On Learning Domain-Invariant Representations for Transfer Learning with Multiple Sources

Anonymous Author(s)
Affiliation
Address
email

Abstract

Domain adaptation (DA) benefits from the rigorous theoretical works that study its insightful characteristics and various aspects, e.g., learning domain-invariant representations and its trade-off. However, it seems not the case for the multiple source DA and domain generalization (DG) settings which are remarkably more complicated and sophisticated due to the involvement of multiple source domains and potential unavailability of target domain during training. In this paper, we develop novel upper-bounds for the target general loss which appeal us to define two kinds of domain-invariant representations. We further study the pros and cons as well as the trade-offs of enforcing learning each domain-invariant representation. Finally, we conduct experiments to inspect the trade-off of these representations for offering practical hints regarding how to use them in practice and explore other interesting properties of our developed theory.

1 Introduction

Although annotated data has been shown to be really precious and valuable to deep learning (DL), in many real-world applications, annotating a sufficient amount of data for training qualified DL models is prohibitively labour-expensive, time-consuming, and error-prone. Transfer learning is a vital solution for the lack of labeled data. Additionally, in many situations, we are only able to collect limited number of annotated data from multiple source domains. Therefore, it is desirable to train a qualified DL model primarily based on multiple source domains and possibly together with a target domain. Depending on the availability of the target domain during training, we encounter either *multiple source domain adaptation* (MSDA) or *domain generalization* (DG) problem.

Domain adaptation (DA) is a specific case of MSDA when we need to transfer from a single source domain to an another target domain available during training. For the DA setting, the pioneering work [3] and other following work [39, 48, 49, 21] are abundant to study its insightful characterizations and aspects, notably what *domain-invariant (**DI**) representations* are, how to learn this kind of representations, and the trade-off of enforcing learning DI representations. Those well-grounded studies lay the foundation for developing impressive practical DA methods [14, 43, 40, 48, 14, 7, 47].

Due to the appearance of multiple source domains and possibly unavailability of target domain, establishing theoretical foundation and characterizing *DI representations* for the MSDA and especially DG settings are significantly more challenging. A number of state-of-the-art works in MSDA and DG [52, 35, 19, 11, 37, 26, 27, 33, 15, 51, 45] have implicitly exploited and used DI representations in some sense to achieve impressive performances, e.g., [27, 27, 11] minimize representation divergence during training, [37] decomposes representation to obtain invariant feature, [51] maximizes domain prediction entropy to learn invariance. Despite the successes, the notion DI representations in MSDA and DG is not fully-understood, and rigorous studies to theoretically characterize DI representations for these settings are still very crucial and imminent. In this paper, we provide theoretical answers to

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the following questions: (1) what are DI representations in MSDA and DG, and (2) what one should expect when learning them. Overall, our contributions in this work can be summarized as:

- In Section 2.2, we first develop two upper-bounds for the general target loss in the MSDA and DG settings, whose proofs can be found in Appendix A. We then base on these bounds to characterize and define two kinds of DI representations: *general DI representations* and *compressed DI representations*.
- We further develop theory to inspect the pros and cons of two kinds of DI representations in Section 2.3, aiming to shed light on how to use those representations in practice. Particularly, two types of DI representations optimize different types of divergence on feature space, hence serving different purposes. Appendix B contains proofs regarding these characteristics.
- Finally, we study in Section 2.4 the trade-off of two kinds of DI representations which theoretically answers the question whether and how the target performance is hurt when enforcing learning each kind of representation. We refer reader to Appendix C for its proof.
- We conduct experiments to investigate the trade-off of two kinds of representations for giving practical hints regarding how to use those representations in practice as well as exploring other interesting properties of our developed theory.

