TAROT: Towards Essentially Domain-Invariant Robustness with Theoretical Justification

Supplementary Material

7. Novelty Summarization

Our work goes beyond a simple theoretical extension of the existing MDD in three key aspects. **First**, we expand upon MDD by introducing a newly defined robust divergence to derive an upper bound of the target domain robust risk that does not use adversarial samples in the source domain. Instead, the robust divergence measures the distance between distributions of clean examples in the source domain and adversarial examples in the target domain and thus we can save computation times for generating adversarial samples in the source domain. Moreover, an interesting property of TAROT is that the trained model is robust not only on the target domain but also on the source domain (see Proposition 3 in Section 7.3 in Appendix for theoretical evidence and Table 3 in Section 5.1.2 for empirical evidence) even if no adversarial samples in the source domain are used in the training phase. Note that the replacing $\mathcal{R}_{\mathcal{S}}$ with $\mathcal{R}_{\mathcal{S}}^{\text{rob}}$ also becomes an upper bound, but it is a looser bound than ours since $\mathcal{R}_{\mathcal{S}}(f) \leq \mathcal{R}_{\mathcal{S}}^{\text{rob}}(f)$. **Second**, the local Lipschitz constant, a newly introduced term, provides a theoretical bridge to existing research on adversarial robustness [42, 43]. **Finally**, by offering a partial theoretical explanation of the existing robust UDA algorithms, we integrate the previous works and pave a new direction for robust UDA, highlighting novel approaches and potential advancements in the field. Hence we believe that our bound is cleverly devised for robust UDA beyond a simple extenstion of MDD.

8. Theoretical Results

8.1. Auxiliary Lemmas

Lemma 1 (Lemma C.4 from Zhang et al. [45], Theorem 8.1 from Mehryar Mohri and Talwalkar [23])**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set of score functions where $\mathcal{Y} = \{1, \dots, C\}$. Define*

$$\Pi_1\mathcal{F} = \{\boldsymbol{x} \mapsto f(\boldsymbol{x}, y) | y \in \mathcal{Y}, f \in \mathcal{F}\}$$

and fix the margin parameter $\rho > 0$. Then for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{F}$.

$$|\mathcal{R}_{\mathcal{D}}^{(\rho)}(f) - \mathcal{R}_{\widehat{\mathcal{D}}}^{(\rho)}(f)| \leq \frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_1\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Lemma 2 (Talagrand's lemma [23, 33])**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an l-Lipschitz function. Then for any hypothesis set $\mathcal{H}$ of real-valued functions and any samples $\widehat{\mathcal{D}}$ of size n, the following inequality holds:*

$$\widehat{\mathfrak{R}}_{\widehat{\mathcal{D}}}(\Phi \circ H) \leq l\widehat{\mathfrak{R}}_{\widehat{\mathcal{D}}}(H)$$

Lemma 3 (Lemma 8.1 from Mehryar Mohri and Talwalkar [23])**.** *Consider $k > 1$ hypothesis sets $\mathcal{F}_1, \dots, \mathcal{F}_k$ in $\mathbb{R}^{\mathcal{X}}$. Let $\mathcal{G} = \{\max\{h_1, \dots, h_l\} : h_i \in \mathcal{F}_j, j \in [1, l]\}$. Then for any sample $\widehat{\mathcal{D}}$ size of n, the following holds:*

$$\widehat{\mathfrak{R}}_{\widehat{\mathcal{D}}}(\mathcal{G}) \leq \sum_{j=1}^{l}\widehat{\mathfrak{R}}_{\widehat{\mathcal{D}}}(\mathcal{F}_j) \tag{21}$$

8.2. Proofs

Lemma 4. *For any distribution $\mathcal{D}$ which $(\mathbf{X}, Y)$ follows, and score functions f, f', the following inequality holds.*

$$\text{disp}_{\mathcal{D}_{\mathbf{X}}}^{(\rho)}(f', f) \leq \mathcal{R}_{\mathcal{D}}^{(\rho)}(f') + \mathcal{R}_{\mathcal{D}}^{(\rho)}(f) \tag{22}$$

Proof. We first show that the following holds.

$$\Phi_\rho \circ \mathcal{M}_f(\mathbf{X}, h_f(\mathbf{X})) \leq \Phi_\rho \circ \rho_{f'}(\mathbf{X}', Y) + \Phi_\rho \circ \mathcal{M}_f(\mathbf{X}, Y)$$

If $h_{f'}(\mathbf{X}') \neq Y$ or $h_f(\mathbf{X}) \neq Y$ holds, then the right-hand side is bigger than 1, consequently the inequality holds. Now consider the case $h_{f'}(\mathbf{X}') = h_f(\mathbf{X}) = Y$. Since $\Phi_\rho \circ \mathcal{M}_f(\mathbf{X}', h_f(\mathbf{X})) = \Phi_\rho \circ \mathcal{M}_f(\mathbf{X}', Y)$ holds, the wanted inequality holds.

Therefore, following inequality holds.

$$\begin{aligned}
&\text{disp}_{\mathcal{D}_{\mathbf{X}}}^{(\rho)}(f', f) \\
&= \mathbb{E}_{\mathbf{X}\sim\mathcal{D}}\Phi_\rho \circ \mathcal{M}_f(\mathbf{X}, h_f(\mathbf{X})) \\
&\leq \mathbb{E}_{\mathbf{X}\sim\mathcal{D}}\Phi_\rho \circ \rho_{f'}(\mathbf{X}, Y) + \mathbb{E}_{\mathbf{X}\sim\mathcal{D}}\Phi_\rho \circ \mathcal{M}_f(\mathbf{X}, Y) \\
&= \mathcal{R}_{\mathcal{D}}^{(\rho)}(f') + \mathcal{R}_{\mathcal{D}}^{(\rho)}(f)
\end{aligned}$$

□

Proposition 1. *Let $\mathcal{S}$ and $\mathcal{T}$ represent the distributions of the source and target domains, respectively. Similarly, let $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ denote the marginal distributions of the source and target domains over $\mathbf{X}$, respectively. For every score function $f \in \mathcal{F}$, the following inequality holds:*

$$\begin{aligned}
&\mathcal{R}_{\mathcal{T}}^{rob}(f) \leq \\
&\mathcal{R}_{\mathcal{S}}^{(\rho)}(f) + \left\{\text{disp}_{\mathcal{T}_{\mathbf{X}}}^{rob,(\rho)}(f^*, f) - \text{disp}_{\mathcal{S}_{\mathbf{X}}}^{(\rho)}(f^*, f)\right\} + \lambda,
\end{aligned} \tag{7}$$

where $f^* = \underset{f\in\mathcal{F}}{\text{argmin}}\{\mathcal{R}_{\mathcal{T}}^{(\rho)}(f) + \mathcal{R}_{\mathcal{S}}^{(\rho)}(f)\}$ is ideal hypothesis and $\lambda = \lambda(\mathcal{F}, \mathcal{S}, \mathcal{T}, \varepsilon, \rho) = \mathcal{R}_{\mathcal{T}}^{(\rho)}(f^*) + \mathcal{R}_{\mathcal{S}}^{(\rho)}(f^*)$ is constant of f.

