A. Proof and Analysis

A.1. Rademacher Complexity

Let $\mathcal{F}$ be a class of real-valued functions mapping $\mathcal{X} \to \mathbb{R}$ and $\hat{P} = \{x_1, ..., x_m\}$ a finite sample drawn i.i.d. according to a distribution P, the empirical Rademacher Complexity of $\mathcal{F}$ is defined as follows:

$$\hat{\Re}_{\hat{P}}(\mathcal{F}) = \frac{1}{m}\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(x_i)\right].$$

The expectation is taken over $\sigma = (\sigma_1, ..., \sigma_m)$ where σ_i is an independent uniform random variable taking values in $\{-1, +1\}$. Following the established theory proposed by Mansour et al. [35], we denote the empirical disagreement between hypotheses $h \in \mathcal{H}, h' \in \mathcal{H}' : \mathcal{X} \to \mathcal{K}$ by $\epsilon_{\hat{P}}(h, h')$ and its expectation over samples drawn according to the distribution by $\epsilon_P(h, h')$. According to Bartlett and Mendelson [2], Koltchinskii and Panchenko [25], for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\hat{P}$ of size m, the following inequality holds for $\forall h \in \mathcal{H}, \forall h' \in \mathcal{H}'$ when the loss function ϵ is bounded with M:

$$\epsilon_P(h, h') \le \epsilon_{\hat{P}}(h, h') + 2\hat{\Re}_{\hat{P}}(\mathcal{F}^{\epsilon}_{\mathcal{H},\mathcal{H}'}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

where $\mathcal{F}^{\epsilon}_{\mathcal{H},\mathcal{H}'} = \{f(x) = \epsilon(h(x), h'(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}, h' \in \mathcal{H}'\}$. Similarly, the expected error is bounded, with probability at least $1 - \delta$, for $\forall h \in \mathcal{H}$,

$$\epsilon_P(h) = \epsilon_P(h, f_P) \le \epsilon_{\hat{P}}(h) + 2\hat{\Re}_{\hat{P}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_P}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

where $\mathcal{F}^{\epsilon}_{\mathcal{H},f_P} = \{f(x) = \epsilon(h(x), f_P(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$. Alternatively, $\mathcal{F}^{\epsilon}_{\mathcal{H},f_P}$ can be regarded as a set of functions $\{f(x, k) = \epsilon(h(x), k) : (\mathcal{X} \times \mathcal{K}_P) \to [0, M] | h \in \mathcal{H}\}$ where $(\mathcal{X} \times \mathcal{K}_P) = \{(x, f_P(x)) | x \in \mathcal{X}\}$.

A.2. Proof of Theorem 3.1

Let $f_S : \mathcal{X} \to \mathcal{K}$, $f_T : \mathcal{X} \to \mathcal{K}$, $f_V : \mathcal{X} \to \mathcal{K}$ denote the true labeling functions on the source, unlabeled and labeled target domains. Given a distance metric ϵ that satisfies the triangle inequality, for $\forall h \in \mathcal{H} : \mathcal{X} \to \mathcal{K}$, the expected target error is bounded as

$$\begin{aligned}
\epsilon_T(h, f_T) &= \frac{1}{2}[2\epsilon_T(h, f_T) - \epsilon_V(h, f_V) - \epsilon_S(h, f_S) + \epsilon_V(h, f_V) \\
&\quad + \epsilon_S(h, f_S) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) - \epsilon_T(h, f_S) \\
&\quad - \epsilon_T(h, f_V) + \epsilon_V(h, f_S) + \epsilon_S(h, f_V) - \epsilon_V(h, f_S) - \epsilon_S(h, f_V)] \\
&= \frac{1}{2}([\epsilon_T(h, f_T) - \epsilon_T(h, f_S)] + [\epsilon_T(h, f_T) - \epsilon_T(h, f_V)] \\
&\quad + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + [\epsilon_V(h, f_S) - \epsilon_V(h, f_V)] \\
&\quad + [\epsilon_S(h, f_V) - \epsilon_S(h, f_S)] - \epsilon_V(h, f_S) - \epsilon_S(h, f_V) \\
&\quad + \epsilon_V(h, f_V) + \epsilon_S(h, f_S)) \\
&\le \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) \\
&\quad + \epsilon_V(f_S, f_V) + \epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) - \epsilon_S(h, f_V)] \\
&\quad + [\epsilon_V(h) + \epsilon_S(h)]) \\
&= D_{S,T,V}(f_S, f_T, f_V, h) + \frac{1}{2}[\epsilon_V(h) + \epsilon_S(h)].
\end{aligned}$$

A.3. Proof of Corollary 3.2

Given a distance metric ϵ that satisfies the triangle inequality, for $\forall h \in \mathcal{H}$, the following holds:

$$\begin{aligned}
U(h) &= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + \epsilon_V(f_S, f_V) \\
&\quad + \epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) - \epsilon_S(h, f_V)] + [\epsilon_V(h) + \epsilon_S(h)]) \\
&= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + \epsilon_V(f_S, f_V) \\
&\quad + \epsilon_S(f_V, f_S) - \epsilon_S(h, f_V) + \epsilon_S(h)] + [\epsilon_V(h, f_V) - \epsilon_V(h, f_S)]) \\
&\le \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) \\
&\quad + 2\epsilon_V(f_S, f_V) + \epsilon_S(f_V, f_S) - \epsilon_S(h, f_V) + \epsilon_S(h)],
\end{aligned}$$

$$\begin{aligned}
U(h) &= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + \epsilon_V(f_S, f_V) \\
&\quad + \epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) - \epsilon_S(h, f_V)] + [\epsilon_V(h) + \epsilon_S(h)]) \\
&= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + \epsilon_V(f_S, f_V) \\
&\quad + \epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) + \epsilon_V(h)] + [\epsilon_S(h, f_S) - \epsilon_S(h, f_V)]) \\
&\le \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) \\
&\quad + \epsilon_V(f_S, f_V) + 2\epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) + \epsilon_V(h)],
\end{aligned}$$

$$\begin{aligned}
U(h) &= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + \epsilon_V(f_S, f_V) \\
&\quad + \epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) - \epsilon_S(h, f_V)] + [\epsilon_V(h) + \epsilon_S(h)]) \\
&= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) - \epsilon_V(h, f_S) \\
&\quad - \epsilon_S(h, f_V)] + [\epsilon_V(h, f_V) + \epsilon_V(f_S, f_V) + \epsilon_S(h, f_S) + \epsilon_S(f_V, f_S)]) \\
&\ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V)].
\end{aligned}$$

