A Theory of Learning Unified Model via Knowledge Integration from Label Space Varying Domains

Supplementary Material

6. Proof

6.1. Proof of Theorem 2.1

Let $f_{S_i}, f_T, f_V : \mathcal{X} \to \mathcal{K}$ 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

$$
\begin{aligned}
\epsilon_T(h, f_T) &= \frac{1}{2}[2\epsilon_T(h, f_T) - \epsilon_V(h, f_V) - \epsilon_{S_i}(h, f_{S_i}) + \epsilon_V(h, f_V) \\
&+ \epsilon_{S_i}(h, f_{S_i}) + \epsilon_T(h, f_{S_i}) + \epsilon_T(h, f_V) - \epsilon_T(h, f_{S_i}) \\
&- \epsilon_T(h, f_V) + \epsilon_V(h, f_{S_i}) + \epsilon_{S_i}(h, f_V) - \epsilon_V(h, f_{S_i}) - \epsilon_{S_i}(h, f_V)] \\
&= \frac{1}{2}([\epsilon_T(h, f_T) - \epsilon_T(h, f_{S_i})] + [\epsilon_T(h, f_T) - \epsilon_T(h, f_V)] \\
&+ \epsilon_T(h, f_{S_i}) + \epsilon_T(h, f_V) + [\epsilon_V(h, f_{S_i}) - \epsilon_V(h, f_V)] \\
&+ [\epsilon_{S_i}(h, f_V) - \epsilon_{S_i}(h, f_{S_i})] - \epsilon_V(h, f_{S_i}) - \epsilon_{S_i}(h, f_V) \\
&+ \epsilon_V(h, f_V) + \epsilon_{S_i}(h, f_{S_i})) \\
&\leq \frac{1}{2}([\epsilon_T(f_{S_i}, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_{S_i}) + \epsilon_T(h, f_V) \\
&+ \epsilon_V(f_{S_i}, f_V) + \epsilon_{S_i}(f_V, f_{S_i}) - \epsilon_V(h, f_{S_i}) - \epsilon_{S_i}(h, f_V)] \\
&+ [\epsilon_V(h) + \epsilon_{S_i}(h)]) = B(h) \\
&\leq \frac{1}{2}[\epsilon_T(f^*_{S_i}, f^*_T) + \epsilon_T(f^*_V, f^*_T) + \epsilon_T(h, f^*_{S_i}) + \epsilon_T(h, f^*_V) \\
&+ \epsilon_V(f^*_{S_i}, f^*_V) + \epsilon_{S_i}(f^*_V, f^*_{S_i}) - \epsilon_V(h, f^*_{S_i}) - \epsilon_{S_i}(h, f^*_V)] \\
&+ \frac{1}{2}[\epsilon_V(h) + \epsilon_{S_i}(h)] \\
&+ \underbrace{\frac{1}{2}\epsilon_{S_i}(f_{S_i}, f^*_{S_i}) + \epsilon_V(f_{S_i}, f^*_{S_i}) + \epsilon_T(f_{S_i}, f^*_{S_i})}_{\theta^i_S} \\
&+ \underbrace{\frac{1}{2}\epsilon_V(f_V, f^*_V) + \epsilon_{S_i}(f_V, f^*_V) + \epsilon_T(f_V, f^*_V)}_{\theta^i_V} + \underbrace{\epsilon_T(f_T, f^*_T)}_{\theta^i_T} \\
&= D_{S_i,V,T}(f^*_{S_i}, f^*_T, f^*_V, h) + \frac{1}{2}[\epsilon_V(h) + \epsilon_{S_i}(h)] + \theta_i
\end{aligned} \tag{19}
$$

We introduce a weight parameter α_i and sum up the upper bound w.r.t. all source domains $S_i, i = 1, .., N$ leading to:

$$
\begin{aligned}
\alpha_i \epsilon_T(h) &\leq \alpha_i \frac{1}{2}[\epsilon_V(h) + \epsilon_{S_i}(h)] + \alpha_i D_{S_i,V,T}(f^*_{S_i}, f^*_V, f^*_T, h) + \alpha_i \theta_i \\
\epsilon_T(h) &\leq \frac{1}{2}\epsilon_V(h) + \frac{1}{2}\sum_i \alpha_i[\epsilon_{S_i}(h) + 2D_{S_i,V,T}(f^*_{S_i}, f^*_V, f^*_T, h) + 2\theta_i] \\
&\leq \frac{1}{2}\epsilon_V(h) + \frac{1}{2}\sum_i \alpha_i U_i(h), \quad s.t. \quad \sum_i \alpha_i = 1
\end{aligned} \tag{20}
$$

6.2. Proof of Theorem 2.3

Definition 6.1 (Covering Number)**.** Let (A, d_p) be a metric space. Set C is an γ-cover of A if for $\forall x \in A, \exists y \in C$ such that $d_p(x, y) = \|x - y\|_p < \gamma$. The covering number $\mathcal{N}(\gamma, A, d_p)$ is the size of the smallest γ-cover.

$$
\mathcal{N}(\gamma, A, d_p) = \min\{|C| \text{ s.t. C is a } \gamma\text{-cover of } A\} \tag{21}
$$

Definition 6.2 (Uniform Covering Number)**.** Let $\mathcal{F}$ be a hypothesis space of real-valued functions. For any $\gamma > 0$, and m the d_p uniform covering number $\mathcal{N}_p(\gamma, F, m)$ is defined as

$$
\mathcal{N}_p(\gamma, \mathcal{F}, m) = \max_{C:|C|=m} \mathcal{N}(\gamma, \mathcal{F}|_C, d_p), \tag{22}
$$

where $\mathcal{F}|_C = \{[f(x_1), .., f(x_m)] \in \mathbb{R}^m | C = \{x_1, .., x_m\}, f \in \mathcal{F}\}$.