Proof. Since

$$\max_{\mathbf{X}'\in\mathcal{B}_p(\mathbf{X},\varepsilon)} \mathbb{1}\{Y \neq h_f(\mathbf{X}')\}$$

$$\leq \max_{\mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon)} \mathbb{1}\{h_{f^*}(\mathbf{X}) \neq h_f(\mathbf{X}')\} + \mathbb{1}\{h_{f^*}(\mathbf{X}) \neq Y\}$$

holds, we can start to derive the following inequalities.

$$\begin{aligned}
&\mathcal{R}^{\text{rob}}_{\mathcal{T}}(f) \\
&\leq \text{disp}^{\text{rob}}_{\mathcal{T}_{\mathbf{X}}}(f^*, f) + \mathcal{R}_{\mathcal{T}}(f^*) \\
&\leq \text{disp}^{\text{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f^*, f) + \mathcal{R}^{(\rho)}_{\mathcal{T}}(f^*) \\
&= \text{disp}^{\text{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f^*, f) + \mathcal{R}^{(\rho)}_{\mathcal{T}}(f^*) + \mathcal{R}^{(\rho)}_{\mathcal{S}}(f) - \mathcal{R}^{(\rho)}_{\mathcal{S}}(f) \\
&\leq \text{disp}^{\text{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f^*, f) + \mathcal{R}^{(\rho)}_{\mathcal{T}}(f^*) + \mathcal{R}^{(\rho)}_{\mathcal{S}}(f) \\
&\quad + \mathcal{R}^{(\rho)}_{\mathcal{S}}(f^*) - \text{disp}^{(\rho)}_{\mathcal{S}_{\mathbf{X}}}(f^*, f) \\
&= \mathcal{R}^{(\rho)}_{\mathcal{S}}(f) + \text{disp}^{\text{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f^*, f) - \text{disp}^{(\rho)}_{\mathcal{S}_{\mathbf{X}}}(f^*, f) + \lambda
\end{aligned}$$

Here, the third inequality holds from Lemma 4 and $\lambda = \mathcal{R}^{(\rho)}_{\mathcal{T}}(f^*) + \mathcal{R}^{(\rho)}_{\mathcal{S}}(f^*)$. □

Lemma 5 (Part of Theorem C.7 from Zhang et al. [45]). *For a given distribution $\mathcal{D}$, corresponding empirical distribution $\widehat{\mathcal{D}}$, and any $\delta > 0$, with probability at least $1 - \delta$, the following holds for $\forall f, f' \in \mathcal{F}$ simultaneously.*

$$\begin{aligned}
&\left|\text{disp}^{(\rho)}_{\mathcal{D}_{\mathbf{X}}}(f', f) - \text{disp}^{(\rho)}_{\widehat{\mathcal{D}}_{\mathbf{X}}}(f', f)\right| \\
&\leq \frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}
\end{aligned} \tag{23}$$

Lemma 6. *For a given distribution $\mathcal{D}$, its marginal distribution $\mathcal{D}_{\mathbf{X}}$, corresponding empirical distribution $\widehat{\mathcal{D}}_{\mathbf{X}}$, and any $\delta > 0$, with probability at least $1 - \delta$, the following holds for $\forall f, f' \in \mathcal{F}$ simultaneously.*

$$\begin{aligned}
&\left|\text{disp}^{rob,(\rho)}_{\mathcal{D}_{\mathbf{X}}}(f', f) - \text{disp}^{rob,(\rho)}_{\widehat{\mathcal{D}}_{\mathbf{X}}}(f', f)\right| \\
&\leq \frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho}
\end{aligned} \tag{24}$$

Proof. Denote $f(\boldsymbol{x}) = (f(\boldsymbol{x}, 1), \ldots, f(\boldsymbol{x}, C))^T \in \mathbb{R}^C$ for arbitrary $\boldsymbol{x} \in \mathcal{X}$. Note that Φ_ρ is $1/\rho$-Lipschitz, and the margin operator is 2-Lipschitz [5]. Hence, for $\forall \boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)$, the following inequality holds.

$$\begin{aligned}
&|\Phi_\rho \circ \mathcal{M}_f(\boldsymbol{x}', y) - \Phi_\rho \circ \mathcal{M}_f(\boldsymbol{x}, y)| \\
&\leq \frac{1}{\rho}\,|\mathcal{M}_f(\boldsymbol{x}', y) - \mathcal{M}_f(\boldsymbol{x}', y)| \\
&\leq \frac{2}{\rho}\,\|f(\boldsymbol{x}') - f(\boldsymbol{x})\|_1 \\
&\leq \frac{2}{\rho}\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)
\end{aligned}$$

Here we utilize the proof technique from Zhang et al. [45]. For $\forall f, f' \in \mathcal{F}$, define the $\tau_{f'}$-transform of f as follows:

$$\tau_{f'} f(\boldsymbol{x}, y) = \begin{cases} f(\boldsymbol{x}, 1) & \text{if } y = h_{f'}(\boldsymbol{x}) \\ f(\boldsymbol{x}, h_{f'}(\boldsymbol{x})) & \text{if } y = 1 \\ f(\boldsymbol{x}, y) & \text{o.w.} \end{cases}$$

where h_f is the induced classifier from f. Let $\mathcal{G} = \{\tau_{f'} f | f, f' \in \mathcal{F}\}$ and $\tilde{\mathcal{G}} = \{(x, y) \mapsto \rho_g(x, y) | g \in \mathcal{G}\}$. Now using these sets, we can represent the disparity terms into risk terms. For any $f, f' \in \mathcal{F}$, let $g = \tau_{f'} f$.

Then,

$$\begin{aligned}
&\rho_g(\boldsymbol{x}, 1) \\
&= \rho_{\tau_{f'} f}(\boldsymbol{x}, 1) \\
&= \tau_{f'} f(\boldsymbol{x}, 1) - \max_{y' \neq 1} \tau_{f'} f(\boldsymbol{x}, y') \\
&= f(\boldsymbol{x}, h_{f'}(\boldsymbol{x})) - \max\left\{\max_{y' \neq 1, h_{f'}(\boldsymbol{x})} f(\boldsymbol{x}, y'), f(\boldsymbol{x}, 1)\right\} \\
&= f(\boldsymbol{x}, h_{f'}(\boldsymbol{x})) - \max_{y' \neq h_{f'}(\boldsymbol{x})} f(\boldsymbol{x}, y) \\
&= \mathcal{M}_f(\boldsymbol{x}, h_{f'}(\boldsymbol{x}))
\end{aligned}$$

holds.

Hence,

$$\begin{aligned}
&\text{disp}^{\text{rob},(\rho)}_{\mathcal{D}_{\mathbf{X}}}(f', f) \\
&= \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_{\mathbf{X}}} \max_{\mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon)} \Phi_\rho \circ \mathcal{M}_f(\mathbf{X}', h_{f'}(\mathbf{X})) \\
&\leq \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_{\mathbf{X}}} \Phi_\rho \circ \mathcal{M}_f(\mathbf{X}, h_{f'}(\mathbf{X})) + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho} \\
&= \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_{\mathbf{X}}} \Phi_\rho \circ \rho_g(\mathbf{X}, 1) + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho}
\end{aligned}$$

holds for $g = \tau_{f'} f$.

For arbitrary set of score functions $\mathcal{U}$, we define following term:

$$\mathfrak{R}^0_{n,\mathcal{D}}(\mathcal{U}) := \mathbb{E}_{(\boldsymbol{x}_i, 1), \boldsymbol{x}_i \sim \mathcal{D}^n} \hat{\mathfrak{R}}_{\widehat{\mathcal{D}}}(\mathcal{U})$$

Regard all the data as from the same class 1. Then by using Lemma 1, the following inequality holds simultaneously for any $g \in \mathcal{G}$, with probability at least $1 - \delta$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X} \sim \mathcal{D}_{\mathbf{X}}} \Phi_\rho \circ \rho_g(\mathbf{X}, 1) \\
&\leq \mathbb{E}_{\mathbf{X} \sim \widehat{\mathcal{D}}_{\mathbf{X}}} \Phi_\rho \circ \rho_g(\mathbf{X}, 1) + 2\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}
\end{aligned}$$

Hence,

$$\begin{aligned}
&\text{disp}^{\text{rob},(\rho)}_{\mathcal{D}_{\mathbf{X}}}(f', f) \\
&\leq \mathbb{E}_{\mathbf{X} \sim \widehat{\mathcal{D}}} \Phi_\rho \circ \rho_g(\mathbf{X}, 1) + 2\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}}) \\
&\quad + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho} \\
&= \mathbb{E}_{\mathbf{X} \sim \widehat{\mathcal{D}}} \Phi_\rho \circ \rho_{f'}(\mathbf{X}, h_f(\mathbf{X})) + 2\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}})] \\
&\quad + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho} \\
&\leq \mathbb{E}_{\mathbf{X} \sim \widehat{\mathcal{D}}} \max_{\mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon)} \Phi_\rho \circ \rho_{f'}(\mathbf{X}', h_f(\mathbf{X})) + 2\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}})
\end{aligned}$$

$$+ \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho}$$
$$= \mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{D}}_{\mathbf{X}}}(f', f) + 2\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}})$$
$$+ \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho}$$

holds. Now, we want to bound the term $\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}})$.