To the best of our knowledge, ours is the first work that rigorously formulates and provides theoretical analysis for DI representations in MSDA and DG. Our theory is developed in a general setting of multi-class classification with a sufficiently general loss, which can be viewed as a non-trivial generalization of existing works in DA [29, 3, 49, 30, 18], each of which considers binary classification with specific loss functions. Specifically, our bounds developed in Theorems 1 and 3 interweaving both input and latent spaces are novel and benefit theoretical analyses in deep learning. Moreover, our theoretical bounds developed in Theorem 8 is the first theoretical analysis of the trade-off of learning DI representations in MSDA and DG. Particularly, by considering the MSDA setting with single source and target domains, we achieve the same trade-off nature discovered in [49], but again our setting is more general than binary classification setting in that work.

2 Our Main Theory

2.1 Notations

Let $\mathcal{X}$ be a data space on which we endow a data distribution $\mathbb{P}$ with a corresponding density function $p(x)$. We consider the multi-class classification problem with the label set $\mathcal{Y} = [C]$, where C is the number of classes and $[C] := \{1, \ldots, C\}$. Denote $\mathcal{Y}_\Delta := \left\{\alpha \in \mathbb{R}^C : \|\alpha\|_1 = 1 \wedge \alpha \geq \mathbf{0}\right\}$ as the $C-$simplex, let $f : \mathcal{X} \mapsto \mathcal{Y}_\Delta$ be a probabilistic labeling function returning a C-tuple $f(x) = [f(x,i)]_{i=1}^C$, whose element $f(x,i) = p(y = i \mid x)$ is the probability to assign a data sample $x \sim \mathbb{P}$ to the class i (i.e., $i \in \{1, ..., C\}$). Moreover, a domain is denoted compactly as pair of data distribution and labeling function $\mathbb{D} := (\mathbb{P}, f)$. We note that given a data sample $x \sim \mathbb{P}$, its categorical label $y \in \mathcal{Y}$ is sampled as $y \sim Cat(f(x))$ which a categorical distribution over $f(x) \in \mathcal{Y}_\Delta$.

Let $l : \mathcal{Y}_\Delta \times \mathcal{Y} \mapsto \mathbb{R}$ be a loss function, where $l\left(\hat{f}(x), y\right)$ with $\hat{f}(x) \in \mathcal{Y}_\Delta$ and $y \in \mathcal{Y}$ specifies the loss (e.g., cross-entropy, Hinge, L1, or L2 loss) to assign a data sample x to the class y by the hypothesis $\hat{f}$. Moreover, given a prediction probability $\hat{f}(x)$ w.r.t. the ground-truth prediction $f(x)$, we define the loss $\ell\left(\hat{f}(x), f(x)\right) = \mathbb{E}_{y \sim f(x)}\left[l\left(\hat{f}(x), y\right)\right] = \sum_{y=1}^C l\left(\hat{f}(x), y\right) f(x,y)$. We further define the general loss caused by using a classifier $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}_\Delta$ to predict $\mathbb{D} \equiv (\mathbb{P}, f)$ as

$$\mathcal{L}\left(\hat{f}, f, \mathbb{P}\right) = \mathcal{L}\left(\hat{f}, \mathbb{D}\right) := \mathbb{E}_{x \sim \mathbb{P}}\left[\ell\left(\hat{f}(x), f(x)\right)\right].$$

We inspect the multiple source setting in which we are given multiple source domains $\{\mathbb{D}^{S,i}\}_{i=1}^K$ over the common data space $\mathcal{X}$, each of which consists of data distribution and its own labeling function $\mathbb{D}^{S,i} := \left(\mathbb{P}^{S,i}, f^{S,i}\right)$. Based on these source domains, we need to work out a learner or classifier that requires to evaluate on a target domain $\mathbb{D}^T := \left(\mathbb{P}^T, f^T\right)$. Depending on the *knownness* or *unknownness* of a target domain during training, we experience either *multiple source DA* (MSDA) [30, 28, 38, 18] or *domain generalization* (DG) [23, 27, 26, 33] setting.

One typical approach in MSDA and DG is to combine the source domains together [14, 27, 26, 33, 1, 11, 42] to learn *DI representations* with hope to generalize well to a target domain. When combining the source domains, we obtain a mixture of multiple source distributions denoted as $\mathbb{D}^{\pi} = \sum_{i=1}^{K} \pi_i \mathbb{D}^{S,i}$, where the mixing coefficients $\pi = [\pi_i]_{i=1}^{C}$ can be conveniently set to $\pi_i = \frac{N_i}{\sum_{j=1}^{K} N_j}$ with N_i being the training size of the $i-$th source domain.