Lemma 6.3. *For space $\mathcal{F}$ of real-valued functions:$\mathcal{X} \to [0, M]$, for any distribution D, samples $\hat{D} = \{x_1, .., x_m\}$ i.i.d $\sim D$, $\gamma > 0$ and $m \geq 2/\gamma^2$, Definition 6.2, the following holds given Hoeffding's inequality:*

$$
\begin{aligned}
&P(\sup_{f \in \mathcal{F}} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \sim D} f(x)| \geq \gamma) \\
&\leq 4\mathcal{N}_1(\frac{\gamma}{8}, \mathcal{F}, 2m) exp(-\frac{m\gamma^2}{32M^2})
\end{aligned} \tag{23}
$$

For the proof, we first show $P(\sup_{f \in \mathcal{F}} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \sim D} f(x)| \geq \gamma) \leq 2P(\sup_{f \in \mathcal{F}} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \geq \frac{\gamma}{2})$ if $m \geq 2/\gamma^2$. For a fixed $\hat{D}$ and $f \in \mathcal{F}$ that satisfies $|\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \sim D} f(x)| \geq \gamma$, the following holds given Jensen's inequality:

$$
\mathbb{E}_{\hat{D}' \sim D} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \geq |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \sim D} f(x)| \geq \gamma \tag{24}
$$

For $m \geq 2/\gamma^2$, the following holds given Hoeffding's inequality:

$$
\begin{aligned}
&P(\mathbb{E}_{\hat{D}' \sim D} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \\
&\quad - |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \leq \frac{\gamma}{2}) \geq \frac{1}{2}
\end{aligned} \tag{25}
$$

$$
\begin{aligned}
&P(\mathbb{E}_{\hat{D}' \sim D} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| - \frac{\gamma}{2} \\
&\quad \leq |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)|) \geq \frac{1}{2}
\end{aligned} \tag{26}
$$

$$
P(|\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \geq \frac{\gamma}{2}) \geq \frac{1}{2}, \tag{27}
$$

such that $P(\sup_{f \in \mathcal{F}} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \sim D} f(x)| \geq \gamma) \leq 2P(\sup_{f \in \mathcal{F}} |\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \geq \frac{\gamma}{2})$ for $\hat{D}, \hat{D}' \sim D$.

Let σ_i denote an independent uniform random variable taking values in $\{-1, +1\}$. $f(x_i) - f(x'_i)$ and $\sigma_i(f(x_i) - f(x'_i))$ follows the same distribution as $x_i \in \hat{D} \sim D, x'_i \in \hat{D}' \sim D$ such that for $\forall f \in \mathcal{F}, P(|\mathbb{E}_{x \in \hat{D}} f(x) - \mathbb{E}_{x \in \hat{D}'} f(x)| \geq \frac{\gamma}{2}) = P(\frac{1}{m}|\sum_{i=1}^m \sigma_i(f(x_i) - f(x'_i))| \geq$

$\frac{\gamma}{2}) = \mathbb{E}_{\hat{D},\hat{D}'\sim D} P(\frac{1}{m}|\sum_{i=1}^m \sigma_i(f(x_i) - f(x_i'))| \geq \frac{\gamma}{2}|\hat{D},\hat{D}')$.

For fixed $\hat{D},\hat{D}'$, let $\mathcal{G}(\frac{\gamma}{8})$ denote the smallest $\frac{\gamma}{8}$-cover of $\mathcal{F}|_{\hat{D}\cup\hat{D}'}$ such that $\mathcal{F} \subset \cup_{g\in\mathcal{G}(\frac{\gamma}{8})}\mathcal{F}(g,\frac{\gamma}{8})$ where $\mathcal{F}(g,\frac{\gamma}{8})|_{\hat{D}\cup\hat{D}'}$ is $\frac{\gamma}{8}$-covered by $[g(x_1),..,g(x_m')] \in \mathbb{R}^{2m}$. Note that for $\forall f \in \mathcal{F}(g,\frac{\gamma}{8})$,

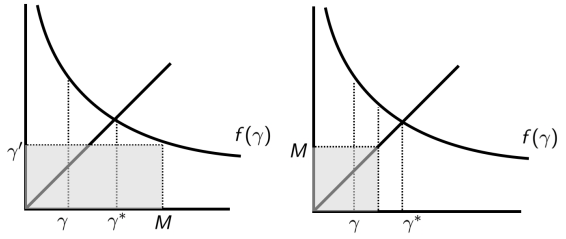
$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m |\sigma_i(f(x_i) - g(x_i))| + \frac{1}{m}\sum_{i=1}^m |\sigma_i(f(x_i') - g(x_i'))| \leq \frac{\gamma}{4}\\ \Rightarrow &\frac{1}{m}|\sum_{i=1}^m \sigma_i(f(x_i) - f(x_i'))| - \frac{1}{m}|\sum_{i=1}^m \sigma_i(g(x_i) - g(x_i'))| \leq \frac{\gamma}{4} \quad (28)\end{aligned}$$

