Supplementary Material for “LAMDA: Label Matching Deep Domain Adaptation”

In this supplementary material, we provide complete detail for all proofs presented in our main paper together with the related background so that it can be as self-contained as possible. In the following part, we present the experiment on a synthetic dataset to verify our theory, followed by the experimental settings and datasets for our LAMDA.

1 Related Background

In this section, we present the related background for our paper. We depart with the introduction of pushforward measure followed by the definition of optimal transport and the introduction of a standard machine learning setting.

1.1 Pushforward Measure

Given two probability spaces $(\mathcal{X},\mathcal{F},\mu)$ and $(\mathcal{Y},\mathcal{G})$ where $\mathcal{X},\mathcal{Y}$ are two sample spaces, $\mathcal{F},\mathcal{G}$ are two σ-algebras over $\mathcal{X},\mathcal{Y}$ respectively, and μ is a probability measure, a map $T:\ \mathcal{X}\rightarrow\mathcal{Y}$ is said to be $(\mathcal{Y},\mathcal{G})$- $(\mathcal{X},\mathcal{F})$ measurable if for every $A\in\mathcal{G}$, the inverse $T^{-1}\left(A\right)\in\mathcal{F}$. The $(\mathcal{Y},\mathcal{G})$- $(\mathcal{X},\mathcal{F})$ measurable map T when applied to $(\mathcal{X},\mathcal{F},\mu)$ induces a distribution ν over $(\mathcal{Y},\mathcal{G})$ which is defined as:

$$\nu\left(A\right)=\mu\left(T^{-1}\left(A\right)\right),\,\forall A\in\mathcal{G}$$

We also say that the map T transport the probability measure μ to ν and denote as $\nu=T_{\#}\mu$. Furthermore, if μ and ν are two given atomless probability measures over $(\mathcal{X},\mathcal{F})$ and $(\mathcal{Y},\mathcal{G})$, there exists a bijection $T:\ \mathcal{X}\rightarrow\mathcal{Y}$ that transports μ to ν. This is known as measurable isomorphism and formally stated in [32] (Chapter 1, Page 19).

Theorem 1. *Given two probability spaces $(\mathcal{X},\mathcal{F},\mu)$ and $(\mathcal{Y},\mathcal{G},\nu)$ with two atomless probability μ, ν over two Polish spaces $\mathcal{X}$, $\mathcal{Y}$ (i.e., separably complete metric spaces), there exist a bijection $T:\ \mathcal{X}\rightarrow\mathcal{Y}$ that transports μ to ν, i.e., $T_{\#}\mu=\nu$.*

1.2 Optimal Transport

Given two probability measures $(\mathcal{X},\,\mu)$ and $(\mathcal{Y},\,\nu)$ and a cost function $c\left(\mathbf{x},\,\mathbf{x}'\right)$, under the conditions stated in Theorems 1.32 and 1.33 [26], two following definitions of Wasserstein (WS) distance are equivalent:

$$W_{c,p}\left(\mu,\,\nu\right)=\inf_{T_{\#}\mu=\nu}\mathbb{E}_{\mathbf{x}\sim\mu}\left[c\left(\mathbf{x},T\left(\mathbf{x}\right)\right)^{p}\right]^{1/p}$$

$$W_{c,p}\left(\mu,\,\nu\right)=\inf_{\pi\in\Gamma\left(\mu,\nu\right)}\mathbb{E}_{\left(\mathbf{x},\mathbf{x}'\right)\sim\pi}\left[c\left(\mathbf{x},\mathbf{x}'\right)^{p}\right]^{1/p}$$

where $p>0$ and $\Gamma\left(\mu,\nu\right)$ specifies the set of joint distributions over $\mathcal{X}\times\mathcal{Y}$ which admits μ and ν as marginals. The first definition is known as Monge problem (MP), while the second one is known as Kantorovich problem (KP).

We now restate the sufficient conditions for which (MP) and (KP) are equivalent (cf. Theorems 1.32 and 1.33 [26]).

Theorem 2. *If $\mathcal{X}$ and $\mathcal{Y}$ are compact, Polish metric spaces, μ and ν are atomless, and c is a lower semi-continuous function, (KP) is equivalent to (MP) in the sense that two infima are equal.*

In what follows, we assume that the relevant conditions are satisfied and use (KP) and (MP) interchangeably depending on the contexts. More specifically, we use (MP) in Theorem (9), while using (KP) in the rest.

1.3 Machine Learning Setting and General Loss

According to [31], a standard machine learning system consists of three components: the generator, the supervisor, and the hypothesis class.

Generator The generator is the mechanism to generate data examples $\mathbf{x} \in \mathbb{R}^d$ and is mathematically formulated by an existed but unknown distribution $p\left(\mathbf{x}\right)$.

Supervisor The supervisor is the mechanism to assign labels y (e.g., $y \in \{1, 2, \ldots, C\}$ for the classification problem and $y \in \mathbb{R}$ for the regression problem) to a data example $\mathbf{x}$ and is mathematically formulated as a conditional distribution $p\left(y \mid \mathbf{x}\right)$.

Hypothesis class This specifies the hypothesis set $\mathcal{H} = \{h_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta\}$ parameterized by $\boldsymbol{\theta}$ which is used to predict label for the data examples $\mathbf{x}$.

Given a loss function $l\left(x, y; \boldsymbol{\theta}\right) = \ell\left(y, \, h_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)$ where $\ell : \Delta_C \to \mathbb{R}$ (Δ_C is the C-simplex) and $\ell\left(y, y'\right)$ specifies the loss suffered if predicting the data example $\mathbf{x}$ with the label y' while its true label is y, the general loss of the hypothesis $h_{\boldsymbol{\theta}}$ is defined as the expected loss caused by $h_{\boldsymbol{\theta}}$:

$$R\left(\boldsymbol{\theta}\right) = \mathbb{E}_{p(\mathbf{x},y)}\left[\ell\left(y, \, h_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)\right] = \int \ell\left(y, \, h_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right) p\left(\mathbf{x}, y\right) d\mathbf{x} dy$$

The optimal parameter $\boldsymbol{\theta}^* \in \Theta$ is sought by minimizing the general loss as:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\text{argmin}} \, R\left(\boldsymbol{\theta}\right)$$

2 Theoretical Results

2.1 Gap between target and source domains

In this section, we investigate the variance $\Delta R\left(h^s, h^t\right)$ between the expected loss in target domain $R^t\left(h^t\right)$ and the expected loss in source domain $R^s\left(h^s\right)$ where $h^t = h^s \circ T$. We embark on with the following simple yet key proposition indicating the connection between $R^t\left(h^t\right)$ and $R^{\#}\left(h^s\right)$.

Proposition 3. *As long as $h^t = h^s \circ T$, we have $R^t\left(h^t\right) = R^{\#}\left(h^s\right)$.*

Proof. The proof of the proposition is directly from the definitions of h^t, h^s, and expected losses. In particular, we find that

$$R^{\#}\left(h^s\right) = \int \ell\left(y, h^s\left(\mathbf{x}\right)\right) p^{\#}\left(y \mid \mathbf{x}\right) p^{\#}\left(\mathbf{x}\right) d\mathbf{x} dy = \mathbb{E}_{\mathbb{P}^{\#}}\left[\int \ell\left(y, h^s\left(\mathbf{x}\right)\right) p^{\#}\left(y \mid \mathbf{x}\right) dy\right].$$

Recall that, T transports the target distribution $\mathbb{P}^t$ to the source distribution $\mathbb{P}^\#$, we achieve that

$$\begin{aligned} R^\# \left(h^s\right) &= \mathbb{E}_{\mathbb{P}^t}\left[\int \ell\left(y, h^s\left(T\left(\mathbf{x}\right)\right)\right) p^\#\left(y \mid T\left(\mathbf{x}\right)\right) dy\right] \\ &= \mathbb{E}_{\mathbb{P}^t}\left[\int \ell\left(y, h^t\left(\mathbf{x}\right)\right) p^t\left(y \mid \mathbf{x}\right) dy\right] = R^t\left(h^t\right), \end{aligned}$$

where the second equality is due to the connection $h^t = h^s \circ T$. As a consequence, we reach the conclusion of the proposition. □

Theorem 4. *(**Theorem 1 in the main paper**) Given Assumption (A.1), then for any hypothesis $h^s \in \mathcal{H}^s$, the following inequality holds:*

$$\Delta R\left(h^s, h^t\right) \leq M\left(\mathrm{W}_{c_{0/1}}\left(\mathbb{P}^s, \mathbb{P}^\#\right) + \mathbb{E}_{\mathbb{P}^t}\left[\left\|\Delta p\left(\cdot \mid \mathbf{x}\right)\right\|_1\right]\right),$$

where $\Delta p\left(\cdot \mid \mathbf{x}\right) := \left\|\left[p^t\left(y = i \mid \mathbf{x}\right) - p^s\left(y = i \mid T\left(\mathbf{x}\right)\right)\right]_{i=1}^C\right\|_1$, and $W_{c_{0/1}}\left(\cdot, \cdot\right)$ is the Wasserstein distance with respect to the cost function $c_{0/1}\left(\mathbf{x}, \mathbf{x}'\right) = \mathbf{1}_{\mathbf{x} \neq \mathbf{x}'}$, returning 1 if $\mathbf{x} \neq \mathbf{x}'$ and 0 otherwise.

Proof. Invoking the result from Proposition 3 and the basic triangle inequality, we obtain that

$$\begin{aligned} \Delta R\left(h^s, h^t\right) =& \left|R^t\left(h^t\right) - R^s\left(h^s\right)\right| = \left|R^\#\left(h^s\right) - R^s\left(h^s\right)\right| \\ =& \left|R^\#\left(h^s\right) - R^{\#,s}\left(h^s\right) + R^{\#,s}\left(h^s\right) - R^s\left(h^s\right)\right| \\ \leq& \left|R^\#\left(h^s\right) - R^{\#,s}\left(h^s\right)\right| + \left|R^{\#,s}\left(h^s\right) - R^s\left(h^s\right)\right|. \end{aligned}$$

