Learning Invariant Representations and Risks for Semi-supervised Domain Adaptation

Bo Li[13*], Yezhen Wang[2*], Shanghang Zhang[1*]
Dongsheng Li[3], Trevor Darrell[1], Kurt Keutzer[1], Han Zhao[4†]

[1]BAIR, UC Berkeley
[2]UC San Diego
[3]Microsoft Research Asia
[4]D. E. Shaw & Co.

Abstract

The success of supervised learning hinges on the assumption that the training and test data come from the same underlying distribution, which is often not valid in practice due to potential distribution shift. In light of this, most existing methods for unsupervised domain adaptation focus on achieving domain-invariant representations and small source domain error. However, recent works have shown that this is not sufficient to guarantee good generalization on the target domain, and in fact, is provably detrimental under label distribution shift. Furthermore, in many real-world applications it is often feasible to obtain a small amount of labeled data from the target domain and use them to facilitate model training with source data. Inspired by the above observations, in this paper we propose the first method that aims to simultaneously learn invariant representations and risks under the setting of semi-supervised domain adaptation (Semi-DA). First, we provide a finite sample bound for both classification and regression problems under Semi-DA. The bound suggests a principled way to obtain target generalization, i.e., by aligning both the marginal and conditional distributions across domains in feature space. Motivated by this, we then introduce the LIRR algorithm for jointly **L**earning **I**nvariant **R**epresentations and **R**isks. Finally, extensive experiments are conducted on both classification and regression tasks, which demonstrate that LIRR consistently achieves state-of-the-art performance and significant improvements compared with the methods that only learn invariant representations or invariant risks.

1 Introduction

The success of supervised learning hinges on the key assumption that test data should share the same distribution with the training data. Unfortunately, in most of the real-world applications, data are dynamic, meaning that there is often a distribution shift between the training (source) and test (target) domains. To this end, unsupervised domain adaptation (UDA) methods aim to approach this problem by adapting the predictive model from labeled source data to the unlabeled target data. Recent advances in UDA focus on learning domain-invariant representations that also lead to a small error on the source domain. The goal is to learn representations, along with the source predictor, that can generalize to the target domain (Long

[*]Equal contribution
[†]Work done while at Carnegie Mellon University

et al., 2015; Ganin et al., 2016; Tzeng et al., 2017; Long et al., 2018; Chen et al., 2019; Zhao et al., 2018). However, recent works (Zhao et al., 2019a; Wu et al., 2019; Combes et al., 2020) have shown that the above conditions are not sufficient to guarantee good generalizations on the target domain. In fact, if the marginal label distributions are distinct across domains, the above method provably hurts target generalization (Zhao et al., 2019a).

On the other hand, while labeled target data is usually more difficult or costly to obtain than labeled source data, it can lead to better accuracy (Hanneke & Kpotufe, 2019). Furthermore, in many practical applications, e.g., vehicle counting, object detection, speech recognition, etc., it is often feasible to at least obtain a small amount of labeled data from the target domain so that it can facilitate model training with source data (Li & Zhang, 2018; Saito et al., 2019). Motivated by these observations, in this paper we focus on a more realistic setting of semi-supervised domain adaptation (Semi-DA). In Semi-DA, in addition to the large amount of labeled source data, the learner also has access to a small amount of labeled data from the target domain. Again, the learner's goal is to produce a hypothesis that well generalizes to the target domain, under the potential shift between the source and the target. Semi-DA is both a more-realistic and generalizable setting that allows practitioners to design better algorithms that can overcome the aforementioned limitations in UDA. The key question in this scenario is: *how to maximally exploit the labeled target data for better model training?*

In this paper, we address the above question under the Semi-DA setting. In order to first understand how performance discrepancy occurs, we derive a finite-sample generalization bound for both classification and regression problems under Semi-DA. Our theory shows that, for a given predictor, the accuracy discrepancy between two domains depends on two terms: (i) the distance between the marginal feature distributions, and (ii) the distance between the optimal predictors from source and target domains. Our observation naturally leads to a principled way of learning invariant representations (to minimize discrepancy between marginal feature distributions) and risks (to minimize discrepancy between conditional distributions over the features) across domains simultaneously for a better generalization on the target. In light of this, we introduce our novel bound minimization algorithm LIRR, a model of jointly **L**earning **I**nvariant **R**epresentations and **R**isks for such purposes. As a comparison, existing works focus on either learning invariant representations only (Ganin et al., 2016; Tzeng et al., 2017; Zhao et al., 2018; Chen et al., 2019), or learning invariant risks only (Arjovsky et al., 2019; Chang et al., 2020), which are not sufficient to reduce the accuracy discrepancy for good generalizations on the target. Different from these methods, LIRR jointly learns invariant representations and risks, and as a result, better mitigates the accuracy discrepancy across domains. To better understand our method, we illustrate the proposed algorithm, LIRR, in Fig. 1.

Our Contributions In summary, our work provides the following contributions:

- Theoretically, we provide finite-sample generalization bounds for Semi-DA on both classification (Theorem 3.1) and regression (Theorem 3.2) problems. Our bounds inform new directions for simultaneously optimizing both marginal and conditional distributions across domains for better generalization on the target. To the best of our knowledge, this is the first generalization analysis in the Semi-DA setting.
- To bridge the gap between theory and practice, we provide an information-theoretic interpretation of our theoretical results. Based on this perspective, we propose a bound minimization algorithm, LIRR, to jointly learn invariant representations and invariant optimal predictors, in order to mitigate the accuracy discrepancy across domains for better generalizations.
- We systematically analyze LIRR with extensive experiments on both classification and regression tasks. Compared with methods that only learn invariant representations or invariant risks, LIRR demonstrates

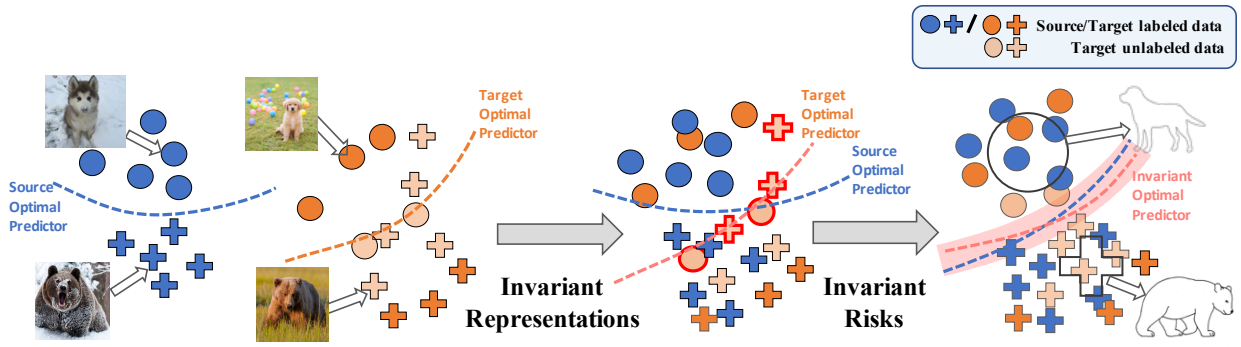

Figure 1: Overview of the proposed model. Learning invariant representations induces indistinguishable representations across domains, but there can still be mis-classified samples (as stated in red circle) due to misaligned optimal predictors. Besides learning invariant representations, LIRR model jointly learns invariant risks to better align the optimal predictors across domains.

significant improvements on Semi-DA. We also analyze the adaptation performance with increasing labeled target data, which shows LIRR even surpasses oracle method *Full Target* trained only on labeled target data, suggesting that LIRR can successfully exploit the structure in source data to improve generalization on the target domain.

