Supplementary Material

1. Proof

In this section, we provide a detailed proof of Thm. 4.1 and Thm. 4.2 in sequence.

1.1. Proof of classification bound

Before we reach the proof to the main theorem, we first prove the following lemmas for each theorem. With the notations introduced in Sec. 4, we introduce the following lemmas that will be used in proving the main theorem:

Lemma 1.1. [[1]] Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0, 1\}\}$, where $VCdim(\mathcal{H}) = d$, and for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$, then

$$|\epsilon_S(h) - \epsilon_T(h)| \leq d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z))$$

Lemma 1.2. Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0, 1\}\}$, where $VCdim(\mathcal{H}) = d$, and for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$. Let the noises on the source and target are defined as $n_S := \mathbb{E}_S[|Y - f_S(Z)|]$ and $n_T := \mathbb{E}_T[|Y - f_T(Z)|]$, where $f : \mathcal{Z} \to [0, 1]$ is the conditional mean function, i.e., $f(Z) = \mathbb{E}[Y|Z]$ then we have:

$$\begin{aligned}\big|\varepsilon_S(h) - \varepsilon_T(h)\big| &\leq |n_S + n_T| + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) \\ &+ \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}\end{aligned}$$

Proof. To begin with, we first show that for the source domain, $\varepsilon_S(h)$ cannot be too large if h is close to the optimal classifier f_S on source domain for $\forall h \in \mathcal{H}$:

$$\begin{aligned}&|\varepsilon_S(h) - \mathbb{E}_S[|h(Z) - f_S(Z)|]| \\ &= \big|\mathbb{E}_S[|h(Z) - Y|] - \mathbb{E}_S[|h(Z) - f_S(Z)|]\big| \\ &\leq \mathbb{E}_S\left[\big||h(Z) - Y| - |f_S(Z) - h(Z)|\big|\right] \\ &\leq \mathbb{E}_S[|Y - f_S(Z)|] \\ &= n_S\end{aligned}$$

Similarly, we also have an analogous inequality hold on the target domain:

$$|\varepsilon_T(h) - \mathbb{E}_T[|h(Z) - f_T(Z)|]| \leq n_T.$$

Combining both inequalities above, yields:

$$\begin{aligned}\varepsilon_S(h) \in [&\mathbb{E}_S[|h(Z) - f_S(Z)|] - n_S, \\ &\mathbb{E}_S[|h(Z) - f_S(Z)|] + n_S], \\ -\varepsilon_T(h) \in [&-\mathbb{E}_T[|h(Z) - f_T(Z)|] - n_T, \\ &-\mathbb{E}_T[|h(Z) - f_T(Z)|] + n_T]\end{aligned}$$

Hence,

$$\begin{aligned}&\big|\varepsilon_S(h) - \varepsilon_T(h)\big| \leq |n_S + n_T| + \\ &\big|\mathbb{E}_S[|h(Z) - f_S(Z)|] - \mathbb{E}_T[|h(Z) - f_T(Z)|]\big|\end{aligned}$$

Now to simplify the notation, for $e \in \{S, T\}$, define $\epsilon_e(h, h') = \mathbb{E}_e[|h(Z) - h'(Z)|]$, so that

$$\begin{aligned}&\big|\mathbb{E}_S[|h(Z) - f_S(Z)|] - \mathbb{E}_T[|h(Z) - f_T(Z)|]\big| \\ &= \big|\epsilon_S(h, f_S) - \epsilon_T(f_T, h)\big|.\end{aligned}$$