To achieve the conclusion of the theorem, it is sufficient to upper bound the two terms $\left|R^\#\left(h^s\right) - R^{\#,s}\left(h^s\right)\right|$ and $\left|R^{\#,s}\left(h^s\right) - R^s\left(h^s\right)\right|$. For the first term, according the definition of expected losses, we find that

$$\begin{aligned} \left|R^\#\left(h^s\right) - R^{\#,s}\left(h^s\right)\right| &= \left|\int \ell\left(h^s\left(\mathbf{x}\right), y\right)\left(p^\#\left(y \mid \mathbf{x}\right) - p^s\left(y \mid \mathbf{x}\right)\right) p^\#\left(\mathbf{x}\right) d\mathbf{x}dy\right| \\ &= \left|\sum_{y=1}^C \int \ell\left(h^s\left(\mathbf{x}\right), y\right)\left(p^\#\left(y \mid \mathbf{x}\right) - p^s\left(y \mid \mathbf{x}\right)\right) p^\#\left(\mathbf{x}\right) d\mathbf{x}\right| \\ &\leq \sum_{y=1}^C \int \ell\left(h^s\left(\mathbf{x}\right), y\right)\left|p^\#\left(y \mid \mathbf{x}\right) - p^s\left(y \mid \mathbf{x}\right)\right| p^\#\left(\mathbf{x}\right) d\mathbf{x} \\ &\leq M \sum_{y=1}^C \int \left|p^\#\left(y \mid \mathbf{x}\right) - p^s\left(y \mid \mathbf{x}\right)\right| p^\#\left(\mathbf{x}\right) d\mathbf{x} \\ &= M \int \sum_{y=1}^C \left|p^\#\left(y \mid \mathbf{x}\right) - p^s\left(y \mid \mathbf{x}\right)\right| p^\#\left(\mathbf{x}\right) d\mathbf{x} \\ &\leq M\mathbb{E}_{\mathbb{P}^\#}\left[\left\|\left[p^\#\left(y \mid \mathbf{x}\right) - p^s\left(y \mid \mathbf{x}\right)\right]_{y=1}^C\right\|_1\right] \\ &\overset{(1)}{=} M\mathbb{E}_{\mathbb{P}^t}\left[\left\|\left[p^\#\left(y \mid T\left(\mathbf{x}\right)\right) - p^s\left(y \mid T\left(\mathbf{x}\right)\right)\right]_{y=1}^C\right\|_1\right] \\ &= M\mathbb{E}_{\mathbb{P}^t}\left[\left\|\left[p^t\left(y \mid \mathbf{x}\right) - p^s\left(y \mid T\left(\mathbf{x}\right)\right)\right]_{y=1}^C\right\|_1\right], \end{aligned} \tag{1}$$

where (1) is from the fact that $T_\# \mathbb{P}^t = \mathbb{P}^\#$.

For the second term, similar argument as the above argument leads to

$$\begin{aligned}
\left|R^{\#,s}\left(h^{s}\right)-R^{s}\left(h^{s}\right)\right| &= \left|\sum_{y=1}^{C}\int \ell\left(h^{s}\left(\mathbf{x}\right),y\right)p^{s}\left(y\mid\mathbf{x}\right)\left[p^{\#}\left(\mathbf{x}\right)-p^{s}\left(\mathbf{x}\right)\right]d\mathbf{x}\right| \\
&\leq \sum_{y=1}^{C}\int \ell\left(h^{s}\left(\mathbf{x}\right),y\right)p^{s}\left(y\mid\mathbf{x}\right)\left|p^{\#}\left(\mathbf{x}\right)-p^{s}\left(\mathbf{x}\right)\right|d\mathbf{x} \\
&\leq M\sum_{y=1}^{C}\int p^{s}\left(y\mid\mathbf{x}\right)\left|p^{\#}\left(\mathbf{x}\right)-p^{s}\left(\mathbf{x}\right)\right|d\mathbf{x} \\
&= M\int\sum_{y=1}^{C}p^{s}\left(y\mid\mathbf{x}\right)\left|p^{\#}\left(\mathbf{x}\right)-p^{s}\left(\mathbf{x}\right)\right|d\mathbf{x} \\
&= M\int\left|p^{\#}\left(\mathbf{x}\right)-p^{s}\left(\mathbf{x}\right)\right|d\mathbf{x} = MW_{c_{0/1}}\left(\mathbb{P}^{s},\mathbb{P}^{\#}\right), \qquad (2)
\end{aligned}$$

where the final equality is from the fact that cost matrix $c_{0/1}$ is given by $c_{0/1}\left(\mathbf{x},\mathbf{x}'\right)=\mathbf{1}_{\mathbf{x}\neq\mathbf{x}'}$, which returns 1 if $\mathbf{x}\neq\mathbf{x}'$ and 0 otherwise (for the second equality, please refer to [12], Page 7 and the coupling characterization of total variance distance).

Combining the results from (1) and (2), we arrive at the bound that

$$\begin{aligned}
\Delta R\left(h^{s},h^{t}\right) &\leq M\Bigg(W_{c_{0/1}}\left(\mathbb{P}^{s},\mathbb{P}^{\#}\right)+\mathbb{E}_{\mathbb{P}^{t}}\left[\left\|p^{t}\left(y\mid\mathbf{x}\right)-p^{s}\left(y\mid T\left(\mathbf{x}\right)\right)\right\|_{1}\right]\Bigg) \\
&= M\left(W_{c_{0/1}}\left(\mathbb{P}^{s},\mathbb{P}^{\#}\right)+\mathbb{E}_{\mathbb{P}^{t}}\left[\left\|\Delta p\left(y\mid\mathbf{x}\right)\right\|_{1}\right]\right).
\end{aligned}$$

As a consequence, we reach the conclusion of the theorem. □

Remark 5. If the following assumptions hold:

(i) The transformation mapping $T\left(\mathbf{x}\right)=\mathbf{x}$, i.e., we use the same hypothesis set for both the source and target domains,

(ii) The loss $\ell\left(y,h\left(\mathbf{x}\right)\right)=\frac{1}{2}\left|y-h\left(\mathbf{x}\right)\right|$ where we restrict to consider hypothesis $h:\mathcal{X}\rightarrow\{-1,1\}$,

then we recover the theoretical result obtained in [2].

Remark 6. When $W_{c_{0/1}}\left(\mathbb{P}^{s},\mathbb{P}^{\#}\right)=0$, i.e., $T_{\#}\mathbb{P}^{t}=\mathbb{P}^{s}$, and there is a harmony between two supervisors of source and target domain, i.e., $p^{t}\left(y\mid\mathbf{x}\right)=p^{s}\left(y\mid T\left(\mathbf{x}\right)\right)$), Theorem 4 suggests that we can perfectly do transfer learning without loss of performance. This fact is summarized in the following corollary.

Corollary 7. *Assume that $T_{\#}\mathbb{P}^{t}=\mathbb{P}^{s}$ and the source and target supervisor distributions are harmonic in the sense that $p^{s}\left(y\mid T\left(\mathbf{x}\right)\right)=p^{t}\left(y\mid\mathbf{x}\right)$ for $\mathbf{x}\sim\mathbb{P}_{t}$. Then, we can do a perfect transfer learning between the source and target domains.*

Proof. For any $h^{s}\in\mathcal{H}^{s}$, denote $h^{t}=h^{s}\circ T$, we have

$$\begin{aligned}
R^{s}\left(h^{s}\right) &= \mathbb{E}_{\mathbb{P}^{s}}\left[\int\ell\left(y,h^{s}\left(\mathbf{x}\right)\right)p^{s}\left(y\mid\mathbf{x}\right)dy\right] \\
&= \mathbb{E}_{\mathbb{P}^{t}}\left[\int\ell\left(y,h^{s}\left(T\left(\mathbf{x}\right)\right)\right)p^{s}\left(y\mid T\left(\mathbf{x}\right)\right)dy\right] \\
&= \mathbb{E}_{\mathbb{P}^{t}}\left[\int\ell\left(y,h^{t}\left(\mathbf{x}\right)\right)p^{t}\left(y\mid\mathbf{x}\right)dy\right]=R^{t}\left(h^{t}\right),
\end{aligned}$$

where the second equality is from the fact T transport $\mathbb{P}^{t}$ to $\mathbb{P}^{s}$. □

Furthermore, given a decreasing function $\phi : \mathbb{R} \to [0,1]$, a hypothesis h^s is said to be ϕ-Lipschitz transferable [6] w.r.t. a joint distribution $\gamma \in \Gamma\left(\mathbb{P}^s, \mathbb{P}^{\#}\right)$, the metric c, and the norm $\|\cdot\|$ if for all $\lambda > 0$, we have

$$\mathbb{P}_{(\mathrm{x}_s, \mathbf{x}_{\#}) \sim \gamma}\left[\left\|h^s\left(\mathbf{x}_s\right) - h^s\left(\mathbf{x}_{\#}\right)\right\| > \lambda c\left(\mathbf{x}_s, \mathbf{x}_{\#}\right)\right] \leq \phi\left(\lambda\right).$$

Theorem 8. ***(Theorem 3 in the main paper)*** *Assume that Assumptions (A.1) and (A2) hold, the hypothesis h^s satisfies ϕ-Lipschitz transferable w.r.t the optimal joint distribution (transport plan) $\gamma^* \in \Gamma\left(\mathbb{P}^s, \mathbb{P}^{\#}\right)$, c and $\|\cdot\|$, the following inequality holds for all $\lambda > 0$:*

$$\begin{aligned}\Delta R\left(h^s, h^t\right) \leq\, & M\left(\mathbb{E}_{\mathbb{P}^t}\left[\left\|\Delta p\left(\cdot \mid \mathbf{x}\right)\right\|_1\right] + 2\phi\left(\lambda\right)\right) \\ & + kC\lambda \mathrm{W}_{c,p}\left(\mathbb{P}^s, \mathbb{P}^{\#}\right).\end{aligned}$$

Proof. We have

$$\begin{aligned}\Delta R\left(h^s, h^t\right) = & \left|R^t\left(h^t\right) - R^s\left(h^s\right)\right| = \left|R^{\#}\left(h^s\right) - R^s\left(h^s\right)\right| \\ = & \left|R^{\#}\left(h^s\right) - R^{\#,s}\left(h^s\right) + R^{\#,s}\left(h^s\right) - R^s\left(h^s\right)\right| \\ \leq & \left|R^{\#}\left(h^s\right) - R^{\#,s}\left(h^s\right)\right| + \left|R^{\#,s}\left(h^s\right) - R^s\left(h^s\right)\right|.\end{aligned}$$

We know that the first term can be bounded as

$$\left|R^{\#}\left(h^s\right) - R^{\#,s}\left(h^s\right)\right| \leq M\mathbb{E}_{\mathbb{P}^t}\left[\left\|\Delta p\left(\cdot \mid \mathbf{x}\right)\right\|_1\right].$$