2 Preliminaries

Unsupervised Domain Adaptation We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the input and output space, respectively. Similarly, $\mathcal{Z}$ stands for the representation space induced from $\mathcal{X}$ by a feature transformation $g : \mathcal{X} \mapsto \mathcal{Z}$. Accordingly, we use X, Y, Z to denote random variables which take values in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. Throughout the paper, a domain corresponds to a joint distribution on the input space $\mathcal{X}$ and output space $\mathcal{Y}$. We use $\mathcal{D}_S$ ($\mathcal{D}_T$) to denote the source (target) domain and subsequently we also use $\mathcal{D}_S(Z)(\mathcal{D}_T(Z))$ to denote the marginal distributions of $\mathcal{D}_S(\mathcal{D}_T)$ over Z. Furthermore, let D be a categorical variable that corresponds to the index of domain, i.e., $D \in \{S, T\}$. The overall sampling process for our data can then be specified by first drawing a value of D, and then depending on the value of D, we sample from the corresponding distribution $\mathcal{D}_D$. Under this setting, the probabilities of $\Pr(D = T)$ and $\Pr(D = S)$ then determine the relative sample sizes of our target and source data.

A hypothesis over the feature space $\mathcal{Z}$ is a function $h : \mathcal{Z} \to [0, 1]$. The *error* of a hypothesis h under distribution $\mathcal{D}_S$ and feature transformation g is defined as: $\varepsilon_S(h, f) := \mathbb{E}_{\mathcal{D}_S}[|h(g(X)) - f(X)|]$. In classification setting, in which f and h are binary classification functions, above definition reduces to the probability that h disagrees with f under $\mathcal{D}_S$: $\mathbb{E}_{\mathcal{D}_S}[|h(g(X)) - f(X)|] = \Pr_{\mathcal{D}_S}(h(g(X)) \neq Y)$. In regression, the above error is then the usual mean absolute error, i.e., the ℓ_1 loss. As a common notation, we also use $\widehat{\varepsilon}_S(h)$ to denote the empirical risk of h on the source domain. Similarly, $\varepsilon_T(h)$ and $\widehat{\varepsilon}_T(h)$ are the true risk and the empirical risk on the target domain. For a hypothesis class $\mathcal{H}$, we use $VCdim(\mathcal{H})$ and $Pdim(\mathcal{H})$ to denote the VC-dimension and pseudo-dimension of $\mathcal{H}$, respectively.

Semi-supervised Domain Adaptation Formally, in Semi-DA the learner is allowed to have access to a small amount of labeled data in target domain $\mathcal{D}_T$. Let $S = \{(\mathbf{x}_i^{(S)}, y_i^{(S)})\}_{i=1}^n$ be a set of labeled data

sampled i.i.d. from $\mathcal{D}_S$. Similarly, we have $T = \{(\mathbf{x}_j^{(T)})\}_{j=1}^k$ as the set of target unlabeled data sampled from $\mathcal{D}_T$, and we let $\widetilde{T} = \{(\mathbf{x}_j^{(\widetilde{T})}, y_j^{(\widetilde{T})})\}_{j=1}^m$ be the small set of labeled data where $m \leq k$. Usually, we also have $m \ll n$, and the goal of the learner is to find a hypothesis $h \in \mathcal{H}$ by learning from S, T and $\widetilde{T}$ so that h has a small target error $\varepsilon_T(h)$.

Clearly, with the additional small amount of labeled data $\widetilde{T}$, one should expect a better generalization performance than what the learner could hope to achieve in the setting of unsupervised domain adaptation. To this end, we first state the following generalization upper bound from Zhao et al. (2019a) in the setting of unsupervised domain adaptation:

Theorem 2.1. (Zhao et al., 2019a) Let $\langle \mathcal{D}_S(X), f_S \rangle$ and $\langle \mathcal{D}_T(X), f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\mathcal{H}}(\mathcal{D}_S(X), \mathcal{D}_T(X)) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}. \tag{1}$$

The $d_{\mathcal{H}}(\cdot,\cdot)$ is known as the $\mathcal{H}$-divergence (Ben-David et al., 2010), a pseudo-metric parametrized by $\mathcal{H}$ to measure the discrepancy between two distributions. It should be noted that the above theorem is a *population result*, hence it does not give a *finite sample bound*. Furthermore, the setting above is *noiseless*, where f_S and f_T correspond to the groundtruth labeling functions in source and target domains. Nevertheless, it provides an insight on achieving domain adaptation through bounding the error difference on source and target domains: to simultaneously minimize the distances between feature representations and between the optimal labeling functions.

3 Generalization Bounds for Semi-supervised Domain Adaptation

In this section, we derive a finite-sample generalization bound for Semi-DA, where the model has access to both a large amount of labeled data S from the source domain, and a small amount of labeled data $\widetilde{T}$ from the target domain. For this purpose, we first introduce the definition of $\mathcal{H}$ on both classification and regression settings, and then present our theoretical results of the generalization upper bounds for Semi-DA.

Definition 3.1. Let $\mathcal{H}$ be a family of binary functions from $\mathcal{Z}$ to $\{0,1\}$, and $\mathcal{A}_{\mathcal{H}}$ be the collection of subsets of $\mathcal{Z}$ defined as $\mathcal{A}_{\mathcal{H}} := \{h^{-1}(1) \mid h \in \mathcal{H}\}$. The distance between two distributions $\mathcal{D}$ and $\mathcal{D}'$ based on $\mathcal{H}$ is: $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') := \sup_{A \in \mathcal{A}_{\mathcal{H}}} |\Pr_{\mathcal{D}}(A) - \Pr_{\mathcal{D}'}(A)|$.

