Information-theoretic regularization for Multi-source Domain Adaptation

Geon Yeong Park Sang Wan Lee
KAIST
Daejeon, South Korea
{pky3436, sangwan}@kaist.ac.kr

Abstract

Adversarial learning strategy has demonstrated remarkable performance in dealing with single-source Domain Adaptation (DA) problems, and it has recently been applied to Multi-source DA (MDA) problems. Although most existing MDA strategies rely on a multiple domain discriminator setting, its effect on the latent space representations has been poorly understood. Here we adopt an information-theoretic approach to identify and resolve the potential adverse effect of the multiple domain discriminators on MDA: disintegration of domain-discriminative information, limited computational scalability, and a large variance in the gradient of the loss during training. We examine the above issues by situating adversarial DA in the context of information regularization. This also provides a theoretical justification for using a single and unified domain discriminator. Based on this idea, we implement a novel neural architecture called a Multi-source Information-regularized Adaptation Networks (MIAN). Large-scale experiments demonstrate that MIAN, despite its structural simplicity, reliably and significantly outperforms other state-of-the-art methods.

1. Introduction

Although a large number of studies have demonstrated the ability of deep learning to solve challenging tasks, the problems are mostly confined to a similar type or a single domain. One remaining challenge is the problem known as domain shift [15], where a direct transfer of information gleaned from a single source domain to unseen target domains may lead to significant performance impairment. Domain adaptation (DA) approaches aim to mitigate this problem by learning to map data of both domains onto a common feature space. Whereas several theoretical results [3, 45] and algorithms for DA [23, 25, 11] have focused on the case in which only a single-source domain dataset is given, we consider a more challenging and generalized problem of knowledge transfer, referred to as Multi-source unsupervised DA (MDA). Following a seminal theoretical result on MDA [2], many deep MDA approaches have been proposed, mainly depend on the adversarial framework.

Most of existing adversarial MDA works [44, 46, 21, 48, 47, 43] have focused on approximating all combinations of pairwise domain discrepancy between each source and the target, which inevitably necessitates training of multiple binary domain discriminators. While substantial technical advances have been made in this regard, the pitfalls of using multiple domain discriminators have not been fully studied. This paper focuses on the potential adverse effects of using multiple domain discriminators on MDA in terms of both quantity and quality. First, the domain-discriminative information is inevitably distributed across multiple discriminators. For example, such discriminators primarily focus on domain shift between each source and the target, while the discrepancies between the source domains are neglected. Moreover, the multiple source-target discriminator setting often makes it difficult to approximate the combined $\mathcal{H}$-divergence between mixture of sources and the target domain because each discriminator is deemed to utilize the samples only from the corresponding source and target domain as inputs. Compared to a bound using combined divergence, a bound based on pairwise divergence is not sufficiently flexible to accommodate domain structures [2]. Second, the computational load of the multiple domain discriminator setting rapidly increases with the number of source domains ($\mathcal{O}(N)$), which significantly limits scalability. Third, it could undermine the stability of training, as earlier works solve multiple adversarial min-max problems.

To overcome such limitations, instead of relying on multiple pairwise domain discrepancy, we constrain the mutual information between latent representations and domain labels. The contribution of this study is summarized as follows. First, we show that such mutual information regularization is closely related to the explicit optimization of the $\mathcal{H}$-divergence between the source and target domains. This affords the theoretical insight that the conventional adversarial DA can be translated into an information-theoretic regularization problem. Second, from these theoretical findings we derive a new optimization problem for MDA: minimizing adversarial loss over multiple domains with a single domain discriminator. The algorithmic solution to this problem is called Multi-source Information regularized Adaptation Networks (**MIAN**). Third, we show that our single domain discriminator setting serves to penalize every pairwise combined domain discrepancy between the given domain and the mixture of the others. Moreover, by analyzing existing studies in terms of information regularization, we found another negative effect of the multiple discriminators setting: significant increase in the variance of the stochastic gradients.

Despite its structural simplicity, we demonstrated that **MIAN** works efficiently across a wide variety of MDA scenarios, including the DIGITS-Five [30], Office-31 [32], and Office-Home datasets [41]. Intriguingly, **MIAN** reliably and significantly outperformed several state-of-the-art methods, including ones that employ a domain discriminator separately for each source domain [44] and that align the moments of deep feature distribution for every pairwise domain [30].

2. Related works

Several DA methods have been used in attempt to learn domain-invariant representations. Along with the increasing use of deep neural networks, contemporary work focuses on matching deep latent representations from the source domain with those from the target domain. Several measures have been introduced to handle domain shift, such as maximum mean discrepancy (MMD) [24, 23], correlation distance [36, 37], and Wasserstein distance [7]. Recently, adversarial DA methods [11, 40, 19, 34, 33] have become mainstream approaches owing to the development of generative adversarial networks [14]. However, the above-mentioned single-source DA approaches inevitably sacrifice performance for the sake of multi-source DA.

Some MDA studies [3, 2, 28, 18] have provided the theoretical background for algorithm-level solutions. [3, 2] explore the extended upper bound of true risk on unlabeled samples from the target domain with respect to a weighted combination of multiple source domains. Following these theoretical studies, MDA studies with shallow models [9, 8, 5] as well as with deep neural networks [27, 30, 21] have been proposed. Recently, some adversarial MDA methods have also been proposed. [44] implemented a k-way domain discriminator and classifier to battle both domain and category shifts. [46] also used multiple discriminators to optimize the average case generalization bounds. [48] chose relevant source training samples for the DA by minimizing the empirical Wasserstein distance between the source and target domains. Instead of using separate encoders, domain discriminators or classifiers for each source domain as in earlier works, our approach uses unified networks, thereby improving reliability, resource-efficiency and scalability. To the best of our knowledge, this is the first study to bridge the gap between MDA and information regularization, and show that a single domain-discriminator is sufficient for the adaptation. Moreover, compared to the proposed methods without robust theoretical justifications [21, 46, 30], our analysis does not require any assumption or estimation for the domain coefficients. In our framework, the representations are distilled to be independent of the domain, thereby rendering the performance relatively insensitive to explicit weighting strategies.