We manipulate the second term as

$$\begin{aligned}\left|R^{\#,s}\left(h^s\right) - R^s\left(h^s\right)\right| &= \left|\int \ell\left(h^s\left(\mathbf{x}\right), y\right) p^s\left(y \mid \mathbf{x}\right)\left[p^{\#}\left(\mathbf{x}\right) - p^s\left(\mathbf{x}\right)\right] d\mathbf{x} dy\right| \\ &= \left|\sum_{y=1}^{C} \int \ell\left(h^s\left(\mathbf{x}\right), y\right) p^s\left(y \mid \mathbf{x}\right)\left(p^{\#}\left(\mathbf{x}\right) - p^s\left(\mathbf{x}\right)\right) d\mathbf{x}\right| \\ &\leq \sum_{y=1}^{C}\left|\int \ell\left(h^s\left(\mathbf{x}\right), y\right) p^s\left(y \mid \mathbf{x}\right) p^{\#}\left(\mathbf{x}\right) d\mathbf{x} - \int \ell\left(h^s\left(\mathbf{x}\right), y\right) p^s\left(y \mid \mathbf{x}\right) p^s\left(\mathbf{x}\right) d\mathbf{x}\right| \\ &= \sum_{y=1}^{C}\left|\int \ell\left(h^s\left(\mathbf{x}^{\#}\right), y\right) p^s\left(y \mid \mathbf{x}^{\#}\right) d\mathbb{P}^{\#}\left(\mathbf{x}^{\#}\right) - \int \ell\left(h^s\left(\mathbf{x}^s\right), y\right) p^s\left(y \mid \mathbf{x}^s\right) d\mathbb{P}^s\left(\mathbf{x}^s\right)\right| \\ &= \sum_{y=1}^{C}\left|\int\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right), y\right) p^s\left(y \mid \mathbf{x}^{\#}\right) - \ell\left(h^s\left(\mathbf{x}^s\right), y\right) p^s\left(y \mid \mathbf{x}^s\right)\right] d\gamma^*\left(\mathbf{x}^{\#}, \mathbf{x}^s\right)\right| \\ &\leq \sum_{y=1}^{C}\left(\left|\int_A\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right), y\right) p^s\left(y \mid \mathbf{x}^{\#}\right) - \ell\left(h^s\left(\mathbf{x}^s\right), y\right) p^s\left(y \mid \mathbf{x}^s\right)\right] d\gamma^*\left(\mathbf{x}^{\#}, \mathbf{x}^s\right)\right|\right. \\ &\left. + \left|\int_{A^c}\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right), y\right) p^s\left(y \mid \mathbf{x}^{\#}\right) - \ell\left(h^s\left(\mathbf{x}^s\right), y\right) p^s\left(y \mid \mathbf{x}^s\right)\right] d\gamma^*\left(\mathbf{x}^{\#}, \mathbf{x}^s\right)\right|\right),\end{aligned}$$

where we denote $A = \left\{\left(\mathbf{x}^{\#}, \mathbf{x}^s\right) : \left\|h\left(\mathbf{x}^{\#}\right) - h\left(\mathbf{x}^s\right)\right\| \leq \lambda c\left(\mathbf{x}^{\#}, \mathbf{x}^s\right)\right\}$, hence $\gamma^*\left(A^c\right) \leq \phi\left(\lambda\right)$.

We manipulate the second term as:

$$
\begin{aligned}
&\sum_{y=1}^{C}\left|\int_{A^c}\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)p^s\left(y\mid\mathbf{x}^{\#}\right)-\ell\left(h^s\left(\mathbf{x}^s\right),y\right)p^s\left(y\mid\mathbf{x}^s\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)\right|\\
&\leq\sum_{y=1}^{C}\int_{A^c}\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)p^s\left(y\mid\mathbf{x}^{\#}\right)+\ell\left(h^s\left(\mathbf{x}^s\right),y\right)p^s\left(y\mid\mathbf{x}^s\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)\\
&\leq M\sum_{y=1}^{C}\int_{A^c}\left[p\left(y\mid\mathbf{x}^{\#}\right)+p^s\left(y\mid\mathbf{x}^s\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)\\
&=M\int_{A^c}\left[\sum_{y=1}^{C}p\left(y\mid\mathbf{x}^{\#}\right)+\sum_{y=1}^{C}p^s\left(y\mid\mathbf{x}^s\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)\\
&=2M\gamma^*\left(A^c\right)\leq 2M\phi\left(\lambda\right).
\end{aligned}
$$

We derive the first term as:

$$
\begin{aligned}
U=&\sum_{y=1}^{C}\left|\int_{A}\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)p^s\left(y\mid\mathbf{x}^{\#}\right)-\ell\left(h^s\left(\mathbf{x}^s\right),y\right)p^s\left(y\mid\mathbf{x}^s\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)\right|\\
=&\sum_{y=1}^{C}\left|\int_{A}\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)p^s\left(y\mid\mathbf{x}^{\#}\right)d\mathbb{P}^{\#}\left(\mathbf{x}^{\#}\right)-\int_{A}\ell\left(h^s\left(\mathbf{x}^s\right),y\right)p^s\left(y\mid\mathbf{x}^s\right)d\mathbb{P}^s\left(\mathbf{x}^s\right)\right|\\
=&\sum_{y=1}^{C}\left|\int_{A}\ell\left(h^s\left(\mathbf{x}\right),y\right)p^s\left(y\mid\mathbf{x}\right)p^{\#}\left(\mathbf{x}\right)d\mathbf{x}-\int_{A}\ell\left(h^s\left(\mathbf{x}\right),y\right)p^s\left(y\mid\mathbf{x}\right)p^s\left(\mathbf{x}\right)d\mathbf{x}\right|\\
\leq&\sum_{y=1}^{C}\int_{A}\ell\left(h^s\left(\mathbf{x}\right),y\right)p^s\left(y\mid\mathbf{x}\right)\left|p^{\#}\left(\mathbf{x}\right)-p^s\left(\mathbf{x}\right)\right|d\mathbf{x}\\
\leq&\sum_{y=1}^{C}\int_{A}\ell\left(h^s\left(\mathbf{x}\right),y\right)\left|p^{\#}\left(\mathbf{x}\right)-p^s\left(\mathbf{x}\right)\right|d\mathbf{x}
\end{aligned}
$$

Denote $A_1=\left\{\mathbf{x}\in A:p^{\#}\left(\mathbf{x}\right)-p^s\left(\mathbf{x}\right)\geq 0\right\}$, we then have

$$
\begin{aligned}
U&\leq\sum_{y=1}^{C}\int_{A}\ell\left(h^s\left(\mathbf{x}\right),y\right)\left|p^{\#}\left(\mathbf{x}\right)-p^s\left(\mathbf{x}\right)\right|d\mathbf{x}\\
&=\sum_{y=1}^{C}\left[\int_{A_1}\ell\left(h^s\left(\mathbf{x}\right),y\right)\left(p^{\#}\left(\mathbf{x}\right)-p^s\left(\mathbf{x}\right)\right)d\mathbf{x}+\int_{A\setminus A_1}\ell\left(h^s\left(\mathbf{x}\right),y\right)\left(p^s\left(\mathbf{x}\right)-p^{\#}\left(\mathbf{x}\right)\right)d\mathbf{x}\right]\\
&=\sum_{y=1}^{C}\left[\int_{A_1}\left[\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)-\ell\left(h^s\left(\mathbf{x}^s\right),y\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)+\int_{A\setminus A_1}\left[\ell\left(h^s\left(\mathbf{x}^s\right),y\right)-\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)\right]d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right)\right]\\
&\leq\sum_{y=1}^{C}\int_{A}\left|\ell\left(h^s\left(\mathbf{x}^{\#}\right),y\right)-\ell\left(h^s\left(\mathbf{x}^s\right),y\right)\right|d\gamma^*\left(\mathbf{x}^{\#},\mathbf{x}^s\right).
\end{aligned}
$$

$$\begin{aligned}
U &\leq \sum_{y=1}^{C}\int_{A}\left|\ell\left(h^{s}\left(\mathbf{x}^{\#}\right),y\right)-\ell\left(h^{s}\left(\mathbf{x}^{s}\right),y\right)\right|d\gamma^{*}\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)\\
&\overset{(1)}{\leq} k\sum_{y=1}^{C}\int_{A}\left\Vert h^{s}\left(\mathbf{x}^{\#}\right)-h^{s}\left(\mathbf{x}^{s}\right)\right\Vert_{1}d\gamma^{*}\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)\\
&= kC\int_{A}\left\Vert h^{s}\left(\mathbf{x}^{\#}\right)-h^{s}\left(\mathbf{x}^{s}\right)\right\Vert_{1}d\gamma^{*}\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)\\
&\overset{(2)}{\leq} \lambda kC\int_{A}c\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)d\gamma^{*}\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)\leq\lambda kC\int c\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)d\gamma^{*}\left(\mathbf{x}^{\#},\mathbf{x}^{s}\right)=\lambda kCW_{c}\left(\mathbb{P}^{\#},\mathbb{P}^{s}\right)\\
&\overset{(3)}{\leq} \lambda kCW_{c,p}\left(\mathbb{P}^{\#},\mathbb{P}^{s}\right).
\end{aligned}$$

Here we note that we have (1) due to ℓ is k-Lipschitz w.r.t $\left\Vert \cdot\right\Vert$, (2) due to the definition of A, and (3) due to $p\geq1$ hence $W_{c}\left(\mathbb{P}^{\#},\mathbb{P}^{s}\right)\leq W_{c,p}\left(\mathbb{P}^{\#},\mathbb{P}^{s}\right)$ (see Section 5.1 in [26]). □

2.2 Data shift via Wasserstein metric

Let $\mathcal{Z}$ be an intermediate space (i.e., the joint space $\mathcal{Z}=\mathbb{R}^{m}$). We consider the composite mappings: $T\left(\mathbf{x}\right)=T^{2}\left(T^{1}\left(\mathbf{x}\right)\right)$ where T^{1} is a mapping from the target domain $\mathcal{X}^{t}$ to the joint space $\mathcal{Z}$ and T^{2} maps from the joint space $\mathcal{Z}$ to the source domain $\mathcal{X}^{s}$ (note that if $\mathcal{Z}=\mathcal{X}^{s}$ then $T^{2}=id$ is the identity function). Based on this structure, we consider the following optimization problem:

$$\min_{T^{1},T^{2}}W_{c,p}\left(\left(T^{2}\circ T^{1}\right)_{\#}\mathbb{P}^{t},\mathbb{P}^{s}\right). \tag{3}$$

In the following theorem, we demonstrate that the above optimization problem can be equivalently transformed into another form involving the joint space (see Figure 1 for an illustration of that theorem).

Theorem 9. ***(Theorem 4 in the main paper)*** *The optimal objective value of the OP* (3) *is equal to that of the OP (4), that is*

$$\begin{aligned}
&\min_{T^{1},T^{2}}\mathrm{W}_{c,p}\left(\left(T^{2}\circ T^{1}\right)_{\#}\mathbb{P}^{t},\mathbb{P}^{s}\right)=\\
&\min_{T^{1},T^{2}}\min_{G^{1}:T_{\#}^{1}\mathbb{P}^{t}=G_{\#}^{1}\mathbb{P}^{s}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[c\left(\mathbf{x},T^{2}\left(G^{1}\left(\mathbf{x}\right)\right)\right)^{p}\right]^{1/p}
\end{aligned} \tag{4}$$

where G^{1} is a map from $\mathcal{X}^{s}$ to $\mathcal{Z}$.