With the definition, we have the symmetric difference w.r.t. itself as: $\mathcal{H}\Delta\mathcal{H} = \{h(z) \oplus h'(z) \mid h, h' \in \mathcal{H}\}$, where $\oplus$ is the **XOR** operation. Next, considering that for a joint distribution $\mathcal{D}$ over $\mathcal{Z} \times \mathcal{Y}$ in our setting, there may be noise in the conditional distribution $\Pr_{\mathcal{D}}(Y \mid Z)$. It is then necessary to define a term to measure the noise level of each domain. To this end, in classification, we define the noise on the source domain $n_S := \mathbb{E}_S[|Y - f_S(Z)|]$, where $f_S : \mathcal{Z} \to [0,1]$ is the conditional mean function, i.e., $f_S(Z) = \mathbb{E}_S[Y \mid Z]$. Similar definition also applies to the target domain, where we use n_T to denote the noise in target. In regression, with ℓ_1 loss, we define $f_S : \mathcal{Z} \to \mathbb{R}$ to be the conditional median function of $\Pr(Y \mid Z)$, i.e. $f_S(Z) := \inf_y\{y \in \mathbb{R} : 1/2 \leq \Pr(Y \leq y \mid Z)\}$. Now we are ready to state the main results in this section:

Theorem 3.1. (Classification generalization bound in Semi-DA). Let $\mathcal{H}$ be a hypothesis set with functions $h : \mathcal{Z} \to \{0,1\}$ and $VCdim(\mathcal{H}) = d$. For $0 < \delta < 1$, then w.p. at least $1 - \delta$ over the n samples in S and

m samples in $\widetilde{T}$, for all $h \in \mathcal{H}$, we have:

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

Theorem 3.2. (Regression generalization bound in Semi-DA). Let $\mathcal{H}$ be a hypothesis set with functions $h : \mathcal{Z} \to [0, 1]$ and $Pdim(\mathcal{H}) = d$. Then we define $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. For $0 < \delta < 1$, then w.p. at least $1 - \delta$ over the n samples in S and m samples in $\widetilde{T}$, for all $h \in \mathcal{H}$, we have:

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

Remark It is worth pointing out that both n_S and n_T are constants that only depend on the underlying source and target domains, respectively. Hence $|n_S + n_T|$ essentially captures the the amplitude of noise. The last two terms of the bound come from standard concentration analysis for uniform convergence.

Comparing with previous results (Ben-David et al., 2010; Zhao et al., 2019a), our bounds here contain empirical error terms from *both* the source and target domains. Furthermore, the relative importance of these two terms is naturally controlled by the relative number of data we have from S and T, which also explains the importance of the availability of the target labeled data. More importantly, these bounds imply a natural and principled way for a better generalization to the target domain by learning invariant representations and risks simultaneously. Note that this is in sharp contrast to previous works where only invariant representations are pursued (Ganin et al., 2016; Zhao et al., 2018).

4 Learning Invariant Representations and Risks

Motivated by the generalization error bounds in Theorem 3.1 and Theorem 3.2 in Sec. 3, in this section we propose our bound minimization algorithm LIRR. Since the last two terms reflect the noise level, complexity measures and error caused by finite samples, respectively, we then hope to optimize the upper bound by minimizing the first four terms. The first two terms are the convex combination of empirical errors of h on S and T, which can be optimized with the labeled source and target data. The third term measures the distance of representations between the source and target domains, which is a good inspiration for us to learn the *invariant representation* (Ganin et al., 2016) across domains. The fourth term corresponds to the distance of the optimal classifiers between S and T. To minimize this term, the model is forced to learn the data representations that induce the same optimal predictors for both source and target domains, which exactly corresponds to the principle of *invariant risk minimization* (Arjovsky et al., 2019).

4.1 Information Theoretic Interpretation

To better understand why the bound minimization strategy can solve the intrinsic problems of Semi-DA, in what follows we provide interpretations from an information-theoretic perspective.

Invariant Representations Learning invariant representations corresponds to minimizing the third term of the bound Theorem 3.1 and bound Theorem 3.2. We consider a feature transformation $Z = g(X)$ that can obtain the invariant representation Z from input X. The invariance on representations can be described as achieving statistical independence $D \perp Z$, where D stands for the domain index. This independence is equivalent to the minimization of mutual information $I(D; Z)$. To see this, if $I(D; Z) = 0$, then $\mathcal{D}_S(Z) = \mathcal{D}_T(Z)$, so the third term in the bounds will vanish. Intuitively, this means that by looking at the representations Z, even a well-trained domain classifier $\mathcal{C}(\cdot)$ cannot correctly guess the domain index D. We also call the learned feature transformation $g : \mathcal{X} \to \mathcal{Z}$ the invariant encoder.

Invariant Risks Learning invariant risks corresponds to minimizing the fourth term of the bound Theorem 3.1 and bound Theorem 3.2. Inspired by Arjovsky et al. (2019), we want to identify a subset of feature representation through feature transformation $Z = g(X)$ that best supports an invariant optimal predictor for source and target domains. That means the identified feature representation $Z = g(X)$ can induce the same optimal predictors. This objective can be interpreted with a conditional independence $D \perp Y \mid Z$, which is equivalent to minimizing $I(D; Y \mid Z)$. To see this, when the conditional mutual information of $I(D; Y \mid Z)$ equals 0, the two conditional distributions $\Pr_S(Y \mid Z)$ and $\Pr_T(Y \mid Z)$ coincide with each other. As a result, the Bayes optimal predictors, which only depend on the conditional distributions of $Y \mid Z$, become the same across domains, so the fourth term in our bounds Theorem 3.1, Theorem 3.2 will vanish.

In summary, our learning objective on invariant representations and invariant risks are achievable with the joint minimization of $I(D; Z)$ and $I(D; Y \mid Z)$. It is instructive to present the integrated form as in Eq. 2. In words, the integrated form suggests the independence of $D \perp (Y, Z)$. We regard the independence as an intrinsic objective for domain adaptation since it implies an alignment of the joint distributions over (Y, Z) across domains, as opposed to only the marginal distributions over Z in existing works.

$$I(D; Y, Z) = \underbrace{I(D; Z)}_{\text{Invariant Representation}} + \underbrace{I(D; Y \mid Z)}_{\text{Invariant Risk}} \tag{2}$$

4.2 Algorithm Design

To learn invariant representations, we adopt the adversarial training method as in Ganin et al. (2016). The invariant representation objective focuses on learning the feature transformation $g(\cdot)$ to obtain the invariant feature representations from input X, which can fool the domain classifier $\mathcal{C}$. This part of the objective function can be described as in Eq. 3.

$$\mathcal{L}_{\text{rep}}(g, \mathcal{C}) = \mathbb{E}_{X \sim \mathcal{D}_S(X)}[\log(\mathcal{C}(g(X)))] + \mathbb{E}_{X \sim \mathcal{D}_T(X)}[\log(1 - \mathcal{C}(g(X)))]. \tag{3}$$

To learn invariant risks, we convert the conditional mutual information $I(Y; D \mid Z)$ to the difference of the two conditional entropies, as in Eq. 4.

$$I(Y; D \mid Z) = H(Y \mid Z) - H(Y \mid D, Z) \tag{4}$$

The following proposition gives a variational form of the conditional entropy as infimum over a family of croess-entropies, where L denotes the cross-entropy loss.