By Lemma 2,

$$\mathfrak{R}^0_{n,\mathcal{D}}(\Phi \circ \tilde{\mathcal{G}}) \le \frac{1}{\rho}\mathfrak{R}^0_{n,\mathcal{D}}(\tilde{\mathcal{G}})$$

holds. Also,

$$\begin{aligned}
&\mathfrak{R}^0_{n,\mathcal{D}}(\tilde{\mathcal{G}}) \\
&= \frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i \rho_g(\boldsymbol{x}_i, 1) \\
&\le \frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \mathcal{F}, h \in \mathcal{H}} \sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i, h(\boldsymbol{x}_i)) \\
&\quad + \frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \mathcal{F}, h \in \mathcal{H}} \sum_{i=1}^n \sigma_i \left(- \max_{y \neq h(\boldsymbol{x}_i)} f(\boldsymbol{x}_i, y)\right) \\
&= \mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F}) + \frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \mathcal{F}, h \in \mathcal{H}} \sum_{i=1}^n \sigma_i \max_{y \neq h(\boldsymbol{x}_i)} f(\boldsymbol{x}_i, y)
\end{aligned}$$

holds.

Define the permutation

$$\xi(i) = \begin{cases} i+1 & i = 1, \dots, C-1 \\ 1 & i = C \end{cases}$$

As we assumed that $\mathcal{H}$ is permutation-invariant, we know that for $\forall h \in \mathcal{H}$ and $j = 1, \dots, k-1$, $\xi^j h \in \mathcal{H}$ holds. Let $\Pi_{\mathcal{H}}\mathcal{F}^{(C-1)} = \{\max\{f_1, \dots, f_{C-1}\} | f_i \in \Pi_{\mathcal{H}}\mathcal{F}, i = 1, \dots, C-1\}$.

Then,

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \mathcal{F}, h \in \mathcal{H}} \sum_{i=1}^n \sigma_i \max_{y \neq h(\boldsymbol{x}_i)} f(\boldsymbol{x}_i, y) \\
&= \frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \mathcal{F}, h \in \mathcal{H}} \sum_{i=1}^n \sigma_i \max_{j \in \{1,\dots,k-1\}} f(\boldsymbol{x}_i, \xi^j h(\boldsymbol{x}_i)) \\
&= \frac{1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \Pi_{\mathcal{H}}\mathcal{F}^{(C-1)}} \sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i) \\
&\le \frac{C-1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \Pi_{\mathcal{H}}\mathcal{F}} \sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i)
\end{aligned}$$

holds, where the last inequality holds from Lemma 3. Hence,

$$\begin{aligned}
&\mathfrak{R}^0_{n,\mathcal{D}}(\tilde{\mathcal{G}}) \\
&\le \mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F}) + \frac{C-1}{n}\mathbb{E}_{\widehat{\mathcal{D}},\sigma} \sup_{f \in \Pi_{\mathcal{H}}\mathcal{F}} \sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i) \\
&\le C\mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F})
\end{aligned}$$

holds.

Combining above inequalities, we have the following inequality

$$\begin{aligned}
&\mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{D}_{\mathbf{X}}}(f', f) \\
&\le \mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{D}}_{\mathbf{X}}}(f', f) + \frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F}) \\
&\quad + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho}
\end{aligned}$$

holds simultaneously for $\forall f, f' \in \mathcal{F}$ with probability at least $1 - \delta$.

In the same way, we have the opposite direction by exchanging $\mathcal{D}$ and $\widehat{\mathcal{D}}$. Therefore, the following holds simultaneously for $\forall f, f' \in \mathcal{F}$ with probability at least $1 - \delta$,

$$\begin{aligned}
&\left| \mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{D}_{\mathbf{X}}}(f', f) - \mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{D}}_{\mathbf{X}}}(f', f) \right| \\
&\le \frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho}
\end{aligned}$$

concluding the proof. □

Lemma 7. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$.*

$$\begin{aligned}
&|\mathcal{R}^{rob,(\rho)}_{\mathcal{D}}(f) - \mathcal{R}^{rob,(\rho)}_{\widehat{\mathcal{D}}}(f)| \\
&\le \frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_1\mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{X}, \varepsilon)}{\rho}
\end{aligned} \quad (25)$$

Proof. We know that

$$|\Phi_\rho \circ \mathcal{M}_f(\boldsymbol{x}', y) - \Phi_\rho \circ \mathcal{M}_f(\boldsymbol{x}, y)| \le \frac{2}{\rho}\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)$$

holds for $\forall \boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)$.

Then, the following holds with probability at least $1 - \delta$.

$$\begin{aligned}
&\mathcal{R}^{\mathrm{rob},(\rho)}_{\mathcal{D}}(f) \\
&= \mathbb{E}_{\mathcal{D}} \max_{\mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon)} \Phi_\rho \circ \mathcal{M}_f(\mathbf{X}', \mathbf{Y}) \\
&\le \mathbb{E}_{\mathcal{D}}\Phi_\rho \circ \mathcal{M}_f(\mathbf{X}, \mathbf{Y}) + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho} \\
&\le \mathcal{R}^{(\rho)}_{\widehat{\mathcal{D}}}(f) + \frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_1\mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}}, \varepsilon)}{\rho} \\
&= \frac{1}{n}\sum_{i=1}^n \Phi_\rho \circ \mathcal{M}_f(\boldsymbol{x}_i, y_i) + \frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_1\mathcal{F})
\end{aligned}$$

$$+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}},\varepsilon)}{\rho}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\max_{\boldsymbol{x}'_i\in\mathcal{B}_p(\boldsymbol{x}_i,\varepsilon)}\Phi_\rho\circ\mathcal{M}_f(\boldsymbol{x}'_i,y_i)+\frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_1\mathcal{F})$$

$$+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}},\varepsilon)}{\rho}$$

$$=\mathcal{R}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{D}}}(f)+\frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{D}}(\Pi_1\mathcal{F})+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{2\varepsilon L_f(\mathcal{D}_{\mathbf{X}},\varepsilon)}{\rho}$$

□

Lemma 8. *For any $\delta>0$, with probability $1-2\delta$, the following holds simultaneously for any score function f,*

$$\begin{aligned}&\left|d^{rob,(\rho)}_{f,\mathcal{F}}(\widehat{\mathcal{S}}_{\mathbf{X}},\widehat{\mathcal{T}}_{\mathbf{X}})-d^{rob,(\rho)}_{f,\mathcal{F}}(\mathcal{S}_{\mathbf{X}},\mathcal{T}_{\mathbf{X}})\right|\\&\leq\frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{S}}\left(\Pi_{\mathcal{H}}\mathcal{F}\right)+\frac{2k}{\rho}\mathfrak{R}_{m,\mathcal{T}}\left(\Pi_{\mathcal{H}}\mathcal{F}\right)\\&\quad+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}+\frac{2\varepsilon L_f(\mathcal{T}_{\mathbf{X}},\varepsilon)}{\rho}\end{aligned}\tag{26}$$

Proof. From Lemma 6, we have

$$\begin{aligned}&\left|\mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f',f)-\mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{T}}_{\mathbf{X}}}(f',f)\right|\\&\leq\frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{T}}(\Pi_{\mathcal{H}}\mathcal{F})+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{2\varepsilon L_f(\mathcal{T}_{\mathbf{X}},\varepsilon)}{\rho}\end{aligned}$$

