\mathcal{X} + d_\mathcal{Y}$. Proposition 1 is crucial for us to develop an imitation learning view based on optimal transport.

Proposition 1. *The WS distance of interest can be expressed as:*

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

Proof. Observe first that for any $U_A \subset \mathcal{X}^A \times \mathcal{Y}_\triangle$, we have $\mathbb{P}_{A,h^A}(U_A) = \mathbb{P}^A(V_A)$ where $V_A := \left\{\mathbf{x} \in \mathcal{X}^A \mid \left(\mathbf{x}, h^A(\mathbf{x})\right) \in U_A\right\}$. Similarly, we have for any $U_B \subset \mathcal{X}^B \times \mathcal{Y}_\triangle$ that $\mathbb{P}_{B,h^B}(U_B) = \mathbb{P}^B(V_B)$ where $V_B := \left\{\mathbf{x} \in \mathcal{X}^B \mid \left(\mathbf{x}, h^B(\mathbf{x})\right) \in U_B\right\}$.

Let $K : \text{supp}\left(\mathbb{P}_{A,h^A}\right) \to \text{supp}\left(\mathbb{P}_{B,h^B}\right)$ (i.e., supp indicates the support of a distribution) be such that $K_\#\mathbb{P}_{A,h^A} = \mathbb{P}_{B,h^B}$. We can express K as

$$K\left(\mathbf{x}, h^A(\mathbf{x})\right) := \left(K_1\left(\mathbf{x}, h^A(\mathbf{x})\right), K_2\left(\mathbf{x}, h^A(\mathbf{x})\right)\right),$$

with $K_1\left(\mathbf{x}, h^A(\mathbf{x})\right) \in \mathcal{X}^B$ and $K_2\left(\mathbf{x}, h^A(\mathbf{x})\right) \in \mathcal{Y}_\triangle$. Define $L(\mathbf{x}) := K_1\left(\mathbf{x}, h^A(\mathbf{x})\right)$. We claim that $L_\#\mathbb{P}^A = \mathbb{P}^B$. Indeed, let $V_B \subset \mathcal{X}^B$ be any measurable set and take $U_B := V_B \times \mathcal{Y}_\triangle$. Then by using the observation above and the fact $K_\#\mathbb{P}_{A,h^A} = \mathbb{P}_{B,h^B}$, we obtain

$$\mathbb{P}^B(V_B) = \mathbb{P}_{B,h_B}(U_B) = \mathbb{P}_{A,h^A}\left(K^{-1}(U_B)\right) = \mathbb{P}_{A,h^A}\left(L^{-1}(V_B) \times \mathcal{Y}_\triangle\right) = \mathbb{P}_A\left(L^{-1}(V_B)\right).$$

Thus the claim is proved.

It also follows from $K_\#\mathbb{P}_{A,h^A} = \mathbb{P}_{B,h^B}$ and the claim that $K_2\left(\mathbf{x}, h_A(\mathbf{x})\right) = h_B\left(L(\mathbf{x})\right)$, which gives

$$d\left(\left(\mathbf{x}, h^A(\mathbf{x})\right), K\left(\mathbf{x}, h^A(\mathbf{x})\right)\right) = \lambda d_\mathcal{X}\left(\mathbf{x}, L(\mathbf{x})\right) + d_\mathcal{Y}\left(h^A(\mathbf{x}), h^B(L(\mathbf{x}))\right). \tag{1}$$

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Therefore, we deduce that

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

In order to prove the reverse inequality, let us consider any map L satisfying $L_{\#}\mathbb{P}^A=\mathbb{P}^B$. Define $K\left(\mathbf{x},h^A\left(\mathbf{x}\right)\right)\coloneqq\left(L(\mathbf{x}),h^B\left(L\left(\mathbf{x}\right)\right)\right)$. Then (1) holds and $K_{\#}\mathbb{P}_{A,h^A}=\mathbb{P}_{B,h^B}$. To verify the latter, let $U_B\subset\mathcal{X}^B\times\mathcal{Y}_{\triangle}$ be any measurable set and take $V_B\coloneqq\left\{\mathbf{x}\in\mathcal{X}^B\mid\left(\mathbf{x},h^B\left(\mathbf{x}\right)\right)\in U_B\right\}$. Then as

$$K^{-1}\left(U_B\right)=\left\{\left(\mathbf{x},h^A\left(\mathbf{x}\right)\right)\mid L\left(\mathbf{x}\right)\in V_B\right\}=\left\{\left(\mathbf{x},h^A\left(\mathbf{x}\right)\right)\mid\mathbf{x}\in L^{-1}\left(V_B\right)\right\},$$

we have

$$\mathbb{P}_{A,h^A}\left(K^{-1}\left(U_B\right)\right)=\mathbb{P}^A\left(L^{-1}\left(V_B\right)\right)=\mathbb{P}^B\left(V_B\right)=\mathbb{P}_{B,h^B}\left(U_B\right).$$

Thus it follows that

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

By combining the above two inequalities, we obtain the equality

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

Symmetrically, we achieve the other equality. □

Theorem 2. *The following statements hold*

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

ii) 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* $h^A=h^B$.