Proof. From the definition of Wasserstein metric, we obtain that

$$W_{c,p}\left(\left(T^{2}\circ T^{1}\right)_{\#}\mathbb{P}^{t},\mathbb{P}^{s}\right)=\inf_{L:L_{\#}\mathbb{P}^{s}=T_{\#}\mathbb{P}^{t}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[c\left(\mathbf{x},L\left(\mathbf{x}\right)\right)^{p}\right]^{1/p}.$$

Therefore, we can rewrite the optimization problem in the left side of (4) as follows:

$$\min_{T^{1},T^{2}}W_{c,p}\left(\left(T^{2}\circ T^{1}\right)_{\#}\mathbb{P}^{t},\mathbb{P}^{s}\right)=\min_{T^{1},T^{2}}\inf_{L:L_{\#}\mathbb{P}^{s}=T_{\#}\mathbb{P}^{t}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[c\left(\mathbf{x},L\left(\mathbf{x}\right)\right)^{p}\right]^{1/p}.$$

We first prove that

$$\min_{T^{1},T^{2}}\min_{G^{1}:T_{\#}^{1}\mathbb{P}^{t}=G_{\#}^{1}\mathbb{P}^{s}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[c\left(\mathbf{x},T^{2}\left(G^{1}\left(\mathbf{x}\right)\right)\right)^{p}\right]^{1/p}\geq\min_{T^{1},T^{2}}W_{c,p}\left(\left(T^{2}\circ T^{1}\right)_{\#}\mathbb{P}^{t},\mathbb{P}^{s}\right).$$

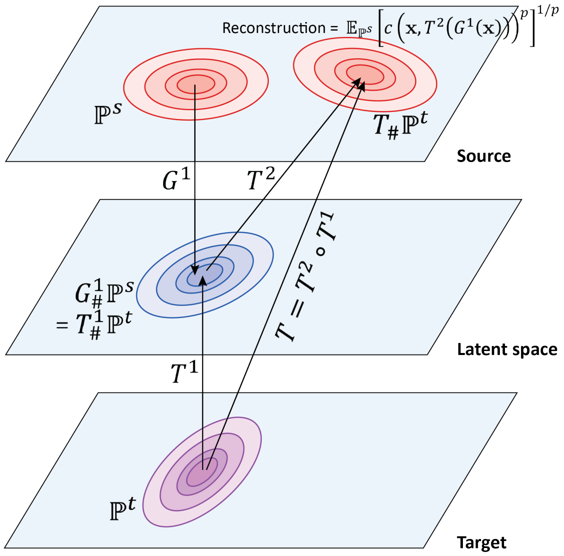

Figure 1: The mapping $T = T^2 \circ T^1$ maps from the target to source domains. We minimize $D\left(G^1_\#\mathbb{P}^s, T^1_\#\mathbb{P}^t\right)$ to close the discrepancy gap of the source and target domains in the joint space.

Given the mappings T^1, T^2, for any mapping G^1 satisfying the equation $T^1_\#\mathbb{P}^t = G^1_\#\mathbb{P}^s$, we let $T' = T^2 \circ G^1$. Then, we arrive at $T'_\#\mathbb{P}^s = T_\#\mathbb{P}^t$. Hence, we find that

$$\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p} = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T'\left(\mathbf{x}\right)\right)^p\right]^{1/p} \geq \inf_{L:L_\#\mathbb{P}^s = T_\#\mathbb{P}^t} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p}.$$

The above inequality directly leads to

$$\min_{G^1:T^1_\#\mathbb{P}^t = G^1_\#\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p} \geq \inf_{L:L_\#\mathbb{P}^s = T_\#\mathbb{P}^t} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p}.$$

As a consequence, we achieve the following inequality

$$\begin{aligned}\min_{T^1,T^2} \min_{G^1:T^1_\#\mathbb{P}^t = G^1_\#\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p} &\geq \min_{T^1,T^2} \inf_{L:L_\#\mathbb{P}^s = T_\#\mathbb{P}^t} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} \\ &= \min_{T^1,T^2} W_{c,p}\left(\left(T^2 \circ T^1\right)_\# \mathbb{P}^t, \mathbb{P}^s\right).\end{aligned}$$

We now prove that

$$\min_{T^1,T^2} W_{c,p}\left(\left(T^2 \circ T^1\right)_\# \mathbb{P}^t, \mathbb{P}^s\right) \geq \min_{T^1,T^2} \min_{G^1:T^1_\#\mathbb{P}^t = G^1_\#\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.$$

Given the mapping T^1, we consider the distribution $\mathbb{Q}$ over the source domain such that there exists a map T^2 for which $T^2_\#\left(T^1_\#\mathbb{P}^t\right) = \mathbb{Q}$. For any mapping L satisfying the equation $L_\#\mathbb{P}^s = \mathbb{Q}$, we can find mappings U, V such that $U_\#\mathbb{P}^s = T^1_\#\mathbb{P}^t$ and $L = V \circ U$. To this end, there exists a bijective mapping V satisfying $V_{\#}\left(T^1_{\#}\mathbb{P}^t\right) = \mathbb{Q}$ since these two distributions are atomless (see Theorem 1). Additionally, we can set $U = V^{-1} \circ L$. It is obvious that $U_{\#}\mathbb{P}^s = T^1_{\#}\mathbb{P}^t$ and $L = V \circ U$ from the definitions of U and V. Therefore, we have that

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} =& \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, V\left(U\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p} \\ &\geq \min_{T^2: T^2_{\#}\left(T^1_{\#}\mathbb{P}^t\right)=\mathbb{Q}} \ \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.\end{aligned}$$

Invoking the above equality, we find that

$$\inf_{L: L_{\#}\mathbb{P}^s=\mathbb{Q}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} \geq \min_{T^2: T^2_{\#}\left(T^1_{\#}\mathbb{P}^t\right)=\mathbb{Q}} \ \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.$$

With that inequality, we directly achieve the following inequality

$$\begin{aligned}&\min_{\mathbb{Q}} \inf_{L: L_{\#}\mathbb{P}^s=\mathbb{Q}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} \\ &\geq \min_{\mathbb{Q}} \min_{T^2: T^2_{\#}\left(T^1_{\#}\mathbb{P}^t\right)=\mathbb{Q}} \ \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.\end{aligned}$$

$$\min_{T^2} \inf_{L: L_{\#}\mathbb{P}^s=T_{\#}\mathbb{P}^t} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} \geq \min_{T^2} \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.$$

Note that from the definitions of $\mathbb{Q}$ and T^2, it is obvious that

$$\min_{\mathbb{Q}} \inf_{L: L_{\#}\mathbb{P}^s=\mathbb{Q}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} = \min_{T^2} \inf_{L: L_{\#}\mathbb{P}^s=T_{\#}\mathbb{P}^t} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p}.$$

$$\begin{aligned}&\min_{T^2} \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p} \\ &= \min_{\mathbb{Q}} \min_{T^2: T^2_{\#}\left(T^1_{\#}\mathbb{P}^t\right)=\mathbb{Q}} \ \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.\end{aligned}$$

By varying the mapping T^1 in both sides of the above inequality, we arrive at the following inequality

$$\min_{T^1, T^2} \inf_{L: L_{\#}\mathbb{P}^s=T_{\#}\mathbb{P}^t} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, L\left(\mathbf{x}\right)\right)^p\right]^{1/p} \geq \min_{T^1, T^2} \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.$$

Hence, we obtain that

$$\min_{T^1, T^2} W_{c,p}\left(\left(T^2 \circ T^1\right)_{\#}\mathbb{P}^t, \mathbb{P}^s\right) \geq \min_{T^1, T^2} \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.$$

Finally, we reach the conclusion as:

$$\min_{T^1, T^2} W_{c,p}\left(\left(T^2 \circ T^1\right)_{\#}\mathbb{P}^t, \mathbb{P}^s\right) = \min_{T^1, T^2} \min_{G^1: T^1_{\#}\mathbb{P}^t = G^1_{\#}\mathbb{P}^s} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x}, T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}.$$

□

It is interesting to interpret G^1 and T^1 as two generators that map the source and target domains to the common joint space $\mathcal{Z}$ respectively. The constraint $T^1_\#\mathbb{P}^t = G^1_\#\mathbb{P}^s$ further indicates that the gap between the source and target distributions is closed in the joint space via two generators G^1 and T^1. Furthermore, T^2 maps from the joint space to the source domain and aims to reconstruct G^1. Similar to [29], we do relaxation and arrive at the optimization problem:

$$\min_{T^1,T^2,G^1}\left(\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[c\left(\mathbf{x},T^2\left(G^1\left(\mathbf{x}\right)\right)\right)^p\right]^{1/p}+\alpha D\left(G^1_\#\mathbb{P}^s,T^1_\#\mathbb{P}^t\right)\right),\tag{5}$$

where $D\left(\cdot,\cdot\right)$ specifies a divergence between two distributions over the joint space and $\alpha > 0$.

2.3 Label shift via Wasserstein metric

Since G^1 and T^1 are two maps from the source and target domains to the joint space, we can further define two source and target supervisor distributions on the joint space as $p^{\#,s}\left(y\mid G^1\left(\mathbf{x}\right)\right) = p^s\left(y\mid\mathbf{x}\right)$ and $p^{\#,t}\left(y\mid T^1\left(\mathbf{x}\right)\right) = p^t\left(y\mid\mathbf{x}\right)$. With respect to the joint space, the second term of the upper bound in Theorem 4 can be rewritten as in the following corollary.

Corollary 10. ***(Corollary 5 in the main paper)*** *The second term of the upper bound in Theorem 4 can be rewritten as*

$$\mathbb{E}_{\mathbb{P}^t}\left[\left\|p^{\#,s}\left(y\mid G^1\left(T^2\left(T^1\left(\mathbf{x}\right)\right)\right)\right)-p^{\#,t}\left(y\mid T^1\left(\mathbf{x}\right)\right)\right\|_1\right].\tag{6}$$

Proof. The proof is trivial from the definitions of $p^{\#,s}\left(y\mid G^1\left(\mathbf{x}\right)\right) = p^s\left(y\mid\mathbf{x}\right)$ and $p^{\#,t}\left(y\mid T^1\left(\mathbf{x}\right)\right) = p^t\left(y\mid\mathbf{x}\right)$. □

Corollary 11. ***(Corollary 6 in the main paper)*** *Under the ideal scenario, the label mismatch term in (6) has a lower-bound*

$$\left\|\left[p^s\left(y=i\right)-p^t\left(y=i\right)\right]_{i=1}^C\right\|_1.$$

Proof. Under the ideal scenario, the label mismatch term becomes

$$\mathbb{E}_{\mathbb{P}^t}\left[\left\|p^{\#,s}\left(y\mid\left(T^1\left(\mathbf{x}\right)\right)\right)-p^{\#,t}\left(y\mid T^1\left(\mathbf{x}\right)\right)\right\|_1\right].$$