Then we can proceed to the upper bounded by

$$\begin{aligned}&P(\sup_{f\in\mathcal{F}}\frac{1}{m}|\sum_{i=1}^m \sigma_i(f(x_i) - f(x_i'))| \geq \frac{\gamma}{2})\\ &= P(\sup_{g\in\mathcal{G}_{\frac{\gamma}{8}}}\sup_{f\in\mathcal{F}(g,\frac{\gamma}{8})}\frac{1}{m}|\sum_{i=1}^m \sigma_i(f(x_i) - f(x_i'))| \geq \frac{\gamma}{2})\\ &= P(\exists g \in \mathcal{G}_{\frac{\gamma}{8}} : \sup_{f\in\mathcal{F}(g,\frac{\gamma}{8})}\frac{1}{m}|\sum_{i=1}^m \sigma_i(f(x_i) - f(x_i'))| \geq \frac{\gamma}{2})\\ &\leq \sum_{g\in\mathcal{G}_{\frac{\gamma}{8}}} P(\frac{1}{m}|\sum_{i=1}^m \sigma_i(g(x_i) - g(x_i'))| \geq \frac{\gamma}{4})\\ &\leq 2\mathcal{N}_1(\frac{\gamma}{8},\mathcal{F},2m)exp\{-\frac{m^2\gamma^2}{32\sum_{i=1}^m[g(x_i) - g(x_i')]^2}\}\\ &\leq 2\mathcal{N}_1(\frac{\gamma}{8},\mathcal{F},2m)exp(-\frac{m\gamma^2}{32M^2}) \quad (29)\end{aligned}$$

For any $0 < \delta < 1$, we solve the following for γ to complete the generalization bound:

$$\begin{aligned}&\delta = 4\mathcal{N}_1(\frac{\gamma}{8},\mathcal{F},2m)exp(-\frac{m\gamma^2}{32M^2})\\ \Rightarrow &\gamma = M\sqrt{\frac{32}{m}\log\frac{2\mathcal{N}_1(\frac{\gamma}{8},\mathcal{F},2m)}{\delta}} = f(\gamma) \quad (30)\end{aligned}$$



(a) $c_1 = \gamma'/\gamma^*$ (b) $c_2 = M/\gamma^*$

Figure 8

Corollary 6.4. *According to Fig. 8, there exists universal constant $c = \min(c_1, c_2) < 1$ such that:*

$$cM\gamma^* \leq \inf_{\gamma\leq M}(\int_\gamma^M f(\xi)d\xi + M\gamma) \quad (31)$$

Corollary 6.5. *Let $\hat{D}$ denote a finite set with size m sampled i.i.d. from a distribution D. For hypotheses $f, f' \in \mathcal{H} : \mathcal{X} \to \mathcal{K}$, let $\mathcal{F} = \{f(x) = \epsilon(h(x), h'(x)) : \mathcal{X} \to [0, M]|h, h' \in \mathcal{H}\}$ be a real-valued function space. Given Theorem 6.3, Corollary 6.4, for $m \geq 2/\gamma^2$ and $0 < \delta < 1$, with probability at least $1 - \frac{\delta}{11N+6}$,*

$$\begin{aligned}&\epsilon_D(f, f') \leq \epsilon_{\hat{D}}(f, f')\\ &+ \mathcal{O}\left(\inf_{\sqrt{\frac{2}{m}}\leq\gamma\leq M}(\gamma + \int_\gamma^M \sqrt{\frac{1}{m}\log\frac{2(11N+6)\mathcal{N}_1(\frac{\xi}{8},\mathcal{F},2m)}{\delta}}d\xi)\right) \quad (32)\end{aligned}$$

Lemma 6.6. *For hypothesis space $\mathcal{F} = \{f(x) = \epsilon(h(x), h'(x)) : \mathcal{X} \to [0, M]|h, h' \in \mathcal{H}\}$, $\mathcal{F}' = \{f(x) = \epsilon(h(x), f_{S_i}^*(x)) : \mathcal{X} \to [0, M]|h \in \mathcal{H}\}$, $\mathcal{F}'' = \{f(x) = \epsilon(h(x), f_{S_i}(x)) : \mathcal{X} \to [0, M]|h \in \mathcal{H}\}$, the following holds:*

$$\mathcal{N}_1(\gamma,\mathcal{F},m) \geq \mathcal{N}_1(\gamma,\mathcal{F}',m) \geq \mathcal{N}_1(\gamma + \Delta\gamma,\mathcal{F}'',m) \quad (33)$$

The left inequality is trivial as $\mathcal{F}' \subset \mathcal{F}$ such that for any $C = \{x_1, .., x_m\}$, a smallest γ-cover of $\mathcal{F}|_C$ is also a γ-cover for $\mathcal{F}'|_C$. For the right inequality, let $\mathcal{G}|_C$ denote the γ-cover of $\mathcal{F}'|_C$ such that for $\forall h \in \mathcal{H}$, there exists $g \in \mathcal{G}$:

$$\begin{aligned}&\sum_{j=1}^m |\epsilon(h(x_j), f_{S_i}^*(x_j)) - \epsilon(g(x_j), f_{S_i}^*(x_j))| \leq \gamma\\ &\Rightarrow \sum_{j=1}^m |\epsilon(h(x_j), f_{S_i}(x_j)) - \epsilon(g(x_j), f_{S_i}(x_j))|\\ &\leq \gamma + 2\sum_{j=1}^m \epsilon(f_{S_i}^*(x_j), f_{S_i}(x_j)) = \gamma + \Delta\gamma|_C \quad (34)\end{aligned}$$

Let $\Delta\gamma = \max_{C:|C|=m}\Delta\gamma|_C$ such that for any $C = \{x_1, .., x_m\}$, a smallest γ-cover of $\mathcal{F}'|_C$ is also a $\gamma + \Delta\gamma$-cover for $\mathcal{F}''|_C$. Given Assumption 2.2, for any finite set C, there exist $f_{S_i}^* \in \mathcal{H}$ such that $\epsilon_C(f_{S_i}^*, f_{S_i}) \to 0 \Rightarrow \Delta\gamma \approx 0$ and we can conclude $\mathcal{N}_1(\gamma,\mathcal{F}',m) \gtrapprox \mathcal{N}_1(\gamma,\mathcal{F}'',m)$.