3. Theoretical insights

We first introduce the notations for the MDA problem in classification. A set of source domains and the target domain are denoted by $\{D_{S_i}\}_{i=1}^N$ and D_T, respectively. Let $X_{S_i} = \left\{\mathbf{x}_{S_i}^j\right\}_{j=1}^m$ and $Y_{S_i} = \left\{\mathbf{y}_{S_i}^j\right\}_{j=1}^m$ be a set of m i.i.d. samples from D_{S_i}. Let $X_T = \left\{\mathbf{x}_T^j\right\}_{j=1}^m \sim (D_T^X)^m$ be the set of m i.i.d. samples generated from the marginal distribution D_T^X. The domain label and its probability distribution are denoted by V and $P_V(\mathbf{v})$, where $\mathbf{v} \in \mathcal{V}$ and $\mathcal{V}$ is the set of domain labels. In line with prior works [17, 12, 27, 13], domain label can be generally treated as a stochastic latent random variable in our framework. However, for simplicity, we take the empirical version of the true distributions with given samples assuming that the domain labels for all samples are known. The latent representation of the sample is given by Z, and the encoder is defined as $F : \mathcal{X} \to \mathcal{Z}$, with $\mathcal{X}$ and $\mathcal{Z}$ representing data space and latent space, respectively. Accordingly, Z_{S_i} and Z_T refer to the outputs of the encoder $F(X_{S_i})$ and $F(X_T)$, respectively. For notational simplicity, we will omit the index i from D_{S_i}, X_{S_i} and Z_{S_i} when $N = 1$. A classifier is defined as $C : \mathcal{Z} \to \mathcal{Y}$ where $\mathcal{Y}$ is the class label space.

3.1. Problem formulation

For comparison with our formulation, we recast single-source DA as a constrained optimization problem. The true risk $\epsilon_T(h)$ on unlabeled samples from the target domain is bounded above the sum of three terms [2]: (1) true risk $\epsilon_S(h)$ of hypothesis h on the source domain; (2) $\mathcal{H}$-divergence $d_{\mathcal{H}}(D_S, D_T)$ between a source and a target domain distribution; and (3) the optimal joint risk λ^*.

Theorem 1 ([2]). *Let hypothesis class $\mathcal{H}$ be a set of binary classifiers $h : \mathcal{X} \to \{0, 1\}$. Then for the given domain distributions D_S and D_T,*

$$\forall h \in \mathcal{H}, \epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}(D_S, D_T) + \lambda^*, \quad (1)$$

where $d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim D_S^X} \left[\mathbb{I}(h(\mathbf{x}) = 1)\right] - \mathop{\mathbb{E}}_{\mathbf{x} \sim D_T^X} \left[\mathbb{I}(h(\mathbf{x}) = 1)\right] \right|$ *and* $\mathbb{I}(a)$ *is an indicator function* *whose value is* 1 *if a is true, and* 0 *otherwise.*

The empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}(X_S, X_T)$ can be computed as follows [2]:

Lemma 1.

$$\hat{d}_{\mathcal{H}}(X_S, X_T) = 2\Big(1 - \min_{h \in \mathcal{H}} \Big[\frac{1}{m} \sum_{\mathbf{x} \in X_S} \mathbb{I}[h(\mathbf{x}) = 0] + \frac{1}{m} \sum_{\mathbf{x} \in X_T} \mathbb{I}[h(\mathbf{x}) = 1]\Big]\Big) \tag{2}$$

Following Lemma 1, a domain classifier $h : \mathcal{Z} \to \mathcal{V}$ can be used to compute the empirical $\mathcal{H}$-divergence. Suppose the optimal joint risk λ^* is sufficiently small as assumed in most adversarial DA studies [33, 6]. Thus, one can obtain the ideal encoder and classifier minimizing the upper bound of $\epsilon_T(h)$ by solving the following min-max problem:

$$\begin{aligned} F^*, C^* &= \arg\min_{F,C} L(F,C) + \beta \hat{d}_{\mathcal{H}}(Z_S, Z_T) \\ &= \arg\min_{F,C} \max_{h \in \mathcal{H}} L(F,C) + \\ &\frac{\beta}{m}\Big(\sum_{i:\mathbf{z}_i \in Z_S} \mathbb{I}[h(\mathbf{z}_i) = 1] + \sum_{j:\mathbf{z}_j \in Z_T} \mathbb{I}[h(\mathbf{z}_j) = 0]\Big), \end{aligned} \tag{3}$$

where $L(F, C)$ is the loss function on samples from the source domain, β is a Lagrangian multiplier, $\mathcal{V} = \{0, 1\}$ such that each source instance and target instance are labeled as 1 and 0, respectively, and h is the binary domain classifier.

3.2. Information-regularized min-max problem for MDA

Intuitively, it is not highly desirable to adapt the learned representation in the given domain to the other domains, particularly when the representation itself is not sufficiently domain-independent. This motivates us to explore ways to learn representations independent of domains. Inspired by a contemporary fair model training study [31], the mutual information between the latent representation and the domain label $I(Z; V)$ can be expressed as follows:

Theorem 2. *Let $P_Z(\mathbf{z})$ be the distribution of Z where $\mathbf{z} \in \mathcal{Z}$. Let h be a domain classifier $h : \mathcal{Z} \to \mathcal{V}$, where $\mathcal{Z}$ is the feature space and $\mathcal{V}$ is the set of domain labels. Let $h_{\mathbf{v}}(\mathbf{z})$ be a conditional probability of V where $\mathbf{v} \in \mathcal{V}$ given $Z = \mathbf{z}$, defined by h. Then the following holds:*

$$I(Z; V) = \max_{h_{\mathbf{v}}(\mathbf{z}) : \sum_{\mathbf{v} \in \mathcal{V}} h_{\mathbf{v}}(\mathbf{z}) = 1, \forall \mathbf{z}} \sum_{\mathbf{v} \in \mathcal{V}} P_V(\mathbf{v}) \mathbb{E}_{\mathbf{z} \sim P_{Z|\mathbf{v}}} \big[\log h_{\mathbf{v}}(\mathbf{z})\big] + H(V) \tag{4}$$