In term of representation learning, input is mapped into a latent space $\mathcal{Z}$ by a feature map $g : \mathcal{X} \mapsto \mathcal{Z}$ and then a classifier $\hat{h} : \mathcal{Z} \mapsto \mathcal{Y}_{\Delta}$ is trained based on the representations $g(\mathcal{X})$. Let $f : \mathcal{X} \mapsto \mathcal{Y}_{\Delta}$ be the original labeling function. To facilitate the theory developed for latent space, we introduce representation distribution being the pushed-forward distribution $\mathbb{P}_g := g_{\#}\mathbb{P}$, and the labeling function $h : \mathcal{Z} \mapsto \mathcal{Y}_{\Delta}$ induced by g as $h(z) = \frac{\int_{g^{-1}(z)} f(x)p(x)dx}{\int_{g^{-1}(z)} p(x)dx}$ [21]. Going back to our multiple source setting, the source mixture becomes $\mathbb{D}_g^{\pi} = \sum_i \pi_i \mathbb{D}_g^{S,i}$, where each source domain is $\mathbb{D}_g^{S,i} = \left(\mathbb{P}_g^{S,i}, h^{S,i}\right)$, and the target domain is $\mathbb{D}_g^{T} = \left(\mathbb{P}_g^{T}, h^{T}\right)$.

Finally, in our theory development, we use Hellinger divergence between two distributions defined as $D_{1/2}\left(\mathbb{P}^1, \mathbb{P}^2\right) = 2\int \left(\sqrt{p^1(x)} - \sqrt{p^2(x)}\right)^2 dx$, whose squared $d_{1/2} = \sqrt{D_{1/2}}$ is a proper metric.

2.2 Two Types of Domain-Invariant Representations

Hinted by the theoretical bounds developed in [3], *DI representations learning*, in which feature extractor g maps source and target data distributions to a common distribution on the latent space, is well-grounded for the DA setting. However, this task becomes significantly challenging for the MSDA and DG settings due to multiple source domains and potential unknownness of target domain. Similar to the case of DA, it is desirable to develop theoretical bounds that directly motivate definitions of DI representations for the MSDA and DG settings, as being done in the next theorem.

Theorem 1. *Consider a mixture of source domains $\mathbb{D}^{\pi} = \sum_{i=1}^{K} \pi_i \mathbb{D}^{S,i}$ and the target domain $\mathbb{D}^T$. Let ℓ be any loss function upper-bounded by a positive constant L. For any hypothesis $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}_{\Delta}$ where $\hat{f} = \hat{h} \circ g$ with $g : \mathcal{X} \to \mathcal{Z}$ and $\hat{h} : \mathcal{Z} \to \mathcal{Y}_{\Delta}$, the target loss on input space is upper bounded*

$$\mathcal{L}\left(\hat{f}, \mathbb{D}^T\right) \leq \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{f}, \mathbb{D}^{S,i}\right) + L \max_{i \in [K]} \mathbb{E}_{\mathbb{P}^{S,i}}\left[\|\Delta p^i(y|x)\|_1\right] + L\sqrt{2}\, d_{1/2}\left(\mathbb{P}_g^T, \mathbb{P}_g^{\pi}\right), \qquad (1)$$

where $\Delta p^i(y|x) := \left[\left|f^T(x,y) - f^{S,i}(x,y)\right|\right]_{y=1}^{C}$ is the absolute of single point label shift on input space between source domain $\mathbb{D}^{S,i}$, the target domain $\mathbb{D}^T$, and $[K] := \{1, 2, ..., K\}$.