iii) 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.*

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

Proof. i) Using Proposition 1, we have

$$\begin{aligned}\mathcal{W}_d\left(\mathbb{P}_{A,h^A},\mathbb{P}_{B,h^B}\right)&=\min_{L:L_{\#}\mathbb{P}^A=\mathbb{P}^B}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^A}\left[\lambda d_{\mathcal{X}}\left(\mathbf{x},L\left(\mathbf{x}\right)\right)+d_{\mathcal{Y}}\left(h^A\left(\mathbf{x}\right),h^B\left(L\left(\mathbf{x}\right)\right)\right)\right]\\&\geq\min_{L:L_{\#}\mathbb{P}^A=\mathbb{P}^B}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^A}\left[\lambda d_{\mathcal{X}}\left(\mathbf{x},L\left(\mathbf{x}\right)\right)\right]=\lambda\mathcal{W}_{d_{\mathcal{X}}}\left(\mathbb{P}^A,\mathbb{P}^B\right).\end{aligned}$$

ii) It is obvious that if $\mathbb{P}^A=\mathbb{P}^B$ and $h^A=h^B$, then $\mathcal{W}_d\left(\mathbb{P}_{A,h^A},\mathbb{P}_{B,h^B}\right)=0$. For the converse, note that

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

by statement (i). Thus, $\mathcal{W}_d\left(\mathbb{P}_{A,h^A},\mathbb{P}_{B,h^B}\right)=0$ implies that $\mathcal{W}_{d_{\mathcal{X}}}\left(\mathbb{P}^A,\mathbb{P}^B\right)=0$ and hence $\mathbb{P}^A=\mathbb{P}^B$. Next, let L^* be the optimal transport:

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

Then thanks to Proposition 1, we have

$$\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^A}\left[\lambda d_{\mathcal{X}}\left(\mathbf{x},L^*\left(\mathbf{x}\right)\right)+d_{\mathcal{Y}}\left(h^A\left(\mathbf{x}\right),h^B\left(L^*\left(\mathbf{x}\right)\right)\right)\right]=0$$

which yield $L^*\left(\mathbf{x}\right)=\mathbf{x}$ and hence $h^A\left(\mathbf{x}\right)=h^B\left(\mathbf{x}\right)$ almost everywhere w.r.t the measure $\mathbb{P}^A$.

iii) We have

$$\min_{h_A} \mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{A,f^A}\right) = 0,$$

and the minimum value is attained at $h_*^A = f^A$ and $L^* = id$. From ii, this optimal solution is unique.

iv) Using Proposition 1, we have:

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

Then by choosing L as the identity map (i.e., $L(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x}$), we obtain:

$$\mathcal{W}_d\left(\mathbb{P}_{A,h^A}, \mathbb{P}_{A,f^A}\right) \leq \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^A}\left[d_{\mathcal{Y}}\left(h^A\left(\mathbf{x}\right), f^A\left(\mathbf{x}\right)\right)\right] = \mathcal{L}\left(h^A, f^A, \mathbb{P}^A\right).$$

□

2 OUR PROPOSED METHOD

2.1 PROBLEM FORMULATION

In multi-source domain adaptation, we have K multiple source domains with collected data and labels, and single target domain with only collected data. We wish to transfer a model learned on labeled source domains to an unlabeled target domain. Let us denote the collected data and labels for the source domains by $\mathcal{D}_k^S = \left\{\left(\boldsymbol{s}\mathbf{x}_i^k, y_i^k\right)\right\}_{i=1}^{N_k^S}$ with label $y_i^k \in \{1, 2, ..., M\}$ 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 corresponding transformations to map them to a common space. We further equip source domains with data distribution $\mathbb{P}_{1:K}^S$ whose density functions are $p_{1:K}^S\left(\mathbf{x}\right)$. Let us denote the ground-truth labeling functions for source domains by $f_{1:K}^S\left(\cdot\right) \in \mathcal{Y}_\Delta$, implying that $p_k^S\left(y \mid \mathbf{x}\right) = f_k^S\left(\mathbf{x}, y\right)$ (i.e., $f_k^S\left(\mathbf{x}, y\right)$ represents the y-th value of $f_k^S\left(\mathbf{x}\right)$). Therefore, the joint distribution to generate data instance $\mathbf{x}$ and categorical label $y \in \{1, ..., M\}$ is $p_k^S\left(\mathbf{x}, y\right) = p_k^S\left(\mathbf{x}\right) f_k^S\left(\mathbf{x}, y\right)$.