Given Theorem 2.1, for any labeling functions $f_{S_i}^* \in \mathcal{H}_{S_i} \subseteq \mathcal{H}, f_T^* \in \mathcal{H}_T \subseteq \mathcal{H}, f_V^* \in \mathcal{H}_V \subseteq \mathcal{H}$, the expected target error is bounded for $\forall h \in \mathcal{H}$:

$$\epsilon_T(h, f_T) = \frac{1}{2}[\epsilon_V(h) + \epsilon_{S_i}(h)] + D_{S_i,T,V}(f_{S_i}^*, f_T^*, f_V^*, h) + \theta_i \quad (35)$$

Corollary 6.7. *According to Assumption 2.2, there exist $f_{S_i}^* \in \mathcal{H}_{S_i} \subseteq \mathcal{H}, f_V^* \in \mathcal{H}_V \subseteq \mathcal{H}, f_T^* \in \mathcal{H}_T \subseteq \mathcal{H}$ such that $\sum_i \alpha_i\hat{\theta}_i \approx 0$. Given Eqs. (32) and (35), Lemma 6.6, for $0 < \delta < 1$, with probability at least $1 - \delta$, for $\forall h \in \mathcal{H}$:*

$$2\epsilon_T(h) \le \epsilon_{\hat{V}}(h) + \sum_{i=1}^{N} \alpha_i \hat{U}_i(h) + \mathcal{O}\left(\inf_{\sqrt{\frac{2}{m}} \le \gamma \le M} (\gamma + \int_{\gamma}^{M} \sqrt{\frac{1}{m} \log \frac{2(11N+6)\mathcal{N}_1(\frac{\xi}{8}, \mathcal{F}, 2m)}{\delta}} d\xi) \right), \tag{36}$$

$$\hat{U}_i(h) = \epsilon_{\hat{S}_i}(h) + 2D_{\hat{S}_i,\hat{V},\hat{T}}(f^*_{S_i}, f^*_V, f^*_T, h) \tag{37}$$

Let $\alpha_i = \frac{exp(\nu \hat{U}_i(h))}{\sum_j exp(\nu \hat{U}_j(h))}$ denote the log-sum-exp trick. for $\nu > 0$, given Jensen's & Cauchy's inequality, we can derive:

$$\begin{aligned} \frac{1}{\nu} \sum_i \alpha_i \nu \hat{U}_i(h) &\le \frac{1}{\nu} \log \mathbb{E}_\alpha [exp(\nu \hat{U}_i(h))] \\ &= \frac{1}{\nu} \log \frac{\sum_i exp^2(\nu \hat{U}_i(h))}{\sum_i exp(\nu \hat{U}_i(h))} \\ &\le \frac{1}{\nu} \log \sum_i exp(\nu \hat{U}_i(h)) \end{aligned} \tag{38}$$

Combing Corollary 6.7, Eq. (38), Theorem 2.3 can be proved.

6.3. Proof of Lemma 3.4

Corollary 6.8. *Given Open-set Margin Discrepancy (Definition 3.1) and Unknown Predictive Discrepancy (Definition 3.2), ϵ measured on unknown class K can be related to υ for $\forall f \in \mathcal{H}$,*

$$\epsilon_{S_i^K}(f, f^*_{S_i}) = \upsilon_{S_i^K}(f, f^*_{S_i}) \tag{39}$$

$$\epsilon_{V^K}(f, f^*_V) = \upsilon_{V^K}(f, f^*_V) \tag{40}$$

$$\epsilon_{T^K}(f, f^*_T) = \upsilon_{T^K}(f, f^*_T) \tag{41}$$

For the proof, the Open-set Margin Discrepancy between $f, f^*_T \in \mathcal{H}$ over T^K is defined by,

$$\epsilon_{T^K}(f, f^*_T) = \mathbb{E}_{x \sim T^K}[\text{omd}(f(x), f^*_T(x))] \tag{42}$$

$$\begin{aligned} \text{omd}(f(x), f^*_T(x)) = \max(&|\log(1 - f(x)[y]) - \log(1 - f^*_T(x)[y])|, \\ &|\log(1 - f^*_T(x)[y^*]) - \log(1 - f(x)[y^*])|), \end{aligned} \tag{43}$$

where $y = l(f(x)), y^* = l(f^*_T(x))$. When measuring on $x \sim T^K$, $y^* = K$ and $f^*_T(x)[K] \approx 1$ since f^*_T is the approximated labeling function of target domain. Therefore, we can derive that,

$$\begin{aligned} \epsilon_{T^K}(f, f^*_T) &= \mathbb{E}_{x \sim T^K} |\log(1 - f^*_T(x)[K]) - \log(1 - f(x)[K])| \\ &= \upsilon_{T^K}(f, f^*_T), \end{aligned} \tag{44}$$

where the rest can be proved analogously. Given the label distribution $\pi^k_{S_i}, \pi^k_V, \pi^k_T, k = \{1, 2, ..., K\}$, Assumption 3.3 and Corollary 6.8, we can reformulate the following terms with $g : \mathcal{X} \to \mathcal{Z}$ and $h, f^*_V : \mathcal{Z} \to \mathcal{K}$:

$$\begin{aligned} &\epsilon_{S_i}(f^*_V \circ g, h \circ g) \\ &= \sum_{k=1}^{K-1} \pi^k_{S_i} \epsilon_{S_i^k}(f^*_V \circ g, h \circ g) + \pi^K_{S_i} \epsilon_{S_i^K}(f^*_V \circ g, h \circ g) \\ &= \sum_{k=1}^{K-1} \pi^k_{S_i} \epsilon_{S_i^k}(f^*_V \circ g, h \circ g) + \pi^K_{S_i} \upsilon_{S_i^K}(f^*_V \circ g, h \circ g) \end{aligned} \tag{45}$$

$$\begin{aligned} &\upsilon_T(f^*_V \circ g, h \circ g) \\ &= \sum_{k=1}^{K-1} \pi^k_T \upsilon_{T^k}(f^*_V \circ g, h \circ g) + \pi^K_T \upsilon_{T^K}(f^*_V \circ g, h \circ g) \\ &= \sum_{k=1}^{K-1} \pi^k_T \upsilon_{T^k}(f^*_V \circ g, h \circ g) + \pi^K_T \upsilon_{S_i^K}(f^*_V \circ g, h \circ g) \end{aligned} \tag{46}$$