Also, from Lemma 5, the following holds with probability at least $1-\delta$,

$$\begin{aligned}&\left|\mathrm{disp}^{(\rho)}_{\mathcal{S}_{\mathbf{X}}}(f',f)-\mathrm{disp}^{(\rho)}_{\widehat{\mathcal{S}}_{\mathbf{X}}}(f',f)\right|\\&\leq\frac{2C}{\rho}\mathfrak{R}_{m,\mathcal{S}}(\Pi_{\mathcal{H}}\mathcal{F})+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\end{aligned}$$

Hence,

$$\begin{aligned}&\left|d^{\mathrm{rob},(\rho)}_{f,\mathcal{F}}(\mathcal{S}_{\mathbf{X}},\mathcal{T}_{\mathbf{X}})-d^{\mathrm{rob},(\rho)}_{f,\mathcal{F}}(\widehat{\mathcal{S}}_{\mathbf{X}},\widehat{\mathcal{T}}_{\mathbf{X}})\right|\\&=\left|\sup_{f'\in\mathcal{F}}\left\{\mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f',f)-\mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{S}_{\mathbf{X}}}(f',f)\right\}\right.\\&\quad\left.-\sup_{f'\in\mathcal{F}}\left\{\mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{T}}_{\mathbf{X}}}(f',f)-\mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{S}}_{\mathbf{X}}}(f',f)\right\}\right|\\&\leq\sup_{f'\in\mathcal{F}}\left|\mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{S}_{\mathbf{X}}}(f',f)-\mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{S}}_{\mathbf{X}}}(f',f)\right|\\&\quad+\sup_{f'\in\mathcal{F}}\left|\mathrm{disp}^{\mathrm{rob},(\rho)}_{\mathcal{T}_{\mathbf{X}}}(f',f)-\mathrm{disp}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{T}}_{\mathbf{X}}}(f',f)\right|\end{aligned}$$

holds, concluding the proof. □

Theorem 1. ***(Generalization Bound on the Robust Risk of Target Distribution)**. For any $\delta>0$, with probability $1-3\delta$, we have the following uniform generalization bound for any score function f in $\mathcal{F}$:*

$$\begin{aligned}&\mathcal{R}^{rob}_{\mathcal{T}}(f)\\&\leq\mathcal{R}^{(\rho)}_{\widehat{\mathcal{S}}}(f)+d^{rob,(\rho)}_{f,\mathcal{F}}(\widehat{\mathcal{S}}_{\mathbf{X}},\widehat{\mathcal{T}}_{\mathbf{X}})+\lambda\\&+\frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{S}}(\Pi_1\mathcal{F})+\frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{S}}(\Pi_{\mathcal{H}}\mathcal{F})+2\sqrt{\frac{\log 2/\delta}{2n}}\\&+\frac{2C}{\rho}\mathfrak{R}_{m,\mathcal{T}}(\Pi_{\mathcal{H}}\mathcal{F})+\sqrt{\frac{\log 2/\delta}{2m}}+\frac{2\varepsilon L_f(\mathcal{T}_{\mathbf{X}},\varepsilon)}{\rho},\end{aligned}\tag{16}$$

where $\lambda=\min_{f\in\mathcal{F}}\{\mathcal{R}^{(\rho)}_{\mathcal{T}}(f)+\mathcal{R}^{(\rho)}_{\mathcal{S}}(f)\}$.

Proof. From Eq. (9),

$$\begin{aligned}&\mathcal{R}^{\mathrm{rob}}_{\mathcal{T}}(h_f)\\&\leq\mathcal{R}^{(\rho)}_{\mathcal{S}}(f)+\mathrm{d}^{\mathrm{rob},(\rho)}_{f,\mathcal{F}}(\mathcal{S}_{\mathbf{X}},\mathcal{T}_{\mathbf{X}})+\lambda\\&\leq\mathcal{R}^{(\rho)}_{\widehat{\mathcal{S}}}(f)+\frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{S}}(\Pi_1\mathcal{F})+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\\&\quad+\mathrm{d}^{\mathrm{rob},(\rho)}_{f,\mathcal{F}}(\mathcal{S}_{\mathbf{X}},\mathcal{T}_{\mathbf{X}})+\lambda\\&\leq\mathcal{R}^{(\rho)}_{\widehat{\mathcal{S}}}(f)+\frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{S}}(\Pi_1\mathcal{F})+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\\&\quad+\mathrm{d}^{\mathrm{rob},(\rho)}_{f,\mathcal{F}}(\widehat{\mathcal{S}}_{\mathbf{X}},\widehat{\mathcal{T}}_{\mathbf{X}})+\frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{S}}\left(\Pi_{\mathcal{H}}\mathcal{F}\right)\\&\quad+\frac{2C}{\rho}\mathfrak{R}_{m,\mathcal{T}}\left(\Pi_{\mathcal{H}}\mathcal{F}\right)+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\\&\quad+\frac{2\varepsilon L_f(\mathcal{T}_{\mathbf{X}},\varepsilon)}{\rho}+\lambda\\&=\mathcal{R}^{\mathrm{rob},(\rho)}_{\widehat{\mathcal{S}}}(f)+\mathrm{d}^{\mathrm{rob},(\rho)}_{f,\mathcal{F}}(\widehat{\mathcal{S}},\widehat{\mathcal{T}})+\lambda\\&\quad+\frac{2C^2}{\rho}\mathfrak{R}_{n,\mathcal{S}}(\Pi_1\mathcal{F})+\frac{2C}{\rho}\mathfrak{R}_{n,\mathcal{S}}\left(\Pi_{\mathcal{H}}\mathcal{F}\right)\\&\quad+\frac{2C}{\rho}\mathfrak{R}_{m,\mathcal{T}}\left(\Pi_{\mathcal{H}}\mathcal{F}\right)+2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\\&\quad+\frac{2\varepsilon L_f(\mathcal{T}_{\mathbf{X}},\varepsilon)}{\rho}\end{aligned}$$

Here, the second inequality holds from Lemma 1 and the third inequality holds from Lemma 8. □

8.3. Source Robusk Risk of TAROT

In this section, we derive an upper bound for the robust risk on the source domain. The components of following upper bound — standard source risk and robust disparity — correspond to the upper bound in Proposition 1, suggesting that our algorithm can effectively improve adversarial robustness on the source domain.

Proposition 3. *Consider a source domain $\mathcal{S}$, a target domain $\mathcal{T}$ and their marginal distributions $\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}$ on $\mathbf{X}$. For every score function $f \in \mathcal{F}$, the following inequality holds:*

$$\mathcal{R}_{\mathcal{S}}^{rob}(f) \le \mathcal{R}_{\mathcal{S}}^{(\rho)}(f)+2d_{f,\mathcal{F}}^{rob,(\rho)}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})+\frac{2\varepsilon L_f(\mathcal{S}_{\mathbf{X}}, \varepsilon)}{\rho}+\lambda \tag{27}$$

where $\lambda = \min_{f\in\mathcal{F}}\{\mathcal{R}_{\mathcal{T}}^{(\rho)}(f) + \mathcal{R}_{\mathcal{S}}^{(\rho)}(f)\}$.

Proof.

$$\begin{aligned}
&\mathcal{R}_{\mathcal{S}}^{\text{rob}}(f) \\
&\le \mathcal{R}_{\mathcal{T}}^{(\rho)}(f) + \text{disp}_{\mathcal{S}_{\mathbf{X}}}^{\text{rob},(\rho)}(f^*, f) - \text{disp}_{\mathcal{T}_{\mathbf{X}}}^{(\rho)}(f^*, f) + \lambda \\
&\le \mathcal{R}_{\mathcal{T}}^{(\rho)}(f) + \text{disp}_{\mathcal{S}_{\mathbf{X}}}^{(\rho)}(f^*, f) - \text{disp}_{\mathcal{T}_{\mathbf{X}}}^{(\rho)}(f^*, f) \\
&\quad + \frac{2\varepsilon L_f(\mathcal{S}_{\mathbf{X}}, \varepsilon)}{\rho} + \lambda \\
&\le \mathcal{R}_{\mathcal{S}}^{(\rho)}(f) + 2\,\text{disp}_{\mathcal{S}_{\mathbf{X}}}^{(\rho)}(f^*, f) - \text{disp}_{\mathcal{T}_{\mathbf{X}}}^{(\rho)}(f^*, f) \\
&\quad + \frac{2\varepsilon L_f(\mathcal{S}_{\mathbf{X}}, \varepsilon)}{\rho} + \lambda \\
&\le \mathcal{R}_{\mathcal{S}}^{(\rho)}(f) + 2d_{f,\mathcal{F}}^{\text{rob},(\rho)}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \frac{2\varepsilon L_f(\mathcal{S}_{\mathbf{X}}, \varepsilon)}{\rho} + \lambda
\end{aligned}$$

Here, the first inequality holds by replacing $\mathcal{S}$ and $\mathcal{T}$ from Proposition 1. □

Note that since the upper-bound considers the local Lipschitz constant on the source domain, this is a partial explanation for the source robust risk.