To bound $\big|\epsilon_S(h, f_S) - \epsilon_T(f_T, h)\big|$, on one hand, we have:

$$\begin{aligned}&\big|\epsilon_S(h, f_S) - \epsilon_T(f_T, h)\big| = \\ &\big|\epsilon_S(h, f_S) - \epsilon_S(h, f_T) + \epsilon_S(h, f_T) - \epsilon_T(f_T, h)\big| \\ &\leq \big|\epsilon_S(h, f_S) - \epsilon_S(h, f_T)\big| + \big|\epsilon_S(h, f_T) - \epsilon_T(f_T, h)\big| \\ &\leq \mathbb{E}_S[|f_S(Z) - f_T(Z)|] + \big|\epsilon_S(h, f_T) - \epsilon_T(f_T, h)\big|\end{aligned}$$

From 1.1, we have:

$$\leq \mathbb{E}_S[|f_S(Z) - f_T(Z)|] + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)).$$

Similarly, by the same trick of subtracting and adding back $\epsilon_T(h, f_S)$ above, the following inequality also holds:

$$\begin{aligned}&\big|\epsilon_S(h, f_S) - \epsilon_T(f_T, h)\big| \leq \\ &\mathbb{E}_T[|f_S(Z) - f_T(Z)|] + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)).\end{aligned}$$

Combine all the inequalities above, we know that:

$$\begin{aligned}\big|\varepsilon_S(h) - \varepsilon_T(h)\big| &\leq |n_S + n_T| + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) \\ &+ \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}\end{aligned}$$

■

Lemma 1.3. [[2], Corollary 3.19] Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0,1\}\}$, where $VCdim(\mathcal{H}) = d$. Then $\forall h \in \mathcal{H}, \forall \delta > 0$, w.p.b. at least $1-\delta$ over the choice of a sample size m and natural exponential e, the following inequality holds:

$$\varepsilon(h) \leq \widehat{\varepsilon}(h) + \sqrt{\frac{2d}{m}\log\frac{em}{d}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}.$$

Lemma 1.4. Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0,1\}\}$, where $VCdim(\mathcal{H}) = d$. Then $\forall h \in \mathcal{H}, \forall \delta > 0$, then w.p.b. at least $1-\delta$ over the choice of a sample size m and natural exponential e, the following inequality holds:

$$\begin{aligned}\varepsilon_T(h) \leq\ & \widehat{\varepsilon}_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\} \\ & + |n_S + n_T| + \sqrt{\frac{2d}{n}\log\frac{en}{d}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}\end{aligned}$$

Proof. Invoking the upper bound in 1.2, we have w.p.b at least $1-\delta$:

$$\begin{aligned}\varepsilon_T(h) \leq\ & \varepsilon_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\} \\ & + |n_S + n_T| \\ \leq\ & \widehat{\varepsilon}_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\} \\ & + |n_S + n_T| + \sqrt{\frac{2d}{n}\log\frac{en}{d}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}\end{aligned}$$

■

Theorem 1.1. Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0,1\}\}$, where $VCdim(\mathcal{H}) = d$. For $0 < \delta < 1$, then w.p.b. at least $1-\delta$ over the draw of samples S and T, for all $h \in \mathcal{H}$, we have:

$$\begin{aligned}\varepsilon_T(h) \leq\ & \frac{m}{n+m}\widehat{\varepsilon}_T(h) + \frac{n}{n+m}\widehat{\varepsilon}_S(h) \\ & + \frac{n}{n+m}(d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}) \\ & + \frac{n}{n+m}|n_S + n_T| \\ & + O\left(\sqrt{(\frac{1}{m}+\frac{1}{n})\log\frac{1}{\delta} + \frac{d}{n}\log\frac{n}{d} + \frac{d}{m}\log\frac{m}{d}}\right).\end{aligned}$$

Proof. Having 1.3, 1.4, we can use a union bound to combine them with coefficients $m/(n+m)$ and $n/(n+m)$ respectively, we have:

$$\begin{aligned}\varepsilon_T(h) \leq\ & \frac{m}{n+m}\left(\widehat{\varepsilon}_T(h) + \sqrt{\frac{2d}{m}\log\frac{em}{d}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}\right) \\ & + \frac{n}{n+m}(\widehat{\varepsilon}_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}) \\ & + \frac{n}{n+m}\left(|n_S + n_T| + \sqrt{\frac{2d}{n}\log\frac{en}{d}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}\right).\end{aligned}$$

From Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}\varepsilon_T(h) \leq\ & \frac{m}{n+m}\left(\widehat{\varepsilon}_T(h) + \sqrt{\frac{4d}{m}\log\frac{em}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right) \\ & + \frac{n}{n+m}(\widehat{\varepsilon}_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}) \\ & + \frac{n}{n+m}\left(|n_S + n_T| + \sqrt{\frac{4d}{n}\log\frac{en}{d} + \frac{1}{n}\log\frac{1}{\delta}}\right).\end{aligned}$$

As $m \ll n$ and applying Cauchy-Schwartz inequality one more time, we have

$$\begin{aligned}& \leq \frac{m}{n+m}\widehat{\varepsilon}_T(h) + \frac{n}{n+m}\widehat{\varepsilon}_S(h) \\ & + \frac{n}{n+m}(d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \\ & \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}) \\ & + \frac{n}{n+m}(|n_S + n_T| \\ & + \sqrt{\frac{8d}{m}\log\frac{em}{d} + \frac{2}{m}\log\frac{1}{\delta} + \frac{8d}{n}\log\frac{en}{d} + \frac{2}{n}\log\frac{1}{\delta}}). \\ \leq\ & \frac{m}{n+m}\widehat{\varepsilon}_T(h) + \frac{n}{n+m}\widehat{\varepsilon}_S(h) \\ & + \frac{n}{n+m}(d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}) \\ & + \frac{n}{n+m}(|n_S + n_T|) \\ & + O(\sqrt{(\frac{1}{m}+\frac{1}{n})\log\frac{1}{\delta} + \frac{d}{m}\log\frac{m}{d} + \frac{d}{n}\log\frac{n}{d}})\end{aligned}$$

■

1.2. Proof of regression bound

For regression generalization bound, we follow the proof strategy in previous section, but with slight change of definitions. We let $\mathcal{H} = \{h : \mathcal{Z} \to [0,1]\}$ be a set of bounded real-valued functions from the input space $\mathcal{Z}$ to $[0,1]$. We use $Pdim(\mathcal{H})$ to denote the pseudo-dimension of $\mathcal{H}$, and let $Pdim(\mathcal{H}) = d$. We first prove the following lemmas that will be used in proving the main theorem:

Lemma 1.5. [3] For $h, h' \in \mathcal{H} := \{h : \mathcal{Z} \to [0,1]\}$, where $Pdim(\mathcal{H}) = d$, and for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$,

$$|\epsilon_S(h,h') - \epsilon_T(h,h')| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z))$$

where $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$.

Lemma 1.6. For $h, h' \in \mathcal{H} := \{h : \mathcal{Z} \to [0,1]\}$, where $Pdim(\mathcal{H}) = d$, and for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$, we define $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} := h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. Then $\forall h \in \mathcal{H}$, the following inequality holds:

$$\begin{aligned}\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq |n_S + n_T| + d_{\tilde{\mathcal{H}}}(\mathcal{D}_T(Z), \mathcal{D}_S(Z)) \\ + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}\end{aligned}$$

Lemma 1.7. Thm.11.8 [2] Let $\mathcal{H}$ be the set of real-valued function from $\mathcal{Z}$ to $[0,1]$. Assume that $Pdim(\mathcal{H}) = d$. Then $\forall h \in \mathcal{H}, \forall \delta > 0$, with probability at least $1-\delta$ over the choice of a sample size m and natural exponential e, the following inequality holds:

$$\varepsilon(h) \leq \widehat{\varepsilon}(h) + \sqrt{\frac{2d}{m}\log\frac{em}{d}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}.$$