A.4. Proof of Corollary 3.3

According to Corollary 3.2, for $\forall h \in \mathcal{H}$,

$$U(h) \ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V)].$$

Given $h^* = \arg\min_{h\in\mathcal{H}} U(h)$, we can further derive:

$$\begin{aligned}
\min_{h\in\mathcal{H}} U(h) &= U(h^*) \\
&\ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h^*, f_S) + \epsilon_T(h^*, f_V)] \\
&\ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(f_V, f_S)].
\end{aligned}$$

Similarly, according to Corollary 3.2, for $\forall h \in \mathcal{H}$,

$$U(h) \leq \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + 2\epsilon_V(f_S, f_V) + \epsilon_S(h) - \epsilon_S(h, f_V) + \epsilon_S(f_V, f_S)].$$

Let $h = f_S$ if $f_S \in \mathcal{H}$, we can further derive:

$$\frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(f_V, f_S)] + \epsilon_V(f_S, f_V) \geq U(f_S) \geq \min_{h\in\mathcal{H}} U(h).$$

Otherwise, we can always derive:

$$\begin{aligned}&\epsilon_T(h, f_V) + \epsilon_T(h, f_S) + \epsilon_S(h) + \epsilon_S(f_V, f_S)\\ &\leq \epsilon_T(f_S, f_V) + \epsilon_T(h, f_S) + \epsilon_T(h, f_S^*) + \epsilon_T(f_S^*, f_S)\\ &+ \epsilon_S(h, f_S^*) + 2\epsilon_S(f_S^*, f_S) + \epsilon_S(f_V, f_S^*).\end{aligned}$$

Let $\hat{S}, \hat{T}$ denote a finite set with size n, m from domain S, T. According to Assumption 3.8, there exist hypotheses $f_S^* \in \mathcal{H}$ such that we can ignore $\epsilon_{\hat{T}}(f_S^*, f_S), \epsilon_{\hat{S}}(f_S^*, f_S)$ and upper bound $U(h)$ with empirical Rademacher Complexity (A.1). Given function space $\mathcal{F}^{\epsilon}_{\mathcal{H},f_S} = \{f(x) = \epsilon(h(x), f_S(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$ and Corollary 3.2, for any $\delta > 0$, with probability at least $1 - \delta$, for $\forall h \in \mathcal{H}$:

$$\begin{aligned}U(h) \leq &\frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S^*) + \epsilon_T(f_S, f_V)\\ &+ 2\epsilon_V(f_S, f_V) + \epsilon_S(h, f_S^*) - \epsilon_S(h, f_V) + \epsilon_S(f_V, f_S^*)]\\ &+ \frac{1}{2}[\epsilon_{\hat{T}}(h, f_S) + \underbrace{\epsilon_{\hat{T}}(f_S^*, f_S) + 2\epsilon_{\hat{S}}(f_S^*, f_S)}_{\text{zero}}]\\ &+ 2\hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) + 2\hat{\mathfrak{R}}_{\hat{S}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) + 3M(\sqrt{\frac{\log\frac{6}{\delta}}{2m}} + \sqrt{\frac{\log\frac{6}{\delta}}{2n}}) = U^*(h),\end{aligned}$$

where the minimum is achieved at $h = f_S^*$ such that $\min_{h\in\mathcal{H}} U(h) \leq U^*(f_S^*)$

A.5. Proof of Lemma 3.4

if $f_S \in \mathcal{H}$, we can derive:

$$\epsilon_T(f_S, f_T) = \epsilon_T(f_S, f_T) + \epsilon_S(f_S, f_S) \geq \min_{h\in\mathcal{H}}(\epsilon_T(h) + \epsilon_S(h)) = \lambda_{S,T}.$$

Otherwise, we can always derive:

$$\epsilon_T(f_S, f_T) \geq \epsilon_T(f_S^*, f_T) - \epsilon_T(f_S^*, f_S) + \epsilon_S(f_S^*, f_S) - \epsilon_S(f_S^*, f_S).$$

Let $\hat{S}, \hat{T}$ denote a finite set with size n, m from domain S, T. According to Assumption 3.8, there exist hypotheses $f_S^* \in \mathcal{H}$ such that we can ignore $\epsilon_{\hat{T}}(f_S^*, f_S), \epsilon_{\hat{S}}(f_S^*, f_S)$ and lower bound $\epsilon_T(f_S, f_T)$ with empirical Rademacher Complexity (A.1). Given function space $\mathcal{F}^{\epsilon}_{\mathcal{H},f_S} = \{f(x) = \epsilon(h(x), f_S(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned}\epsilon_T(f_S, f_T) \geq \lambda_{S,T} - &\underbrace{[\epsilon_{\hat{S}}(f_S^*, f_S) + \epsilon_{\hat{T}}(f_S^*, f_S)]}_{\text{zero}}\\ &- 2\hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) - 2\hat{\mathfrak{R}}_{\hat{S}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) - 3M(\sqrt{\frac{\log\frac{4}{\delta}}{2m}} + \sqrt{\frac{\log\frac{4}{\delta}}{2n}}).\end{aligned}$$

A.6. Proof of Lemma 3.5

Let $\hat{S}$ be a random sample of size n from domain S, let $\hat{T}$ be a random sample of size m from domain T, and let $\hat{V}$ be a random sample of size l from domain V. Given the empirical Rademacher Complexity $\hat{\mathfrak{R}}$ (A.1) of function space, e.g., $\mathcal{F}^{\epsilon}_{\mathcal{H},f_S} = \{f(x) = \epsilon(h(x), f_S(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for $\forall h \in \mathcal{H}$ according to Theorem 3.1:

$$\begin{aligned}\epsilon_T(h) &\leq \frac{1}{2}[\epsilon_V(h) + \epsilon_S(h)] + D_{S,T,V}(f_S, f_T, f_V, h)\\ &\leq \frac{1}{2}[\epsilon_{\hat{V}}(h) + \epsilon_{\hat{S}}(h)] + D_{\hat{S},\hat{T},\hat{V}}(f_S, f_T, f_V, h)\\ &+ \hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) + \hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_V}) + \hat{\mathfrak{R}}_{\hat{V}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) + \hat{\mathfrak{R}}_{\hat{S}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_V})\\ &+ \hat{\mathfrak{R}}_{\hat{V}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_V}) + \hat{\mathfrak{R}}_{\hat{S}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) + \frac{3M}{2}(3\sqrt{\frac{\log\frac{16}{\delta}}{2n}}\\ &+ 3\sqrt{\frac{\log\frac{16}{\delta}}{2l}} + 4\sqrt{\frac{\log\frac{16}{\delta}}{2m}})\\ &= \frac{1}{2}[\epsilon_{\hat{V}}(h) + \epsilon_{\hat{S}}(h)] + D_{\hat{S},\hat{T},\hat{V}}(f_S, f_T, f_V, h) + RC\end{aligned}$$

$$+ \frac{3M}{2}(3\sqrt{\frac{\log\frac{16}{\delta}}{2n}} + 3\sqrt{\frac{\log\frac{16}{\delta}}{2l}} + 4\sqrt{\frac{\log\frac{16}{\delta}}{2m}}).$$