We derive as follows:

$$\begin{aligned}
U =&\mathbb{E}_{\mathbb{P}^t}\left[\left\|p^{\#,s}\left(y\mid\left(T^1\left(\mathbf{x}\right)\right)\right)-p^{\#,t}\left(y\mid T^1\left(\mathbf{x}\right)\right)\right\|_1\right]\\
=&\sum_{i=1}^C\int\left|p^{\#,s}\left(y=i\mid\left(T^1\left(\mathbf{x}\right)\right)\right)-p^{\#,t}\left(y=i\mid T^1\left(\mathbf{x}\right)\right)\right|p^t\left(\mathbf{x}\right)d\mathbf{x}\\
\geq&\sum_{i=1}^C\left|\int\left(p^{\#,s}\left(y=i\mid\left(T^1\left(\mathbf{x}\right)\right)\right)-p^{\#,t}\left(y=i\mid T^1\left(\mathbf{x}\right)\right)\right)p^t\left(\mathbf{x}\right)d\mathbf{x}\right|\\
=&\sum_{i=1}^C\left|\int p^{\#,s}\left(y=i\mid\left(T^1\left(\mathbf{x}\right)\right)\right)p^t\left(\mathbf{x}\right)d\mathbf{x}-\int p^{\#,t}\left(y=i\mid T^1\left(\mathbf{x}\right)\right)p^t\left(\mathbf{x}\right)d\mathbf{x}\right|\\
=&\sum_{i=1}^C\left|\int p^{\#,s}\left(y=i\mid\left(T^1\left(\mathbf{x}\right)\right)\right)d\mathbb{P}^t-\int p^{\#,t}\left(y=i\mid T^1\left(\mathbf{x}\right)\right)d\mathbb{P}^t\right|\\
=&\sum_{i=1}^C\left|\int p^{\#,s}\left(y=i\mid\mathbf{z}\right)dT^1_\#\mathbb{P}^t-\int p^{\#,t}\left(y=i\mid T^1\left(\mathbf{x}\right)\right)d\mathbb{P}^t\right|\\
=&\sum_{i=1}^C\left|\int p^{\#,s}\left(y=i\mid\mathbf{z}\right)dG^1_\#\mathbb{P}^s-\int p^{\#,t}\left(y=i\mid T^1\left(\mathbf{x}\right)\right)d\mathbb{P}^t\right|
\end{aligned}$$

$$
\begin{aligned}
U &\geq \sum_{i=1}^{C} \left| \int p^{\#,s}\left(y=i \mid \mathbf{z}\right) dG^{1}_{\#}\mathbb{P}^{s} - \int p^{\#,t}\left(y=i \mid T^{1}\left(\mathbf{x}\right)\right) d\mathbb{P}^{t} \right| \\
&= \sum_{i=1}^{C} \left| \int p^{\#,s}\left(y=i \mid G^{1}\left(\mathbf{x}\right)\right) d\mathbb{P}^{s} - \int p^{\#,t}\left(y=i \mid T^{1}\left(\mathbf{x}\right)\right) d\mathbb{P}^{t} \right| \\
&= \sum_{i=1}^{C} \left| \int p^{s}\left(y=i \mid \mathbf{x}\right) d\mathbb{P}^{s} - \int p^{t}\left(y=i \mid \mathbf{x}\right) d\mathbb{P}^{t} \right| \\
&= \sum_{i=1}^{C} \left| \int p^{s}\left(y=i \mid \mathbf{x}\right) p^{s}\left(\mathbf{x}\right) d\mathbf{x} - \int p^{t}\left(y=i \mid \mathbf{x}\right) p^{t}\left(\mathbf{x}\right) d\mathbf{x} \right| \\
&= \sum_{i=1}^{C} \left| p^{s}\left(y=i\right) - p^{t}\left(y=i\right) \right| \\
&= \left\| \left[p^{s}\left(y=i\right) - p^{t}\left(y=i\right) \right]_{i=1}^{C} \right\|_{1}.
\end{aligned}
$$

□

It is also worth mentioning that with regard to the latent space and the above equipment for $T = T^{2} \circ T^{1}$, we have the following formulations for the source classifier (i.e., h^{s}) and target classifier (i.e., h^{t}) now become:

$$
h^{s}\left(\mathbf{x}\right) = \mathcal{A}\left(G^{1}\left(\mathbf{x}\right)\right) \text{ and } h^{t}\left(\mathbf{x}\right) = \mathcal{A}\left(G^{1}\left(T\left(\mathbf{x}\right)\right)\right). \tag{7}
$$

We define a new metric $\tilde{c}$ w.r.t. the family $\mathcal{H}^{a}$ of the classifier $\mathcal{A}$ as:

$$
\tilde{c}\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right) = \sup_{\mathcal{A} \in \mathcal{H}^{a}} \left\| \mathcal{A}\left(\mathbf{z}_{1}\right) - \mathcal{A}\left(\mathbf{z}_{2}\right) \right\|_{1},
$$

where $\mathbf{z}_{1}$ and $\mathbf{z}_{2}$ lie on the latent space. The following lemma states under which conditions, $\tilde{c}$ is a proper metric on the latent space.

Lemma 12. ***(Lemma 7 in the main paper)*** *For any $\mathbf{z}_{1}$ and $\mathbf{z}_{2}$, if $\mathcal{A}\left(\mathbf{z}_{1}\right) = \mathcal{A}\left(\mathbf{z}_{2}\right), \forall \mathcal{A} \in \mathcal{H}^{a}$ leads to $\mathbf{z}_{1} = \mathbf{z}_{2}$, $\tilde{c}$ is a proper metric.*

Proof. First, $\tilde{c}\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right) \geq 0$ and $\tilde{c}\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right) = 0$ means $\mathcal{A}\left(\mathbf{z}_{1}\right) = \mathcal{A}\left(\mathbf{z}_{2}\right), \forall \mathcal{A} \in \mathcal{H}^{a}$, which leads to $\mathbf{z}_{1} = \mathbf{z}_{2}$. Second, it is obvious that $\tilde{c}\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right) = \tilde{c}\left(\mathbf{z}_{2}, \mathbf{z}_{1}\right), \forall \mathbf{z}_{1}, \mathbf{z}_{2}$.

Given any $\mathbf{z}_{1}, \mathbf{z}_{2}, \mathbf{z}_{3}$, we have

$$
\begin{aligned}
\tilde{c}\left(\mathbf{z}_{1}, \mathbf{z}_{3}\right) &= \sup_{\mathcal{A} \in \mathcal{H}^{a}} \left\| \mathcal{A}\left(\mathbf{z}_{1}\right) - \mathcal{A}\left(\mathbf{z}_{3}\right) \right\|_{1} \leq \sup_{\mathcal{A} \in \mathcal{H}^{a}} \left(\left\| \mathcal{A}\left(\mathbf{z}_{1}\right) - \mathcal{A}\left(\mathbf{z}_{2}\right) \right\|_{1} + \left\| \mathcal{A}\left(\mathbf{z}_{2}\right) - \mathcal{A}\left(\mathbf{z}_{3}\right) \right\|_{1} \right) \\
&\leq \sup_{\mathcal{A} \in \mathcal{H}^{a}} \left(\left\| \mathcal{A}\left(\mathbf{z}_{1}\right) - \mathcal{A}\left(\mathbf{z}_{2}\right) \right\|_{1} \right) + \sup_{\mathcal{A} \in \mathcal{H}^{a}} \left(\left\| \mathcal{A}\left(\mathbf{z}_{2}\right) - \mathcal{A}\left(\mathbf{z}_{3}\right) \right\|_{1} \right) \\
&= \tilde{c}\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right) + \tilde{c}\left(\mathbf{z}_{2}, \mathbf{z}_{3}\right).
\end{aligned}
$$

Therefore, $\tilde{c}$ is a proper metric. □

It turns out that the necessary (also sufficient) condition in Lemma 12 is realistic and not hard to be satisfied (e.g., the family $\mathcal{H}^{a}$ contains any bijection). We now can define a WS distance $W_{\tilde{c},p}$ that involves in Theorem 14 whose proof needs the following lemma.

Lemma 13. *Let p^{h^s} be the density of the distribution $\mathbb{P}^{h^s}$ formed by pushing forward $\mathbb{P}^s$ via h^s and p^{h^t} be the density of the distribution $\mathbb{P}^{h^t}$ formed by pushing forward $\mathbb{P}^t$ via h^t. If $\gamma \in \Gamma\left(\mathbb{P}^{h^s}, \mathbb{P}^{h^t}\right)$, there exists $\gamma' \in \Gamma\left(\mathbb{P}^s, \mathbb{P}^t\right)$ such that $\left(h^s, h^t\right)_{\#} \gamma' = \gamma$.*

Proof. Let denote γ^s as the joint distribution of the samples $\left(\mathbf{x}^s, h^s\left(\mathbf{x}^s\right)\right)$ where $\mathbf{x}^s \sim \mathbb{P}^s$ and γ^t as the joint distribution of the samples $\left(\mathbf{x}^t, h^t\left(\mathbf{x}^t\right)\right)$ where $\mathbf{x}^t \sim \mathbb{P}^t$. It is obvious that γ^s is a joint distribution of $\mathbb{P}^s$ and $\mathbb{P}^{h^s}$ and γ^t is a joint distribution of $\mathbb{P}^t$ and $\mathbb{P}^{h^t}$. According to the gluing lemma (see Lemma 5.5 in [26]), there exists a joint distribution μ such that for any draw $\left(\mathbf{x}^s, \boldsymbol{\tau}^s, \boldsymbol{\tau}^t, \mathbf{x}^t\right) \sim \mu$ then $\left(\mathbf{x}^s, \boldsymbol{\tau}^s\right) \sim \gamma^s$, $\left(\boldsymbol{\tau}^s, \boldsymbol{\tau}^t\right) \sim \gamma$, and $\left(\mathbf{x}^t, \boldsymbol{\tau}^t\right) \sim \gamma^t$.

Let γ' be the distribution of samples $\left(\mathbf{x}^s, \mathbf{x}^t\right)$ (i.e., the projection of μ onto the first and fourth dimensions). This follows that γ' is a joint distribution of $\mathbb{P}^s$ and $\mathbb{P}^t$ (i.e., $\gamma' \in \Gamma\left(\mathbb{P}^s, \mathbb{P}^t\right)$). In addition, since $\left(\mathbf{x}^s, \boldsymbol{\tau}^s\right) \sim \gamma^s$, $\boldsymbol{\tau}^s = h^s\left(\mathbf{x}^s\right)$, since $\left(\mathbf{x}^t, \boldsymbol{\tau}^t\right) \sim \gamma^t$, $\boldsymbol{\tau}^t = h^t\left(\mathbf{x}^t\right)$, and $\left(\boldsymbol{\tau}^s, \boldsymbol{\tau}^t\right) \sim \gamma$. Therefore, we reach $\left(h^s, h^t\right)_{\#} \gamma' = \gamma$.

We note that in the above proof, we employ a general form of the gluing lemma for 4 distributions and spaces. The proof is mainly based on the gluing lemma for 3 distributions and spaces and trivial. □

Theorem 14. ***(Theorem 8 in the main paper)*** *If $\tilde{c}$ is a proper metric and $p \geq 1$, the quantity $\left\|\left[p^s\left(y=i\right) - p^t\left(y=i\right)\right]_{i=1}^C\right\|_1$ has the upper-bounds:*

i) $R_1^s\left(h^s\right) + R_1^t\left(h^t\right) + W_{\tilde{c},p}\left(G^1_{\#}\mathbb{P}^s, T^1_{\#}\mathbb{P}^t\right)$ if $h^s := \mathcal{A}\left(G^1\left(\mathbf{x}\right)\right)$ and $h^t := \mathcal{A}\left(T^1\left(\mathbf{x}\right)\right)$.

ii) $R_1^s\left(h^s\right) + R_1^t\left(h^t\right) + W_{\tilde{c},p}\left(G^1_{\#}\mathbb{P}^s, T^1_{\#}\mathbb{P}^t\right) + W_{\tilde{c},p}\left(L_{\#}\mathbb{P}^t, T^1_{\#}\mathbb{P}^t\right)$ where $L := T \circ G^1$, and h^s and h^t are defined in (7).