By excluding the intractable term $\upsilon_{S^K}(f^*_V \circ g, h \circ g)$ due to the shortage of source data in class K, we can approximate the expected discrepancy on S_i by $S_i^{\setminus K}, T$ with a mild condition that $\pi^K_{S_i} = \pi^K_T = 1 - \alpha$:

$$\begin{aligned} &\epsilon_{S_i}(f^*_V \circ g, h \circ g) \\ &= \sum_{k=1}^{K-1} \pi^k_{S_i} \epsilon_{S_i^k}(f^*_V \circ g, h \circ g) + \frac{\pi^K_{S_i}}{\pi^K_T} [\upsilon_T(f^*_V \circ g, h \circ g) \\ &- \sum_{k=1}^{K-1} \pi^k_T \upsilon_{T^k}(f^*_V \circ g, h \circ g)] \\ &= \alpha[\epsilon_{S_i^{\setminus K}}(f^*_V \circ g, h \circ g) - \upsilon_{S_i^{\setminus K}}(f^*_V \circ g, h \circ g)] + \upsilon_T(f^*_V \circ g, h \circ g), \end{aligned} \tag{47}$$

where the rest can be proved analogously.

7. Towards Joint Error

In this section, we prove that our proposal is an upper bound of joint error. For simplicity, we consider a single source domain S. Given Eq. (19), $\forall h \in \mathcal{H}$, our upper bound is further lower bounded by:

$$\begin{aligned} B(h) &= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) + \epsilon_V(f_S, f_V) \\ &\quad + \epsilon_S(f_V, f_S) - \epsilon_V(h, f_S) - \epsilon_S(h, f_V)] + [\epsilon_V(h) + \epsilon_S(h)]) \\ &= \frac{1}{2}([\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V) - \epsilon_V(h, f_S) \\ &\quad - \epsilon_S(h, f_V)] + [\epsilon_V(h, f_V) + \epsilon_V(f_S, f_V) + \epsilon_S(h, f_S) + \epsilon_S(f_V, f_S)]) \\ &\ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h, f_S) + \epsilon_T(h, f_V)] \end{aligned} \tag{48}$$

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

$$\begin{aligned} \min_{h \in \mathcal{H}} B(h) &= B(h^*) \\ &\ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(h^*, f_S) + \epsilon_T(h^*, f_V)] \\ &\ge \frac{1}{2}[\epsilon_T(f_S, f_T) + \epsilon_T(f_V, f_T) + \epsilon_T(f_V, f_S)] \end{aligned} \tag{49}$$

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

$$\epsilon_T(f_S, f_T) = \epsilon_T(f_S, f_T) + \epsilon_S(f_S, f_S) \geq \min_{h\in\mathcal{H}}(\epsilon_T(h) + \epsilon_S(h)) = \lambda_{S,T} \tag{50}$$

Otherwise, we can always derive:

$$\epsilon_T(f_S, f_T) \geq \epsilon_T(f_S^*, f_T) - \epsilon_T(f_S^*, f_S) + \epsilon_S(f_S^*, f_S) - \epsilon_S(f_S^*, f_S) \tag{51}$$

Let $\hat{S}, \hat{T}$ denote a finite set with size m from domain S, T. According to Assumption 2.2, 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 Uniform Covering Number (Definition 6.2). Given function space $\mathcal{F}_S = \{f(x) = \epsilon(h(x), f_S(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$, for any $\delta > 0$, with probability at least $1 - 2\delta$,

$$\epsilon_T(f_S, f_T) \geq \lambda_{S,T} - \underbrace{[\epsilon_{\hat{S}}(f_S^*, f_S) + \epsilon_{\hat{T}}(f_S^*, f_S)]}_{\text{zero}} - \mathcal{O}\left(\inf_{\sqrt{\frac{2}{m}}\leq\gamma\leq M}(\gamma + \int_\gamma^M \sqrt{\frac{1}{m}\log\frac{2\mathcal{N}_1(\frac{\xi}{8}, \mathcal{F}_S, 2m)}{\delta}}d\xi)\right) \tag{52}$$

8. Approximated Labeling Function Assumption vs. Joint Error Assumption

E.g., let $\hat{T}$ denote a finite set with size m from target domain T. According to Corollary 6.5, given the Uniform Covering Number (Definition 6.2) of function space $\mathcal{F}_T = \{f(x) = \epsilon(h(x), f_T(x)) : \mathcal{X} \to [0, M] | h \in \mathcal{H}\}$, 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 $\forall f_T^* \in \mathcal{H}$:

$$\epsilon_T(f_T^*, f_T) \leq \epsilon_{\hat{T}}(f_T^*, f_T) + \mathcal{O}\left(\inf_{\sqrt{\frac{2}{m}}\leq\gamma\leq M}(\gamma + \int_\gamma^M \sqrt{\frac{1}{m}\log\frac{2\mathcal{N}_1(\frac{\xi}{8}, \mathcal{F}_T, 2m)}{\delta}}d\xi)\right) \tag{53}$$