A.7. Proof of Theorem 3.7

For certain labeling functions $f_S^* \in \mathcal{H}_S \subseteq \mathcal{H}, f_T^* \in \mathcal{H}_T \subseteq \mathcal{H}, f_V^* \in \mathcal{H}_V \subseteq \mathcal{H}$, the empirical discrepancy $D_{\hat{S},\hat{T},\hat{V}}(f_S, f_T, f_V, h)$ is bounded for $\forall h \in \mathcal{H}$:

$$\begin{aligned}&D_{\hat{S},\hat{T},\hat{V}}(f_S, f_T, f_V, h)\\ &= \frac{1}{2}[\epsilon_{\hat{T}}(f_S, f_T) + \epsilon_{\hat{T}}(f_V, f_T) + \epsilon_{\hat{T}}(h, f_S) + \epsilon_{\hat{T}}(h, f_V) + \epsilon_{\hat{V}}(f_S, f_V)\\ &+ \epsilon_{\hat{S}}(f_V, f_S) - \epsilon_{\hat{V}}(h, f_S) - \epsilon_{\hat{S}}(h, f_V)]\\ &\leq \frac{1}{2}[\epsilon_{\hat{T}}(f_S, f_S^*) + \epsilon_{\hat{T}}(f_S^*, f_T^*) + \epsilon_{\hat{T}}(f_T^*, f_T) + \epsilon_{\hat{T}}(f_V, f_V^*)\\ &+ \epsilon_{\hat{T}}(f_V^*, f_T^*) + \epsilon_{\hat{T}}(f_T^*, f_T) + \epsilon_{\hat{T}}(h, f_S^*) + \epsilon_{\hat{T}}(f_S, f_S^*)\\ &+ \epsilon_{\hat{T}}(h, f_V^*) + \epsilon_{\hat{T}}(f_V^*, f_V) + \epsilon_{\hat{V}}(f_S, f_S^*) + \epsilon_{\hat{V}}(f_S^*, f_V^*)\\ &+ \epsilon_{\hat{V}}(f_V^*, f_V) + \epsilon_{\hat{S}}(f_S, f_S^*) + \epsilon_{\hat{S}}(f_S^*, f_V^*) + \epsilon_{\hat{S}}(f_V^*, f_V)\\ &+ \epsilon_{\hat{V}}(f_S, f_S^*) - \epsilon_{\hat{V}}(h, f_S^*) + \epsilon_{\hat{S}}(f_V, f_V^*) - \epsilon_{\hat{S}}(h, f_V^*)]\\ &= D_{\hat{S},\hat{T},\hat{V}}(f_S^*, f_T^*, f_V^*, h) + \hat{\theta},\\ &\hat{\theta} = \underbrace{\epsilon_{\hat{S}}(f_S, f_S^*)/2 + \epsilon_{\hat{V}}(f_S, f_S^*) + \epsilon_{\hat{T}}(f_S, f_S^*)}_{\hat{\theta}_{f_S}} + \underbrace{\epsilon_{\hat{T}}(f_T, f_T^*)}_{\hat{\theta}_{f_T}}\\ &+ \underbrace{\epsilon_{\hat{V}}(f_V, f_V^*)/2 + \epsilon_{\hat{S}}(f_V, f_V^*) + \epsilon_{\hat{T}}(f_V, f_V^*)}_{\hat{\theta}_{f_V}}\end{aligned}$$

Given Lemma 3.5 and Corollary 3.6, Theorem 3.7 can be proved.

A.8. Feasibility of Assumption 3.8

E.g., let $\hat{T}$ denote a finite set with size m from target domain T. Given the empirical Rademacher Complexity (A.1) of function space $\mathcal{F}^{\epsilon}_{\mathcal{H},f_T} = \{f(x) = \epsilon(h(x), f_T(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$ over $\hat{T}$ as $\hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_T})$, for any $\delta > 0$, with probability at least $1 - \delta$, the expected disagreement $\theta_{f_T} = \epsilon_T(f_T^*, f_T)$ is bounded by the empirical disagreement $\hat{\theta}_{f_T} =$

$\epsilon_{\hat{T}}(f_T^*, f_T)$ for $\forall f_T^* \in \mathcal{H}$:

$$\epsilon_T(f_T^*, f_T) \leq \epsilon_{\hat{T}}(f_T^*, f_T) + 2\hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_T}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

We assume there exists $f_T^* \in \mathcal{H}$ such that in Theorem 3.7, $\hat{\theta}_{f_T} \to 0$ thus can be ignored during the practical learning process. Note that this assumption is more feasible than assuming empirical joint error $\lambda_{\hat{S},\hat{T}} \to 0$ in Ben-David et al. [4], especially when the domain shift is large. To facilitate the analysis, let $g : \mathcal{X} \subseteq \mathbb{R}^I \to \mathcal{Z} \subseteq \mathbb{R}^F$ be injective on $\hat{S}, \hat{T}$ with the size $n = m$ respectively, such that true labeling functions f_S, f_T can be decomposed as $f_S^F \circ g, f_T^F \circ g$. Let $\hat{S} \xrightarrow{g} \hat{Z}_S \cup \hat{Z}_C$ and $\hat{T} \xrightarrow{g} \hat{Z}_T \cup \hat{Z}_C$ denote the feature space that overlaps at $\hat{Z}_C$ with size c. For $h \in \mathcal{H}^F : \mathcal{Z} \to \mathcal{K}$,

$$\begin{aligned}
&\min_{h\in\mathcal{H}^F}[\epsilon_{\hat{S}}(h\circ g, f_S) + \epsilon_{\hat{T}}(h\circ g, f_T)] = \lambda_{\hat{S},\hat{T}} \\
&= \min_{h\in\mathcal{H}^F}[\frac{m-c}{m}\epsilon_{\hat{Z}_S}(h, f_S^F) + \frac{m-c}{m}\epsilon_{\hat{Z}_T}(h, f_T^F) \\
&+ \frac{c}{m}\epsilon_{\hat{Z}_C}(h, f_S^F) + \frac{c}{m}\epsilon_{\hat{Z}_C}(h, f_T^F)] \\
&\geq \frac{m-c}{m}\min_{h\in\mathcal{H}^F}\epsilon_{\hat{Z}_S}(h, f_S^F) + \frac{m-c}{m}\min_{h\in\mathcal{H}^F}\epsilon_{\hat{Z}_T}(h, f_T^F) + \frac{c}{m}\epsilon_{\hat{Z}_C}(f_T^F, f_S^F),
\end{aligned}$$