Lemma 1.8. Let $\mathcal{H}$ be a set of real-valued functions from $\mathcal{Z}$ to $[0,1]$ with $Pdim(\mathcal{H}) = d$, and $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. For $0 < \delta < 1$, then w.p.b. at least $1-\delta$ over the draw of samples S and T, for all $h \in \mathcal{H}$, we have:

$$\begin{aligned}\varepsilon_T(h) \leq \widehat{\varepsilon}_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) \\ + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\} \\ + |n_S + n_T| + \sqrt{\frac{2d}{n}\log\frac{en}{d}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}\end{aligned}$$

Proof. Invoking the upper bound in 1.6 and 1.7, we have w.p.b at least $1-\delta$:

$$\begin{aligned}\varepsilon_T(h) \leq\ & \widehat{\varepsilon}_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\} \\ & + |n_S + n_T| \\ \leq\ & \widehat{\varepsilon}_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\} \\ & + |n_S + n_T| + \sqrt{\frac{2d}{n}\log\frac{en}{d}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}\end{aligned}$$

■

Theorem 1.2. Let $\mathcal{H}$ be a set of real-valued functions from $\mathcal{Z}$ to $[0,1]$ with $Pdim(\mathcal{H}) = d$, and $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. For $0 < \delta < 1$, then w.p.b. at least $1-\delta$ over the draw of samples S and T, for all $h \in \mathcal{H}$, we have:

$$\begin{aligned}\varepsilon_T(h) \leq\ & \frac{m}{n+m}\widehat{\varepsilon}_T(h) + \frac{n}{n+m}\widehat{\varepsilon}_S(h) \\ & + \frac{n}{n+m}(d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) \\ & + \min\{\mathbb{E}_S[|f_S(Z) - f_T(Z)|], \mathbb{E}_T[|f_S(Z) - f_T(Z)|]\}) \\ & + \frac{n}{n+m}|n_S + n_T| + \\ & O\left(\sqrt{(\frac{1}{m}+\frac{1}{n})\log\frac{1}{\delta} + \frac{d}{n}\log\frac{n}{d} + \frac{d}{m}\log\frac{m}{d}}\right).\end{aligned}$$

Proof. Having 1.5, 1.6, 1.7, 1.8, we can use a union bound to combine them with coefficients $m/(n+m)$ and $n/(n+m)$ respectively, and replace the $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ with $d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T)$ in the proof of Thm. 1.1. Obviously, we have Thm. 1.2. ■

1.3. Discussions

From binary to multi-class We leave the extension from binary to multi-class in the future work. Definition 4.1 is essentially a relaxed version of the classic total variation distance, so it can also be used in multi-class classification problems where we have more than 2 conditional distributions. However, the results in Theorem 4.1 indeed can only be applied in the binary classification setting, due to the use of VC-dimension, which only makes sense for binary classification problems. That being said, extension to the multi-class classification problem shouldn't be hard, although its presentation would be much more involved. At a high level, the proof essentially boils down to replace the VC dimension with the the Natarajan dimension [4].

Generalizing to ℓ_1 loss The bound can indeed be straightforwardly generalized to ℓ_p loss. We use the ℓ_1 loss in the presentation in order to be consistent with our experiments of vehicle counting, where we use the mean absolute error. A similar variational formulation of the conditional mutual information for continuous variables with ℓ_p loss also holds [5]. We use the cross-entropy loss in practice for classification problems because the 0-1 loss is computationally intractable to optimize, and the cross-entropy loss serves as a convex proxy for the binary loss.

Bridging theory and practice The bounds serve as justification to design our algorithms but they contain terms that are not directly available in practice, e.g., the optimal hypothesis of the two domains. On the other hand, using information-theoretic tools, we can close the gap between theory and practice, since the minimization of the conditional mutual information implies the minimization of the distance between the two optimal hypothesis.