9. Further Details on Experiments

Loss The exact forms of Eq. (20) loss function are as follows:

$$\ell_{\text{ce}}\left((\pi \circ \psi)(\boldsymbol{x}), y)\right) := -\log \sigma_y(\pi \circ \psi(\boldsymbol{x})),$$

$$\ell_{\text{mod-ce}}^{\text{rob}}\left((\pi \circ \psi)(\boldsymbol{x}), y)\right) := \log(1 - \sigma_y(\pi \circ \psi(\boldsymbol{x}^{\text{adv}}))),$$

$$\ell_{\text{ce}}^{\text{rob}}\left((\pi \circ \psi)(\boldsymbol{x}), y)\right) := -\log \sigma_y(\pi \circ \psi(\boldsymbol{x}^{\text{adv}})),$$

where σ_y denotes the predictive confidence for class y, i.e., the y-th component of the softmax output and $\boldsymbol{x}^{\text{adv}}$ is the adversarial example.

Datasets Office-31 consists of 4,110 images from three domains — Amazon (A), Webcam (W), and DSLR (D) — considered to be classical data for domain adaptation due to the differences in image quality and capture methods. Office-Home is more diverse, with 15,588 images across four domains — Art (Ar), Clipart (Cl), Product (Pr), and Realworld (Rw) — covering different styles, from artistic drawings to real photos. VisDA2017 features over 280,000 images, focusing on the domain gap between synthetic and real images, providing a challenge for algorithms to handle synthetic (S) to real (R) adaptation. DomainNet is the largest and challenging dataset, containing around 600,000 images from six domains, including Clipart (C), Infograph (I), Sketch (S), Painting (P), Quickdraw (Q) and Real (R).

Hyperparameters We follow the default experimental settings of *TLlib* [18]. We conduct experiments using the following training configuration. Models are trained for 20 epochs with a weight decay of 5×10^{-4}. Robust pretraining is set at $\varepsilon = \frac{1}{255}$ for TAROT, PL, ARTUDA, and SRoUDA, while RFA uses models trained with different ε values identical to evaluation ε, as it does not directly generate adversarial examples during training. We conduct experiments using the following training configuration. Models are trained for 20 epochs with a weight decay of 5×10^{-4}. Robust pretraining is set at $\varepsilon = \frac{1}{255}$ for TAROT, PL, ARTUDA, and SRoUDA. In contrast, RFA utilizes models trained with different ε values matching the evaluation ε, as it does not directly generate adversarial examples during training. For TAROT, PL, ARTUDA, and SRoUDA, the step size during training is defined as $\frac{\varepsilon}{4\times255}$, with 10 steps per iteration. For model selection, we evaluate using PGD20 with ε and the same step size of $\frac{\varepsilon}{4\times255}$, using a batch size of 32. Optimization is performed using SGD with a momentum of 0.9, a weight decay of 5×10^{-4}, and an initial learning rate of 0.005. These settings ensure consistency and robustness across all algorithms under evaluation.

10. Additional Experimental Results

Here, we present experimental results that were not included in the manuscript. Additionally, we perform supplementary experiments to further support the effectiveness of our proposed method, TAROT.

10.1. Essentially Domain-Invariant Robustness

In Table 3, we present partial performance results of PL and TAROT on the source and unseen domains on Office-Home dataset, when $\varepsilon = 8/255$. Here, we present the unreported values in Table 6. In Table 6, we observe that TAROT consistently outperforms its competitors in terms of robust accuracy, as shown in Table 3. The only notable competitor in terms of standard accuracy is RFA. However, its robust accuracy is significantly lower than that of TAROT. Furthermore, TAROT outperforms other methods, across all metrics except ith only a few exceptions. In summary, TAROT demonstrates superior performance on both source and unseen domains compared to its competitors, owing to its ability to learn essentially domain-invariant robust features.

10.2. Effect of Robust-PT on Various ε

We present the previously unreported values from Figure 3 for the OfficeHome dataset. In Table 10, we provide the standard and robust accuracies of PL and TAROT across varying values of ε, both with and without Robust-PT. Notably, TAROT with Robust-PT consistently outperforms

Table 6. **Performances of PL and TAROT on Source Domain and Unseen Domain, on OfficeHome.** Standard accuracy (%) / Robust accuracy (%) for AA with $\varepsilon = 8/255$. Bold numbers indicate the best performance.