where f_T^F, f_S^F tend to disagree on $\hat{Z}_C$ in large domain shift such that $\lambda_{\hat{S},\hat{T}}$ increases as c grows. In addition, even if $\epsilon_{\hat{Z}_C}(f_T^F, f_S^F) \to 0$, the solution for $\lambda_{\hat{S},\hat{T}} \to 0$ is likely to be more complex, which can be outside the hypothesis space $\mathcal{H}^F$. E.g., let $f_S^F = |z|$ and $f_T^F = -|z-1|+1$. For $\hat{Z}_S \subset (-\infty, 0], \hat{Z}_C \subset (0,1), \hat{Z}_T \subset [1,\infty)$, the optimal solution for h is

$$\left.\begin{array}{ll} -z, & z \in \hat{Z}_S \\ z, & z \in \hat{Z}_C \\ -z+2, & z \in \hat{Z}_T \end{array}\right\} = h(z) \notin \mathcal{H}^F = \{z \mapsto a|z-b|+c | a,b,c \in \mathbb{R}\}.$$

A.9. Bound of Gap

Given Theorem 3.1, for $f_S = f_V \notin \mathcal{H}$,

$$U(h) = \epsilon_T(f_T, f_S) + \epsilon_T(h, f_S).$$

For any $f_S^* \in \mathcal{H}$, let $h^* = \arg\min_{h\in\mathcal{H}} U(h)$,

$$\epsilon_T(h^*, f_S) \leq \epsilon_T(f_S^*, f_S).$$

The gap between the expected target error and the upper bound at h^* follows:

$$\begin{aligned}
U(h^*) - \epsilon_T(h^*) &= \epsilon_T(f_T, f_S) + \epsilon_T(h^*, f_S) - \epsilon_T(h^*, f_T) \\
&\leq 2\epsilon_T(h^*, f_S) \leq 2\epsilon_T(f_S^*, f_S).
\end{aligned}$$

Let $\hat{T}$ denote a finite set with size m from target domain T. According to Assumption 3.8, there exist hypotheses $f_S^* \in \mathcal{H}$ such that we can ignore $\epsilon_{\hat{T}}(f_S^*, f_S)$ and upper bound the gap with empirical Rademacher Complexity (A.1). Given function space $\mathcal{F}^{\epsilon}_{\mathcal{H},f_S} = \{f(x) = \epsilon(h(x), f_S(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$, for any $\delta > 0$, with probability at least $1-\delta$,

$$U(h^*) - \epsilon_T(h^*) \leq \underbrace{2\epsilon_{\hat{T}}(f_S^*, f_S)}_{\text{zero}} + 4\hat{\mathfrak{R}}_{\hat{T}}(\mathcal{F}^{\epsilon}_{\mathcal{H},f_S}) + 6M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

B. Implementation

For a fair comparison, we use the same labeled source data, labeled target data, and unlabeled target data as Saito et al. [46]. We provide details of our implementation in Algorithm 1, where we introduce a gradient reversal layer [16] to train the overall objective together. Note that we do not optimize f_T' on semi-supervised regularization losses in practice as it can lead to an early convergence to bad local optimum. The pre-trained model (e.g., ResNet34) except the last layer combined with a single-layer bottleneck [65] is used as feature extractor g and randomly initialized 2-layer fully-connected networks are used for classifiers f_S', f_T', f_V', h. We introduce smoothed cross-entropy loss [31, 39] to prevent the network from becoming over-confident on labeled data. We adopt SGD with momentum 0.9 for optimization, where the learning rate is set to α for all fully connected layers, whereas it is set to 0.1α for the other convolution layers. The initial learning rate is set to 0.01 for DomainNet, 0.004 for Office-Home, and 0.001 for VisDA according to Saito et al. [46], Zhang et al. [62, 65]. We employ the learning rate annealing strategy proposed in Ganin et al. [16]. We use RandomFlip, RandomCrop, and RandAugment as data augmentation, and the batch size is fixed to 32. The results of adaptation scenarios from all three benchmarks, DomainNet, Office-Home, and VisDA, are given by 50k iterations run on Tesla V100.

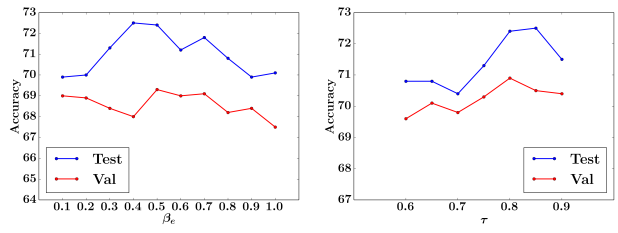
C. Hyper-parameter Selection

We set $\beta_d = 0.01, \beta_c = 30$ in all benchmarks according to Yang et al. [58], Zhang et al. [62]. We fine-tune β_e, τ to improve performance based on a validation set containing 3 labeled target samples per class from DomainNet dataset C to S scenario. In Fig.5. , we show the performance when varying the hyper-parameters β_e, τ. Given the validation accuracy, we set $\beta_e = 0.5, \tau = 0.8$ in all benchmarks.

D. Additional Experiments

D.1. Feature Visualization on DomainNet

We plot learned features of the Real to Sketch task from DomainNet with t-SNE [54] in Fig.6. Fig.6d shows features of unlabeled target data, where each color represents a different class. In our method, most of the target samples are well-clustered and do not have a large variance within the class. In Fig.6h, our method almost perfectly matches



(a) $\beta_e = 0.5$ (b) $\tau = 0.8$

Figure 5. Sensitivity w.r.t hyper-parameters β_e, τ tested on C→S scenario in DomainNet. The hyper-parameters are set to the same values for all benchmarks based on the validation accuracy .

Stats	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Avg
avg.	81.5	76.7	80.9	72.5	75.3	74.3	84.6	78.0
std.	0.3	0.14	0.23	0.14	0.26	0.13	0.15	0.19

Table 5. Statistics (%) on DomainNet under the setting of 3-shot using ResNet34.

conditional distributions of the feature space as we expect. We also plot features of the source (red) and target domains (blue) in Fig.6l to show that our proposal can align marginal distributions. In our method, each cluster is separated while others sometimes merge different clusters.