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

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

2.2 MULTI-SOURCE EXPERT TEACHER

Using the labeled source training sets $\mathcal{D}_{1:K}^S$, we can train qualified domain expert classifier $h_{1:K}^S$ (i.e., $h_k^S\left(\mathbf{x}\right) \in \mathcal{Y}_\Delta$ represents the prediction probability of h_k^S for a data instance $\mathbf{x}$) with good generalization capacity (e.g., $\mathcal{L}\left(h_k^S, f_k^S, \mathbb{P}_k^S\right) \leq \epsilon$ for some small $\epsilon > 0$). The next arising question is how to combine those domain experts to achieve a multi-source expert teacher h^S that can work well on $\mathbb{P}_{\boldsymbol{\pi}}^S$ (i.e., $\mathcal{L}\left(h^S, f^S, \mathbb{P}_{\boldsymbol{\pi}}^S\right) \leq \epsilon$). We leverage the domain experts to achieve a more powerful multi-source expert teacher by a weighted ensembling as follows:

$$h^S\left(\mathbf{x}, y\right) = \sum_{k=1}^K \frac{\pi_k p_k^S\left(\mathbf{x}, y\right)}{\sum_{j=1}^K \pi_j p_j^S\left(\mathbf{x}, y\right)} h_k^S\left(\mathbf{x}, y\right), \tag{2}$$

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

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

Theorem 3. *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., $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) \leq \epsilon$).

Proof. i) We have

$$
\begin{aligned}
\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) &= \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^S_{\boldsymbol{\pi}}}\left[d_{\mathcal{Y}}\left(h^S(\mathbf{x}), f^S(\mathbf{x})\right)\right] = \sum_{k=1}^{K} \pi_k \int d_{\mathcal{Y}}\left(h^S(\mathbf{x}), f^S_k(\mathbf{x})\right) d\mathbb{P}^S_k \\
&= \sum_{k=1}^{K} \pi_k \int d_{\mathcal{Y}}\left(h^S(\mathbf{x}), f^S_k(\mathbf{x})\right) d\mathbb{P}^S_k = \sum_{k=1}^{K} \pi_k \int \sum_{i=1}^{M} f^S_k(\mathbf{x}, i)\, \ell\left(h^S(\mathbf{x}, i)\right) d\mathbb{P}^S_k \\
&= \sum_{i=1}^{M} \int \sum_{k=1}^{K} \pi_k p^S_k(y = i \mid \mathbf{x})\, \ell\left(h^S(\mathbf{x}, i)\right) d\mathbb{P}^S_k = \sum_{i=1}^{M} \int \ell\left(h^S(\mathbf{x}, i)\right) \sum_{k=1}^{K} \pi_k p^S_k(y = i \mid \mathbf{x})\, p^S_k(\mathbf{x})\, d\mathbf{x} \\
&= \sum_{i=1}^{M} \int \ell\left(h^S(\mathbf{x}, i)\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y = i)\, d\mathbf{x} = \sum_{i=1}^{M} \int \ell\left(\sum_{k=1}^{K} \frac{\pi_k p^S_k(\mathbf{x}, y = i)}{\sum_{j=1}^{K} \pi_j p^S_j(\mathbf{x}, y = i)} h^S_k(\mathbf{x}, i)\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y = i)\, d\mathbf{x} \\
&\leq \sum_{i=1}^{M} \int \sum_{k=1}^{K} \frac{\pi_k p^S_k(\mathbf{x}, y = i)}{\sum_{j=1}^{K} \pi_j p^S_j(\mathbf{x}, y = i)} \ell\left(h^S_k(\mathbf{x}, i)\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y = i)\, d\mathbf{x} \quad \text{(thanks to the convexity of } \ell\text{)} \\
&= \sum_{i=1}^{M} \int \sum_{k=1}^{K} \frac{\pi_k p^S_k(\mathbf{x}, y = i)}{p^S_{\boldsymbol{\pi}}(\mathbf{x}, y = i)} \ell\left(h^S_k(\mathbf{x}, i)\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y = i)\, d\mathbf{x} \\
&= \sum_{i=1}^{M} \int \sum_{k=1}^{K} \pi_k p^S_k(\mathbf{x}, y = i)\, \ell\left(h^S_k(\mathbf{x}, i)\right) d\mathbf{x} = \sum_{k=1}^{K} \pi_k \sum_{i=1}^{M} \int \ell\left(h^S_k(\mathbf{x}, i)\right) p^S_k(\mathbf{x}, y = i)\, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \sum_{i=1}^{M} \int \ell\left(h^S_k(\mathbf{x}, i)\right) p^S_k(y = i \mid \mathbf{x})\, p^S_k(\mathbf{x})\, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \int \sum_{i=1}^{M} \ell\left(h^S_k(\mathbf{x}, i)\right) f^S_k(\mathbf{x}, i)\, d\mathbb{P}^S_k = \sum_{k=1}^{K} \pi_k \int d_{\mathcal{Y}}\left(h^S_k(\mathbf{x}), f^S_k(\mathbf{x})\right) d\mathbb{P}^S_k \\
&= \sum_{k=1}^{K} \pi_k \mathcal{L}\left(h^S_k, f^S_k, \mathbb{P}^S_k\right) \leq \max_{1\leq k\leq K} \mathcal{L}\left(h^S_k, f^S_k, \mathbb{P}^S_k\right).
\end{aligned}
$$