The bound in (1) implies that the target loss in the input or latent space depends on three terms: (i) *the source-target data shift*: $d_{1/2}\left(\mathbb{P}_g^T, \mathbb{P}_g^{\pi}\right)$, (ii) *the label shift*: $\max_{i \in [K]} \mathbb{E}_{\mathbb{P}^{S,i}}\left[\|\Delta p^i(y|x)\|_1\right]$, and (iii) *the general source loss*: $\sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{f}, \mathbb{D}^{S,i}\right)$. To minimize the target loss in the left side, we need to minimize the three aforementioned terms. First, the *label shift* term is a natural characteristics of domains, hence almost impossible to tackle. Secondly, the *source-target data shift* term can be explicitly tackled for the MSDA setting, while almost impossible for the DG setting. Finally, the *general source loss* term is convenient to tackle, where its minimization results in a feature extractor g and a classifier $\hat{h}$. The objective of g is to map samples onto representation space in a way that the common classifier $\hat{h}$ can effectively and correctly classifies them, regardless of which domain the data comes from. This observation motivates us to formulate the first type of DI representations for the MSDA and DG settings.

Definition 2. *Consider a class $\mathcal{G}$ of feature maps and a class $\mathcal{H}$ of hypotheses. Let $\{\mathbb{D}^{S,i}\}_{i=1}^{K}$ be a set of source domains.*

*i) (**DG with unknown target data**) A feature map $g^* \in \mathcal{G}$ is said to be a **DG** general domain-invariant (**DI**) feature map if g^* is the solution of the optimization problem (OP): $\min_{g \in \mathcal{G}} \min_{\hat{h} \in \mathcal{H}} \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,i}\right)$. Moreover, the latent representations $z = g^*(x)$ induced by g^* is called general DI representations for the DG setting.*

*ii) (**MSDA with known target data**) A feature map $g^* \in \mathcal{G}$ is said to be a **MSDA** general DI feature map if g^* is the solution of the optimization problem (OP): $\min_{g\in\mathcal{G}} min_{\hat{h}\in\mathcal{H}} \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,i}\right)$ which satisfies $\mathbb{P}_{g^*}^T = \mathbb{P}_{g^*}^\pi$ (i.e., $\min_{g\in\mathcal{G}} d_{1/2}\left(\mathbb{P}_g^T, \mathbb{P}_g^\pi\right)$). Moreover, the latent representations $z = g^*(x)$ induced by g^* is called general DI representations for the MSDA setting.*

The definition of general DI representations for the DG setting in Definition 2 is transparent in light of Theorem 1, wherein we aim to find g and $\hat{h}$ to minimize the general source loss $\sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{f}, \mathbb{D}^{S,i}\right)$ due to the unknownness of $\mathbb{P}^T$. Meanwhile, regarding the general DI representations for the MSDA setting, we aim to find g^* satisfying $\mathbb{P}_{g^*}^T = \mathbb{P}_{g^*}^\pi$ and $\hat{h}$ to minimize the general source loss $\sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{f}, \mathbb{D}^{S,i}\right)$. Practically, to find general representations for MSDA, we solve

$$\min_{g\in\mathcal{G}} \min_{\hat{h}\in\mathcal{H}} \left\{ \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,i}\right) + \lambda D\left(\mathbb{P}_g^T, \mathbb{P}_g^\pi\right) \right\},$$

where $\lambda > 0$ is a trade-off parameter and D can be any divergence (e.g., Jensen Shannon (JS) divergence [16], f-divergence [34], MMD distance [26], or WS distance [38]).

In addition, the general DI representations have been exploited in some works [53, 25] from the practical perspective and previously discussed in [2] from the theoretical perspective. Despite the similar definition to our work, general DI representation in [2] is not motivated from minimization of a target loss bound.

As pointed out in the next section, learning general DI representations increases the span of latent representations $g(x)$ on the latent space which might help to reduce $d_{1/2}\left(\mathbb{P}_g^T, \mathbb{P}_g^\pi\right)$ for a general target domain. On the other hand, many other works [26, 27, 15, 45, 11] have also explored the possibility of enhancing generalization by finding common representation among source distributions. The following theorem motivates the latter, which is the second kind of DI representations.