D.2. Varying Number of Labeled Samples

Fig.8 shows the behavior of different methods when the number of labeled examples in the target domain varies from 0 to 20 per class on DomainNet using ResNet34 backbone. Cluster based methods like MME [46] will finally exceeded by a simple entropy minimization when the sample size grows. On the contrary, our method maintains a high level of performance for various sizes of the labeled target data.

D.3. Quantitative Analysis

We further conducted the quantitative analysis of other tasks regarding Offie-Home and VisDA datasets (Fig.7) and the conclusion remains the same, where the feature obtained by our method is more discriminative with less domain divergence.

D.4. Statistics

For each sub-task in DomainNet, Office-Home, and VisDA benchmarks, we ran the algorithm three times with different seeds. This section provides the average accuracy and the standard deviations of our method. Tab.5, 6, 7 show the details of the statistics about model performance in DomainNet, Office-Home, and VisDA datasets respectively. Tab.8 shows the robustness of our method against different selections of labeled target data.

D.5. One-shot Results

We report the comparison with baselines in the one-shot setting on DomainNet in Tab.10 and Office-Home in Tab.9. Our method consistently outperforms other baselines, e.g., ECB by 0.6% in DomainNet and 1.4% in Office-Home with less labeled target data.

D.6. Consistency

In this section, we tackle a general problem associated with the consistency between the algorithm and theory in domain adaptation. The triangle inequality is essential to build the theory, and the measurement of the source error and terms related to the discrepancy should be the same. These requirements should be satisfied by any method that introduces an upper bound to approximate the target error. However, most upper bound-based methods violate these rules, known as the gap between the algorithm and theory. For instance, MCD [45] chooses cross entropy for the source error but replaces the discrepancy with a L_1 norm between the predictions of two classifiers. As for DANN [16], it uses logistic loss as a surrogate to approximate 0-1 loss, which is no longer an upper bound of $\mathcal{H}\Delta\mathcal{H}$-distance in Ben-David et al. [4]. Even though our proposal does not serve as a perfect cure to this problem, we can prove that the proposed MCMD asymptotically satisfies the consistency.

First of all, we show that MCMD obeys the triangle inequality under the following circumstances. For the case where two hypotheses agree on the point x ($y = l(h_1(x)) = l(h_2(x)), l(h_3(x)) = y'$; this condition is met when we use triangle inequality to derive the upper bound in Theorem 3.1 except for f_T, h), given Definition 3.12:

$$
\begin{aligned}
&\mathrm{mcmd}(h_1(x), h_3(x)) + \mathrm{mcmd}(h_2(x), h_3(x)) \\
&\quad = \max(|\log h_1(x)[y] - \log h_3(x)[y]|, |\log h_1(x)[y'] - \log h_3(x)[y']|) \\
&\quad + \max(|\log h_2(x)[y] - \log h_3(x)[y]|, |\log h_2(x)[y'] - \log h_3(x)[y']|) \\
&\quad \geq |\log h_1(x)[y] - \log h_3(x)[y]| + |\log h_2(x)[y] - \log h_3(x)[y]| \\
&\quad \geq |\log h_1(x)[y] - \log h_2(x)[y]| = \mathrm{mcmd}(h_1(x), h_2(x)).
\end{aligned}
$$

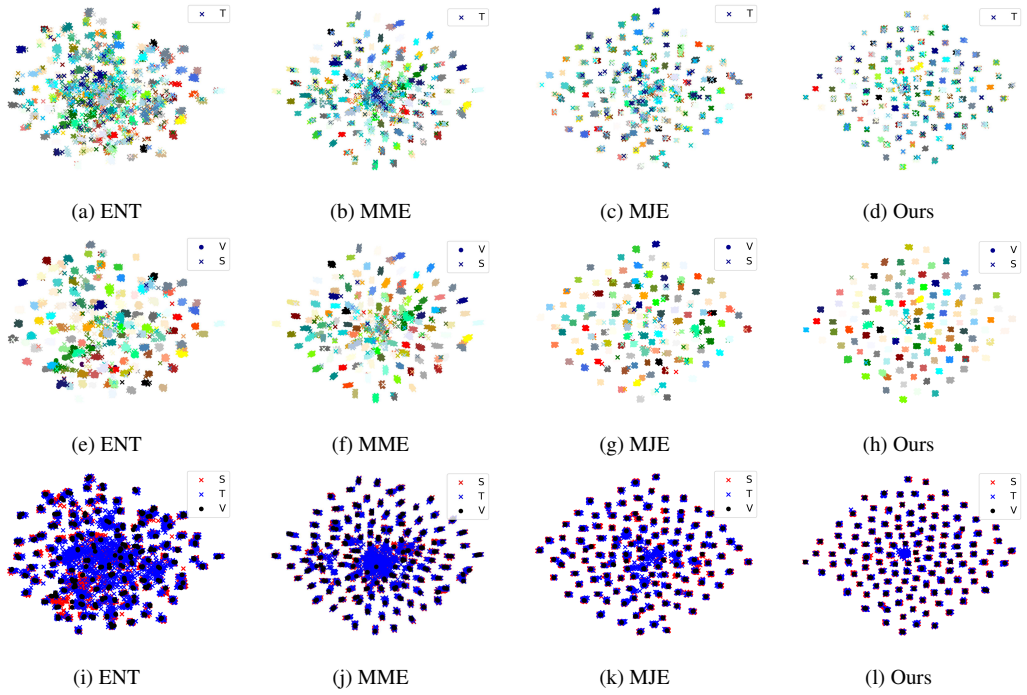
As the training proceeds, the target error of h will be minimized, which means the discrepancy between h, f_T over domain T is constantly reduced. Given the assumption that f_T and h gradually agree on T, we can conclude that our proposal asymptotically satisfies the triangle inequality.

Then we prove that the cross-entropy loss is a special case of MCMD by reasonably assuming $f_S(x)[y] = 1$ and $l(f_S(x)) = l(h(x)) = y$ for $(x, y) \sim S$. According to Definition 3.12, the expected source error of h defined based on MCMD can be written as (the same goes for V):

$$
\begin{aligned}
\epsilon_S(h) &= \mathbb{E}_{(x,y)\sim S}[\mathrm{mcmd}(h(x), f_S(x))] \\
&= \mathbb{E}_{(x,y)\sim S}|\log f_S(x)[y] - \log h(x)[y]| \\
&= -\mathbb{E}_{(x,y)\sim S}[\log h(x)[y]].
\end{aligned}
$$

D.7. Interpretability

In this section, we explain the relation between the objective function in Theorem 3.7 and CGAN [36] and prove that our