2. More Experimental Results

2.1. Comparison with State-of-the-art Methods

Here we should address that we are not strictly in the same setting with MME [6] and its related works [7]. In all datasets, MME addresses in the few-shot setting and uses 1 or 3-shot as target labeled data. We argue that 1 or 3-shot is only suitable for few-shot learning and prototypical representations learning. The data selection will bias the model to a large extent. While we use more data to alleviate the selection bias and to mimic the usage in real domain adaptation application, such as 1%, 5% and more proportion of target labeled data. In DomainNet, MME use 126 out of the total 345 classes data, while we keep use the 345 classes for better evaluate our model performance in a more challenging setting. Although not strictly under the same setting, we also compare ours with the current state-of-the-art Semi-DA methods in our setting. MME [6] and other related works (such as APE [7]) assumes that there exists class-wise prototypical representations between source and target domain, and exploit Cosine Classifier [8] to help learn the prototypical representations. However, our method can also seamlessly combine with Cosine Classifier for better performance, even comparing with the methods above as in Table 2.

Table 1: The weights trade off between invariant representation part and invariant risk part under OfficeHome: Art to Real scenarios.(mean $\pm$ std)

	λ_{risk}		
λ_{rep}	1	0.1	0.01
1	70.23±0.18	70.96±0.17	70.55±0.18
0.1	71.20±0.14	72.66±0.16	72.31±0.19
0.01	72.65±0.15	**73.12**±0.19	72.97±0.20

2.2. Hyper-parameters

There are two fundamental part in our proposed LIRR loss. One is the invariant representation item, the other is the invariant risk item. We use λ_{rep} and λ_{risk} represent the weights of invariant representation item and invariant risk item respectively. In order to explore the best trade off between this two items, we conduct extra experiments on Art to Real scenario in OfficeHome dataset. All other hyper-parameters settings are set as same as Sec.6.3. The results can be found in Table. 1. From which, we can see that the optimal performance is achieved when $\lambda_{\text{risk}} = 0.1$ and $\lambda_{\text{rep}} = 0.01$.

2.3. Implementation Details

For image classification task: we use ResNet34 as backbone networks. We adopt SGD with learning rate of 1e-3, momentum of 0.9 and weight decay factor of 5e-4. We decay the learning rate with a multiplier 0.1 when training process reach three quarters of the total iterations. The batch size is set as 128 for VisDA2017 and Domainnet, 64 for officehome. For adversarial training, we use gradient reversal layer (GRL) to flip gradient in the backpropagation between feature encoder $g(\cdot)$ and domain discriminator $\mathcal{C}(\cdot)$ to obtain domain-invariant representation w.r.t. source labeled data and target unlabeled data. For min-max training objective for $\mathcal{L}_i$ and $\mathcal{L}_d$ in Eq.7, we implement it with the difference on two losses, $L(y, h(z))$ and $L(y, h(z, d))$. $h(z)$ is realized by a common predictor which only takes feature z as input. $h(z, d)$ indicates an additional predictor which takes the combination of feature z and domain index d, e.g. we concatenate original feature z with an additional full 0 (or 1) channel to represent source(or target) domain. It's worth noting that according to [6], the utilization of entropy minimization hurts the performance. Thus, we implement the CDAN method without entropy minimization. Our results are all obtained without heavy engineering tricks. All code is implemented in Pytorch and will be made available upon acceptance.

For traffic counting regression task: we use VGG16 as encoder and FCN8s [9] as decoder. The model will output a density map as the regression result for input images. The optimizing goal is a joint loss including both the euclidean loss between the groundtruth density map and the predicted one, and the mean absolute counting error loss between the total predicted count and groundtruth count. We use mean absolute error (MAE) metric for evaluation, which measure the absolute difference between the output count and the ground-truth count. We adopt Adam optimizer with learning rate set to 1e-6. The batch size is set as 24.

2.4. Discussions

LIRR with cosine classifier It should be noted that LIRR alone achieves favorable performance over other baselines. With the cosine classifier, a useful technique adopted by many previous works [6], LIRR's performance can be further improved. We perform this additional experiment mainly to show that our framework is also compatible with existing techniques, e.g., the cosine classifier.