	Source	**Unseen**		
Method	Ar → Pr(Ar)	Ar → Pr(Cl)	Ar → Pr(Rw)	Avg.
ARTUDA	62.59 / 8.53	29.46 / 11.02	32.29 / 7.30	41.45 / 8.95
RFA	**99.63** / 37.33	40.21 / 18.05	**60.29** / 19.05	**66.71** / 24.81
SRoUDA	22.46 / 5.11	31.32 / 15.92	41.57 / 15.91	31.78 / 12.31
PL	24.68 / 10.88	35.51 / 24.72	43.84 / 25.25	34.68 / 20.28
TAROT	98.31 / **43.02**	**43.05** / **27.15**	56.14 / **27.77**	65.83 / **32.65**
	Pr → Ar(Pr)	Pr → Ar(Cl)	Pr → Ar(Rw)	Avg.
ARTUDA	66.16 / 23.00	20.87 / 7.45	24.08 / 5.90	37.04 / 12.12
RFA	**96.71** / 67.20	40.02 / 17.82	**59.79** / 19.37	**65.51** / 34.80
SRoUDA	59.21 / 50.89	**40.82** / 22.12	41.91 / 21.00	47.31 / 31.34
PL	45.21 / 27.01	33.47 / 22.09	44.60 / 23.07	41.09 / 24.05
TAROT	96.33 / **78.69**	40.12 / **25.68**	54.35 / **28.21**	63.60 / **44.19**
	Cl → Rw(Cl)	Cl → Rw(Ar)	Cl → Rw(Pr)	Avg.
ARTUDA	84.77 /55.81	14.34 / 3.21	27.26 / 13.99	42.12 / 24.34
RFA	**95.35** / 84.01	**40.38** / 8.82	54.86 / 22.19	**63.53** / 38.34
SRoUDA	48.29 / 35.58	33.87 / 15.62	48.19 / 33.09	43.45 / 28.10
PL	48.75 / 35.95	36.30 / 16.69	50.80 / 35.66	45.28 / 29.43
TAROT	93.65 / **84.35**	39.72 / **17.18**	**55.46** / **37.67**	62.95 / **46.40**
	Rw → Cl(Rw)	Rw → Cl(Ar)	Rw → Cl(Pr)	Avg.
ARTUDA	85.40 / 22.08	31.64 / 5.15	51.59 / 18.72	56.21 / 15.32
RFA	**99.59** / 38.54	**46.90** / 7.50	**64.79** / 23.36	**70.42** / 23.13
SRoUDA	38.72 / 15.84	21.18 / 5.85	37.08 / 18.86	32.33 / 13.51
PL	45.95 / 24.00	25.67 / 9.52	46.14 / 27.10	39.25 / 20.21
TAROT	97.68 / **51.55**	42.73 / **11.83**	63.39 / **34.67**	67.93 / **32.68**
	Ar → Cl(Ar)	Ar → Cl(Pr)	Ar → Cl(Rw)	Avg.
ARTUDA	78.78 / 11.00	29.80 / 7.37	34.08 / 8.40	47.56 / 8.92
RFA	**99.63** / 33.79	**49.83** / 17.59	**60.50** / 17.37	**69.99** / 22.92
SRoUDA	30.70 / 7.87	34.47 / 15.68	35.85 / 13.54	33.67 / 12.36
PL	32.51 / 11.83	38.30 / **24.24**	39.89 / 19.99	36.90 / 18.69
TAROT	99.59 / **40.38**	45.73 / 22.35	55.64 / **22.42**	66.98 / **28.38**
	Ar → Rw(Ar)	Ar → Rw(Cl)	Ar → Rw(Pr)	Avg.
ARTUDA	18.83 / 2.64	7.70 / 0.89	6.28 / 0.45	10.94 / 1.33
RFA	**99.63** / 46.89	42.52 / 21.47	51.52 / 21.47	64.56 / 29.94
SRoUDA	39.39 / 17.47	39.54 / 28.64	50.76 / 35.84	43.23 / 27.32
PL	41.78 / 19.41	41.97 / 31.39	52.42 / 37.24	45.39 / 29.34
TAROT	98.35 / **64.24**	**47.86** / **35.19**	**58.32** / **39.54**	**68.18** / **46.32**
	Cl → Ar(Cl)	Cl → Ar(Pr)	Cl → Ar(Rw)	Avg.
ARTUDA	72.60 / 18.67	13.90 / 1.10	8.40 / 0.64	31.63 / 6.81
RFA	**95.37** / **84.81**	**48.61** / 19.49	50.52 / 20.11	**65.84** / 42.47
SRoUDA	39.86 / 25.98	33.61 / 18.14	39.55 / 21.23	37.67 / 21.78
PL	42.45 / 28.94	37.33 / 22.39	42.92 / 24.81	40.90 / 25.38
TAROT	94.27 / 83.78	46.27 / **27.96**	**50.72** / **26.76**	63.76 / **46.17**
	Cl → Pr(Cl)	Cl → Pr(Ar)	Cl → Pr(Rw)	Avg.
ARTUDA	72.99 / 38.10	13.14 / 3.05	19.92 / 5.90	35.35 / 15.68
RFA	**98.35** / 84.70	**34.91** / 7.95	**51.58** / 17.08	**61.61** / 36.57
SRoUDA	35.72 / 24.35	18.09 / 6.88	37.16 / 19.92	30.32 / 17.05
PL	41.44 / 29.07	21.92 / 9.19	41.54 / 23.96	34.97 / 20.74
TAROT	95.65 / **86.09**	30.20 / **11.54**	51.41 / **27.34**	59.09 / **41.66**
	Pr → Cl(Pr)	Pr → Cl(Ar)	Pr → Cl(Rw)	Avg.
ARTUDA	97.30 / 47.94	23.28 / 3.79	45.12 / 11.59	55.23 / 21.11
RFA	**99.75** / 59.88	**34.91** / 5.93	**55.87** / 15.14	**63.51** / 26.99
SRoUDA	61.48 / 41.41	18.87 / 6.55	31.95 / 15.06	37.43 / 21.00
PL	52.20 / 33.03	22.42 / 9.31	38.74 / 19.74	37.78 / 20.69
TAROT	98.60 / **79.43**	30.70 / **9.48**	53.27 / **22.88**	60.86 / **37.26**
	Pr → Rw(Pr)	Pr → Rw(Ar)	Pr → Rw(Cl)	Avg.
ARTUDA	98.29 / 60.17	25.67 / 4.62	36.24 / 16.24	53.40 / 27.01
RFA	**99.75** / 68.48	37.67 / 7.83	41.79 / 19.70	59.74 / 32.01
SRoUDA	61.48 / 41.41	34.82 / 16.07	42.09 / 30.68	46.13 / 29.38
PL	61.05 / 43.34	36.09 / **17.10**	42.11 / 30.91	46.42 / 30.45
TAROT	96.13 / **82.43**	**39.14** / 16.15	**46.87** / **33.01**	**60.71** / **43.86**
	Rw → Ar(Rw)	Rw → Ar(Cl)	Rw → Ar(Pr)	Avg.
ARTUDA	90.66 / 19.72	41.19 / 15.46	53.17 / 16.99	61.67 / 17.39
RFA	**99.56** / 51.02	**47.24** / 22.25	**64.09** / 27.33	**70.30** / 33.53
SRoUDA	49.53 / 25.20	32.21 / 19.89	36.34 / 20.43	39.36 / 21.84
PL	51.32 / 27.50	35.40 / 24.26	40.39 / 23.56	42.37 / 25.11
TAROT	95.50 / **64.82**	46.35 / **32.21**	59.63 / **37.37**	67.16 / **44.80**
	Rw → Pr(Rw)	Rw → Pr(Ar)	Rw → Pr(Cl)	Avg.
ARTUDA	66.63 / 18.16	25.30 / 4.33	39.04 / 16.06	43.66 / 12.85
RFA	**99.59** / 41.52	**48.26** / 7.99	44.35 / 20.18	**64.07** / 23.23
SRoUDA	44.27 / 21.57	20.03 / 6.84	33.49 / 20.89	32.60 / 16.44
PL	48.80 / 26.60	23.65 / 9.60	36.98 / 24.86	36.47 / 20.35
TAROT	97.89 / **61.85**	42.23 / **14.30**	**47.65** / **31.50**	62.59 / **35.88**

other methods. Is it worth emphasizing that Robust-PT is crucial for enhancing the performance of both PL and TAROT. As discussed in Sec. 5.2, the performance gap between TAROT and PL widens as ε increases.

10.3. Sensitivity Analysis of α

We present the previously unreported values from Figure 2. In Table 7, the results for DomainNet are reported. We can observe that the target performance is highest when $\alpha = 1.0$. Additionally, $\alpha = 1.0$ yields the best performance on both source and unseen (average) domains.

In Table 8, the results for VisDA2017 results are reported. As shown in Figure 2, the standard and robust accuracies on the target domain exhibit minimal variation across different values of α. However, the performances on the source domain exhibit relatively large variations. We choose $\alpha = 0.1$, since it shows highest robust accuracy among the candidate values of α.

10.4. Evidence on Local Lipschitz Surrogate

In constructing the objective for TAROT, we employ adversarial training to reduce the local Lipschitz constant. Here, we empirically demonstrate that combining adversarial training with pseudo labeling effectively reduces the local Lipschitz constant. Following the approach of Yang et al. [42], we compute the empirical local Lipschitz constant using the following formula:

$$\frac{1}{n}\sum_{i=1}^{n}\max_{\boldsymbol{x}_i' \in \mathcal{B}_\infty(\boldsymbol{x}_i,\epsilon)} \frac{\|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i')\|_1}{\|\boldsymbol{x}_i - \boldsymbol{x}_i'\|_\infty} \tag{28}$$

Table 9 illustrates the training dynamics of the local Lipschitz constants, showing that adversarial training with pseudo labels effectively reduces these constants during training phase of PL. We evaluate the empirical local Lipschitz constant under various settings, considering four cases: with or without Robust-PT, and with or without adversarial training. We observe that when conducting an adversarial training, the empirical local Lipschitz constant significantly decreases across all tasks.

10.5. Performance with Lower Perturbation Budgets ε, on Office31 and OfficeHome.

We also conduct experiments with smaller values of ε than those used in the main experiment in Sec. 5.1.1. Specifically, we evaluate $\varepsilon \in \{8/255, 4/255\}$ on the Office31 and OfficeHome datasets, aligning with the experimental settings described in the original works [3, 20, 46]. Tables 13, 14, 15 and 16 demonstrate that TAROT also outperforms existing methods under small perturbation budgets. Compared to the other methods presented in Tables 1 and 2, which experience significant performance degradation at larger perturbation budgets ($\varepsilon = 16/255$), TAROT maintains its robustness even under these larger perturbation budgets.