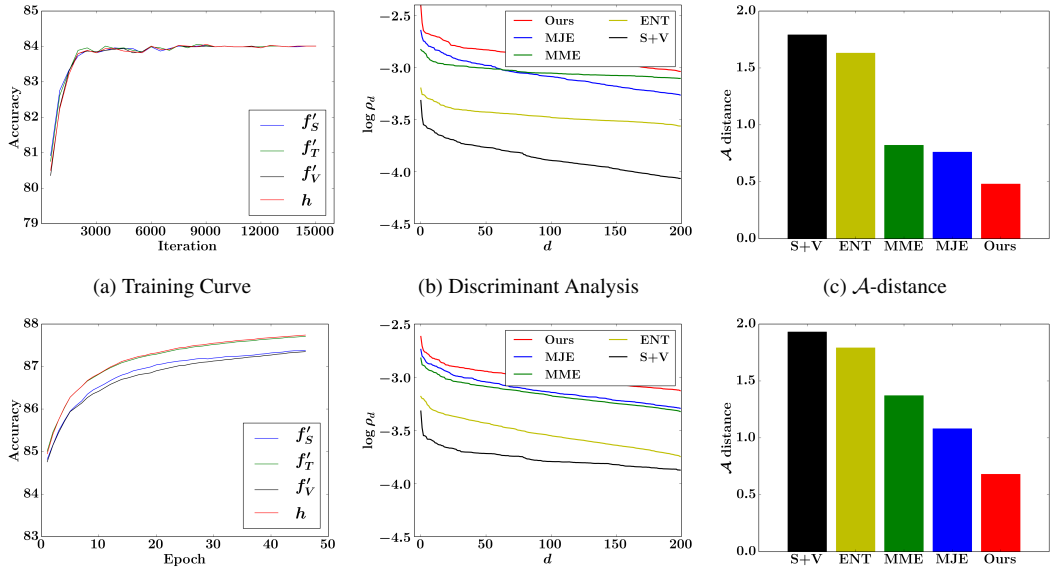


(a) ENT (b) MME (c) MJE (d) Ours

(e) ENT (f) MME (g) MJE (h) Ours

(i) ENT (j) MME (k) MJE (l) Ours

Figure 6. Comparisons of the feature space visualized by t-SNE after the adaptation from Real to Sketch; (a)-(d) show the alignment between source and labeled target domains where ours achieves a perfect conditional distribution alignment; (e)-(h) show the alignment between marginal distributions where ours gives a tighter match.



(a) Training Curve (b) Discriminant Analysis (c) $\mathcal{A}$-distance

(d) Training Curve (e) Discriminant Analysis (f) $\mathcal{A}$-distance

Figure 7. (a,d) Training procedure is stable, and all classifiers will reach a reliable convergence; (b,e) Ratio between inter-class and intra-class distance for each dimension of extracted features from target data. A high ρ_d means a more discriminative feature; (c,f) Our method significantly reduces the domain divergence; (a-c) Quantitative analysis on A →P task from **Office-Home** dataset; (d-f) Quantitative analysis on **VisDA** dataset. For a fair comparison, we only deploy L_{ent} w/o strong data augmentation in (b,c,e,f).

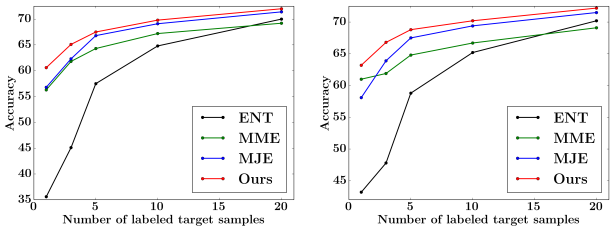
proposal can reduce the conditional discrepancy between domains. According to the constraints of hypothesis space

Stats	R→C	R→P	R→A	P→R	P→C	P→A	A→P	A→C	A→R	C→R	C→A	C→P	Avg
avg.	72.2	88.3	76.2	84.5	69.1	73.2	84.5	69.9	83.2	82.7	72.1	85.5	78.5
std.	0.14	0.22	0.30	0.13	0.34	0.26	0.34	0.29	0.15	0.25	0.17	0.35	0.25

Table 6. Statistics (%) on Office-Home under the setting of 3-shot using ResNet34.

Stats	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
avg.	96.7	88.7	85.1	85.3	96.1	96.3	92.8	86.4	96.3	94.3	88.5	46.0	87.7
std.	0.34	1.42	1.41	1.96	0.25	0.29	0.18	1.31	0.56	1.09	1.25	2.44	0.31

Table 7. Statistics (%) on VisDA under the setting of 3-shot using ResNet34.



(a) C to S (b) R to S

Figure 8. Accuracy vs the number of labeled target samples on DomainNet using ResNet34 backbone. Our method maintains a high level of performance for different sample sizes of the labeled target domain. To be fair, we only introduce L_{ent} as semi-supervised regularization w/o strong data augmentation.

(Definition 3.10), f'_S and f'_V must both classify the source (S) and labeled target domains (V). In addition, f'_S tends to be more confident about the predictions on V, and f'_V is supposed to be more confident about S based on the formula of $\hat{\theta}$. Given a feature extractor $g : \mathcal{X} \subseteq \mathbb{R}^I \to \mathcal{Z} \subseteq \mathbb{R}^F$ and MCMD (Definition 3.12), we can derive that a part of the objective function (we take expectation here to be consistent with GAN) can be reformed as CGAN, where f'_S, f'_V are two discriminators which regard V, S as the real data distributions respectively (in practice, we optimize $\log(1 - f(x)[y])$ instead of $-\log f(x)[y]$ to avoid exploding or vanishing gradient [18]):

$$
\begin{aligned}
&\max_{f'_S, f'_V \mathcal{H}^F} [\epsilon_V(f'_S \circ g, f'_V \circ g) + \epsilon_S(f'_S \circ g, f'_V \circ g)] \\
&\quad = \max_{f'_S, f'_V \in \mathcal{H}^F} \{\mathbb{E}_{(x,y)\sim V}[\log f'_S(g(x))[y] + \log(1 - f'_V(g(x))[y])] \\
&\quad + \mathbb{E}_{(x,y)\sim S}[\log f'_V(g(x))[y] + \log(1 - f'_S(g(x))[y])]\} \\
&\quad = \max_{f'_S \in \mathcal{H}^F} [\mathbb{E}_{(x,y)\sim V} \log f'_S(g(x))[y] + \mathbb{E}_{(x,y)\sim S} \log(1 - f'_S(g(x))[y])] \\
&\quad + \max_{f'_V \in \mathcal{H}^F} [\mathbb{E}_{(x,y)\sim S} \log f'_V(g(x))[y] + \mathbb{E}_{(x,y)\sim V} \log(1 - f'_V(g(x))[y])].
\end{aligned}
$$