Proposition 4.1. (Farnia & Tse, 2016) $H(Y \mid Z) = \inf_f \mathbb{E}[L(Y; f(Z))]$.

Using the above variational form, the minimization of the conditional entropy over g could be transformed to a minimization of the cross-entropy over both f and g. Hence, the learning objective of Eq. 2 can be achieved with the following loss functions.

$$\begin{aligned}\min_{g,f_i} \max_{f_d} \mathcal{L}_{\text{risk}}(g, f_i, f_d) = \mathbb{E}_{(x,y)\sim\mathcal{D}_S,\mathcal{D}_{\widetilde{T}}}[L(y, f_i(g(x)))] \\ - \mathbb{E}_{d\sim D}\, \mathbb{E}_{(x,y)\sim\mathcal{D}_S,\mathcal{D}_{\widetilde{T}}}[L(y, f_d(g(x), d))]\end{aligned} \tag{5}$$

Note that in Eq. 5, the difference between the discriminators f_i and f_d is that besides the features given by $g(\cdot)$, f_d also takes the domain index D as its input whereas f_i can only have access to the features $Z = g(X)$. This difference is due to the conditioning variables in $H(Y \mid Z)$ and $H(Y \mid D, Z)$ respectively.

In general, as the factorization in Eq. 2 suggests, in order to achieve improved adaptation performance by minimizing the accuracy discrepancy between domains, we need to enforce the joint independence of $(Y, Z) \perp D$ by learning feature transformation g. To achieve it, we propose our learning objective of LIRR as in Eq. 6, where λ_{risk} and λ_{rep} are set to 1 by default.

$$\min_{g,f_i} \max_{\mathcal{C},f_d} \mathcal{L}_{\text{LIRR}}(g, f_i, f_d, \mathcal{C}) := \lambda_{\text{risk}}\mathcal{L}_{\text{risk}}(g, f_i, f_d) + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}}(g, \mathcal{C}) \tag{6}$$

At a high level, the first term $\mathcal{L}_{\text{risk}}(g, f_i, f_d)$ in the above optimization formulation stems from the minimization of $I(Y; D \mid Z)$, and the second term $\mathcal{L}_{\text{rep}}(g, \mathcal{C})$ is designed to minimize $I(D; Z)$.

5 Experiments

To empirically corroborate the effectiveness of LIRR, in this section we conduct experiments on both classification and regression tasks under the setting of Semi-DA and compare LIRR to existing methods. We first introduce the experimental settings, and then present analysis to the experimental results. We also provide ablation study for the experiments on both classification and regression tasks. More experimental settings, implementation details, and results are discussed in the Appendix.

5.1 Image Classification

Datasets To verify the effectiveness of LIRR on image classification problems, we conduct experiments on NICO (He et al., 2020), VisDA2017 (Peng et al., 2017), OfficeHome (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019) datasets. **NICO** is dedicatedly designed for **O.O.D.** (out-of-distribution) image classification. It has two superclasses *animal* and *vehicle*, and each superclass contains different environments[1], *e.g.* bear on grass or snow. **VisDA2017** contains Train (T) domain and Validation (V) domain with 12 classes in each domain. **Office-Home** includes four domains: RealWorld (RW), Clipart (C), Art (A), and Product (P), with 65 classes in each domain. **DomainNet** is the largest domain adaptation dataset for image classification with over 600k images from 6 domains: Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R), and Sketch (S), with 345 classes in each domain. For each dataset, we randomly pick source-target pairs for evaluation. To meet the setting of Semi-DA, we randomly select a small ratio (1% or 5%) of the target data as labeled target samples for training.

[1]For *animal*, we sample 8 classes from environments *grass* and *snow* as two domains. For *vehicle*, we sample 7 classes from environments *sunset* and *beach* as two domains.

Table 1: Accuracy (%) comparison (higher means better) on **NICO**, **OfficeHome**, **DomainNet**, and **VisDA2017** 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			Domainnet			VisDA2017
Method	Grass to Snow	Snow to Grass	Sunset to Beach	Beach to Sunset	Art to Real	Real to Prod.	Prod. to Clip.	Real to Clip.	Sketch to Real	Clip. to Sketch	Train to Val.
S+T	70.06±2.14	80.08±1.21	71.37±1.54	70.07±1.28	69.20±0.15	74.63±0.13	48.65±0.12	48.37±0.08	57.44±0.07	44.16±0.05	76.17±0.15
DANN	83.80±1.73	81.57±1.51	72.69±1.35	72.03±1.05	72.20±0.23	78.13±0.26	52.47±0.21	51.53±0.19	60.23±0.15	46.36±0.15	78.91±0.25
CDAN	82.33±0.59	78.25±0.74	75.53±0.55	74.31±0.47	72.98±0.33	79.15±0.31	53.80±0.33	50.67±0.25	60.53±0.23	44.66±0.22	80.23±0.41
ADR	73.06±1.20	76.74±0.89	72.85±0.95	69.47±0.81	70.55±0.27	76.62±0.28	49.47±0.31	49.94±0.21	59.63±0.22	44.73±0.21	80.40±0.36
IRM	78.55±0.34	78.27±0.51	64.58±2.41	69.10±2.36	71.13±0.25	77.60±0.24	51.53±0.21	51.86±0.13	58.04±0.12	46.96±0.15	80.79±0.27
MME	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	51.04±0.12	60.35±0.12	45.09±0.14	80.52±0.35
LIRR	86.80±0.61	84.78±0.53	71.85±0.58	72.04±0.75	73.12±0.19	79.58±0.22	**54.33**±0.24	52.39±0.15	61.20±0.10	47.31±0.11	81.67±0.22
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	**53.42**±0.09	**61.79**±0.11	**47.83**±0.10	**82.31**±0.21
Full T	94.52±0.74	97.98±0.23	99.80±0.87	97.64±0.96	83.67±0.12	91.42±0.05	78.27±0.23	72.40±0.05	77.11±0.07	62.66±0.07	89.56±0.14