Theorem 3. *Consider a mixture of source domains $\mathbb{D}^\pi = \sum_{i=1}^{K} \pi_i \mathbb{D}^{S,i}$ and the target domain $\mathbb{D}^T$. Let ℓ be any loss function upper-bounded by a positive constant L. For any hypothesis $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}_\Delta$ where $\hat{f} = \hat{h} \circ g$ with $g : \mathcal{X} \to \mathcal{Z}$ and $\hat{h} : \mathcal{Z} \to \mathcal{Y}_\Delta$, the target loss on input space is upper bounded*

$$\begin{aligned} \mathcal{L}\left(\hat{f}, \mathbb{D}^T\right) \leq & \sum_{i=1}^{K} \pi_i \mathcal{L}\left(f, \mathbb{D}^{S,i}\right) + L \max_{i\in[K]} \mathbb{E}_{\mathbb{P}^{S,i}}\left[\|\Delta p^i(y|x)\|_1\right] \\ & + \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{L\sqrt{2\pi_j}}{K} d_{1/2}\left(\mathbb{P}_g^T, \mathbb{P}_g^{S,i}\right) + \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{L\sqrt{2\pi_j}}{K} d_{1/2}\left(\mathbb{P}_g^{S,i}, \mathbb{P}_g^{S,j}\right). \end{aligned} \tag{2}$$

Evidently, Theorem 3 suggests another kind of representation learning, where source representation distributions are aligned in order to lower the target loss's bound as concretely defined as follows.

Definition 4. *Consider a class $\mathcal{G}$ of feature maps and a class $\mathcal{H}$ of hypotheses. Let $\{\mathbb{D}^{S,i}\}_{i=1}^K$ be a set of source domains.*

*i) (**DG with unknown target data**) A feature map $g^* \in \mathcal{G}$ is a **DG** compressed DI representations for source domains $\{\mathbb{D}^{S,i}\}_{i=1}^K$ if g^* is the solution of the optimization problem (OP): $\min_{g\in\mathcal{G}} min_{\hat{h}\in\mathcal{H}} \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,i}\right)$ which satisfies $\mathbb{P}_{g^*}^{S,1} = \mathbb{P}_{g^*}^{S,2} = \ldots = \mathbb{P}_{g^*}^{S,K}$ (i.e., the pushed forward distributions of all **source** domains are identical). The latent representations $z = g^*(x)$ is then called compressed DI representations for the DG setting.*

*ii) (**MSDA with known target data**) A feature map $g^* \in \mathcal{G}$ is an **MSDA** compressed DI representations for source domains $\{\mathbb{D}^{S,i}\}_{i=1}^K$ if g^* is the solution of the optimization problem (OP): $\min_{g\in\mathcal{G}} min_{\hat{h}\in\mathcal{H}} \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,i}\right)$ which satisfies $\mathbb{P}_{g^*}^{S,1} = \mathbb{P}_{g^*}^{S,2} = \ldots = \mathbb{P}_{g^*}^{S,K} = \mathbb{P}_{g^*}^T$ (i.e., the pushed forward distributions of all **source** and **target** domains are identical). The latent representations $z = g^*(x)$ is then called compressed DI representations for the MSDA setting.*

2.3 Characteristics of Domain-Invariant Representations

In the previous section, two kinds of DI representations are introduced, where each of them originates from minimization of different terms in the upper bound of target loss. In what follows, we examine

and discuss their benefits and drawbacks. We start with the novel development of *hypothesis-aware divergence for multiple distributions* , which is necessary for our theory developed later.