Then we discuss the case where two hypotheses disagree. By introducing two additional distributions $T^{f'_S \setminus f'_T}, T^{f'_T \setminus f'_S}$, we divide the target domain into two parts labeled by f'_S and f'_T respectively based on the difference of their prediction confidence (for simplicity, let $f'_S(g(x))[y_s] \geq f'_T(g(x))[y_s]$ and $f'_T(g(x))[y_t] \geq f'_S(g(x))[y_t]$):

$$
\begin{cases}
T^{f'_S \setminus f'_T} & = \{(x, y_s) | x \sim T, y_s = (l \circ f'_S \circ g)(x), y_t = (l \circ f'_T \circ g)(x) : \\
& \log f'_S(g(x))[y_s] - \log f'_T(g(x))[y_s] \\
& \geq \log f'_T(g(x))[y_t] - \log f'_S(g(x))[y_t]\}, \\
T^{f'_T \setminus f'_S} & = \{(x, y_t) | x \sim T, y_s = (l \circ f'_S \circ g)(x), y_t = (l \circ f'_T \circ g)(x) : \\
& \log f'_T(g(x))[y_t] - \log f'_S(g(x))[y_t] \\
& > \log f'_S(g(x))[y_s] - \log f'_T(g(x))[y_s]\}.
\end{cases}
$$

Now we can derive that a part of the objective function in Theorem 3.7 can be reformed as CGAN, where f'_S is a discriminator that regards labeled data and a part of pseudo labeled target data as the real data distribution ($S \cup V \cup T^{f'_S \setminus f'_T}$). When combined with the constraint that f'_S must classify S, V, a part of the objective w.r.t f'_S becomes:

$$
\begin{aligned}
&\max_{f'_S \in \mathcal{H}^F} [\epsilon_T(f'_S \circ g, f'_T \circ g) - \epsilon_{S \cup V}(f'_S \circ g)] \\
&\quad = \max_{f'_S \in \mathcal{H}^F} [\mathbb{E}_{(x,y)\sim S \cup V \cup T^{f'_S \setminus f'_T}} \log f'_S(g(x))[y] \\
&\quad + \mathbb{E}_{(x,y)\sim T^{f'_T \setminus f'_S}} \log(1 - f'_S(g(x))[y])] + Const.
\end{aligned}
$$

Owing to the power of semi-supervised regularization, f'_T should be more confident than f'_S on the unlabeled target data, which gives the algorithm enough fake data to optimize. Analogously, f'_V can be regarded as a discriminator of CGAN that tries to align the distributions $T^{f'_T \setminus f'_V}$ and $S \cup V \cup T^{f'_V \setminus f'_T}$.

D.8. Limitation

As for potential concern, we claim that it is possible to build subspace $\mathcal{H}_S, \mathcal{H}_T, \mathcal{H}_V \subseteq H$ according to Definition 3.10 such that:

$$
D_{\hat{S},\hat{T},\hat{V}}(f^*_S, f^*_T, f^*_V, h) \leq \max_{f'_S \in \mathcal{H}_S, f'_T \in \mathcal{H}_T, f'_V \in \mathcal{H}_V} D_{\hat{S},\hat{T},\hat{V}}(f'_S, f'_T, f'_V, h).
$$

A sufficient condition for this would be $f^*_S \in \mathcal{H}_S, f^*_T \in \mathcal{H}_T, f^*_V \in \mathcal{H}_V$. This condition can be easily met for $\mathcal{H}_S, \mathcal{H}_V$ since $\hat{S}, \hat{V}$ are fully labeled. In addition, $D_{\hat{S},\hat{T},\hat{V}}(f'_S, f'_T, f'_V, h)$ computed by MCMD may violate the triangle inequality due to its asymptotic consistency, such that the upper bound can fail during the actual training process. Since the validity is hard to prove theoretically, we validate the inequality by experimental results

Selection	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
S0	96.7	88.7	85.1	85.3	96.1	96.3	92.8	86.4	96.3	94.3	88.5	46.0	87.7
S1	96.6	88.9	87.4	86.5	95.4	96.1	93.1	88.4	96.0	93.8	89.1	46.3	88.1
S2	96.4	87.1	87.4	86.2	96.3	97.1	92.3	86.1	96.3	91.7	88.0	43.1	87.3

Table 8. Results (%) on VisDA under the setting of 3-shot using ResNet34 with different selections of labeled target data.

METHOD	R→C	R→P	R→A	P→R	P→C	P→A	A→P	A→C	A→R	C→R	C→A	C→P	Avg
S+V	52.1	78.6	66.2	74.4	48.3	57.2	69.8	50.9	73.8	70.0	56.3	68.1	63.8
DANN	53.1	74.8	64.5	68.4	51.9	55.7	67.9	52.3	73.9	69.2	54.1	66.8	62.7
ENT	53.6	81.9	70.4	79.9	51.9	63.0	75.0	52.9	76.7	73.2	63.2	73.6	67.9
MME	61.9	82.8	71.2	79.2	57.4	64.7	75.5	59.6	77.8	74.8	65.7	74.5	70.4
APE	60.7	81.6	72.5	78.6	58.3	63.6	76.1	53.9	75.2	72.3	63.6	69.8	68.9
CDAC	61.9	83.1	72.7	80.0	59.3	64.6	75.6	61.2	78.5	75.3	64.5	75.1	71.0
DECOTA	60.9	83.3	65.9	76.7	57.3	61.2	77.6	55.1	75.5	74.0	64.3	77.8	69.1
CDAC+SLA	**66.1**	84.6	72.7	80.5	61.8	67.3	78.0	63.0	79.2	77.0	66.9	77.6	72.9
ECB†	65.4	84.1	72.2	80.9	61.2	69.7	78.0	62.6	78.3	**79.4**	67.8	78.1	73.1
EFTL†	65.8	**87.7**	73.5	82.3	**63.2**	68.5	**81.1**	**65.5**	80.6	78.6	64.2	79.3	74.2
Ours	65.9	85.2	**73.7**	**82.6**	62.3	**70.8**	80.8	64.5	**80.9**	79.0	**69.9**	**79.5**	**74.6**

Table 9. Accuracy (%) on Office-Home under the setting of 1-shot using ResNet34.