We assume there exists $f_T^* \in \mathcal{H}$ such that in Theorem 2.3, $\hat{\theta}_{f_T} \approx 0$ thus can be ignored during the practical learning process. For simplicity, we consider a single source domain S. We show Assumption 2.2 is more feasible than assuming empirical joint error $\lambda_{\hat{S},\hat{T}} \approx 0$ in [1], especially when the domain shift is large. To facilitate the analysis, let $g : \mathcal{X} \subseteq \mathbb{R}^D \to \mathcal{Z} \subseteq \mathbb{R}^F$ 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 $h \in \mathcal{H}^F : \mathcal{Z} \to \mathcal{K}$,

$$\begin{aligned}
&\min_{h\in\mathcal{H}^F}[\epsilon_{\hat{S}}(h \circ g, f_S) + \epsilon_{\hat{T}}(h \circ g, f_T)] = \lambda_{\hat{S},\hat{T}} \\
&= \min_{h\in\mathcal{H}^F}[\frac{m-c}{m}\epsilon_{\hat{Z}_S}(h, f_S^F) + \frac{m-c}{m}\epsilon_{\hat{Z}_T}(h, f_T^F) \\
&+ \frac{c}{m}\epsilon_{\hat{Z}_C}(h, f_S^F) + \frac{c}{m}\epsilon_{\hat{Z}_C}(h, f_T^F)] \\
&\geq \frac{m-c}{m}\min_{h\in\mathcal{H}^F}\epsilon_{\hat{Z}_S}(h, f_S^F) + \frac{m-c}{m}\min_{h\in\mathcal{H}^F}\epsilon_{\hat{Z}_T}(h, f_T^F) \\
&+ \frac{c}{m}\epsilon_{\hat{Z}_C}(f_T^F, f_S^F),
\end{aligned} \tag{54}$$

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

$$\left.\begin{array}{ll} -z, & z \in \hat{Z}_S \\ z, & z \in \hat{Z}_C \\ -z+2, & z \in \hat{Z}_T \end{array}\right\} = h(z) \notin \mathcal{H}^F = \{z \mapsto a|z-b| + c | a, b, c \in \mathbb{R}\} \tag{55}$$

9. Consistency

In this section, we recall a general problem associated with the consistency between the algorithm and theory in domain adaptation. ϵ should be a consistent distance metric across the measurement of source error and discrepancy according to the derivation of any target error upper bound. However, most works violate this consistency as known as the gap between the algorithm and theory. Although our proposal cannot perfectly address this problem, we can prove that Open-set Margin Discrepancy (OMD) in Definition 3.1 asymptotically satisfies the consistency.

Firstly, we show that OMD 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 almost met when we derive the upper bound in Theorem 1 except for f_T, h),

$$\begin{aligned}
&\text{omd}(h_1(x), h_3(x)) + \text{omd}(h_2(x), h_3(x)) \\
&= \max(|\log(1 - h_1(x)[y]) - \log(1 - h_3(x)[y])|, \\
&\quad |\log(1 - h_1(x)[y']) - \log(1 - h_3(x)[y'])|) \\
&+ \max(|\log(1 - h_2(x)[y]) - \log(1 - h_3(x)[y])|, \\
&\quad |\log(1 - h_2(x)[y']) - \log(1 - h_3(x)[y'])|) \\
&\geq |\log(1 - h_1(x)[y]) - \log(1 - h_3(x)[y])| \\
&+ |\log(1 - h_2(x)[y]) - \log(1 - h_3(x)[y])| \\
&\geq |\log(1 - h_1(x)[y]) - \log(1 - h_2(x)[y])| \\
&= \text{omd}(h_1(x), h_2(x))
\end{aligned} \tag{56}$$

As the training proceeds, the target error of h will be minimized such that 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 OMD asymptotically satisfies the triangle inequality.

Then we show that the cross-entropy loss is a special case of OMD by reasonably assuming $f_{S_i}(x)[y] = 1$ and $l(f_{S_i}(x)) = l(h(x)) = y$ for $(x, y) \in S_i$. According to Definition 3.1, the source error of h defined based on OMD can be written as:

$$\begin{aligned}
\min_{h\in\mathcal{H}}\epsilon_{S_i}(h) &= \min_{h\in\mathcal{H}}\mathbb{E}_{x\sim S_i}[\text{omd}(h(x), f_{S_i}(x))] \\
&= \min_{h\in\mathcal{H}}\mathbb{E}_{x,y\sim S_i}|\log(1 - f_{S_i}(x)[y]) - \log(1 - h(x)[y])| \\
&\Rightarrow \min_{h\in\mathcal{H}}\mathbb{E}_{x,y\sim S_i}\log(1 - h(x)[y])
\end{aligned} \tag{57}$$

METHOD	TYPE	STATS	Office-Home				DomainNet			
			→Clipart	→Product	→RealWorld	→Art	→Clipart	→Painting	→Real	→Sketch
UM	Multi-Source	mean	68.0	79.0	79.4	67.7	70.3	66.0	75.1	66.1
		std	0.38	0.49	0.12	0.23	0.09	0.44	0.11	0.17
UM+AFG	Source-Free	mean	61.1	77.0	72.0	60.3	64.8	60.0	67.6	60.0
		std	0.33	0.08	0.35	0.88	0.46	0.09	0.21	0.37

Table 5. Statistics of HOS (%) score with ResNet-50 model fine-tuned under 1-shot setting

METHOD	TYPE	→Clipart	→Product	→RealWorld	→Art
		OS*	OS*	OS*	OS*
$S+V+L_{sim}$	Source-Combine	63.5	83.3	80.1	66.3
$S+V+L_{ssl}$		65.3	86.4	83.7	71.7
$S+V+L_{ssl}+L_{sim}$		66.7	86.9	84.5	72.5
$V+L_{ssl}+L_{sim}$	Source-Free	12.1	63.3	56.8	14.2
$AFG+V+L_{sim}$		54.2	77.8	81.5	67.4
$AFG+V+L_{ssl}$		63.3	80.7	83.0	69.1
$AFG+V+L_{ssl}+L_{sim}$		62.9	81.6	85.1	70.7

Table 6. Accuracy of ResNet-50 model fine-tuned with 1-shot semi-supervised learning on Office-Home dataset

In practice, we optimize $-\log h(x)[y]$ instead to avoid exploding or vanishing gradient.