Here $R_1^s\left(h^s\right) := \int \left\|p^s\left(\cdot \mid \mathbf{x}\right) - h^s\left(\mathbf{x}\right)\right\|_1 p^s\left(\mathbf{x}\right) d\mathbf{x}$ and $R_1^t\left(h^t\right) := \int \left\|p^t\left(\cdot \mid \mathbf{x}\right) - h^t\left(\mathbf{x}\right)\right\|_1 p^t\left(\mathbf{x}\right) d\mathbf{x}$ are the general losses of h^s and h^t w.r.t. $\left\|\cdot\right\|_1$.

Proof. i) We derive as follows:

$$\left\|\left[p^s\left(y=i\right) - p^t\left(y=i\right)\right]_{i=1}^C\right\|_1 \leq \left\|\left[p^s\left(y=i\right) - p^{h^s}\left(y=i\right)\right]_{i=1}^C\right\|_1 + \left\|\left[p^{h^s}\left(y=i\right) - p^{h^t}\left(y=i\right)\right]_{i=1}^C\right\|_1$$
$$+ \left\|\left[p^{h^t}\left(y=i\right) - p^t\left(y=i\right)\right]_{i=1}^C\right\|_1,$$

where p^{h^s} is the density of the distribution $\mathbb{P}^{h^s}$ formed by pushing forward $\mathbb{P}^s$ via h^s and p^{h^t} is the density of the distribution $\mathbb{P}^{h^t}$ formed by pushing forward $\mathbb{P}^t$ via h^t.

We manipulate the first term as:

$$\left\|\left[p^s\left(y=i\right) - p^{h^s}\left(y=i\right)\right]_{i=1}^C\right\|_1 = \sum_{i=1}^C \left|p^s\left(y=i\right) - p^{h^s}\left(y=i\right)\right|$$
$$= \sum_{i=1}^C \left|\int \left(p^s\left(y=i,\mathbf{x}\right) - p^{h^s}\left(y=i,\mathbf{x}\right)\right) p^s\left(\mathbf{x}\right) d\mathbf{x}\right| = \sum_{i=1}^C \left|\int \left(p^s\left(y=i,\mathbf{x}\right) - h_i^s\left(\mathbf{x}\right)\right) p^s\left(\mathbf{x}\right) d\mathbf{x}\right|$$
$$\leq \int \sum_{i=1}^C \left|p^s\left(y=i,\mathbf{x}\right) - h_i^s\left(\mathbf{x}\right)\right| p^s\left(\mathbf{x}\right) d\mathbf{x} = \int \sum_{i=1}^C \left\|\left[p^s\left(y=i,\mathbf{x}\right) - h_i^s\left(\mathbf{x}\right)\right]_{i=1}^C\right\|_1 p^s\left(\mathbf{x}\right) d\mathbf{x} = R_1^s\left(h^s\right).$$

Similarly, we can bound the third term as:

$$\left\|\left[p^{h^t}\left(y=i\right) - p^t\left(y=i\right)\right]_{i=1}^C\right\|_1 \leq R_1^t\left(h^t\right).$$

To handle the second term $\left\|\left[p^{h^s}\left(y=i\right)-p^{h^t}\left(y=i\right)\right]_{i=1}^{C}\right\|_1$, we first prove that

$$\left\|\left[p^{h^s}\left(y=i\right)-p^{h^t}\left(y=i\right)\right]_{i=1}^{C}\right\|_1\leq W_1\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right),$$

where the WS w.r.t. the metric $\left\|\cdot\right\|_1$. Indeed, consider a joint distribution $\gamma\in\Gamma\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)$. According to Lemma 13, there exists $\gamma'\in\Gamma\left(\mathbb{P}^s,\mathbb{P}^t\right)$ such that $\left(h^s,h^t\right)_{\#}\gamma'=\gamma$, we then have

$$\begin{aligned}
&\mathbb{E}_{\left(\mathbf{y}^s,\mathbf{y}^t\right)\sim\gamma}\left[\left\|\mathbf{y}^s-\mathbf{y}^t\right\|_1\right]=\mathbb{E}_{\left(\mathbf{x}^s,\mathbf{x}^t\right)\sim\gamma'}\left[\left\|h^s\left(\mathbf{x}^s\right)-h^t\left(\mathbf{x}^t\right)\right\|_1\right]\\
=&\int\left\|h^s\left(\mathbf{x}^s\right)-h^t\left(\mathbf{x}^t\right)\right\|_1 d\gamma'\left(\mathbf{x}^s,\mathbf{x}^t\right)=\sum_{i=1}^{C}\int\left|h_i^s\left(\mathbf{x}^s\right)-h_i^t\left(\mathbf{x}^t\right)\right|d\gamma'\left(\mathbf{x}^s,\mathbf{x}^t\right)\\
\geq&\sum_{i=1}^{C}\left|\int\left(h_i^s\left(\mathbf{x}^s\right)-h_i^t\left(\mathbf{x}^t\right)\right)d\gamma'\left(\mathbf{x}^s,\mathbf{x}^t\right)\right|=\sum_{i=1}^{C}\left|\int h_i^s\left(\mathbf{x}^s\right)d\mathbb{P}^s\left(\mathbf{x}^s\right)-\int h_i^t\left(\mathbf{x}^t\right)d\mathbb{P}^t\left(\mathbf{x}^t\right)\right|\\
=&\sum_{i=1}^{C}\left|p^{h^s}\left(y=i\right)-p^{h^t}\left(y=i\right)\right|=\left\|\left[p^{h^s}\left(y=i\right)-p^{h^t}\left(y=i\right)\right]_{i=1}^{C}\right\|_1.
\end{aligned}$$

Therefore, we achieve

$$W_1\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)=\inf_{\gamma\in\Gamma\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)}\mathbb{E}_{\left(\mathbf{y}^s,\mathbf{y}^t\right)\sim\gamma}\left[\left\|\mathbf{y}^s-\mathbf{y}^t\right\|_1\right]\geq\left\|\left[p^{h^s}\left(y=i\right)-p^{h^t}\left(y=i\right)\right]_{i=1}^{C}\right\|_1.$$

We now need to prove that $W_1\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)\leq W_{\tilde{c},p}\left(G^1_{\#}\mathbb{P}^s,T^1_{\#}\mathbb{P}^t\right)$ $(p\geq1)$. Indeed, given any $\gamma'\in\Gamma\left(G^1_{\#}\mathbb{P}^s,T^1_{\#}\mathbb{P}^t\right)$, let denote $\gamma=\mathcal{A}_{\#}\gamma'$, then $\gamma\in\Gamma\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)$. We then have:

$$\begin{aligned}
&\mathbb{E}_{\left(\mathbf{y}^s,\mathbf{y}^t\right)\sim\gamma}\left[\left\|\mathbf{y}^s-\mathbf{y}^t\right\|_1\right]=\mathbb{E}_{\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\sim\gamma'}\left[\left\|\mathcal{A}\left(\boldsymbol{\tau}^s\right)-\mathcal{A}\left(\boldsymbol{\tau}^t\right)\right\|_1\right]\\
\leq&\mathbb{E}_{\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\sim\gamma'}\left[\tilde{c}\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\right].
\end{aligned}$$

This follows that

$$W_1\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)\leq\mathbb{E}_{\left(\mathbf{y}^s,\mathbf{y}^t\right)\sim\gamma}\left[\left\|\mathbf{y}^s-\mathbf{y}^t\right\|_1\right]\leq\mathbb{E}_{\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\sim\gamma'}\left[\tilde{c}\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\right],$$

which further implies

$$W_1\left(\mathbb{P}^{h^s},\mathbb{P}^{h^t}\right)\leq\inf_{\gamma'\in\Gamma\left(G^1_{\#}\mathbb{P}^s,T^1_{\#}\mathbb{P}^t\right)}\mathbb{E}_{\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\sim\gamma'}\left[\tilde{c}\left(\boldsymbol{\tau}^s,\boldsymbol{\tau}^t\right)\right]=W_{\tilde{c}}\left(G^1_{\#}\mathbb{P}^s,T^1_{\#}\mathbb{P}^t\right)\leq W_{\tilde{c},p}\left(G^1_{\#}\mathbb{P}^s,T^1_{\#}\mathbb{P}^t\right).$$

ii) Using the same derivation as in (i) in which T^1 is replaced by L, we achieve

$$\begin{aligned}
\left\|\left[p^s\left(y=i\right)-p^t\left(y=i\right)\right]_{i=1}^{C}\right\|_1\leq&R_1^s\left(h^s\right)+R_1^t\left(h^t\right)+W_{\tilde{c},p}\left(G^1_{\#}\mathbb{P}^s,L_{\#}\mathbb{P}^t\right)\\
\leq&R_1^s\left(h^s\right)+R_1^t\left(h^t\right)+W_{\tilde{c},p}\left(G^1_{\#}\mathbb{P}^s,T^1_{\#}\mathbb{P}^t\right)+W_{\tilde{c},p}\left(T^1_{\#}\mathbb{P}^t,L_{\#}\mathbb{P}^t\right).
\end{aligned}$$

□

We note that our proof is still applicable if we generalize $\left\|\cdot\right\|_1$ to any metric d in Δ_C, which can be decomposed $d\left(\mathbf{y}^s,\mathbf{y}^t\right)=\sum_{i=1}^{C}d_i\left(\mathbf{y}_i^s,\mathbf{y}_i^t\right)$.

3 Experiments

3.1 Ablation Study

3.1.1 Experiment on Synthetic Data

Synthetic Dataset for the Source and Target Domains

We generate two synthetic labeled datasets for the source and target domains. We generate the $10,000$ data examples of the source dataset from the mixture of two Gaussian distributions: $p^s(\mathbf{x}) = \pi_1^s \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_1^s, \Sigma_1^s) + \pi_2^s \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_2^s, \Sigma_2^s)$ where $\pi_1^s = \pi_2^s = \frac{1}{2}$, $\boldsymbol{\mu}_1^s = [1, 1, ..., 1] \in \mathbb{R}^{10}$, $\boldsymbol{\mu}_2^s = [2, 2, ..., 2] \in \mathbb{R}^{10}$ and $\Sigma_1^s = \Sigma_2^s = \mathbb{I}_{10}$. Similarly, we generate the another $10,000$ data examples of the target dataset from the mixture of two Gaussian distributions: $p^t(\mathbf{x}) = \pi_1^t \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_1^t, \Sigma_1^t) + \pi_2^t \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_2^t, \Sigma_2^t)$ where $\pi_1^t = \frac{1}{3}$, $\pi_2^t = \frac{2}{3}$, $\boldsymbol{\mu}_1^t = [4, 4, ..., 4] \in \mathbb{R}^{10}$, $\boldsymbol{\mu}_2^t = [5, 5, ..., 5] \in \mathbb{R}^{10}$ and $\Sigma_1^t = \Sigma_2^t = \mathbb{I}_{10}$. For each data example in the source and target domains, we assign label $y = 0$ if this data example is generated from the first Gauss and $y = 1$ if this data example is generated from the second Gauss using Bayes' s rule.