Note that we define $p^S_{\boldsymbol{\pi}}(\mathbf{x}, y = i) := \sum_{k=1}^{K} \pi_k p^S_k(\mathbf{x}, y = i) = \sum_{k=1}^{K} \pi_k p^S_k(y = i \mid \mathbf{x})\, p^S_k(\mathbf{x})$.

ii) It is trivial from (i). □

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

Proof. We have

$$\begin{aligned}\mathbb{E}_{(\mathbf{z},t)\sim\mathcal{D}}\left[CE\left(\mathcal{C}\left(\mathbf{z}\right),t\right)\right] &= \sum_{t=1}^{K}\pi_t\int CE\left(\mathcal{C}(\mathbf{z}),t\right)r_t\left(\mathbf{z}\right)d\mathbf{z}\\ &= -\int\sum_{t=1}^{K}\log\mathcal{C}\left(\mathbf{z},t\right)\pi_t r_t\left(\mathbf{z}\right)d\mathbf{z}.\end{aligned}$$

Given $\mathbf{z}$, we now find $\mathcal{C}^* = [\mathcal{C}_t^*]_{t=1}^K$ subjected to $\|\mathcal{C}^*\|_1 = 1$ and $\mathcal{C}^* \geq \mathbf{0}$ to maximize

$$\max_{\mathcal{C}:\|\mathcal{C}\|_1=1}\sum_{t=1}^{K}\log\mathcal{C}_t\pi_t r_t\left(\mathbf{z}\right).$$

The Lagrange function is as follows:

$$\mathcal{L} = \sum_{t=1}^{K}\log\mathcal{C}_t\pi_t r_t\left(\mathbf{z}\right) - \lambda\left(\sum_{t=1}^{K}\mathcal{C}_t - 1\right).$$

Setting the derivatives to 0, we obtain

$$\frac{\partial\mathcal{L}}{\partial\mathcal{C}_t} = \frac{\pi_t r_t\left(\mathbf{z}\right)}{C_t} - \lambda = 0,\ t = 1, ..., K.$$

Note that $\sum_{t=1}^{K}\mathcal{C}_t = 1$, we arrive at

$$\mathcal{C}_t^* = \frac{\pi_t r_t\left(\mathbf{z}\right)}{\sum_j \pi_j r_j\left(\mathbf{z}\right)},\ t = 1, ..., K.$$

Finally, we reach the conclusion. □

Inspired by the statement (ii) in Theorem 2, recall that f^T is the ground-truth labeling function on the target domain, we propose to learn a classifier 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 $d = \lambda d_{\mathcal{X}} + d_{\mathcal{Y}}$, implying that $\mathcal{W}_d\left(\mathbb{P}_{\cdot,\cdot},\mathbb{P}_{\cdot,\cdot}\right)$ is a proper metric. We can thus bound the quantity of interest $\mathcal{W}_d\left(\mathbb{P}_{T,h^T},\mathbb{P}_{T,f^T}\right)$ as

$$\mathcal{W}_d\left(\mathbb{P}_{T,h^T},\mathbb{P}_{T,f^T}\right) \leq \mathcal{W}_d\left(\mathbb{P}_{T,h^T},\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}\right) + \mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S},\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}\right) + \mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S},\mathbb{P}_{T,f^T}\right), \tag{3}$$

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 $y_\Delta = f^S\left(\mathbf{x}\right)$.

In the upper-bound in (3), $\mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S},\mathbb{P}_{T,f^T}\right)$ is constant. 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 $h^S_{1:K}$. Noting that $\mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S},\mathbb{P}^{\boldsymbol{\pi}}_{S,f^S}\right)$ is upper-bounded by $\mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)$ (thanks to the statement (iii) in Theorem 2), we arrive at the following optimization problem:

$$\min_{h^T}\left\{\mathcal{W}_d\left(\mathbb{P}_{T,h^T},\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}\right) + \mathcal{L}\left(h^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)\right\}. \tag{4}$$

The optimization problem in (4) 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 1 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 **M**ulti-Source Domain Adaptation via **O**ptimal Transport for **S**tudent-**T**eacher Learning (MOST).

Proposition 1 further means that among the transport maps H transporting $\mathbb{P}^T$ to $\mathbb{P}^S_{\boldsymbol{\pi}}$, which incurs a minimal label shift and enables the student h^T easiest to imitate its teacher h^S. Inspired by the statement (iv) in Theorem 2 (i.e., $\mathcal{W}_d\left(\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}, \mathbb{P}_{T,h^T}\right)$ 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:

$$\min_{h^T, G^T} \left\{ \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\pi}\right) + \mathcal{W}_d\left(\mathbb{Q}_{T,h^T}, \mathbb{Q}^{\pi}_{S,h^S}\right) \right\}, \tag{5}$$

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 $\mathbf{x} \sim \mathbb{P}^S$. Note that both h^S and $h^S_{1:K}$ now act on $G^S(\cdot)$.