5% labeled target	NICO Animal		NICO Traffic		OfficeHome			Domainnet			VisDA2017
Method	Grass to Snow	Snow to Grass	Sunset to Beach	Beach to Sunset	Art to Real	Real to Prod.	Prod. to Clip.	Real to Clip.	Sketch to Real	Clip. to Sketch	Train to Val.
S+T	75.83±1.89	83.38±1.23	86.45±1.08	86.13±0.87	72.10±0.13	78.84±0.12	54.51±0.10	59.80±0.13	66.14±0.11	51.71±0.09	82.87±0.12
DANN	76.13±0.73	84.61±1.21	84.13±1.20	87.50±1.09	75.47±0.22	80.41±0.21	59.37±0.20	61.31±0.14	68.21±0.20	52.78±0.22	83.95±0.10
CDAN	82.33±0.59	83.08±2.13	86.97±0.47	87.50±0.56	74.92±0.29	80.57±0.33	59.14±0.31	62.18±0.22	68.49±0.19	53.77±0.21	83.31±0.32
ADR	80.36±0.31	80.97±0.98	84.50±0.91	75.29±0.87	75.47±0.27	79.27±0.26	58.24±0.27	61.22±0.38	67.96±0.37	53.19±0.32	83.57±0.43
IRM	81.57±1.01	84.29±1.10	85.71±2.20	83.61±2.17	74.71±0.21	79.67±0.25	58.98±0.22	60.69±0.30	67.81±0.28	52.31±0.25	82.62±0.29
MME	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	62.31±0.11	69.02±0.18	53.88±0.14	84.12±0.22
LIRR	85.90±0.98	85.24±0.73	90.77±0.42	88.90±0.39	76.14±0.18	**83.64**±0.21	62.61±0.17	62.74±0.21	69.35±0.13	54.05±0.17	84.47±0.19
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	**63.03**±0.17	**69.52**±0.09	**54.44**±0.12	**85.06**±0.17
Full T	94.52±0.74	97.98±0.23	99.80±0.87	97.64±0.96	83.67±0.12	91.42±0.05	78.27±0.23	72.40±0.05	77.11±0.07	62.66±0.07	89.56±0.14

Table 2: Mean absolute error (MAE, lower means better) comparison on **Citycam** with 1% and 5% labeled target data (mean ± std). The best is emphasized in bold.

Method	253 to 398		170 to 398		511 to 398	
	1%	5%	1%	5%	1%	5%
S+T	3.20±0.03	2.42±0.02	3.12±0.02	2.07±0.01	3.45±0.02	2.82±0.04
ADDA	3.13±0.01	2.34±0.03	3.05±0.03	2.05±0.01	2.87±0.03	2.45±0.02
DANN	3.08±0.02	2.38±0.02	3.01±0.04	2.01±0.02	2.95±0.03	2.41±0.04
IRM	3.11±0.02	2.27±0.03	2.91±0.02	2.02±0.01	2.89±0.05	2.33±0.03
LIRR	**2.96**±0.02	**2.13**±0.01	**2.84**±0.01	**1.98**±0.02	**2.80**±0.03	**2.25**±0.01
Full T	1.68±0.01	1.68±0.01	1.68±0.01	1.68±0.01	1.68±0.01	1.68±0.01

Baselines We compare our approach with the following representative domain adaptation methods: **DANN** (Ganin et al., 2016), **CDAN** (Long et al., 2018), **IRM** (Arjovsky et al., 2019), **ADR** (Saito et al., 2017), and **MME** (Saito et al., 2019); **S+T**, a model trained with the labeled source and the few labeled target samples without using unlabeled target samples; and **Full T**, a model trained with the fully labeled target. All these methods are implemented and evaluated under the Semi-DA setting.

5.2 Traffic Counting Regression

Datasets To verify the effectiveness of LIRR on regression problems, we conduct experiments on WebCamT dataset (Zhang et al., 2017) for the Traffic Counting Regression task. WebCamT has 60,000 traffic video frames annotated with vehicle bounding boxes and counts, collected from 16 surveillance cameras with different locations and recording time. We pick three source-target pairs with different visual similarities: 253→398, 170→398, 511→398 (digit denotes camera ID).

Baselines The baseline models for this task are generally aligned with our classification experiments except the methods that can not be applied to the regression task (e.g. **MME**, **ADR**, and **CDAN**). Thus, for the

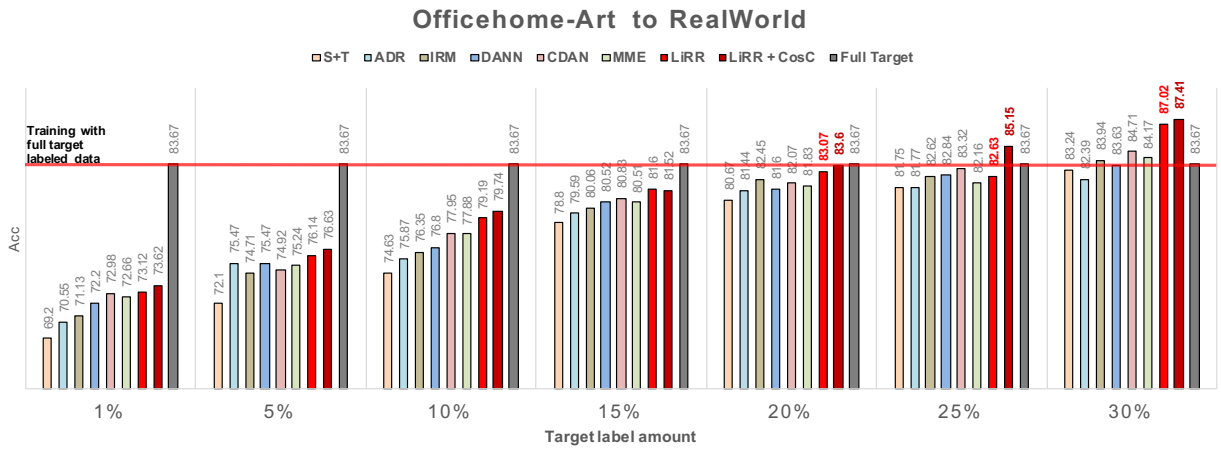

Figure 2: Performance comparison with increasing number of labeled target data, from Domain Art to RealWorld on Officehome dataset. X axis: the ratio of labeled target data; Y axis: accuracy.

traffic counting regression task, we compare with the baseline methods: **ADDA** (Tzeng et al., 2017), **DANN**, **IRM**, **S+T**, and **FullT**.

5.3 Experimental Results Analysis

Classification Tasks The classification results are shown in Table 1 with 1% and 5% labeled target data. LIRR outperforms the baselines on all the five adaptation datasets, which consistently indicates its effectiveness. As our learning objective suggests, LIRR can be viewed as achieving $D \perp (Y, Z)$, which combines the benefits of achieving $D \perp Y$ and $D \perp Y \mid Z$. In contrast, DANN, CDAN, and ADDA can be viewed as only achieving $D \perp Z$ or its variant form; and IRM can be viewed as an approximation to achieve $D \perp Y \mid Z$ using gradient penalty. LIRR outperforms all these methods on different datasets with 1% or 5% labeled target data, demonstrating simultaneously learning invariant representations and risks achieves better generalization for domain adaptation than only learning one of them. Such results are consistent with our theoretical analysis and algorithm design objective. Besides, when applying LIRR along with the cosine classifier (**CosC**) module, which is also used in MME, the performance further outperforms MME by a larger margin.