2.3.1 Hypothesis-Aware Divergence for Multiple Distributions

Let consider multiple distributions $\mathbb{Q}_1, ..., \mathbb{Q}_C$ on the same space, whose density functions are $q_1, ..., q_C$. We are given a mixture of these distributions $\mathbb{Q}^\alpha = \sum_{i=1}^C \alpha_i \mathbb{Q}_i$ with a mixing coefficient $\alpha \in \mathcal{Y}_\Delta$ and desire to measure a divergence between $\mathbb{Q}_1, ..., \mathbb{Q}_C$. One possible solution is employing a hypothesis from an infinite capacity hypothesis class $\mathcal{H}$ to identify which distribution the data come from. In particular, given $z \sim \mathbb{Q}^\alpha$, the function $\hat{h}(z, i), i \in [C]$ outputs the probability that $x \sim \mathbb{Q}_i$. Let $l\left(\hat{h}(z), i\right) \in \mathbb{R}$ be the loss when classifying z using $\hat{h}$ provided the ground-truth label $i \in [C]$. Our motivation is that if the distributions $\mathbb{Q}_1, ..., \mathbb{Q}_C$ are distant, it is easier to distinguish samples from them, hence the minimum classification loss $\min_{\hat{h}\in\mathcal{H}} \mathcal{L}^\alpha_{\mathbb{Q}_{1:C}}\left(\hat{h}\right)$ is much lower than in the case of clutching $\mathbb{Q}_1, ..., \mathbb{Q}_C$. Therefore, we develop the following theorem to connect the loss optimization problem with an f-divergence among $\mathbb{Q}_1, ..., \mathbb{Q}_C$. More discussions about hypothesis-aware divergence can be found in Appendix B of this paper and in [13].

Theorem 5. *Assuming the hypothesis class $\mathcal{H}$ has infinite capacity, we define the hypothesis-aware divergence for multiple distributions as*

$$D^\alpha(\mathbb{Q}_1, ..., \mathbb{Q}_C) = -\min_{\hat{h}\in\mathcal{H}} \mathcal{L}^\alpha_{\mathbb{Q}_{1:C}}\left(\hat{h}\right) + \inf_{\beta\in\mathcal{Y}_\Delta}\left(\sum_{i=1}^C l(\beta, i)\,\alpha_i\right). \tag{3}$$

This divergence is a proper f-divergence among $\mathbb{Q}_1, ..., \mathbb{Q}_C$ in the sense that $D^\alpha(\mathbb{Q}_1, ..., \mathbb{Q}_C) \geq 0, \forall \mathbb{Q}_1, ..., \mathbb{Q}_C$ and $\alpha \in \mathcal{Y}_\Delta$, and $D^\alpha(\mathbb{Q}_1, ..., \mathbb{Q}_C) = 0$ if $\mathbb{Q}_1 = ... = \mathbb{Q}_C$.

2.3.2 General Domain-Invariant Representation

As previously defined, the general DI feature map g^* (cf. Definition 2) is the one that minimizes the total source loss

$$g^* = \arg\min_{g\in\mathcal{G}} \min_{\hat{h}\in\mathcal{H}} \sum_{i=1}^K \pi_i \mathcal{L}\left(\hat{h}, h^{S,i}, \mathbb{P}^i_g\right) = \text{argmin}_{g\in\mathcal{G}} \text{min}_{\hat{h}\in\mathcal{H}} \sum_{i=1}^K \pi_i \mathcal{L}\left(\hat{h}, \mathbb{D}^{S,i}_g\right). \tag{4}$$

Intuitively, for each domain $\mathbb{D}^{S,i}$, the minimal loss $\min_{\hat{h}\in\mathcal{H}} \mathcal{L}\left(\hat{h}, h^{S,i}, \mathbb{P}^i_g\right)$ should be inversely proportional to the divergence between the class-conditional $\mathbb{P}^{S,i,c}_g$. For mixture of source domain, we further define $\mathbb{Q}^{S,c}_g := \sum_{i=1}^K \frac{\pi_i \gamma_{i,c}}{\alpha_c} \mathbb{P}^{S,i,c}_g$ as the mixture of the class c conditional distributions of the source domains on the latent space, where $\alpha_c = \sum_{j=1}^K \pi_j \gamma_{j,c}$ are the mixing coefficients. Then, the inner loop of the objective function in Eq. 4 can be viewed as training the optimal hypothesis $\hat{h} \in \mathcal{H}$ to distinguish the samples from $\mathbb{Q}^{s,c}_g, c \in [C]$ for a given feature map g. Therefore, by linking to the multi-divergence concept developed in Section 2.3.1, we achieve the following theorem.