The detailed proof is provided in the [31] and Supplementary Material. We can derive the empirical version of Theorem 2 as follows:

$$\hat{I}(Z; V) = \max_{h_{\mathbf{v}}(\mathbf{z}) : \sum_{\mathbf{v} \in \mathcal{V}} h_{\mathbf{v}}(\mathbf{z}) = 1, \forall \mathbf{z}} \frac{1}{M} \sum_{\mathbf{v} \in \mathcal{V}} \sum_{i:\mathbf{v}_i = \mathbf{v}} \log h_{\mathbf{v}_i}(\mathbf{z}_i) + H(V), \tag{5}$$

where M is the number of total representation samples, i is the sample index, and $\mathbf{v}_i$ is the corresponding domain label of the ith sample. Using this equation, we combine our information-constrained objective function and the results of Lemma 1. For binary classification $\mathcal{V} = \{0, 1\}$ with Z_S and Z_T of equal size $M/2$, we propose the following information-regularized minimax problem:

$$\begin{aligned} F^*, C^* &= \arg\min_{F,C} \; L(F,C) + \beta \hat{I}(Z; V) \\ &= \arg\min_{F,C} \max_{h \in \mathcal{H}} \; L(F,C) + \\ &\frac{\beta}{M}\Big[\sum_{i:\mathbf{z}_i \in Z_S} \log h(\mathbf{z}_i) + \sum_{j:\mathbf{z}_j \in Z_T} \log(1 - h(\mathbf{z}_j))\Big], \end{aligned} \tag{6}$$

where β is a Lagrangian multiplier, $h(\mathbf{z}_i) \triangleq h_{\mathbf{v}_i=1}(\mathbf{z}_i)$ and $1 - h(\mathbf{z}_i) \triangleq h_{\mathbf{v}_i=0}(\mathbf{z}_i)$, with $h(\mathbf{z}_i)$ representing the probability that $\mathbf{z}_i$ belongs to the source domain. This setting automatically dismisses the condition $\sum_{\mathbf{v} \in \mathcal{V}} h_{\mathbf{v}}(\mathbf{z}) = 1, \forall \mathbf{z}$. Note that we have accommodated a simple situation in which the entropy $H(V)$ remains constant.

3.3. Advantages over other MDA methods

Integration of domain-discriminative information. The relationship between (3) and (6) provides us a theoretical insights that the problem of minimizing mutual information between the latent representation and the domain label is closely related to minimizing the $\mathcal{H}$-divergence using the adversarial learning scheme. This relationship clearly underlines the significance of information regularization for MDA. Compared to the existing MDA approaches [44, 46], which inevitably distribute domain-discriminative knowledge over N different domain classifiers, the above objective function (6) enables us to seamlessly integrate such information with the single-domain classifier h. It will be further discussed in Section 4.

Variance of the gradient. Using a single domain discriminator also helps reduce the variance of gradient. Large variances in the stochastic gradients slow down the convergence, which leads to poor performance [20]. Herein, we analyze the variances of the stochastic gradients of existing optimization constraints. By excluding the weighted source combination strategy, we can approximately express the optimization constraint of existing adversarial MDA methods as sum of the information constraints:

$$\sum_{k=1}^{N} I(Z_k;U_k) = \sum_{k=1}^{N} I_k + \sum_{k=1}^{N} H(U_k), \tag{7}$$

where

$$I_k = \max_{h^k_{\mathbf{u}}(\mathbf{z}):\sum_{\mathbf{u}\in\mathcal{U}} h^k_{\mathbf{u}}(\mathbf{z})=1,\forall\mathbf{z}} \sum_{\mathbf{u}\in\mathcal{U}} P_{U_k}(\mathbf{u})\mathbb{E}_{\mathbf{z}_k\sim P_{Z_k|\mathbf{u}}}\big[\log h^k_{\mathbf{u}}(\mathbf{z}_k)\big], \tag{8}$$

U_k is the kth domain label with $\mathcal{U} = \{0,1\}$, $P_{Z_k|\mathbf{u}=0}(\cdot) = P_{Z|\mathbf{v}=N+1}(\cdot)$ corresponding to the target domain, $P_{Z_k|\mathbf{u}=1}(\cdot) = P_{Z|\mathbf{v}=k}(\cdot)$ corresponding to the kth source domain, and $h^k_{\mathbf{u}}(\mathbf{z}_k)$ being the conditional probability of $\mathbf{u}\in\mathcal{U}$ given $\mathbf{z}_k$ defined by the kth discriminator indicating that the sample is generated from the kth source domain. Again, we treat the entropy $H(U_k)$ as a constant.