Regression Tasks The traffic counting regression results are shown in Table 2 with 1% and 5% labeled target data. The superiority of LIRR over baseline methods is supported by its lowest MAE on all the settings. DANN and ADDA are the representative methods of learning invariant representations, while IRM is the representative method of learning invariant risks. Both DANN, ADDA, and IRM achieve lower error than S+T, which means learning invariant representations or invariant risks can benefit Semi-DA to some extent on the regression task. Similar with the observations from the classification experiments, LIRR outperforms both DANN, ADDA, and IRM, demonstrating simultaneously learning invariant representations and risks achieves better adaptation than only aligning one of them.

5.4 Ablation Study

Comparisons with Optimizing Single Invariant Objective As pointed out in Sec. 5.3, LIRR is simultaneously learning invariant representations and risks, while DANN, CDAN, ADDA can be viewed as only achieving invariant representations or its variant forms, and IRM is an approximation to solely achieve invariant risks. From the results on both classification and regression tasks, we can further acknowledge the importance of simultaneously optimizing these two invariant items together. As shown in Table 1 and 2, all the methods that only minimize one single invariant objective perform worse than LIRR, indicating our method is effective and consistent to the theoretical results.

Increasing Proportions of Labeled Target Data Revisiting Theorem 3.1 and Theorem 3.2, we know that as the proportion of the labeled target data rises, the upper bound of $\epsilon_T(h)$ gets tighter. Accordingly, the margin between LIRR and other methods becomes larger, as shown in Fig. 2. Another riveting observation from Fig. 2 is, LIRR and its variant LIRR+CosC even achieve better performance than the oracle by large margin with 25% or 30% labeled target data. Stunning but plausible, with source and a few labeled target data, LIRR can learn more robust and generalized representations and achieve better performance on the target, comparing with the model trained by the fully labeled target data.

Cosine Classifier As introduced in Saito et al. (2019), cosine classifier is proved to be helpful for improving the model's performance on Semi-DA. As shown in Table. 1, the same phenomenon can be found when comparing the performance of LIRR and LIRR+CosC. For almost all the cases, LIRR plus cosine classifier module achieves higher accuracy than LIRR alone.

6 Related Work

Domain Adaptation Most existing research on domain adaptation focuses on the unsupervised setting, *i.e.* the data from target domain are fully unlabeled. Recent deep unsupervised domain adaptation (UDA) methods usually employ a conjoined architecture with two streams to represent the models for the source and target domains, respectively (Zhuo et al., 2017). Besides the task loss on the labeled source domain, another alignment loss is designed to align the source and target domains, such as discrepancy loss (Long et al., 2015; Sun et al., 2016; Zhuo et al., 2017; Adel et al., 2017; Kang et al., 2019; Chen et al., 2020), adversarial loss (Bousmalis et al., 2017; Tzeng et al., 2017; Shrivastava et al., 2017; Russo et al., 2018; Zhao et al., 2019b), and self-supervision loss (Ghifary et al., 2015, 2016; Bousmalis et al., 2016; Carlucci et al., 2019; Feng et al., 2019; Kim et al., 2020; Mei et al., 2020). Semi-DA deals with the domain adaptation problem where some target labels are available (Donahue et al., 2013; Li et al., 2014; Yao et al., 2015; Ao et al., 2017). Saito et al. (2019) empirically observed that UDA methods often fail in improving accuracy in Semi-DA and proposed a min-max Entropy approach that adversarially optimizes an adaptive few-shot model. Different from these works, our proposed method aims to align *both* the marginal feature distributions as well as the conditional distributions of the label over the features, which can arguably overcome the limitations that exist in UDA methods that only align feature distributions (Zhao et al., 2019a).

Invariant Risk Minimization In a seminal work, Arjovsky et al. (2019) consider the question that data are collected from multiple envrionments with different distributions where spurious correlations are due to dataset biases. This part of spurious correlation will confuse model to build predictions on unrelated correlations (Lake et al., 2017; Janzing & Scholkopf, 2010; Schölkopf et al., 2012) rather than true causal

relations. IRM (Arjovsky et al., 2019) estimates invariant and causal variables from multiple environments by regularizing on predictors to find data represenation matching for all environments. Chang et al. (2020) extends IRM to neural predictions and employ the environment aware predictor to learn a rationale feature encoder. As a comparison, in this work we provably show that IRM is not sufficient to ensure reduced accuracy discrepancy across domains, and we propose to align the marginal features as well simultaneously.

Transferability Transferability of deep networks has been researched in the field of transfer learning (Yosinski et al., 2014), which is normally performed by taking a standard neural architecture along with its pretrained weights on large-scale datasets such as ImageNet, and then fine-tuning the weights on the target task. This method offers little benefit to large-scale tasks but greatly improve the expressive ability of the model on small data sets with light weighted model (Raghu et al., 2019). Existing work (Yosinski et al., 2014) shows an decreasing trend of transferability when going deeper into the deep network. This phenomenon has also been applied in applications such as (Long et al., 2015), which adapts the network to the target domain with multiple layers within the backbone network. Some works (Li et al., 2019; Liu et al., 2019b,a) in few-shot learning also utilize features of multiple appended layers to handle the hierarchy of classes.

7 Conclusion

In this paper, we argue that, compared with UDA, the setting of Semi-DA is more realistic and enjoys broader practical applications with potentially better utility. To this end, in this paper we propose the first finite-sample generalization bounds for both classification and regression problems under Semi-DA. Our results shed new light on Semi-DA by suggesting a principled way of simultaneously learning invariant representations and risks across domains, leading to a bound minimization algorithm - LIRR. Extensive experiments on real-world datasets, including both image classification and traffic counting tasks, demonstrate the effectiveness of LIRR as well as its consistency to our theoretical results. We believe our work takes an important step towards more robust supervised learning methods that resist potential distributional shift between model training and model deployment.

References

Tameem Adel, Han Zhao, and Alexander Wong. Unsupervised domain adaptation with a relaxed covariate shift assumption. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Shuang Ao, Xiang Li, and Charles X Ling. Fast generalized distillation for semi-supervised domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2017.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pp. 129–136, 2008.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2016.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3722–3731, 2017.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S Jaakkola. Invariant rationalization. *arXiv preprint arXiv:2003.09772*, 2020.

Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. *arXiv preprint arXiv:1912.11976*, 2019.

Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2020.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 668–675, 2013.

Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*, pp. 4240–4248, 2016.

Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning from multi-domain data. In *IEEE International Conference on Computer Vision*, pp. 3245–3255, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559, 2015.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 597–613, 2016.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, pp. 9871–9881, 2019.

Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, pp. 107383, 2020.

Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels. *arXiv:2003.08264*, 2020.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *CVPR*, pp. 7212–7220, 2019.

Limin Li and Zhenyue Zhang. Semi-supervised domain adaptation by covariance matching. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2724–2739, 2018.

Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1134–1148, 2014.

Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Prototype propagation networks (ppn) for weakly-supervised few-shot learning on category graph. *arXiv preprint arXiv:1905.04042*, 2019a.

Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Learning to propagate for graph meta-learning. In *Advances in Neural Information Processing Systems*, pp. 1037–1048, 2019b.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. *arXiv preprint arXiv:2008.12197*, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018. ISBN 0262039400.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pp. 3342–3352, 2019.

Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8099–8108, 2018.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *IEEE International Conference on Computer Vision*, pp. 8050–8058, 2019.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 459–466, 2012.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Annals of the History of Computing*, (04):640–651, 2017.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2107–2116, 2017.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*, pp. 2058–2065, 2016.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. *arXiv preprint arXiv:1903.01689*, 2019.

Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2142–2150, 2015.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

Shanghang Zhang, Guanhang Wu, Joao P Costeira, and Jose MF Moura. Understanding traffic density from large-scale web camera data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5898–5907, 2017.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Advances in neural information processing systems*, pp. 8559–8570, 2018.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532, 2019a.

Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pp. 7285–7298, 2019b.

Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *ACM International Conference on Multimedia*, pp. 261–269, 2017.

Appendix A Omitted Proofs

In this section, we provide a detailed proof of Theorem 3.1 and Theorem 3.2 in sequence.

A.1 Proof of Classification Bound

Before we reach the proof to the main theorem, we first introduce and prove the following lemmas, which will be used in proving the main theorem:

Lemma A.1. [Blitzer et al. (2008)] Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0, 1\}\}$. Then for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$, we have

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

Lemma A.2. Let $\mathcal{H} := \{h : \mathcal{Z} \to \{0, 1\}\}$ be a hypothesis class over $\mathcal{Z}$ with $VCdim(\mathcal{H}) = d$. Define the noises on the source and target domains 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 $\forall h \in \mathcal{H}$ and for any distributions $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$, 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,

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

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

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

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 A.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:

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

Combine all the inequalities above, we know that:

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

■

Lemma A.3. [Mohri et al. (2018), Corollary 3.19] Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0, 1\}\}$, where $VCdim(\mathcal{H}) = d$. Then $\forall h \in \mathcal{H}, \forall 0 < \delta < 1$, w.p. at least $1 - \delta$ over the choice of a sample size m, 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 A.4. Let $h \in \mathcal{H} := \{h : \mathcal{Z} \to \{0, 1\}\}$, where $VCdim(\mathcal{H}) = d$. Then $\forall h \in \mathcal{H}, \forall 0 < \delta < 1$, w.p. at least $1 - \delta$ over the choice of a sample size n, 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 A.2, we have w.p.b at least $1 - \delta$:

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

■

With the above tools, now we are ready to prove Theorem 3.1:

Theorem 3.1. (Classification generalization bound in Semi-DA). Let $\mathcal{H}$ be a hypothesis set with functions $h : \mathcal{Z} \to \{0, 1\}$ and $VCdim(\mathcal{H}) = d$. For $0 < \delta < 1$, then w.p. at least $1 - \delta$ over the n samples in S and m samples in $\widetilde{T}$, for all $h \in \mathcal{H}$, we have:

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

Proof. With Lemma A.3, A.4, we can use a union bound to combine them with coefficients $m/(n+m)$ and $n/(n+m)$ respectively:

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

From Cauchy-Schwartz inequality, we obtain

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

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

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

■

A.2 Proof of Regression Bound

For regression generalization bound, we follow the proof strategy in the 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 introduce and prove the following lemmas that will be used in proving the main theorem:

Lemma A.5. (Zhao et al., 2018) For $h,h'\in\mathcal{H}:=\{h:\mathcal{Z}\to[0,1]\}$ with $Pdim(\mathcal{H})=d$, and for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$, the following inequality holds:

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

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

Lemma A.6. For $h, h' \in \mathcal{H} := \{h : \mathcal{Z} \to [0,1]\}$ with $Pdim(\mathcal{H}) = d$, and for any distribution $\mathcal{D}_S(Z)$, $\mathcal{D}_T(Z)$ over $\mathcal{Z}$, we define $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \le t \le 1\}$. Let $f_S(f_T) : \mathcal{Z} \to \mathbb{R}$ be the conditional median function over $\mathcal{D}_S(Z)(\mathcal{D}_T(Z))$, then $\forall h \in \mathcal{H}$, the following inequality holds:

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

Proof. The proof of this lemma is completely symmetric to the one of Lemma A.2 except that in the regression setting we use Lemma A.5 instead of Lemma A.1 as we did for classification problems, so we omit it here. ■

Lemma A.7 (Theorem 11.8 Mohri et al. (2018))**.** Let $\mathcal{H}$ be the set of real-valued function from $\mathcal{Z}$ to $[0,1]$. Assume that $Pdim(\mathcal{H}) = d$. Then $\forall h \in \mathcal{H}, \forall 0 < \delta < 1$, with probability at least $1-\delta$ over the choice of a sample size m, the following inequality holds:

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

Lemma A.8. Let $\mathcal{H}$ be a set of real-valued functions from $\mathcal{Z}$ to $[0,1]$ with $Pdim(\mathcal{H}) = d$, and $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \le t \le 1\}$. For $0 < \delta < 1$, then w.p. at least $1-\delta$ over the draw of samples S and T, for all $h \in \mathcal{H}$, we have:

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

Proof. Invoking the upper bound in A.6 and A.7, we have w.p. at least $1-\delta$:

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

■

Now we proceed to prove Theorem 3.2:

Theorem 3.2. (Regression generalization bound in Semi-DA). Let $\mathcal{H}$ be a hypothesis set with functions $h : \mathcal{Z} \to [0,1]$ and $Pdim(\mathcal{H}) = d$. Then we define $\tilde{\mathcal{H}} := \{\mathbb{I}_{|h(x)-h'(x)|>t} : h, h' \in \mathcal{H}, 0 \le t \le 1\}$. For $0 < \delta < 1$, then w.p. at least $1-\delta$ over the n samples in S and m samples in $\widetilde{T}$, for all $h \in \mathcal{H}$, we have:

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

Proof. The proof is analogous to the one we have in proving Theorem 3.1: with A.5, A.6, A.7, A.8, we can use a union bound to combine them with coefficients $m/(n+m)$ and $n/(n+m)$, respectively. We then 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 Theorem 3.1. ■

Appendix B More Experimental Results

B.1 Hyper-parameters

There are two fundamental parts 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. 5.1. The results can be found in Table. 3. From which, we can see that the optimal performance is achieved when $\lambda_{\text{risk}} = 0.1$ and $\lambda_{\text{rep}} = 0.01$.