Theorem 5. *Let $h^S_* \circ G^S_*$ be the optimal teacher and h^T_*, G^T_* be the optimal solutions of the optimization problem in (5). 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*[1]

$$\min_{h^T, G^T} \mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right) \leq \mathcal{L}\left(h^S_* \circ G^S_*, f^S, \mathbb{P}^S_{\pi}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S_*}_{S,f^S_*}, \mathbb{P}^{G^T_*}_{T,f^T_*}\right), \tag{6}$$

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

Proof. Let $f^{G^T}_T$ be the induced ground-truth labeling function over the joint space for which $f^{G^T}_T$ predicts $G^T(\mathbf{x})$ as same as f^T predicts $\mathbf{x}$ for $\mathbf{x} \sim \mathbb{P}^T$. Let $f^{G^S}_S$ be the induced ground-truth labeling function over the joint space for which $f^{G^S}_S$ predicts $G^S(\mathbf{x})$ as same as f^S predicts $\mathbf{x}$ for $\mathbf{x} \sim \mathbb{P}^S_{\boldsymbol{\pi}}$. We have the following inequality

$$\begin{aligned}
\mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right) &\leq \mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^S}_{S,h^S}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S}_{S,h^S}, \mathbb{P}^{G^S}_{S,f^{G^S}_S}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S}_{S,f^{G^S}_S}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right) \\
&\overset{(1)}{\leq} \mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^S}_{S,h^S}\right) + \mathcal{L}\left(h^S, f^{G^S}_S, \mathbb{P}^{G_S}_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S}_{S,f^{G^S}_S}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right) \\
&\overset{(2)}{=} \mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^S}_{S,h^S}\right) + \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S}_{S,f^{G^S}_S}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right),
\end{aligned} \tag{7}$$

where $\mathbb{P}^{G^S}_{S,h^S}$ consists of the pairs $\left(G^S(\mathbf{x}), h^S\left(G^S(\mathbf{x})\right)\right)$ for which $\mathbf{x} \sim \mathbb{P}^S_{\boldsymbol{\pi}}$, $\mathbb{P}^{G^S}_{S,f^{G^S}_S}$ consists of the pairs $\left(G^S(\mathbf{x}), f^{G^S}_S\left(G^S(\mathbf{x})\right)\right)$ for which $\mathbf{x} \sim \mathbb{P}^S_{\boldsymbol{\pi}}$, $\mathbb{P}^{G^T}_{T,f^{G^T}_T}$ consists of the pairs $\left(G^T(\mathbf{x}), f^{G^T}_T\left(G^T(\mathbf{x})\right)\right)$ for which $\mathbf{x} \sim \mathbb{P}^T$, $\mathbb{P}^{G^T}_{T,h^T}$ consists of the pairs $\left(G^T(\mathbf{x}), h^T\left(G^T(\mathbf{x})\right)\right)$ for which $\mathbf{x} \sim \mathbb{P}^T$, and $\mathbb{P}^{G^S}_{\boldsymbol{\pi}} = G^S \# \mathbb{P}^S_{\boldsymbol{\pi}}$.

We note that the derivation in (1) is from the statement (iv) in Theorem 2 and to reach (2), we derive as follows:

$$\begin{aligned}
\mathcal{L}\left(h^S, f^{G^S}_S, \mathbb{P}^{G^S}_{\boldsymbol{\pi}}\right) &= \int d_{\mathcal{Y}}\left(h^S(\boldsymbol{y}), f^{G^S}_S(\boldsymbol{y})\right) d\mathbb{P}^{G^S}_{\boldsymbol{\pi}} = \int d_{\mathcal{Y}}\left(h^S\left(G^S(\mathbf{x})\right), f^{G^S}_S\left(G_S(\mathbf{x})\right)\right) d\mathbb{P}^S_{\boldsymbol{\pi}} \\
&= \int d_{\mathcal{Y}}\left(h^S\left(G^S(\mathbf{x})\right), f^S(\mathbf{x})\right) d\mathbb{P}^S_{\boldsymbol{\pi}} = \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right),
\end{aligned}$$

where the second identity is due to $G^S \# \mathbb{P}^S_{\boldsymbol{\pi}} = \mathbb{P}^{G^S}_{\boldsymbol{\pi}}$.

To proceed, we first observe that

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

[1] 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 $\mathbf{x}$.

where we consider h^S in the family of expert teachers (i.e., those which are a combination of the domain experts).

Indeed, let $\left(\bar{h}^S, \bar{G}^S\right) \coloneqq \underset{h^S, G^S}{\operatorname{argmin}} \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)$. Thanks to the infinite capacity of the families, we can choose the generator $\bar{G}^T$ and the target classifier $\bar{h}^T$ such that $\bar{G}^T \# \mathbb{P}^T = \bar{G}_S \# \mathbb{P}^S$ and $\bar{h}^T = \bar{h}^S$ (i.e., $\mathcal{W}_d\left(\mathbb{P}^{\bar{G}^T}_{T,\bar{h}^T}, \mathbb{P}^{\bar{G}^S}_{S,\bar{h}^S}\right) = 0$).