Theorem 6. *Assume that $\mathcal{H}$ has infinite capacity, we have the following statements.*

1. *$D^\alpha\left(\mathbb{Q}^{s,1}_g, ..., \mathbb{Q}^{s,C}_g\right) = -\min_{\hat{h}\in\mathcal{H}} \sum_{i=1}^K \pi_i \mathcal{L}\left(\hat{h}, h^{S,i}, \mathbb{P}^{S,i}_g\right) + const$, where $\alpha = [\alpha_c]_{c\in[C]}$ is defined as above.*
2. *Finding the general DI feature map g^* via the OP in (4) is equivalent to solving*

$$g^* = \underset{g\in\mathcal{G}}{argmax}\, D^\alpha\left(\mathbb{Q}^{s,1}_g, ..., \mathbb{Q}^{s,C}_g\right). \tag{5}$$

Theorem 6, especially its second claim, discloses that learning general DI representations maximally expands the coverage of latent representations of the source domains by maximizing $D^\alpha\left(\mathbb{Q}^{s,1}_g, ..., \mathbb{Q}^{s,C}_g\right)$. Hence, the span of source mixture $\mathbb{P}^\pi_g = \sum_{c=1}^C \alpha_c \mathbb{Q}^{s,c}_g$ is also increased. We believe this demonstrates one of the benefits of general DI representations because it implicitly enhance the chance to match a general *unseen target domains* in the DG setting by possibly reducing the source-target data shift term $d_{1/2}\left(\mathbb{P}^T_g, \mathbb{P}^\pi_g\right)$ (cf. Theorem 1). Additionally, in MSDA where we wish to minimize $d_{1/2}\left(\mathbb{P}^T_g, \mathbb{P}^\pi_g\right)$ explicitly, expanding source representation's coverage is also useful.

2.3.3 Compressed Domain-Invariant Representations

It is well-known that generalization gap between the true loss and its empirical estimation affects generalization capability of model [44, 41]. One way to close this gap is increasing sampling density. However, as hinted by our analysis, enforcing general DI representation learning tends to maximize the cross-domain class divergence, hence implicitly increasing the diversity of latent representations from different source domains. This renders learning a classifier $\hat{h}$ on top of those source representations harder due to scattering samples on representation space, leading to higher generalization gap between empirical and general losses. In contrast, compressed DI representations help making the task of learning $\hat{h}$ easier with a lower generalization gap by decreasing the diversity of latent representations from the different source domains, via enforcing $\mathbb{P}_{g^*}^{S,1} = \mathbb{P}_{g^*}^{S,2} = \ldots = \mathbb{P}_{g^*}^{S,K}$. We now develop rigorous theory to examine this observation.

Let the training set be $S = \{(z_i, y_i)\}_{i=1}^{N}$, consisting of independent pairs (z_i, y_i) drawn from the source mixture, i.e., the sampling process starts with sampling domain index $k \sim Cat(\boldsymbol{\pi})$, then sampling $z \sim \mathbb{P}_g^{S,k}$ (i.e., $x \sim \mathbb{P}^{S,k}$ and $z = g(x)$), and finally assigning a label with $y \sim h^{S,k}(z)$ (i.e., $h^{S,k}(z)$ is a distribution over $[C]$). The empirical loss is defined as

$$\mathcal{L}\left(\hat{h}, S\right) = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \ell\left(\hat{h}(z_i), h^{S,k}(z_i)\right), \tag{6}$$

where $N_k,\ k \in [C]$ is the number of samples drawn from the k-th source domain and $N = \sum_{k=1}^{K} N_k$.

Here, $\mathcal{L}\left(\hat{h}, S\right)$ is an unbiased estimation of the general loss $\mathcal{L}\left(\hat{h}, \mathbb{D}_g^{\pi}\right) = \sum_{k=1}^{K} \pi_k \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,k}\right)$. To quantify the quality of the estimation, we investigate the upper bound of the *generalization gap* $\left|\mathcal{L}\left(\hat{h}, S\right) - \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{\pi}\right)\right|$ with the confidence level $1 - \delta$ for two types of DI representations. Our theoretical result indicates that the compressed DI representations offer a tighter upper bound for the generalization gap, hence minimizing $\mathcal{L}\left(\hat{h}, \mathbb{D}_g^{\pi}\right)$ more efficiently.