Table 3: 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

B.2 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 in Eq. (6), 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 Saito et al. (2019), 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 (Shelhamer et al., 2017) 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.

B.3 Visualization

Grad-CAM Results on NICO dataset In order to vividly showcase the learned feature representation which supports the invariant risks across domains. We employ Grad-CAM (Selvaraju et al., 2017) to visualize the most influential part in prediction in Fig 3.

Traffic Counting Examples Visualization Fig. 4 visualizes the counting results of different algorithms on Camera 511 to 398 scenario, WebCamT. The red line represents the LIRR method we proposed while the black line represents the gt count. It's rather clear to see that LIRR have a better ability of cross domain regression fitting than other methods, especially the area within the green bounding box with dot lines.

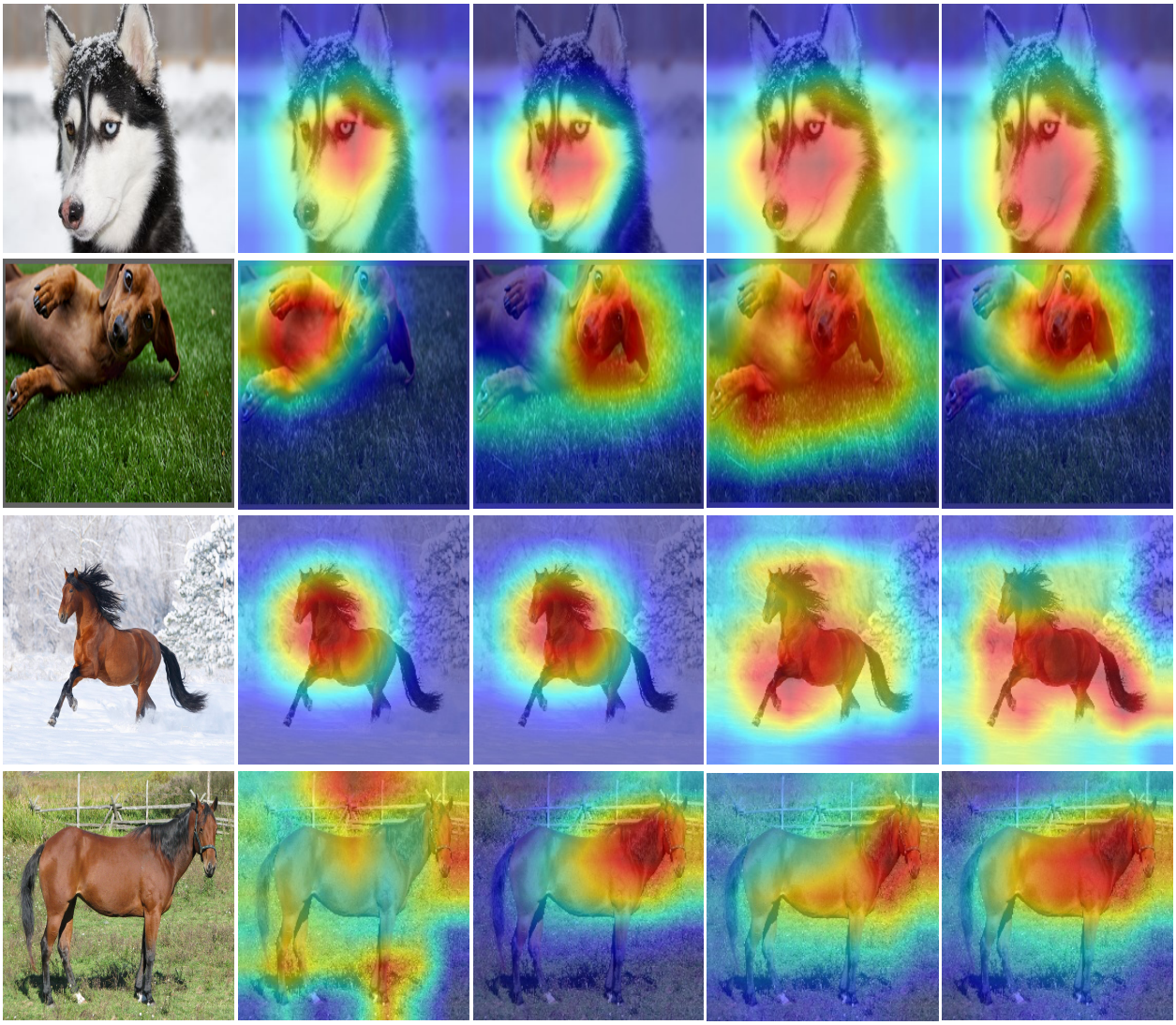

Figure 3: Grad-CAM (Selvaraju et al., 2017) results of different model. LIRR appropriately captures the invariant part of the same object in different domains, *e.g.* the shape of horse and husky leads to invariant prediction across snow and grass domain.

Camera 511

Camera 398

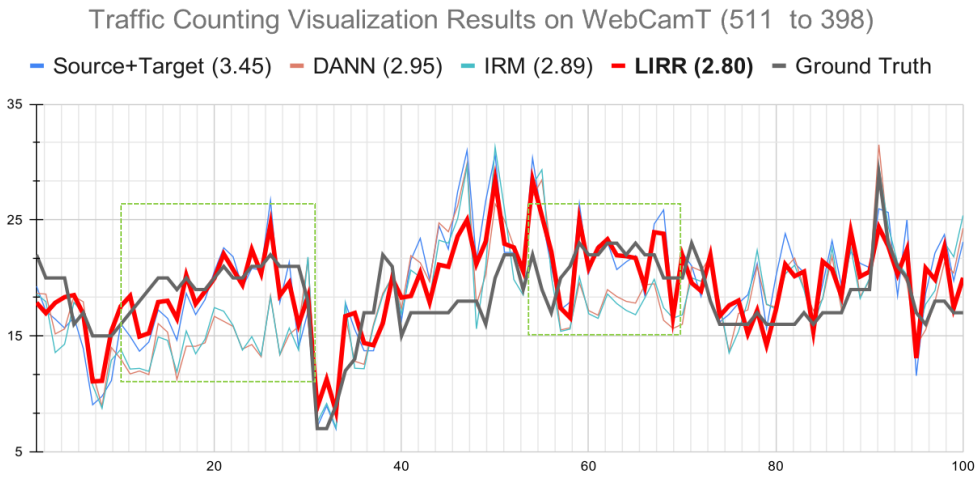

Figure 4: The line chart of the regression results of different DA methods on Camera 511 to 398, WebCamT.