10. Results

10.1. Accuracy

Full tables of the results in the main paper are provided as Tabs. 7 to 10.

10.2. Statistics

For each sub-task in Office-Home and DomainNet datasets, we ran the experiment 3 times with different random seeds. Tab. 5 provides our method's mean accuracy and standard deviation.

10.3. Effectiveness of Attention-based Feature Generation

As suggested in [63], a high OS* score is crucial to improve the model performance in open-set problems as we can always trade the accuracy of known class for more UNK. To briefly demonstrate the effectiveness of AFG, we replace the true source data with generated labeled features under the 1-shot semi-supervised learning setting without any adaptation or unknown separation strategy. Tab. 6 indicates that the labeled features produced by AFG are adequate for learning a reliable classification model of known class in target data.

11. Semi-supervised & Self-supervised Learning

To build a more reliable target function space $\mathcal{H}_T^F$ and facilitate the feature alignment for unknows, we introduce semi-supervised and self-supervised regularization $L_{ssl} = L_{ent} + L_{pse} + L_{con}$ and L_{sim}.

Regularized Entropy Minimization As introduced in [14, 39, 46], we impose a class balance prior that can penalize classifiers with complex decision boundaries on entropy minimization [15] to yield a more sensible solution:

$$L_{ent} = -\mathbb{E}_{x\in\hat{T}} \sum\nolimits_{y\in\mathcal{Y}} f'_T(g(x))[y] \log f'_T(g(x))[y] + \sum\nolimits_{y\in\mathcal{Y}} \mathbb{E}_{x\in\hat{T}} f'_T(g(x))[y] \log \mathbb{E}_{x\in\hat{T}} f'_T(g(x))[y] \quad (58)$$

Pseudo Labeling As introduced in [44, 46], for input $x \in \hat{T}$ and its random augmentation x' [7], we minimize cross entropy for x with pseudo labels of x':

$$L_{pse} = -\mathbb{E}_{x\in\hat{T}} \log f'_T(g(x))[\arg\max_{y\in\mathcal{Y}} h(g(x'))[y]] \quad (59)$$

Consistency Regularization As introduced in [22, 42], we penalize the difference of the outputs for input $x \in \hat{T}$ and its random augmentation x':

$$L_{con} = \mathbb{E}_{x\in\hat{T}} |f'_T(g(x)) - f'_T(g(x'))| \quad (60)$$

Contrastive Regularization Contrastive learning [5] considers every instance as a class of its own and tries to maximize the similarity of features between $x \in \hat{T}$ and its random augmentation x' while pushing different instances far away:

$$L_{sim} = -\mathbb{E}_{x\in\hat{T}} \log \frac{\exp(g(x)\cdot g(x')^\intercal)}{\exp(g(x)\cdot g(x')^\intercal) + \sum_{x''\in\hat{T}:x''\neq x} \exp(g(x)\cdot g(x'')^\intercal)} \quad (61)$$

METHOD	TYPE	→Clipart			→Product			→RealWorld			→Art			Avg.		
		UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS
OSBP	Source-Combine	63.4	57.6	60.4	76.3	64.9	70.1	77.2	63.5	69.7	77.9	49.7	60.7	73.7	58.9	65.2
PGL		68.0	52.1	59.0	71.6	64.3	67.7	73.5	61.0	66.7	69.8	54.5	61.2	70.7	57.9	63.7
ANNA		79.6	56.0	65.8	76.1	66.5	71.0	77.6	64.3	70.3	82.4	48.4	61.0	78.9	58.8	67.0
PUJE		74.8	58.7	65.8	67.6	80.0	73.3	72.1	78.2	75.0	71.2	60.6	65.5	71.4	69.3	69.9
MOSDANET	Multi-Source	63.4	59.6	61.5	67.9	72.3	70.0	71.8	71.0	71.4	67.1	57.0	61.6	67.6	64.9	66.1
HyMOS		59.3	54.1	56.6	67.3	61.8	64.4	75.8	58.8	66.2	70.6	50.7	59.0	68.3	56.4	61.6
UM		78.7	59.9	68.0	78.7	79.3	79.0	82.5	76.5	79.4	72.8	63.3	67.7	78.2	69.8	73.5
MPU*	Source-Free	48.8	44.1	46.3	57.7	61.9	59.7	62.8	53.6	57.8	60.3	56.5	58.3	57.4	54.0	55.5
OSBP*		35.6	59.3	44.5	45.3	72.0	55.6	53.2	66.9	59.3	49.2	63.9	55.6	45.8	65.5	53.8
PUJE*		56.1	48.8	52.2	70.3	60.5	65.0	72.4	61.0	66.2	60.4	57.0	58.7	64.8	56.8	60.5
UM+AFG		70.9	53.8	61.1	83.3	71.6	77.0	73.7	70.4	72.0	66.1	55.5	60.3	73.5	62.8	67.6

Table 7. Accuracy of ResNet-50 model fine-tuned on Office-Home dataset under 1-shot setting