Then we have

$$\begin{aligned}\min_{h^S, G^S}\left\{\mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)\right\} &= \mathcal{W}_d\left(\mathbb{P}^{\bar{G}^T}_{T,\bar{h}^T}, \mathbb{P}^{\bar{G}^S}_{S,\bar{h}^S}\right) + \mathcal{L}\left(\bar{h}^S \circ \bar{G}^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) \\ &\geq \min_{h^S, h^T, G^S, G^T}\left\{\mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^S}_{S,h^S}\right) + \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)\right\}.\end{aligned}$$

Therefore, (8) holds as the reverse inequality is obvious.

Now let $h^S_*, h^T_*, G^S_*, G^T_*$ be the optimal solutions of the left side of (8). Then by (8), we must have

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

This implies that $\mathcal{W}_d\left(\mathbb{P}^{G^T_*}_{T,h^T_*}, \mathbb{P}^{G^S_*}_{S,h^S_*}\right) = 0$ and $\left(h^S_*, G^S_*\right) = \underset{h^S, G^S}{\operatorname{argmin}} \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right)$. From Theorem (2), we see that the first identity means that $G^S_* \# \mathbb{P}^S_{\boldsymbol{\pi}} = G^T_* \# \mathbb{P}^T$ and $h^T_* = h^S_*$.

To further proceed, we refer to the triangle inequality in (7) as

$$\mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right) \leq \mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^S}_{S,h^S}\right) + \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S}_{S,f^{G^S}_S}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right).$$

Taking a minimization and using $\mathcal{W}_d\left(\mathbb{P}^{G^T_*}_{T,h^T_*}, \mathbb{P}^{G^S_*}_{S,h^S_*}\right) = 0$, we obtain

$$\begin{aligned}\min_{h^T, G^T} \mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right) &\leq \min_{h^T, h^S, G^T, G^S}\left\{\mathcal{W}_d\left(\mathbb{P}^{G^T}_{T,h^T}, \mathbb{P}^{G^S}_{S,h^S}\right) + \mathcal{L}\left(h^S \circ G^S, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S}_{S,f^{G^S}_S}, \mathbb{P}^{G^T}_{T,f^{G^T}_T}\right)\right\} \\ &\leq \mathcal{W}_d\left(\mathbb{P}^{G^T_*}_{T,h^T_*}, \mathbb{P}^{G^S_*}_{S,h^S_*}\right) + \mathcal{L}\left(h^S_* \circ G^S_*, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^S_*}_{S,f^{G^S_*}_S}, \mathbb{P}^{G^T_*}_{T,f^{G^T_*}_T}\right) \\ &= \mathcal{L}\left(h^S_* \circ G^S_*, f^S, \mathbb{P}^S_{\boldsymbol{\pi}}\right) + \mathcal{W}_d\left(\mathbb{P}^{G^*_S}_{S,f^{G^*_S}_S}, \mathbb{P}^{G^*_T}_{T,f^{G^*_T}_T}\right).\end{aligned}$$

□

3 CLUSTERING VIEW OF OPTIMAL TRANSPORT

We now present the clustering view of optimal transport which assists us to intuitively explain the training of teacher and student. Let us start with the clustering view of optimal transport.

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete distributions defined as

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

where $\delta_{\mathbf{x}}$ indicates a Dirac measure cantered at x.

Without loss of generality, we can assume that $n \leq m$ and consider the Wasserstein distance $\mathcal{W}_d\left(\mathbb{P}, \mathbb{Q}\right)$ w.r.t a metric d. The following theorem characterizes the clustering view of optimal transport.

Table 1: Small, medium and large network architecture of MOST. The Leaky ReLU (lReLU) parameter a is set to 0.1. *K* and *M* denote the number of source domains and the number of classes respectively.

Architecture	Small	Large
Input size	2048	$32 \times 32 \times 3$
Generator $G_{1:K}^S, G^T$ (Share weights)	256 dense, ReLU dropout, $p = 0.5$ Gaussian noise, σ = 1	instance normalization 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 2×2 max-pool, stride 2 dropout, $p = 0.5$ Gaussian noise, σ = 1 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 2×2 max-pool, stride 2 dropout, $p = 0.5$ Gaussian noise, σ = 1
Classifier $h_{1:K}^S, h^T$	*M* dense, softmax	3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU global average pool *M* dense, softmax
Source domain discriminator $\mathcal{C}$	*K* dense, ReLU	100 dense, ReLU *K* dense, ReLU
ϕ	1 dense, linear	100 dense, ReLU 1 dense, linear

Theorem 6. *Consider the following optimization problem:*

$$\min_{\mathbf{v}_{1:n}} \mathcal{W}_d\left(\mathbb{P}, \mathbb{Q}\right).$$

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*

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

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

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

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

Proof. The proof is quite obvious from the definition of optimal transport and the fact that $T_{\#}\mathbb{P} = \mathbb{Q}$ means $T\left(\mathbf{u}_i\right) = \mathbf{v}_{\sigma(i)}, \forall i = 1,...,n$ for some $\sigma \in \Pi\left(m,n\right)$. □

4 IMPLEMENTATION SPECIFICATION AND ADDITIONAL EXPERIMENTAL RESULTS

4.1 DATA PREPARATION AND PREPROCESSING

Digits-five 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. We resize the resolution of digit images to 32×32, and normalize the value of each pixel to the range of $[-1, 1]$.