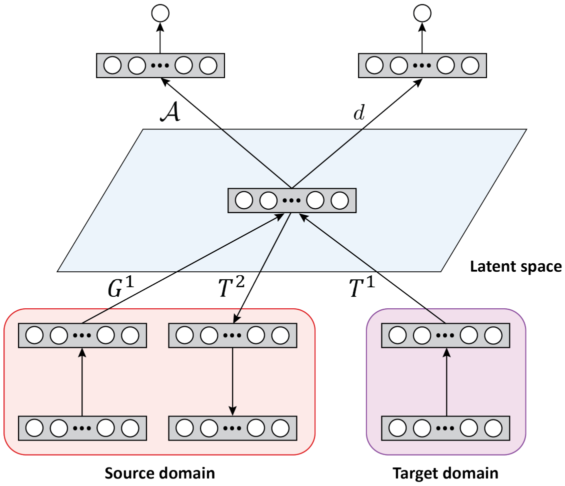

Figure 2: Architecture of networks for deep domain adaptation on the synthetic datasets.

Deep Domain Adaptation on the Synthetic Dataset

Figure 2 shows the architectures of networks used in our experiments on the synthetic datasets. Two generators G^1, T^1 with the same architectures $(10 \to 5\,(\text{ReLu}) \to 5\,(\text{ReLu}))$ map the source and target data to the intermediate joint layer. Note that different from other works in deep domain adaptation, we did not tie G^1 and T^1. The network T^2 with the architecture $(10 \to 5\,(\text{ReLu}) \to 5\,(\text{ReLu}))$ maps from the intermediate joint layer to the source and target domains respectively. To break the gap between the source and target domains in the joint layer, we employ GAN principle [13, 10] wherein we invoke a discriminator network d $(5 \to 5\,(\text{ReLu}) \to 1\,(\text{sigmoid}))$ to discriminate the source and target data examples in the joint space. The classifier network $\mathcal{A}$ $(5 \to 5\,(\text{ReLu}) \to 1\,(\text{sigmoid}))$ is employed to classify the labeled source data examples. To approximate the $0/1$ cost function, we use the modified sigmoid function [23]: $c_\gamma(\mathbf{x}, \mathbf{x}') = 2/\left[1 + \exp\left\{-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|_2\right\}\right] - 1$ with $\gamma = 100$. It can be seen that when $\gamma \to +\infty$, the cost function c_γ approaches the $0/1$ cost function. More specifically, we need to update $G^1, T^1, T^2, \mathcal{A}$, and d as

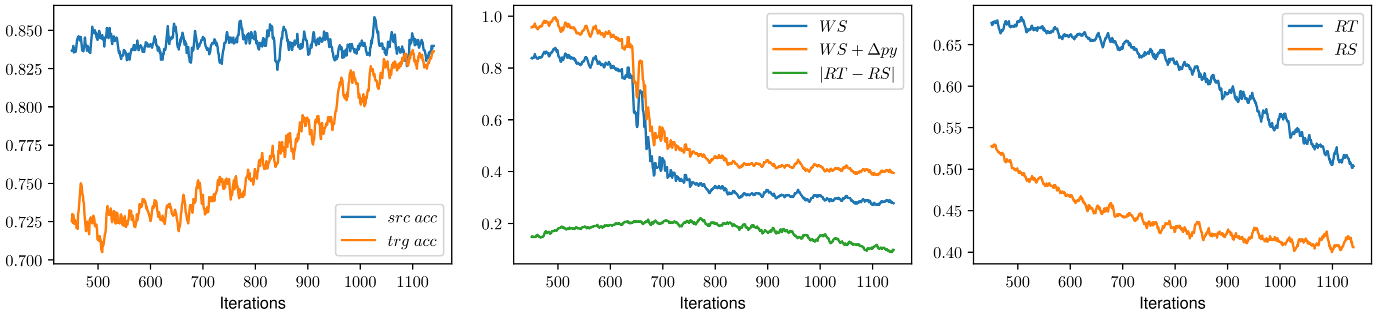

Figure 3: Left: the accuracies on the source and target datasets. Middle: the plots of three terms in Theorem 4. Right: the plot of empirical losses on the source and target datasets.

follows:

$$\left(G^{1},T^{1},T^{2},\mathcal{A}\right)=\underset{G^{1},T^{1},T^{2},\mathcal{A}}{\text{argmin}}\ \mathcal{I}\left(G^{1},T^{1},T^{2},\mathcal{A}\right)\text{ and }d=\underset{d}{\text{argmax}}\,\mathcal{J}\left(d\right),$$

where α is set to 0.1 and we have defined

$$\begin{aligned}\mathcal{I}\left(G^{1},T^{1},T^{2},\mathcal{A}\right)= & \\ & +\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[c_{\gamma}\left(\mathbf{x},T^{2}\left(G^{1}\left(\mathbf{x}\right)\right)\right)\right]+\mathbb{E}_{\left(\mathbf{x},y\right)\sim\mathcal{D}^{s}}\left[\ell\left(y,\mathcal{A}\left(G^{1}\left(\mathbf{x}\right)\right)\right)\right]\\ & +\alpha\left[\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[\log\left(d\left(G^{1}\left(\mathbf{x}\right)\right)\right)\right]+\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{t}}\left[\log\left(1-d\left(T^{1}\left(\mathbf{x}\right)\right)\right)\right]\right]\\ \mathcal{J}\left(d\right)= & \ \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{s}}\left[\log\left(d\left(G^{1}\left(\mathbf{x}\right)\right)\right)\right]+\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{t}}\left[\log\left(1-d\left(T^{1}\left(\mathbf{x}\right)\right)\right)\right].\end{aligned}$$

Based on the classifier $\mathcal{A}$ on the joint space, we can identify the corresponding hypotheses on the source and target domains as: $h^{s}\left(\mathbf{x}\right)=\mathcal{A}\left(G^{1}\left(\mathbf{x}\right)\right)$ and $h^{t}\left(\mathbf{x}\right)=\mathcal{A}\left(T^{1}\left(\mathbf{x}\right)\right)$.

Verification of Our Theory for Unsupervised Domain Adaptation

In this experiment, we assume that none of data example in the target domain has label. We measure three terms, namely $\left|R\left(h^{t}\right)-R\left(h^{s}\right)\right|$, $W\left(\mathbb{P}^{s},\mathbb{P}^{\#}\right)$ and $\mathbb{E}_{\mathbb{P}^{t}}\left[\left\Vert \Delta p\left(y\mid\mathbf{x}\right)\right\Vert _{1}\right]$ ($M=1$ since we are using the logistic loss) as defined in Theorem 4 across the training progress. Actually, we approximate $R\left(h^{t}\right),R\left(h^{s}\right)$ using the corresponding empirical losses. As shown in Figure 3 (middle), the green plot is always above the blue plot and this empirically confirms the inequality in Theorem 4. Furthermore, the fact that three terms consistently decrease across the training progress indicates an improvement when $\mathbb{P}^{\#}$ is shifting toward $\mathbb{P}^{s}$. This improvement is also reflected in Figure 3 (left and right) wherein the target accuracy and empirical loss gradually increase and decrease accordingly.

3.1.2 The Effect of Class Alignment in the Joint Space.

In this experiment, we inspect the influence of the harmony of two labeling assignment mechanisms to the predictive performance. In particular, we assume that a portion ($r=5\%,\ 15\%,\ 25\%,\ 50\%$) of the target domain has label and consider two settings: i) the labels of the target and source domains are totally properly matched in the joint space (i.e., 0 matches 0, 1 matches 1,..., and 9 matches 9) and ii) the labels of the target and source domains are totally improperly matches in the joint space (i.e., 0 matches 1, 1 matches 2,..., and 9 matches 0).

To push a specific labeled portion of the target domain to the corresponding label portion of the source domain in the joint space (the label i to i in the first setting and the label i to $\left(i+1\right)\bmod 10$ in the second setting for $i=0,1,\ldots,9$), we again make use of

GAN principle and employ additional discriminators to push the corresponding labeled portions together. Note that the parameters of the additional discriminators and the primary discriminator (used to push the target data toward source data in the joint space) are tied up to the penultimate layer.

It can be observed from Table 1 that for the case of proper matching, when increasing the ratio of labeled portion, we increase the chance to match the corresponding labeled portions properly, hence significantly improving the predictive performance. In contrast, for the case of improper matching, when increasing the ratio of labeled portion, we increase the chance to match the corresponding labeled portions improperly, hence significantly reducing the predictive performance.

Table 1: The variation of predictive performance in percentage as increasing the ratio of labeled portion when the labels of the target domain are properly or improperly matched to those in the source domain. Note that we emphasize in bold and italic/bold the best and worse performance.

	Proper match				**Improper match**				**Base**
r	5%	15%	25%	50%	5%	15%	25%	50%	0%
MNIST→MNIST-M	86.4	88.8	92.9	**93.2**	75.5	70.2	64.5	***58.4***	81.5
SVHN→MNIST	72.3	74.1	76.2	**77.5**	69.8	60.8	56.8	***56.4***	71.0

3.2 Experimental Setting for our LAMDA

3.2.1 The Objective Function of LAMDA

Note that in our implementation, to reduce the model complexity, we set $G^1 = T^1 = G$, $T^2 = T$, $S = \mathcal{A}$ (S is a transportation probability network which is shared weights with the classifier $\mathcal{A}$). Let us further denote:

$$\mathcal{L}_{\mathcal{A}} := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^s}\left[\ell\left(y,\mathcal{A}\left(G\left(\mathbf{x}\right)\right)\right)\right],$$

$$\begin{aligned}\mathcal{L}_g := {} & \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[\log\,\left(1-d_{C+1}\left(G(\mathbf{x})\right)\right)\right]+\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^t}\left[\log d_{C+1}\left(G\left(\mathbf{x}\right)\right)\right]\\ & +\alpha\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^t}\left[-\sum_{i=1}^{C}\mathcal{A}_i\left(\mathbf{x}\right)\log d_i\left(G\left(\mathbf{x}\right)\right)\right]+\beta R\left(T,G\right)+\mathcal{L}_{\mathcal{A}},\end{aligned}$$

where $R\left(T,G\right)$ is the reconstruction term defined as

$$R\left(T,G\right) := \mathbb{E}_{\mathbb{P}^s}\left[\left\Vert T\left(G\left(\mathbf{x}\right)\right)-\mathbf{x}\right\Vert_2^2\right].$$

$$\begin{aligned}\mathcal{L}_d := \sum_{i=1}^{C}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^s\wedge y=i}\left[\log d_i\left(G\left(\mathbf{x}\right)\right)\right] & +\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^t}\left[\log d_{C+1}\left(G\left(\mathbf{x}\right)\right)\right]\\ & +\mathbb{E}_{\mathbf{x}\sim\mathbb{P}^s}\left[\log\left(1-d_{C+1}\left(G\left(\mathbf{x}\right)\right)\right)\right].\end{aligned}$$

To update G, T and $\mathcal{A}$, we solve:

$$\min_{G,T,\mathcal{A}}\mathcal{L}_g.$$

To update d, we solve:

$$\max_{d}\mathcal{L}_d.$$

Table 2: The full experimental results in percent of our LAMDA and the baselines on digits, traffic sign, and natural image datasets.