METHOD	TYPE	→Clipart			→Product			→RealWorld			→Art			Avg.		
		UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS
OSBP	Source-Combine	78.7	52.0	62.6	78.9	66.8	72.3	69.0	67.7	68.3	75.9	55.7	64.3	75.6	60.6	66.9
PGL		75.6	52.3	61.8	81.7	61.1	69.9	71.9	66.1	68.9	66.3	61.9	64.0	73.9	60.4	66.2
ANNA		70.0	60.0	67.7	79.7	68.1	73.4	71.3	69.3	70.3	75.0	55.3	63.7	74.0	63.2	68.8
PUJE		79.8	65.0	71.7	68.3	81.1	74.2	72.8	84.3	78.1	72.2	63.0	67.3	73.3	73.4	72.8
MOSDANET	Multi-Source	74.3	59.2	65.9	73.5	74.0	73.8	68.3	70.8	69.6	68.2	59.6	63.6	71.1	65.9	68.2
HyMOS		71.7	58.4	64.4	63.1	72.1	67.3	67.9	68.8	68.4	68.4	57.0	62.2	67.8	64.1	65.6
UM		74.6	69.8	72.1	80.5	85.6	83.0	79.8	81.9	80.8	73.9	67.1	70.3	77.2	76.1	76.6
MPU*	Source-Free	62.5	48.2	54.4	64.7	67.9	66.3	72.2	51.6	60.2	77.5	52.3	62.5	69.2	55.0	60.9
OSBP*		58.0	55.1	56.5	56.3	77.3	65.1	57.7	72.7	64.3	60.1	59.8	59.9	58.0	66.2	61.5
PUJE*		60.6	56.4	58.4	74.3	66.7	70.3	75.2	65.4	70.0	66.6	59.3	62.7	69.2	62.0	65.4
UM+AFG		71.2	61.5	66.0	81.1	79.1	80.1	82.3	75.6	78.8	69.4	60.4	64.6	76.0	69.2	72.4

Table 8. Accuracy of ResNet-50 model fine-tuned on Office-Home dataset under 3-shot setting

METHOD	TYPE	→Clipart			→Painting			→Real			→Sketch			Avg.		
		UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS
OSBP	Source-Combine	64.2	46.9	54.2	57.5	43.9	49.8	81.8	50.7	62.6	64.5	40.1	49.5	67.0	45.4	54.0
PGL		73.0	50.6	59.8	67.1	53.2	59.4	71.9	63.3	67.4	68.1	53.1	59.7	70.0	55.1	61.6
ANNA		60.0	51.7	55.6	63.8	46.2	53.6	77.7	59.6	67.5	66.2	51.4	57.9	66.9	52.2	58.7
PUJE		66.3	62.8	64.4	62.4	57.5	59.8	65.2	70.4	67.7	64.3	58.3	61.2	64.6	62.3	63.3
MOSDANET	Multi-Source	72.3	46.2	56.4	65.1	48.5	55.6	69.6	67.5	68.5	65.4	46.1	54.1	68.1	52.1	58.7
HyMOS		63.4	45.6	53.0	65.4	46.1	54.1	78.8	55.5	65.1	61.0	52.5	56.4	67.2	49.9	57.2
UM		77.5	64.5	70.3	77.9	57.3	66.0	74.9	75.2	75.1	80.6	56.0	66.1	77.7	63.3	69.4
MPU*	Source-Free	61.4	49.0	54.5	51.6	58.8	55.0	60.5	64.4	62.4	58.2	41.5	48.4	57.9	53.4	55.1
MOSDANET*		58.7	57.7	58.1	53.1	55.5	54.3	57.2	70.6	63.2	50.2	48.6	49.4	54.8	58.1	56.3
PUJE*		63.8	57.5	60.5	58.9	52.1	55.3	71.1	58.2	64.0	56.3	50.2	53.1	62.5	54.5	58.2
UM+AFG		72.6	58.6	64.8	63.9	56.5	60.0	76.6	60.6	67.6	67.7	53.8	60.0	70.2	57.4	63.1

Table 9. Accuracy of ResNet-50 model fine-tuned on DomainNet dataset under 1-shot setting

METHOD	TYPE	→Clipart			→Painting			→Real			→Sketch			Avg.		
		UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS
OSBP	Source-Combine	64.1	51.9	57.4	67.1	43.9	53.1	80.3	53.2	64.0	56.6	45.0	50.1	67.0	48.5	56.2
PGL		68.3	56.8	62.0	67.5	56.3	61.4	73.6	65.7	69.4	68.2	55.5	61.2	69.4	58.6	63.5
ANNA		69.7	55.0	61.5	63.3	47.6	54.3	74.4	60.1	66.5	66.6	51.5	58.1	68.5	53.6	60.1
PUJE		63.9	68.6	66.2	65.5	58.4	61.7	68.1	70.6	69.3	71.7	58.2	64.2	67.3	64.0	65.4
MOSDANET	Multi-Source	67.3	46.9	55.3	64.5	53.0	58.2	70.5	69.1	69.8	66.4	46.9	54.9	67.2	54.0	59.6
HyMOS		61.0	49.1	54.4	63.3	50.2	56.0	77.8	59.4	67.4	60.1	54.4	57.1	65.6	53.3	58.7
UM		75.0	68.3	71.5	82.1	59.2	68.8	78.3	78.8	78.5	73.2	66.2	69.5	77.2	68.1	72.1
MPU*	Source-Free	64.4	52.1	57.6	74.2	50.5	60.1	67.6	65.3	66.4	64.2	44.9	52.9	67.6	53.2	59.3
MOSDANET*		60.7	60.3	60.5	59.0	59.7	59.3	56.3	70.4	62.5	54.3	54.4	54.3	57.6	61.2	59.2
PUJE*		66.1	58.8	62.2	63.4	59.6	61.4	72.1	64.0	67.8	62.5	51.1	56.2	66.0	58.4	61.9
UM+AFG		72.4	67.1	69.7	66.4	62.3	64.2	76.8	70.4	73.4	68.7	61.4	64.8	71.1	65.3	68.0

Table 10. Accuracy of ResNet-50 model fine-tuned on DomainNet dataset under 3-shot setting