10.6. Evaluation Against Other Attack Methods than AutoAttack

We additionally evaluate TAROT and other existing methods against other attack methods than AutoAttack. We evaluate each methods on OfficeHome with perturbation size of $\varepsilon = 16/255$, against FGSM, MM, CW20, PGD20 and AA. In Table 11, we can observe that TAROT outperforms existing methods in all means.

10.7. On the Use of the Standard Margin Risk of the Source Domain

If replacing $\mathcal{R}_\mathcal{S}(f)$ with $\mathcal{R}_\mathcal{S}^{\text{rob}}(f)$ burdens the computation cost, requiring to generate adversarial examples. Moreover, it would result in a looser bound in theoretical perspective ($\because \mathcal{R}_\mathcal{S}(f) \leq \mathcal{R}_\mathcal{S}^{\text{rob}}(f)$) , making it less desirable. To demonstrate the superiority of the proposed algorithm, we present empirical results obtained by replacing $\mathcal{R}_\mathcal{S}(f)$ with $\mathcal{R}_\mathcal{S}^{\text{rob}}(f)$. As seen in the table below, TAROT with $\mathcal{R}_\mathcal{S}(f)$ shows higher performance in both standard and robust accuracies than TAROT with $\mathcal{R}_\mathcal{S}^{\text{rob}}(f)$. Hence, the use of $\mathcal{R}_\mathcal{S}(f)$ rather than $\mathcal{R}_\mathcal{S}^{\text{rob}}(f)$ is justified both theoretically (a tighter bound) and empirically.

Table 7. **Sensitivity Analysis of α, on DomainNet.** Performance of generalization and robustness when α varies. In each cell, the first number is the standard accuracy (%), while the second number corresponds to the robust accuracy (%) for AA.

	Target	**Source**	**Unseen**			
α	C → R (R)	C → R (C)	C → R (I)	C → R (P)	C → R (S)	Avg.
0.0	43.57 / 28.68	48.41 / 35.89	10.70 / 5.66	24.44 / 11.46	22.84 / 13.55	29.99 / 19.05
0.05	46.24 / 30.66	52.79 / 39.19	26.28 / 12.19	26.28 / 12.19	24.80 / 14.68	32.24 / 20.50
0.1	46.83 / 31.03	55.87 / 41.88	26.65 / 12.49	26.65 / 12.49	26.08 / 15.59	33.42 / 21.41
0.5	49.39 / 31.46	67.71 / 51.57	**30.18 / 13.57**	**30.18 / 13.57**	34.07 / 19.81	38.93 / 24.59
1.0	**49.73 / 31.73**	**71.58 / 54.42**	14.36 / 6.60	31.45 / 13.53	**36.26 / 20.29**	**40.68 / 25.32**

Table 8. **Sensitivity Analysis of α, on VisDA2017.** Performance of generalization and robustness when α varies. In each cell, the first number is the standard accuracy (%), while the second number corresponds to the robust accuracy (%) for AA.

	Target	**Source**	
α	Syn. → Real	Syn.	Avg.
0.0	**67.48 / 38.71**	43.29 / 24.69	55.39 / 31.70
0.05	67.01 / 38.56	78.70 / 47.93	72.86 / 43.25
0.1	66.12 / 37.91	85.18 / **51.21**	75.65 / **44.56**
0.5	66.45 / 36.97	**86.63** / 46.30	**76.54** / 41.64
1.0	64.48 / 35.48	67.63 / 34.32	66.06 / 34.90

Table 9. **Empirical Local Lipschitz Constant in Various Training Settings. Lipschitz** denotes the empirical local Lipschitz constant value. Standard accuracy (%) and the robust accuracy (%) for PGD20 are also described.

Method	Adv. Train.	**Lipschitz**	Ar → Rw	**Lipschitz**	Cl → Rw	**Lipschitz**	Pr → Rw
PL	✗	6653.58	**78.40** / 1.31	6518.85	72.09 / 2.50	6992.54	78.84 / 1.26
PL	$\varepsilon = 8/255$	**1014.95**	77.78 / **70.53**	**981.91**	**72.80 / 64.66**	**1086.23**	**79.30 / 72.14**

Table 10. **Effect of Robust-PT with various ε, on OfficeHome.** In each cell, the first number is the standard accuracy (%), while the second number corresponds to the robust accuracy (%) for AA. Bold numbers indicate the best performance.

ε	Robust-PT	**Method**	Ar → Rw	Cl → Rw	Pr → Rw	Avg.
16/255	✓	PL	73.10 / 40.26	68.14 / 37.37	74.82 / 40.74	72.02 / 39.45
	✗	PL	6.59 / 0.00	3.121 / 0.00	5.92 / 0.161	5.21 / 0.05
	✓	TAROT	**77.78 / 42.62**	**71.31 / 39.22**	**78.72 / 43.13**	**75.94 / 41.66**
	✗	TAROT	22.47 / 0.74	21.71 / 0.90	18.98 / 0.73	21.05 / 0.79
12/255	✓	PL	78.15 / 55.68	71.70 / 50.06	78.29 / 54.17	76.05 / 53.30
	✗	PL	8.54 / 0.05	5.30 / 0.00	5.92 / 0.34	6.59 / 0.13
	✓	TAROT	**78.98 / 56.53**	**72.39 / 52.15**	**79.41 / 57.24**	**76.93 / 55.31**
	✗	TAROT	23.04 / 1.81	37.53 / 3.83	28.80 / 0.62	29.79 / 2.09
8/255	✓	PL	78.70 / 69.43	72.53 / 63.44	78.27 / 69.50	76.50 / 67.45
	✗	PL	10.83 / 1.68	12.65 / 2.32	9.00 / 1.17	10.83 / 1.72
	✓	TAROT	**78.77 / 70.46**	**73.01 / 63.78**	**79.44 / 70.83**	**77.07 / 68.36**
	✗	TAROT	69.70 / 24.54	65.32 / 30.25	64.86 / 21.92	66.63 / 25.57
6/255	✓	PL	79.16 / 73.93	71.52 / 65.87	79.02 / 73.97	76.57 / 71.26
	✗	PL	63.39 / 31.86	59.24 / 29.49	51.85 / 20.82	58.16 / 27.39
	✓	TAROT	**79.41 / 74.36**	**72.48 / 67.27**	**79.57 / 75.24**	**77.16 / 72.29**
	✗	TAROT	77.60 / 52.26	72.05 / 51.11	77.28 / 44.09	75.64 / 49.15
4/255	✓	PL	79.02 / 75.69	71.93 / 67.78	78.59 / 75.56	76.51 / **74.75**
	✗	PL	78.31 / 60.13	71.22 / 55.91	77.83 / 58.05	75.79 / 58.03
	✓	TAROT	78.86 / **76.43**	**73.03 / 69.59**	78.84 / **75.86**	76.91 / 73.96
	✗	TAROT	**79.41** / 69.20	72.53 / 62.57	**79.94** / 70.16	**77.29** / 67.31
2/255	✓	PL	78.08 / 76.11	72.37 / 69.64	79.39 / 77.02	76.61 / 74.25
	✗	PL	77.88 / 73.93	71.72 / 67.23	79.48 / 74.78	76.36 / 71.98
	✓	TAROT	78.36 / **76.70**	**73.42 / 71.24**	79.21 / **77.09**	**77.00 / 75.01**
	✗	TAROT	**78.56** / 72.24	72.37 / 68.65	**79.62** / 75.65	76.85 / 72.18

Table 11. **Performances of ARTUDA, RFA, SRoUDA, PL and TAROT on OfficeHome** ($\varepsilon = 16/255$)**, evaluated with FGSM, MM, CW20, PGD20 and AA.** Bold numbers indicate the best performance.

Method	Dataset	Task	Standard	FGSM	MM	CW20	PGD20	AA
ARTUDA	OfficeHome	All	27.03	11.79	8.55	9.01	9.21	7.86
RFA		All	55.00	20.50	9.81	15.39	16.15	8.49
SRoUDA		All	57.97	46.64	36.21	41.49	42.56	33.42
PL		All	66.00	55.08	47.38	51.08	51.71	44.38
TAROT		All	**68.29**	**57.29**	**49.58**	**53.44**	**54.01**	**46.80**

Table 12. **Performance comparison when using the standard margin risk and the robust margin risk on the source domain.** Bold numbers indicate the best performance.