Given $M = m(N+1)$ samples with m representing the number of samples per domain, an empirical version of (7) is:

$$\sum_{k=1}^{N} \hat{I}(Z_k;U_k) = \frac{1}{M}\sum_{k=1}^{N} \hat{I}_k + \sum_{k=1}^{N} H(U_k), \tag{9}$$

where

$$\hat{I}_k = \max_{h^k_{\mathbf{u}}(\mathbf{z}):\sum_{\mathbf{u}\in\mathcal{U}} h^k_{\mathbf{u}}(\mathbf{z})=1,\forall\mathbf{z}} \sum_{\mathbf{u}\in\mathcal{U}} \sum_{i:\mathbf{u}^i=\mathbf{u}} \log h^k_{\mathbf{u}}(\mathbf{z}^i_k). \tag{10}$$

For the sake of simplicity, we make simplifying assumptions that all $Var[I_k]$ are approximately the same for all k and so are $Cov[I_k, I_j]$ for all pairs. Then the variance of (9) is given by:

$$Var\Big[\sum_{k=1}^{N} \hat{I}(Z_k;U_k)\Big] = \frac{1}{M^2}\Big(\sum_{k=1}^{N} Var[\hat{I}_k] + 2\sum_{k=1}^{N}\sum_{j=k}^{N} Cov[\hat{I}_k,\hat{I}_j]\Big) = \frac{1}{m^2}\Big(\frac{N}{(N+1)^2}Var[\hat{I}_k] + \frac{N(N-1)}{(N+1)^2}Cov[\hat{I}_k,\hat{I}_j]\Big). \tag{11}$$

As earlier works solve N adversarial minimax problems, the covariance term is additionally included and its contribution to the variance does not decrease with increasing N. In other words, the covariance term may dominate the variance of the gradients as the number of domain increases. In contrast, the variance of our constraint (5) is inversely proportional to $(N+1)^2$. Let I_m be a shorthand for the maximization term except $\frac{1}{M}$ in (5). Then the variance of (5) is given by:

$$Var\Big[\hat{I}(Z;V)\Big] = \frac{1}{m^2(N+1)^2}\Big(Var[I_m]\Big). \tag{12}$$

It implies that our framework can significantly improve the stability of stochastic gradient optimization compared to existing approaches, especially when the model is deemed to learn from many domains.

3.4. Situating domain adaptation in context of information bottleneck theory

In this Section, we bridge the gap between the existing adversarial DA method and the information bottleneck (IB) theory [38, 39, 1]. [38] examined the problem of learning an encoding Z such that it is maximally informative about the class Y while being minimally informative about the sample X:

$$\min_{P_{enc}(\mathbf{z}|\mathbf{x})} \beta I(Z;X) - I(Z;Y), \tag{13}$$

where β is a Lagrangian multiplier. Indeed, the role of the bottleneck term $I(Z;X)$ matches our mutual information $I(Z;V)$ between the latent representation and the domain label. We foster close collaboration between two information bottleneck terms by incorporating those into $I(Z;X,V)$.

Theorem 3. *Let $P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})$ be a conditional probabilistic distribution of Z where $\mathbf{z}\in\mathcal{Z}$, defined by the encoder F, given a sample $\mathbf{x}\in\mathcal{X}$ and the domain label $\mathbf{v}\in\mathcal{V}$. Let $R_Z(\mathbf{z})$ denotes a prior marginal distribution of Z. Then the following inequality holds:*

$$I(Z;X,V) \le \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\big[D_{KL}[P_{Z|\mathbf{x},\mathbf{v}} \parallel R_Z]\big] + H(V) + \max_{h_{\mathbf{v}}(\mathbf{z}):\sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z})=1,\forall\mathbf{z}} \sum_{\mathbf{v}\in\mathcal{V}} P_V(\mathbf{v})\mathbb{E}_{P_{\mathbf{z}\sim Z|\mathbf{v}}}\big[\log h_{\mathbf{v}}(\mathbf{z})\big] \tag{14}$$

The proof of Theorem 3 uses the chain rule: $I(Z;X,V) = I(Z;V) + I(Z;X \mid V)$. The detailed proof is provided in the Supplementary Material. Whereas the role of $I(Z;X \mid V)$ is to purify the latent representation generated from the given domain, $I(Z;V)$ serves as a proxy for regularization that aligns the purified representations across different domains. Thus, the existing DA approaches [26, 35] using variational information bottleneck [1] can be reviewed as special cases for Theorem 3 with a single-source domain.

4. Multi-source Information-regularized Adaptation Networks (MIAN)

In this Section, we provide the details of our proposed architecture, referred to as a multi-source information-regularized adaptation network (**MIAN**). **MIAN** addresses the information-constrained min–max problem for MDA (Section 3.2) using the three subcomponents depicted in Figure 1: information regularization, source classification, and Decaying Batch Spectral Penalization (DBSP).

Information regularization. To estimate the empirical mutual information $\hat{I}(Z;V)$ in (5), the domain classifier h

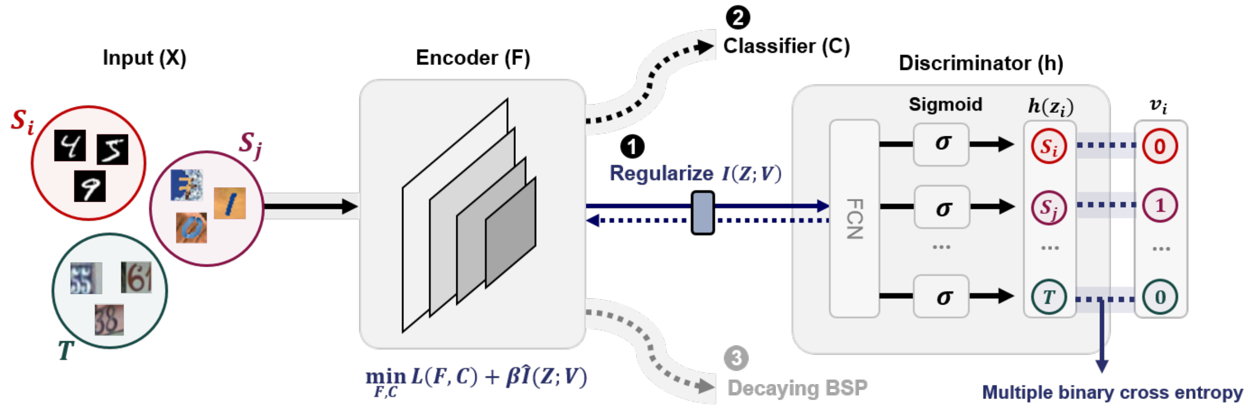


Figure 1: Proposed neural architecture for multi-source domain adaptation: Multi-source Information-regularized Adaptation Network (**MIAN**). Multi-source and target domain input data are fed into the encoder. We denote arbitrary source domains as S_i and S_j. The domain discriminator outputs a logit vector, where each dimension corresponds to each domain.