METHOD	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Avg
S+V	58.1	61.8	57.7	51.5	55.4	49.1	73.1	58.1
DANN	61.2	62.3	56.4	54.0	57.9	55.9	65.6	59.0
CDAN	65.0	64.9	63.7	53.1	63.4	54.5	73.2	62.5
ENT	60.0	60.2	54.9	48.3	55.8	49.4	74.4	57.6
MME	69.5	68.1	64.4	56.7	62.0	59.2	76.9	65.3
APE	70.4	70.8	72.9	56.7	64.5	63.0	76.6	67.6
CDAC	77.4	74.2	75.5	67.6	71.0	69.2	80.4	73.6
DECOTA	79.1	74.9	76.9	65.1	72.0	69.7	79.6	73.9
CLDA	76.1	75.1	71.0	63.7	70.2	67.1	80.1	71.9
CDAC+SLA	**79.8**	75.6	77.4	68.1	71.7	71.7	80.4	75.0
ECB†	77.9	**76.2**	78.4	68.3	73.0	**73.9**	81.0	75.5
EFTL†	79.4	75.2	77.8	68.5	72.5	70.0	**82.9**	75.2
Ours	78.5	76.1	**79.3**	**69.7**	**73.6**	73.8	81.9	**76.2**

Table 10. Accuracy (%) on DomainNet under the setting of 1-shot using ResNet34.

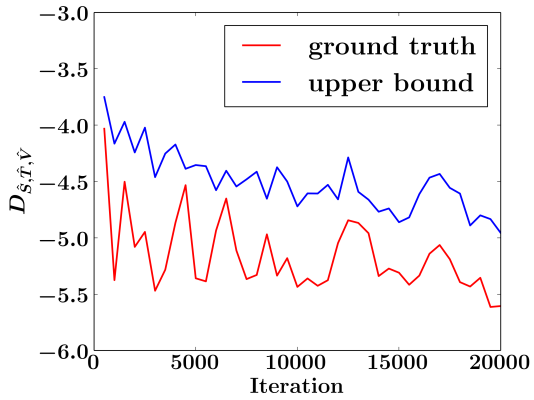


Figure 9. The estimated ground truth and upper bound of $D_{\hat{S},\hat{T},\hat{V}}$ from Product to Clipart scenario in Office-Home dataset.

instead. We choose an adaptation scenario with a large domain shift (Product to Clipart scenario of the Office-Home dataset). We use the full source and target labels to compute $D_{\hat{S},\hat{T},\hat{V}}(f_S^*, f_T^*, f_V^*, h)$ as the ground truth. The upper bound $D_{\hat{S},\hat{T},\hat{V}}(f_S', f_T', f_V', h)$ is given by the maximum inside subspace $\mathcal{H}_S, \mathcal{H}_T, \mathcal{H}_V$ from Definition 3.10. Fig.9 demonstrates that our proposal remains a valid upper bound in practice even if the domain shift is so large that the subspace $\mathcal{H}_T$ we built is not likely to contain f_T^*.

D.9. Label Shift

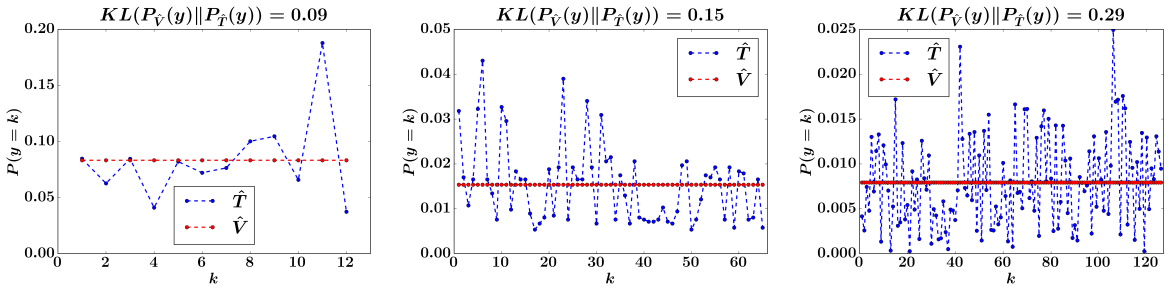
We plot the empirical label distributions of labeled and unlabeled target domains for different datasets in Fig. 10. $P_{\hat{V}}(y)$ is a uniform distribution, and $P_{\hat{T}}(y)$ varies from it in most categories, which characterizes a clear distribution shift. For instance, in Fig. 10a, $P_{\hat{V}}(y = 12)$ is significantly lower than the average, which may cause the decline of the recognition performance for the 12th category "truck" in VisDA (Tab.1). In contrast, our method outperforms others in this category substantially by addressing the overlooked label shift. In addition, we compute the KL-divergence between label distributions as an intuitive measurement of label shift. Note that a bigger value does not necessarily imply an absolutely larger label shift, as the KL-divergence tends to grow for increasing categories under the same level variation.

D.10. Regarding LIRR [29]

LIRR can be considered as an extension of Theorem 1 in Ben-David et al. [4]:

$$\epsilon_T(h, f_T) \leq \frac{|\hat{T}|}{|\hat{V}| + |\hat{T}|}[\epsilon_S(h, f_S) + d_1(S, T) + \mathcal{I}] + \frac{|\hat{V}|}{|\hat{V}| + |\hat{T}|}\epsilon_V(h, f_T),$$

$$\mathcal{I} = \min(\mathbb{E}_{x\sim S}|f_S(x) - f_T(x)|, \mathbb{E}_{x\sim T}|f_S(x) - f_T(x)|),$$



(a) VisDA (b) Office-Home (c) DomianNet

Figure 10. Empirical label distributions of labeled V/unlabeled target T data plotted in red/blue respectively, and the corresponding KL-divergence between them for different datasets: (a) VisDA; (b) Art domain of Office-Home; (c) Painting domain of DomainNet.

which introduces an intractable term $\mathcal{I}$ instead of the joint error. They further show $\min \mathcal{I}$ equals to reduce the distance between $P_S(y|z), P_T(y|z)$ given $g : \mathcal{X} \to \mathcal{Z}$. Since $P_T(y|z)$ is intractable, they replace it with $P_V(y|z)$ with an assumption $P_V(y|z) \approx P_T(y|z)$ (equivalent to $f_V \approx f_T$ and quite close to $\lambda \to 0$ as they both imply small domain shift) that barely holds in practice unless V and T highly overlap, which is why LIRR needs much more labeled data ($|\hat{V}|$>500) in VisDA to enable the algorithm (still 5% behind our accuracy with $|\hat{V}|$=36). In addition, for real-valued function space $\mathcal{H} : \mathcal{Z} \to [0, 1]$, they derive the generalization error bound with pseudo dimension $Pdim(\mathcal{H}) \geq fat_{\mathcal{H}}(\gamma)$ (fat-shattering dimension), which is not a scale-sensitive notion (not informative as Rademacher complexity) [38] and can tend to infinity. Moreover, in LIRR, the derivation is based on L_1 distance as $\epsilon_S(f, f') = \mathbb{E}_{x\sim S}|f(x) - f'(x)|$, which may not suit classification tasks.