Method	Dataset	Task	Stand	AA
TAROT w/ $\mathcal{R}_{\mathcal{S}}^{(\rho)}(f)$ (Ours)	OfficeHome	All	**68.29**	**46.80**
TAROT w/ $\mathcal{R}_{\mathcal{S}}^{\text{rob},(\rho)}(f)$	OfficeHome	All	67.63	44.24

Table 13. **Performances of ARTUDA, RFA, SRoUDA, PL and TAROT on Office31** ($\varepsilon = 8/255$)**.** In each cell, the first number is the standard accuracy (%), while the second number is the robust accuracy (%) for AA. Bold numbers indicate the best performance.

Method	A → D	A → W	D → A	D → W	W → A	W → D	Avg.
ARTUDA	47.79 / 45.58	47.67 / 45.16	42.88 / 33.12	88.81 / 86.54	59.99 / 36.74	94.18 / 91.57	63.55 / 56.45
RFA	78.51 / 45.18	73.84 / 33.08	62.30 / 46.57	98.24 / 79.87	61.02 / 43.95	99.20 / 81.53	78.85 / 55.03
SRoUDA	89.96 / 85.54	91.57 / 90.57	49.38 / 22.36	97.99 / 90.31	71.92 / 65.71	98.59 / 97.99	83.24 / 75.41
PL	**93.37 / 93.37**	**94.72** / 94.34	73.59 / 71.81	98.49 / 98.37	**74.26 / 72.63**	99.80 / 99.60	89.04 / 88.35
TAROT	**93.37** / 92.97	94.47 / **94.47**	**76.32 / 75.19**	**98.62 / 98.49**	72.74 / 71.64	**100.00 / 100.00**	**90.45 / 90.04**

Table 14. **Performances of ARTUDA, RFA, SRoUDA, PL and TAROT on OfficeHome** ($\varepsilon = 8/255$)**.** In each cell, the first number is the standard accuracy (%), while the second number is the robust accuracy (%) for AA. Bold numbers indicate the best performance.

Method	Ar → Cl	Ar → Pr	Ar → Rw	Cl → Ar	Cl → Pr	Cl → Rw	
ARTUDA	47.45 / 32.33	34.94 / 18.00	40.44 / 21.16	21.59 / 12.20	43.23 / 27.06	40.40 / 24.03	
RFA	47.49 / 31.59	53.80 / 29.13	62.98 / 28.44	43.55 / 16.32	59.36 / 32.55	57.20 / 25.78	
SRoUDA	53.61 / 46.64	75.22 / 66.57	78.56 / 69.89	60.07 / 54.68	70.06 / 67.13	70.07 / 62.70	
PL	56.01 / 52.92	72.58 / 68.37	78.63 / 68.99	60.82 / **55.71**	**72.88 / 68.53**	72.64 / 63.19	
TAROT	**56.58 / 53.28**	**75.36 / 71.50**	**79.09 / 70.62**	**61.06** / 55.30	72.52 / 68.17	**73.06 / 63.92**	
	Pr → Ar	Pr → Cl	Pr → Rw	Rw → Ar	Rw → Cl	Rw → Pr	Avg.
ARTUDA	27.73 / 9.81	46.76 / 37.39	49.46 / 28.02	32.18 / 17.18	54.85 / 43.71	68.12 / 40.03	42.26 / 25.91
RFA	42.32 / 14.50	47.61 / 28.34	64.13 / 25.89	54.88 / 19.04	55.62 / 33.31	72.76 / 37.26	55.14 / 26.84
SRoUDA	61.64 / **58.51**	44.74 / 41.51	79.39 / **71.06**	**72.64 / 69.76**	52.28 / 46.30	83.56 / 80.38	60.07 / 54.68
PL	61.10 / 56.20	52.81 / 49.71	78.63 / 69.06	72.60 / 67.74	60.21 / **56.63**	84.14 / 80.42	68.59 / 63.12
TAROT	**61.95** / 55.79	**54.09 / 50.84**	**79.62** / 70.65	72.56 / 68.56	**60.28** / 55.67	**84.66 / 80.74**	**69.23 / 63.75**

Table 15. **Performances of ARTUDA, RFA, SRoUDA, PL and TAROT on Office31** ($\varepsilon = 4/255$)**.** In each cell, the first number is the standard accuracy (%), while the second number is the robust accuracy (%) for AA. Bold numbers indicate the best performance.

Method	A → D	A → W	D → A	D → W	W → A	W → D	Avg.
ARTUDA	71.89 / 71.69	73.71 / 73.33	57.93 / 52.25	93.21 / 93.08	58.93 / 52.68	98.39 / 97.99	75.68 / 73.50
RFA	83.53 / 78.11	81.89 / 72.58	61.38 / 54.03	97.48 / 96.73	63.44 / 56.12	**100.00** / 99.20	81.29 / 76.13
SRoUDA	92.97 / 92.77	**95.22 / 94.21**	74.62 / 65.74	**98.74 / 98.74**	66.45 / 64.57	**100.00 / 100.00**	88.00 / 86.01
PL	89.56 / 89.56	93.46 / 93.33	75.04 / **74.55**	98.49 / 98.49	72.70 / 72.70	**100.00 / 100.00**	88.21 / 88.11
TAROT	**93.37 / 93.17**	93.84 / 93.59	**75.22 / 74.55**	98.49 / 98.49	**74.51 / 73.55**	**100.00 / 100.00**	**91.00 / 90.72**

Table 16. **Performances of ARTUDA, RFA, SRoUDA, PL and TAROT on OfficeHome** ($\varepsilon = 4/255$)**.** In each cell, the first number is the standard accuracy (%), while the second number is the robust accuracy (%) for AA. Bold numbers indicate the best performance.

Method	Ar → Cl	Ar → Pr	Ar → Rw	Cl → Ar	Cl → Pr	Cl → Rw	
ARTUDA	49.44 / 44.38	46.81 / 38.97	57.56 / 44.78	38.53 / 31.23	57.51 / 51.07	55.27 / 45.15	
RFA	49.21 / 40.18	58.80 / 45.28	69.20 / 48.73	50.23 / 29.30	63.11 / 48.46	62.80 / 42.46	
SRoUDA	55.44 / 51.84	**76.48 / 74.34**	79.00 / **77.19**	61.27 / 58.96	68.06 / 66.91	69.38 / 67.32	
PL	**55.79 / 54.18**	75.29 / 72.74	78.01 / 75.08	61.72 / 59.54	71.91 / 69.27	72.16 / 69.02	
TAROT	55.76 / 53.93	75.27 / 73.10	**79.27** / 75.88	**62.65 / 60.55**	**72.51 / 70.51**	**73.01 / 69.58**	
	Pr → Ar	Pr → Cl	Pr → Rw	Rw → Ar	Rw → Cl	Rw → Pr	Avg.
ARTUDA	39.31 / 29.46	52.42 / 48.11	63.23 / 52.33	50.35 / 42.56	54.22 / 54.34	73.71 / 64.90	53.20 / 45.61
RFA	48.04 / 28.88	49.51 / 39.08	70.07 / 47.74	59.37 / 37.78	56.52 / 45.06	75.22 / 57.90	61.21 / 43.49
SRoUDA	**61.60 / 59.15**	48.14 / 46.51	**80.39** / 76.08	**73.79** / 70.76	57.84 / 55.21	83.16 / 81.38	67.88 / 65.47
PL	58.14 / 57.93	52.33 / 50.75	79.34 / **76.59**	72.64 / 70.13	59.59 / 57.73	**83.74 / 82.07**	68.39 / 66.25
TAROT	60.65 / 58.92	**53.08 / 51.34**	79.78 / 76.43	73.05 / **71.74**	**59.92 / 57.92**	83.44 / 81.87	**69.03 / 66.82**