should be trained to minimize softmax cross enropy. Let $\mathcal{V} = \{1, 2, ..., N+1\}$ and denote $h(\mathbf{z})$ as $N+1$ dimensional vector of the conditional probability for each domain given the sample $\mathbf{z}$. Let $\mathbb{1}$ be a $N+1$ dimensional vector of all ones, and $\mathbb{1}_{[k=\mathbf{v}]}$ be a $N+1$ dimensional vector whose $\mathbf{v}$th value is 1 and 0 otherwise. Given $M = m(N+1)$ samples, the objective is:

$$\min_h -\frac{1}{M} \sum_{\mathbf{v}\in\mathcal{V}} \sum_{i:\mathbf{v}_i=\mathbf{v}} \left[\mathbb{1}^T_{[k=\mathbf{v}_i]} \log h(\mathbf{z}_i)\right]. \tag{15}$$

In this study, we slightly modify the softmax cross entropy (15) into multiple binary cross entropy. Specifically, we *explicitly* minimize the conditional probability of the remaining domains excepting the true $\mathbf{v}$th domain. Let $\mathbb{1}_{[k\neq\mathbf{v}]}$ be the flipped version of $\mathbb{1}_{[k=\mathbf{v}]}$. Then the modified objective function for the domain discriminator is:

$$\begin{aligned}\min_h -\frac{1}{M} \sum_{\mathbf{v}\in\mathcal{V}} \sum_{i:\mathbf{v}_i=\mathbf{v}} &\left[\mathbb{1}^T_{[k=\mathbf{v}_i]} \log h(\mathbf{z}_i)\right. \\ &\left. + \mathbb{1}^T_{[k\neq\mathbf{v}_i]} \log(\mathbb{1} - h(\mathbf{z}_i))\right],\end{aligned} \tag{16}$$

where the objective function for encoder training is to maximize (16). Our objective function is also closely related to that of GAN [14], and we experimentally found that using the variant objective function of GAN [29] works slightly better.

Herein, we show that the objective (16) is closely related to optimizing (1) an average of pairwise combined domain discrepancy between the given domain and the mixture of the others $d_{\mathcal{H}}(\mathcal{V})$, and (2) an average of every pairwise $\mathcal{H}$-divergence between each domain. Let each $D_\mathbf{v}$ and $D_{\mathbf{v}^c}$ represent the $\mathbf{v}$th domain and the mixture of the remaining N domains with the same mixture weight $\frac{1}{N}$, respectively. Then we can define $\mathcal{H}$-divergence as $d_{\mathcal{H}}(D_\mathbf{v}, D_{\mathbf{v}^c})$, and an average of such $\mathcal{H}$-divergence for every $\mathbf{v}$ as $d_{\mathcal{H}}(\mathcal{V})$. Assume that the samples of size m, $Z_\mathbf{v}$ and $Z_{\mathbf{v}^c}$, are generated from each $D_\mathbf{v}$ and $D_{\mathbf{v}^c}$, where $Z_{\mathbf{v}^c} = \bigcup_{\mathbf{v}'\neq\mathbf{v}} Z_{\mathbf{v}'}$ with $|Z_{\mathbf{v}'}| = m/N$ for all $\mathbf{v}' \in \mathcal{V}$. Thus the domain label $\mathbf{v}_j \neq \mathbf{v}$ for every jth sample in $Z_{\mathbf{v}^c}$. Then the empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}(\mathcal{V})$ is defined as follows:

$$\begin{aligned}\hat{d}_{\mathcal{H}}(\mathcal{V}) &= \frac{1}{N+1} \sum_{\mathbf{v}\in\mathcal{V}} \hat{d}_{\mathcal{H}}(Z_\mathbf{v}, Z_{\mathbf{v}^c}) \\ &= \frac{1}{N+1} \sum_{\mathbf{v}\in\mathcal{V}} 2\Big(1 - \min_{h\in\mathcal{H}} \Big[\frac{1}{m} \sum_{i:\mathbf{v}_i=\mathbf{v}} \mathbb{I}[h_\mathbf{v}(\mathbf{z}_i) = 0] \\ &\quad + \frac{1}{m} \sum_{j:\mathbf{v}_j\neq\mathbf{v}} \mathbb{I}[h_\mathbf{v}(\mathbf{z}_j) = 1]\Big]\Big),\end{aligned} \tag{17}$$

where $\mathbb{I}[h_\mathbf{v}(\mathbf{z}) = 1]$ corresponds to the $\mathbf{v}$th value of $N+1$ dimensional one-hot classification vector $\mathbb{I}[h(\mathbf{z})]$, unlike the conditional probability vector $h(\mathbf{z})$ in (16). Given the unified domain discriminator h in the inner minimization, we train h to approximate $\hat{d}_{\mathcal{H}}(\mathcal{V})$ as follows:

$$\begin{aligned}h^* &= \arg\max_{h\in\mathcal{H}} \frac{1}{M} \sum_{\mathbf{v}\in\mathcal{V}} \Big(\sum_{i:\mathbf{v}_i=\mathbf{v}} \mathbb{I}[h_\mathbf{v}(\mathbf{z}_i) = 1] \\ &\quad + \sum_{j:\mathbf{v}_j\neq\mathbf{v}} \mathbb{I}[h_\mathbf{v}(\mathbf{z}_j) = 0]\Big) \\ &= \arg\min_{h\in\mathcal{H}} -\frac{1}{M} \sum_{\mathbf{v}\in\mathcal{V}} \sum_{i:\mathbf{v}_i=\mathbf{v}} \Big(\mathbb{1}^T_{[k=\mathbf{v}_i]} \mathbb{I}[h(\mathbf{z}_i)] \\ &\quad + \mathbb{1}^T_{[k\neq\mathbf{v}_i]} \big(\mathbb{1} - \mathbb{I}[h(\mathbf{z}_i)]\big)\Big),\end{aligned} \tag{18}$$

where the latter equality is obtained by rearranging the summation terms in the first equality.