Grad-CAM results In Figure.4, LIRR captures a more complete representation of the husky's face than both DANN and Source+Target, in which they capture only parts of the husky's face. Similarly for puppy (second row), LIRR captures more on the puppy's outline and body shape.

References

[1] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," in *Advances in neural information processing systems*, pp. 129–136, 2008.

Table 2: Accuracy (%) comparison (higher means better) on **NICO**, **OfficeHome** with 1% (above) and 5% (below) labeled target data (mean ± std). Highest accuracies are highlighted in bold.

1% labeled target	NICO Animal		NICO Traffic		OfficeHome		
Method	Grass to Snow	Snow to Grass	Sunset to Beach	Beach to Sunset	Art to Real	Real to Prod.	Prod. to Clip.
MME [6]	87.12±0.76	79.52±0.43	78.69±0.86	74.21±0.78	72.66±0.18	78.07±0.17	52.78±0.16
APE [7]	87.41±0.42	80.22±0.45	79.70±0.80	75.50±0.85	**74.41**±0.49	78.45±0.42	**54.84**±0.43
Ours (LIRR + CosC)	**89.67**±0.72	**89.73**±0.68	**81.00**±0.89	**79.98**±0.95	73.62±0.21	**80.20**±0.23	53.84±0.19

5% labeled target	NICO Animal		NICO Traffic		OfficeHome		
Method	Grass to Snow	Snow to Grass	Sunset to Beach	Beach to Sunset	Art to Real	Real to Prod.	Prod. to Clip.
MME [6]	87.80±0.87	85.50±0.95	92.02±0.85	90.76±0.81	75.24±0.22	82.45±0.18	61.75±0.19
APE [7]	87.85±0.55	87.21±0.80	92.55±0.75	91.05±0.45	75.90±0.50	**83.65**±0.49	62.76±0.51
Ours (LIRR + CosC)	**88.97**±0.45	**88.22**±0.55	**92.70**±0.87	**91.50**±1.05	**76.63**±0.19	83.45±0.22	**62.84**±0.23

Table 3: Accuracy (%) comparison (higher means better) on **DomainNet**, and **VisDA2017** with 1% (above) and 5% (below) labeled target data (mean ± std). Highest accuracies are highlighted in bold.

1% labeled target	Domainnet			VisDA2017
Method	Real to Clip.	Sketch to Real	Clip. to Sketch	Train to Val.
MME [6]	51.04±0.12	60.35±0.12	45.09±0.14	80.52±0.35
APE [7]	52.21±0.37	**62.40**±0.24	46.24±0.42	81.09±0.17
Ours (LIRR + CosC)	**53.42**±0.09	61.79±0.11	**47.83**±0.10	**82.31**±0.21

5% labeled target	Domainnet			VisDA2017
Method	Real to Clip.	Sketch to Real	Clip. to Sketch	Train to Val.
MME [6]	62.31±0.11	69.02±0.18	53.88±0.14	84.12±0.22
APE [7]	63.01±0.23	68.92 ±0.22	53.95±0.27	84.79±0.49
Ours (LIRR + CosC)	**63.03**±0.17	**69.52**±0.09	**54.44**±0.12	**85.06**±0.17

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2nd ed., 2018.

[3] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in neural information processing systems*, pp. 8559–8570, 2018.

[4] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz, "Multiclass learnability and the erm principle," in *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 207–232, JMLR Workshop and Conference Proceedings, 2011.

[5] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances in Neural Information Processing Systems*, pp. 4240–4248, 2016.

[6] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *IEEE International Conference on Computer Vision*, pp. 8050–8058, 2019.

[7] T. Kim and C. Kim, "Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation," *arXiv preprint arXiv:2007.09375*, 2020.

[8] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4367–4375, 2018.

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Annals of the History of Computing*, no. 04, pp. 640–651, 2017.