Source	**MNIST**	**USPS**	**MNIST**	**SVHN**	**MNIST**	**DIGITS**	**SIGNS**	**CIFAR**	**STL**
Target	**USPS**	**MNIST**	**MNIST-M**	**MNIST**	**SVHN**	**SVHN**	**GTSRB**	**STL**	**CIFAR**
MMD [21]	-	-	76.9	71.1	-	88.0	91.1	-	-
DANN [10]	-	-	81.5	71.1	35.7	90.3	88.7	-	-
DRCN [11]	-	-	-	82.0	40.1	-	-	66.4	58.7
DSN [4]	-	-	83.2	82.7	-	91.2	93.1	-	-
kNN-Ad [27]	-	-	86.7	78.8	40.3	-	-	-	-
PixelDA [3]	-	-	98.2	-	-	-	-	-	-
ATT [24]	-	-	94.2	86.2	52.8	92.9	96.2	-	-
Π-model [9]	-	-	-	92.0	71.4	94.2	98.4	76.3	64.2
ADDA [30]	89.4	90.1	-	76.0	-	-	-	-	-
CyCADA [15]	95.6	96.5	-	90.4	-	-	-	-	-
MSTN [33]	92.9	97.6	-	91.7	-	-	-	-	-
CDAN [22]	95.6	98.0	-	89.2	-	-	-	-	-
MCD [25]	94.2	94.1	-	96.2	-	-	94.4	-	-
PFAN [5]	95.0	-	-	93.9	57.6	-	-	-	-
DADA [28]	96.1	96.5	-	95.6	-	-	-	-	-
DeepJDOT[7]	95.7	96.4	92.4	96.7	30.8	84.2	70.0	61.6	49.6
DASPOT [34]	97.5	96.5	94.9	96.2	-	-	-	-	-
GPDA [16]	96.5	96.4	-	98.2	-	-	96.2	-	-
SWD [18]	98.1	97.1	90.9	98.9	49.5	88.7	98.6	65.3	52.1
rRevGrad+CAT [8]	94.0	96.0	-	98.8	-	-	-	-	-
SHOT [20]	98.0	**98.4**	-	98.9	-	-	-	-	-
RWOT [35]	98.5	97.5	-	98.8	-	-	-	-	-
LAMDA	**99.5**	98.3	**98.4**	**99.5**	**82.1**	**95.9**	**99.2**	**78.0**	**71.6**

3.2.2 Experimental Datasets

Digit datasets

MNIST. The dataset is commonly used in domain adaptation literature. To adapt from MNIST to MNIST-M or SVHN, the MNIST images are replicated from single greyscale channel to obtain digit images which has three channels.
MNIST-M. Following by the implementation in [10], we generate the MNIST-M images by replacing the black background of MNIST images by the color ones and obtain the same number of training and test samples as the MNIST dataset.
SVHN. The dataset consists of images obtained by detecting house numbers from Google Street View images. This dataset is a benmark for recognizing digits and numbers in real-world images.
DIGITS. There are roughly 500,000 images are generated using various data augmentation schemes, i.e., varying the text, positioning, orientation, background, stroke color, and the amount of blur.

We compare our LAMDA with renown baselines especially OT-based ones (e.g., SWD [18], DeepJDOT [7], DASPOT [34], ETD [19] and RWOT [35]). As shown in Table 2, LAMDA outperforms other baselines on most of digit datasets. It is noticeable that although the transfer task MNIST→SVHN is extremely challenging in which the source dataset includes grayscale handwritten digits whereas the target dataset is created by real-world digits, our LAMDA is still capable of matching the gap between source and target domains and outperforms the second-best method by a sizeable margin (10.7%).

Traffic sign datasets

SIGNS. A synthetic dataset for traffic sign recognition. Images are collected from Wikipedia and then applied various types of transformations to generate 100,000 images for training and test.
GTSRB. Road sign images are extracted from videos recorded on different road types in Germany. We preprocess the data by croping out the region of interest of each image, and then scale them to a resolution of 32 × 32.

Natural scene datasets

CIFAR. The CIFAR-10 [17] dataset includes 50,000 training images and 10,000 test images. However, to adapt with STL dataset, we base on [9] to remove one non-overlapping class ("frog"). The numbers of training examples and test examples therefore are reduced to 45,000 and 9,000 respectively.
STL. Similar to CIFAR-10, we remove class named "monkey" to obtain a 9-class classification problem. Also, STL-10 images are down-scaled to a resolution of 96 × 96 to 32 × 32.

Object recognition datasets

Office-Home consists of roughly 15,500 images in a total of 65 object classes and belonging to 4 different domains: Artistic (**Ar**), Clip Art (**Cl**), Product (**Pr**) and Real-world (**Rw**).
Office-31 is a popular dataset for domain adaptation that contains 3 domains Amazon (**A**), Webcam (**W**), and DSLR (**D**). There are 31 common classes for all domains and the total number of images is 4,110.
ImageCLEF-DA contains three domains: Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**) and Pascal VOC 2012 (**P**). There are total 600 images in each domain and 12 common classes. We follow the work in [19] to evaluate 6 adaptation tasks.

We resize the resolution of each sample in *digits*, *traffic sign*, and *natural image* datasets to 32 × 32, and normalize the value of each pixel to the range of $[-1, 1]$. For *object recognition* datasets, we use features have 2048 dimensions extracted from ResNet-50 [14] pretrained on ImageNet.

Table 3: Small, medium and large network architecture of LAMDA. We use the small network for object recognition datasets, medium network for digits and traffic sign, and the large one for natural scene datasets. The parameter a for Leaky ReLU (lReLU) activation function is set to 0.1.

Architecture	Small Network	Medium Network	Large Network
Input size	2048	$32 \times 32 \times 3$	$32 \times 32 \times 3$
Generator G		instance normalization	instance normalization
	256 dense, ReLU	3×3 conv. 64 lReLU	3×3 conv. 96 lReLU
	dropout, $p = 0.5$	3×3 conv. 64 lReLU	3×3 conv. 96 lReLU
	Gaussian noise, σ = 1	3×3 conv. 64 lReLU	3×3 conv. 96 lReLU
		2×2 max-pool, stride 2	2×2 max-pool, stride 2
		dropout, $p = 0.5$	dropout, $p = 0.5$
		Gaussian noise, σ = 1	Gaussian noise, σ = 1
		3×3 conv. 64 lReLU	3×3 conv. 192 lReLU
		3×3 conv. 64 lReLU	3×3 conv. 192 lReLU
		3×3 conv. 64 lReLU	3×3 conv. 192 lReLU
		2×2 max-pool, stride 2	3×3 max-pool, stride 2
		dropout, $p = 0.5$	dropout, $p = 0.5$
		Gaussian noise, σ = 1	Gaussian noise, σ = 1
Classifier $\mathcal{A}$	C dense, softmax	3×3 conv. 64 lReLU	3×3 conv. 192 lReLU
		3×3 conv. 64 lReLU	3×3 conv. 192 lReLU
		3×3 conv. 64 lReLU	3×3 conv. 192 lReLU
		global average pool	global average pool
		C dense, softmax	C dense, softmax

3.2.3 Network Architectures

We use small and large network architecture for specific datasets, which are described in Table 3 and 4. Noticeably, batch normalization layers are applied on the top of convolutional layers (6 for the generator and 3 for the classifier) to prevent the overfitting. For *Office-31*, *Office-Home*, and *ImageCLEF-DA*, we removed the dense layers of the pretrained models and replaced by two dense layers (i.e., the first layer has 256 neurons and the second one has C neurons where C is the number of classes).

Table 4: The architecture of discriminator d.

Digits, traffic sign and natural scene datasets	Object recognition datasets
3×3 conv. 64 lReLU	$C + 1$ dense, softmax
3×3 conv. 64 lReLU	
3×3 conv. 64 lReLU	
global average pool	
$C + 1$ dense, softmax	

3.2.4 Hyperparameter setting

We apply Adam Optimizer ($\beta_1 = 0.5$,$\beta_2 = 0.999$) with the learning rate 0.001 *digits*, *traffic sign* and *natural scene* datasets, whereas 0.0001 is the learning rate for *object recognition* datasets. All experiements was trained for 20000 iterations on *Office-31*, *Office-*

Home, and *ImageCLEF-DA* and 80000 for the other datasets. The batch size for each dataset is set to 128. We set $\beta = 0$, $\alpha = 0.5$ as described in the ablation study, and γ is searched in $\{0.1, 0.5\}$. We implement our LAMDA in Python (version 3.5) using Tensorflow (version 1.9.0) [1] and run our experiments on a computer with a CPU named Intel Xeon Processor E5-1660 which has 8 cores at 3.0 GHz and 128 GB of RAM, and a GPU called NVIDIA GeForce GTX Titan X with 12 GB memory.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[4] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016.

[5] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 627–636. Computer Vision Foundation / IEEE, 2019.

[6] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017.

[7] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 467–483, 2018.

[8] Z. Deng, Y. Luo, and J. Zhu. Cluster alignment with a teacher for unsupervised domain adaptation, 2019.

[9] G. French, M. Mackiewicz, and M. Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.

[10] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1180–1189, 2015.

[11] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[12] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *INTERNAT. STATIST. REV.*, pages 419–435, 2002.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] J. Hoffman, E. Tzeng, T. Park, J-Y Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

[16] M. Kim, P. Sahu, B. Gholami, and V. Pavlovic. Unsupervised visual domain adaptation: A deep max-margin gaussian process approach. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4375–4385, 2019.

[17] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[18] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10285–10295. Computer Vision Foundation / IEEE, 2019.

[19] M. Li, Y. Zhai, Y. Luo, P. Ge, and C. Ren. Enhanced transport distance for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, 2020.

[21] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 2015.

[22] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1640–1650. Curran Associates, Inc., 2018.

[23] T. Nguyen and S. Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1085–1093, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[24] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2988–2997. JMLR. org, 2017.

[25] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

[27] O. Sener, H O Song, A. Saxena, and S. Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016.

[28] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation, 2019.

[29] I. O. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. *CoRR*, abs/1711.01558, 2018.

[30] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[31] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, November 1999.

[32] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[33] S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5423–5432. PMLR, 10–15 Jul 2018.

[34] Y. Xie, M. Chen, H. Jiang, T. Zhao, and H. Zha. On scalable and efficient computation of large scale optimal transport. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6882–6892, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

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