Based on the close relationship between (16) and (18), we can make the link between information regularization and $\mathcal{H}$-divergence optimization given multi-source domain; minimizing $\hat{d}_{\mathcal{H}}(\mathcal{V})$ is closely related to implicit regularization of the mutual information between latent representations and domain labels. Because the output classification vector $\mathbb{I}[h(\mathbf{z})]$ often comes from the argmax operation, the objective in (18) is not differentiable w.r.t. $\mathbf{z}$. However, our framework has a differentiable objective for the discriminator as in (16).

There are two additional benefits of minimizing $d_{\mathcal{H}}(\mathcal{V})$. First, it includes $\mathcal{H}$-divergence between the target and a mixture of sources ($\mathbf{v} = N+1$ in (17)). Note that it directly affects the upper bound of the empirical risk on target samples (Theorem 5 in [2]). Moreover, the synergistic penalization of other divergences ($\mathbf{v} \neq N+1$ in (17)) which implicitly include the domain discrepancy between the target and other sources accelerates the adaptation. Second, $d_{\mathcal{H}}(\mathcal{V})$ lower-bounds the average of every pairwise $\mathcal{H}$-divergence between each domain:

Lemma 2. *Let* $d_{\mathcal{H}}(\mathcal{V}) = \frac{1}{N+1}\sum_{\mathbf{v}\in\mathcal{V}} d_{\mathcal{H}}(D_{\mathbf{v}}, D_{\mathbf{v}^c})$*. Let* $\mathcal{H}$ *be a hypothesis class. Then,*

$$d_{\mathcal{H}}(\mathcal{V}) \leq \frac{1}{N(N+1)} \sum_{\mathbf{v},\mathbf{u}\in\mathcal{V}} d_{\mathcal{H}}(D_{\mathbf{v}}, D_{\mathbf{u}}). \tag{19}$$

The detailed proof is provided in the appendix. It implies that not only the domain shift between each source and the target domain, but also the domain shift between each source domain can be indirectly penalized. Note that this characteristic is known to be beneficial to MDA [21, 30]. Unlike our single domain classifier setting, existing methods [21] require a number of about $\mathcal{O}(N^2)$ domain classifiers to approximate all pairwise combinations of domain discrepancy. In this regard, there is no comparison between the proposed method using a single domain classifier and existing approaches in terms of resource efficiency.

Source classification. Along with learning domain-independent latent representations illustrated in the above, we train the classifier with the labeled source domain datasets. To minimize the empirical risk on source domain, we use a generic softmax cross-entropy loss function with labeled source domain samples as $L(F, C)$.

Decaying batch spectral penalization. Applying above information-theoretic insights, we further describe a potential side effect of existing adversarial DA methods. Information regularization may lead to overriding implicit entropy minimization, particularly in the early stages of the training, impairing the richness of latent feature representations. To prevent such a pathological phenomenon, we introduce a new technique called Decaying Batch Spectral Penalization (DBSP), which is intended to control the SVD entropy of the feature space. Our version improves training efficiency compared to original Batch Spectral Penalization [6]. We refer to this version of our model as **MIAN-**γ. Since vanilla **MIAN** is sufficient to outperform other state-of-the-art methods (Section 5), **MIAN-**γ is further discussed in the Supplementary Material.

5. Experiments

To assess the performance of **MIAN**, we ran a large-scale simulation using the following benchmark datasets: Digits-Five, Office-31 and Office-Home. For a fair comparison, we reproduced all the other baseline results using the same backbone architecture and optimizer settings as the proposed method. For the source-only and single-source DA standards, we introduce two MDA approaches [44, 30]: (1) source-combined, i.e., all source-domains are incorporated into a single source domain; (2) single-best, i.e., the best adaptation performance on the target domain is reported. Owing to limited space, details about simulation settings, used baseline models and datasets are presented in the Supplementary Material.

5.1. Simulation results

The classification accuracy for Digits-Five, Office-31, and Office-Home are summarized in Tables 1, 2, and 3, respectively. We found that **MIAN** outperforms most of other state-of-the-art single-source and multi-source DA methods by a large margin. Note that our method demonstrated a significant improvement in difficult task transfer with high domain shift, such as MNIST-M, Amazon or Clipart, which is the key performance indicator of MDA.

5.2. Ablation study and Quantitative analyses

Design of domain discriminator. To quantify the extent to which performance improvement is achieved by unifying the domain discriminators, we compared the performances of the three different versions of **MIAN** (Figure 2a, 2b). *No LS* uses the objective function as in (16), and unlike [29]. *Multi D* employs as many discriminators as the number of source domains which is analogous to the existing approaches. For a fair comparison, all the other experimental settings are fixed. The results illustrate that all the versions with the unified discriminator reliably outperform *Multi D* in terms of both accuracy and reliability. This suggests that unification of the domain discriminators can substantially improves the task performance.

Variance of stochastic gradients. With respect to the above analysis, we compared the variance of the stochastic gradients computed with different available domain discriminators. We trained **MIAN** and *Multi D* using mini-batches of samples. After the early stages of training, we computed the gradients for the weights and biases of both the top and bottom layers of the encoder on the full training set. Figures 2c, 2d show that **MIAN** with the unified discriminator

Table 1: Accuracy (%) on Digits-Five dataset. SYNTH denotes Synthetic Digits [10]. The baseline results for the Digits-Five dataset were taken from [30].