Table 2: The experimental settings in all transfer tasks. The *batch size* column denotes the total batch sizes of all source domains in each iteration. α, β, γ are trade-off parameters of $\mathcal{L}^{WS}$, $\mathcal{L}^{pl}$, $\mathcal{L}^{clus}$ respectively in our final objective function (11).

dataset	domains	classes	pretrained model	input size	batch size	iterations	learning rate	λ	update ϕ (times)	ϵ	α	β	γ
Digits-five	5	10	None	32×32	200	80000	2×10^{-4}	10.0	5	0.1	0.1	$\{0.1, 1.0\}$	$\{0.0, 0.1\}$
Office-Caltech10	4	10	ResNet-101	224×224	60	20000	1×10^{-4}	10.0	5	0.1	0.1	0.1	0.1
Office-31	3	31	AlexNet	227×227	62	20000	1×10^{-4}	10.0	5	0.1	0.1	0.1	0.1

Office-Caltech10 is categorized in four different domains: Amazon (**A**), Caltech (**C**), DSLR (**D**), and Webcam (**W**) with 10 common classes and 2533 images in total. The resolutions of images are resized to 224×224 for finetuning pre-trained ResNet-101.

Office-31 contains 4,110 images with 31 classes, and is categorized into three domains: Amazon (**A**), DSLR (**D**), and Webcam (**W**). To perform finetuning on AlexNet, we resized the resolutions of images in this dataset to 227×227.

4.2 ARCHITECTURE/HYPERPARAMETERS

4.2.1 Network architecture

We use a small network to finetune pretrained AlexNet and ResNet-101, while a larger network is employed to train on digit images. These network types are described in Table 1. We empirically find that using a shared-weight generator improves the imitation learning capability of the student to the teacher since target and source samples are closer in the joint space. To reduce the overfitting, excluding dense layers in ϕ network, we add the batch normalization layers on top of convolutional and dense layers. Finally, we implement our MOST in Python using TensorFlow (version 1.9.0), an open-source software library for Machine Intelligence developed by the Google Brain Team. All experiments are run on a computer with an NVIDIA Tesla V100 SXM2 with 16 GB memory.

4.2.2 Hyperparameters

In the loss $\mathcal{L}^{WS}$, a number of parameters are crucial to make the cross-domain imitation possible such as λ and ϵ. Specifically, λ is set to 10.0 on all transfer tasks, while the coefficient ϵ specifies the shape of joint density for the optimal transport plan and is set to 0.1 in all settings. Finally, we apply Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) with Polyak averaging. The learning rate along with the number of training iterations are depicted in Table 2.

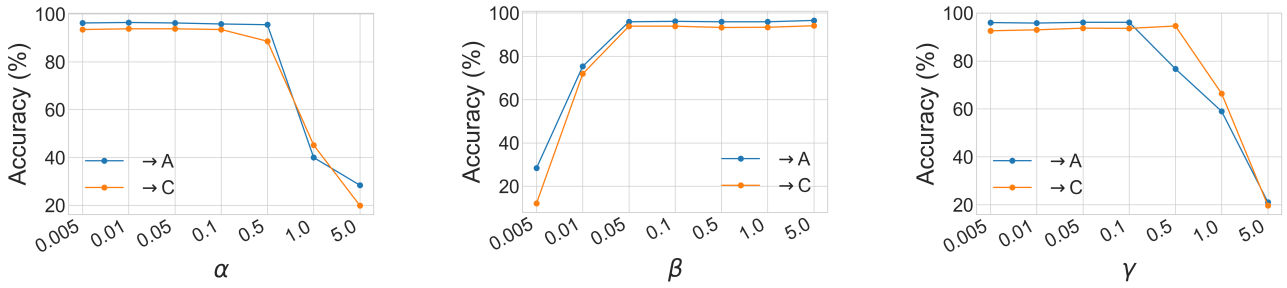
4.3 ADDITIONAL ABLATION STUDY

4.3.1 Parameter Sensitivity

We evaluate the effects of the trade-off parameters α, β, γ in Figure 1. We search all α, β and γ in the grid of $\{0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0\}$ and report the test accuracy on two transfer tasks "→**A**" and "→**C**" on *Office-Caltech10*. The results show that the model yields the stable performance when α from 0.005 to 0.5, β from 0.05 to 5.0, and γ from 0.005 to 0.1. We find that our MOST can achieve high performance when $\alpha = \beta = \gamma = 0.1$, hence we suggest to pick this value on most of our experiments.