Standards	Models	MNIST-M	MNIST	USPS	SVHN	SYNTH	Avg
Source-combined	Source Only [16]	63.70	92.30	90.71	71.51	83.44	80.33
	DAN [23]	67.87	97.50	93.49	67.80	86.93	82.72
	DANN [11]	70.81	97.90	93.47	68.50	87.37	83.61
Single-best	Source Only [16]	63.37	90.50	88.71	63.54	82.44	77.71
	DAN [23]	63.78	96.31	94.24	62.45	85.43	80.44
	DANN [11]	71.30	97.60	92.33	63.48	85.34	82.01
	JAN [25]	65.88	97.21	95.42	75.27	86.55	84.07
	ADDA [40]	71.57	97.89	92.83	75.48	86.45	84.84
	MEDA [42]	71.31	96.47	97.01	78.45	84.62	85.60
	MCD [34]	72.50	96.21	95.33	78.89	87.47	86.10
Multi-source	DCTN [44]	70.53	96.23	92.81	77.61	86.77	84.79
	M^3SDA [30]	69.76	**98.58**	95.23	78.56	87.56	86.13
	M^3SDA-β [30]	72.82	98.43	96.14	81.32	89.58	87.65
	MIAN	**84.36**	97.91	**96.49**	**88.18**	**93.23**	**92.03**

Table 2: Accuracy (%) on Office-31 dataset.

Standards	Models	Amazon	DSLR	Webcam	Avg
Single-best	Source Only [16]	55.23±0.72	95.59±1.37	87.06±1.50	79.29
	DAN [23]	64.19±0.56	**100.00±0.00**	97.45±0.44	87.21
	JAN [25]	69.57±0.27	99.80±0.00	97.4±0.26	88.92
Source-combined	Source Only [16]	60.80±2.00	92.68±0.31	86.91±2.37	80.13
	DSBN [4]	66.82±0.35	97.45±0.22	94.00±0.38	86.09
	JAN [25]	70.15±0.19	95.20±0.36	95.15±0.23	86.83
	DANN [11]	68.15±0.42	97.59±0.60	96.77±0.26	87.50
	DAN [23]	65.77±0.74	99.26±0.23	97.51±0.41	87.51
	DANN+BSP [6]	71.13±0.44	96.65±0.30	98.32±0.26	88.70
	MCD [34]	68.57±1.06	99.49±0.25	**99.30±0.38**	89.12
Multi-source	DCTN [44]	62.74±0.50	99.44±0.25	97.92±0.29	86.70
	M^3SDA [30]	67.19±0.22	99.34±0.19	98.04±0.21	88.19
	M^3SDA-β [30]	69.41±0.82	99.64±0.19	**99.30±0.31**	89.45
	MIAN	**74.65±0.48**	99.48±0.35	98.49±0.59	**90.87**
	MIAN-γ	**76.17±0.24**	99.22±0.35	98.39±0.76	**91.26**

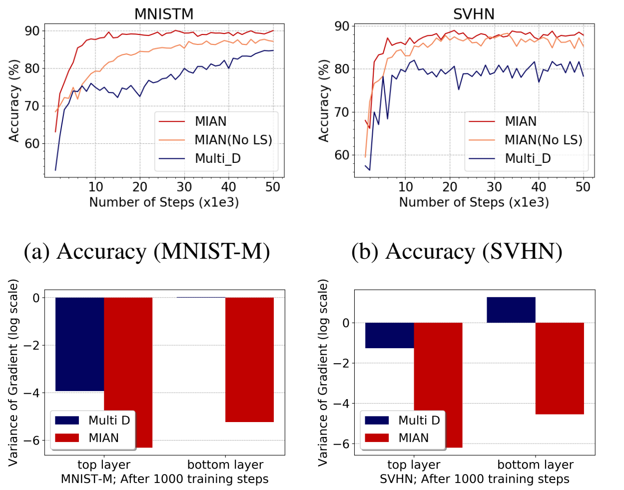
yields exponentially lower variance of the gradients compared to *Multi D*. Thus it is more feasible to use the unified discriminator when a large number of domains are given.

Proxy $\mathcal{A}$-distance. To analyze the performance improvement in depth, we measured Proxy $\mathcal{A}$-Distance (PAD) as an empirical approximation of domain discrepancy [11]. Given the generalization error ϵ on discriminating between the target and source samples, PAD is defined as $\hat{d}_{\mathcal{A}} = 2(1-2\epsilon)$. Figure 3a shows that **MIAN** yields lower PAD between the source and target domain on average, potentially associated with the modified objective of discriminator. To test this conjecture, we conducted an ablation study on the objective of domain discriminator (Figure 3b, 3c). All the other experimental settings were fixed except for using the objective of the unified domain discriminator as (15), or (16). While both cases help the adaptation, using (16) yields lower $\hat{d}_{\mathcal{H}}(\mathcal{V})$ and higher test accuracy.

Estimation of mutual information. We measure the empirical mutual information $\hat{I}(Z;V)$ with the assumption of $H(V)$ as a constant. Figure 3d shows that **MIAN** yields the lowest $\hat{I}(Z;V)$, ensuring that the obtained representation achieves low-level domain dependence. It empirically

Table 3: Accuracy (%) on Office-Home dataset.