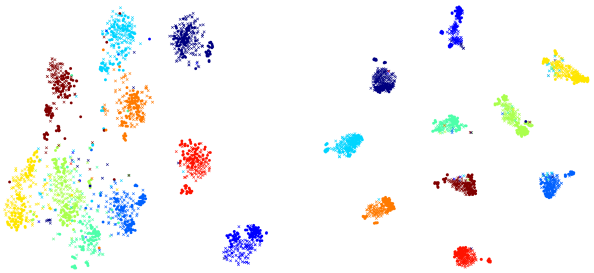
4.3.2 Feature visualization

We visualize the features of ResNet-101 and our methods on "→**C**" tasks by *t*-SNE in Figure 2. Figure 2a shows that ResNet-101 classifies quite well on the mixture of source domains (**A, D, W**) but poorly on the target domain (**C**), while the representation in Figure 2b is generated by our method with better alignment. MOST represents the ability of reducing the data shift and label shift between two domains to exactly achieve 10 clusters corresponding to 10 classes of *Office-Caltech10*.



(a) α ($\beta = \gamma = 0.1$) (b) β ($\alpha = \gamma = 0.1$) (c) γ ($\alpha = \beta = 0.1$)

Figure 1: Test accuracy (%) when tweaking α, β and γ on transfer tasks "→**A**" and "→**C**".



(a) ResNet-101 (b) MOST

Figure 2: The *t*-SNE visualization of the transfer task "→**C**" with label and domain information. Each color denotes a class while the circle and cross markers represent the mixture of source and target data respectively.

4.3.3 Effect of parameters on imitation learning

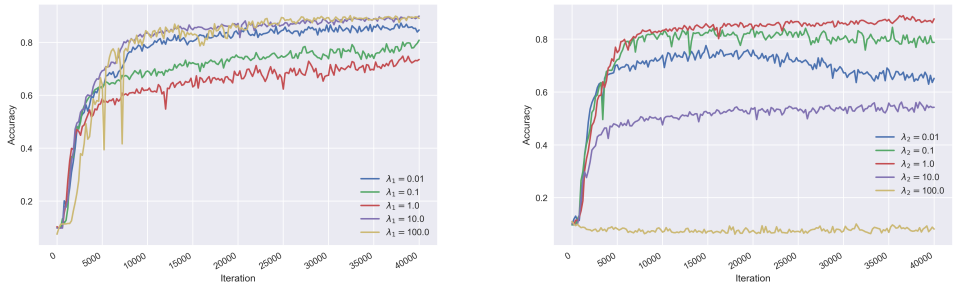
The loss $\mathcal{L}^{WS}$ is introduced in Section 5.4.2 of the main paper. In this section, we further investigate sensitivity of two parameters, λ_1 and λ_2, in $d\left(G^S\left(\mathbf{x}^S\right), G^T\left(\mathbf{x}^T\right)\right)$ to testify the imitation ability of MOST. They are trade-off parameters of $d_X(.,.)$ and $d_{\mathcal{Y}}(.,.)$, respectively as follows:

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

where $d\left(G^S\left(\mathbf{x}^S\right), G^T\left(\mathbf{x}^T\right)\right) = \lambda_1 \left\| G^T\left(\mathbf{x}^T\right) - G^S\left(\mathbf{x}^S\right) \right\| + \lambda_2 d_{\mathcal{Y}}\left(h^T\left(G^T\left(\mathbf{x}^T\right)\right), h^S\left(G^S\left(\mathbf{x}^S\right)\right)\right)$, while $\mathbf{x}^T \sim \mathbb{P}^T, \mathbf{x}^S \sim \mathbb{P}^S_{\boldsymbol{\pi}}$. Adjusting λ_1 and λ_2 helps to mitigate the data shift and label shift effectively. Figure 3a and 3b portray the model performance with a diverse range of values λ_1 and λ_2, respectively. After searching them in the grid of $\{0.01, 0.1, ..., 100.0\}$, we observe from Figure 3 that our model is more sensitive to λ_2 since it directly controls the imitation capability of the student. In general, the training process of MOST is stable with $\lambda_1 = 10.0$ and $\lambda_2 = 1.0$.

4.3.4 Choosing the mixing proportion $\boldsymbol{\pi}$

This ablation study aims to testify the effect of choosing the mixture weights $\boldsymbol{\pi}$. We conduct experiments on two scenarios: i) $\boldsymbol{\pi}$ is a uniform distribution $\left[\frac{1}{K}, ..., \frac{1}{K}\right]$ (**U**), and ii) $\boldsymbol{\pi}$ is proportional to the number of training examples in the source domains (i.e., $N^S_{1:K}$) (**P**). The experimental results in Table 3 show the competitive results between two approaches, which means that the imitation capability is not affected by changing the mixture distribution $\mathbb{P}^{\boldsymbol{\pi}}_{S,h^S}$ in (8).



(a) λ_1 ($\lambda_2 = 1.0$) (b) λ_2 ($\lambda_1 = 10.0$)

Figure 3: Test accuracy (%) when tweaking λ_1 and λ_2 on "→**mm**" task.

Table 3: Results (%) of different choices of $\boldsymbol{\pi}$ on Office-Caltech10.

Settings	→**W**	→**D**	→**C**	→**A**	Avg
P	99.7	100.0	95.1	96.5	97.8
U	100.0	100.0	95.7	96.3	98.0