Standards	Models	Art	Clipart	Product	Realworld	Avg
Source-combined	Source Only [16]	64.58±0.68	52.32±0.63	77.63±0.23	80.70±0.81	68.81
	DANN [11]	64.26±0.59	58.01±1.55	76.44±0.47	78.80±0.49	69.38
	DANN+BSP [6]	66.10±0.27	61.03±0.39	78.13±0.31	79.92±0.13	71.29
	DAN [23]	68.28±0.45	57.92±0.65	78.45±0.05	**81.93±0.35**	71.64
	MCD [34]	67.84±0.38	59.91±0.55	79.21±0.61	80.93±0.18	71.97
Multi-source	M^{3}SDA [30]	66.22±0.52	58.55±0.62	79.45±0.52	81.35±0.19	71.39
	DCTN [44]	66.92±0.60	61.82±0.46	79.20±0.58	77.78±0.59	71.43
	MIAN	**69.39±0.50**	**63.05±0.61**	**79.62±0.16**	80.44±0.24	**73.12**
	MIAN-γ	**69.88±0.35**	**64.20±0.68**	**80.87±0.37**	81.49±0.24	**74.11**



(a) Accuracy (MNIST-M) (b) Accuracy (SVHN)

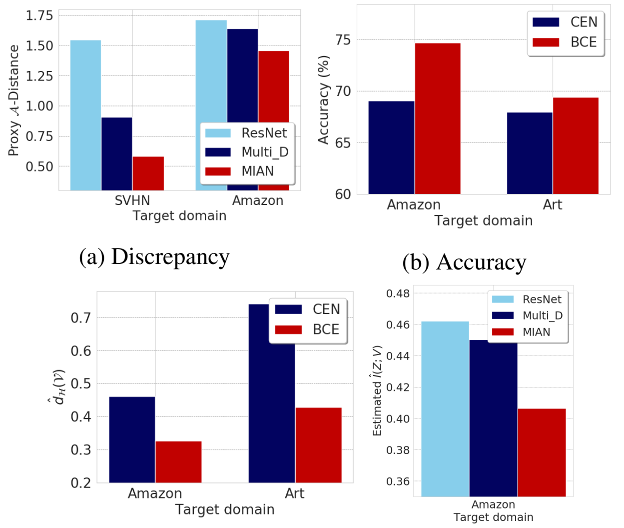
(c) Variance (MNIST-M) (d) Variance (SVHN)

Figure 2: **(a)**~**(b)**: Test accuracies for **(a)** MNIST-M and **(b)** SVHN as target domains. **(c)**~**(d)**: Variance of stochastic gradients after 1000 steps for **(c)** MNIST-M and **(d)** SVHN as target domains in log scale. Less is better.

supports the established bridge between adversarial DA and Information Bottleneck theory in section 3.4.

6. Conclusion

In this paper, we have presented a unified information-regularization framework for MDA. The proposed framework allows us to examine the existing adversarial DA methods and motivated us to implement a novel neural architecture for MDA. Specifically, we provided both theoretical arguments and empirical evidence to justify potential pitfalls of using multiple discriminators: disintegration of domain-discriminative knowledge, limited computational efficiency and high variance in the objective. The proposed model does not require complicated settings such as image generation, pretraining, or multiple networks, which are often adopted in the existing MDA methods [47, 48, 44, 46, 22].



(a) Discrepancy (b) Accuracy

(c) $\hat{d}_{\mathcal{H}}(\mathcal{V})$ (d) $\hat{I}(Z;V)$

Figure 3: **(a)** Proxy $\mathcal{A}$-distance. **(b)**~**(c)** Ablation study on the objective of domain discriminator. *CEN* stands for multi-class cross entropy loss in (15), while BCE stands for binary-class cross entropy losses in (16). **(d)** Empirical information $\hat{I}(Z;V)$. We treat $H(V) = \log|\mathcal{V}|$.

Acknowledgments

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence), National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019M3E5D2A01066267), and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-06.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 4

[2] 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. 1, 2, 3, 6

[3] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008. 1, 2

[4] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. 7

[5] Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–26, 2012. 2

[6] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019. 3, 6, 7, 8

[7] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017. 2

[8] Lixin Duan, Dong Xu, and Shih-Fu Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1338–1345. IEEE, 2012. 2

[9] Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on neural networks and learning systems*, 23(3):504–518, 2012. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 7

[11] 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. 1, 2, 7, 8

[12] Boqing Gong, Kristen Grauman, and Fei Sha. Reshaping visual datasets for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1286–1294, 2013. 2

[13] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 5

[15] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 8

[17] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *European Conference on Computer Vision*, pages 702–715. Springer, 2012. 2

[18] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8246–8256, 2018. 2

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 2

[20] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013. 3

[21] Yitong Li, David E Carlson, et al. Extracting relationships by multi-domain matching. In *Advances in Neural Information Processing Systems*, pages 6798–6809, 2018. 1, 2, 6

[22] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. Multi-source domain adaptation for visual sentiment classification. *arXiv preprint arXiv:2001.03886*, 2020. 8

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 1, 2, 7, 8

[24] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1417, 2014. 2

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017. 1, 7

[26] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6778–6787, 2019. 4

[27] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Boosting domain adaptation by discovering latent domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3771–3780, 2018. 2

[28] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in*

neural information processing systems, pages 1041–1048, 2009. 2

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 5, 6

[30] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 2, 6, 7, 8

[31] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fr-train: A mutual information-based approach to fair and robust training. *arXiv preprint arXiv:2002.10234*, 2020. 3

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2

[33] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 2, 3

[34] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2, 7, 8

[35] Yuxuan Song, Lantao Yu, Zhangjie Cao, Zhiming Zhou, Jian Shen, Shuo Shao, Weinan Zhang, and Yong Yu. Improving unsupervised domain adaptation with variational information bottleneck. *arXiv preprint arXiv:1911.09310*, 2019. 4

[36] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 2

[37] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 2

[38] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 4

[39] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015. 4

[40] 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*, pages 7167–7176, 2017. 2, 7

[41] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 2

[42] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 402–410, 2018. 7

[43] Junfeng Wen, Russell Greiner, and Dale Schuurmans. Domain aggregation networks for multi-source domain adaptation. *arXiv preprint arXiv:1909.05352*, 2019. 1

[44] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3964–3973, 2018. 1, 2, 3, 6, 7, 8

[45] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019. 1

[46] 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*, pages 8559–8570, 2018. 1, 2, 3, 8

[47] 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*, pages 7285–7298, 2019. 1, 8

[48] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. *arXiv preprint arXiv:1911.11554*, 2019. 1, 2, 8