Theorem 7. *For any confident level $\delta \in [0, 1]$ over the choice of S, the estimation of loss is in the ϵ-range of the true loss*

$$Pr\left(\left|\mathcal{L}\left(\hat{h}, S\right) - \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{\pi}\right)\right| \leq \epsilon\right) \geq 1 - \delta,$$

where $\epsilon = \epsilon(\delta) = \left(\frac{A}{\delta}\right)^{1/2}$ is a function of δ for which A is proportional to

$$\frac{1}{N}\left(\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi_i}}{K} \mathcal{L}\left(\hat{h}, \mathbb{D}_g^{S,j}\right) + \sum_{i=1}^{K} \sqrt{\pi_i} L \max_k \mathbb{E}_{\mathbb{P}_g^k}\left[\|\Delta p^{k,i}(y|z)\|_1\right] + \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi_i}}{K} L\sqrt{2}\, d_{1/2}\left(\mathbb{P}_g^{S,j}, \mathbb{P}_g^{S,i}\right)\right)^2.$$

Theorem 7 reveals one benefit of learning compressed DI representations, that is, when enforcing compressed DI representation learning, we minimize $\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi_i}}{K} L\sqrt{2}\, d_{1/2}\left(\mathbb{P}_g^{S,j}, \mathbb{P}_g^{S,i}\right)$, which tends to reduce the generalization gap $\epsilon = \epsilon(\delta)$ for a given confidence level $1 - \delta$. Therefore, compressed DI representations allow us to minimize $\mathcal{L}\left(\hat{h}, \mathbb{D}_g^{\pi}\right)$ more efficiently.

2.4 Trade-off of Learning Domain-Invariant Representations

Similar to the theoretical finding in Zhao et al. [49] developed for DA, we theoretically find that compression does come with a cost for MSDA and DG. We investigate the representation trade-off, typically how compressed DI representation affects classification loss. Specifically, we consider a data processing chain $\mathcal{X} \overset{g}{\longmapsto} \mathcal{Z} \overset{\hat{h}}{\longmapsto} \mathcal{Y}_{\Delta}$, where $\mathcal{X}$ is the common data space, $\mathcal{Z}$ is the latent space induced by a feature extractor g, and $\hat{h}$ is a hypothesis on top of the latent space. We define $\mathbb{P}_{\mathcal{Y}}^{\pi}$ and $\mathbb{P}_{\mathcal{Y}}^{T}$ as two distribution over $\mathcal{Y}$ in which to draw $y \sim \mathbb{P}_{\mathcal{Y}}^{\pi}$, we sample $k \sim Cat(\pi)$, $x \sim \mathbb{P}^{S,k}$, and $y \sim f^{S,k}(x)$, while similar to draw $y \sim \mathbb{P}_{\mathcal{Y}}^{T}$. Our theoretical bounds developed regarding the trade-off of learning DI representations are relevant to $d_{1/2}\left(\mathbb{P}_{\mathcal{Y}}^{\pi}, \mathbb{P}_{\mathcal{Y}}^{T}\right)$.

Theorem 8. *Consider a feature extractor g and a hypothesis $\hat{h}$, the Hellinger distance between two label marginal distributions $\mathbb{P}^{\pi}_{\mathcal{Y}}$ and $\mathbb{P}^{T}_{\mathcal{Y}}$ can be upper-bounded as:*

1. $d_{1/2}\left(\mathbb{P}^{\pi}_{\mathcal{Y}}, \mathbb{P}^{T}_{\mathcal{Y}}\right) \leq d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{\pi}_{g}\right) + \mathcal{L}\left(\hat{h} \circ g, f^{T}, \mathbb{P}^{T}\right) + \sum_{k=1}^{K} \pi_k \mathcal{L}\left(\hat{h} \circ g, f^{S,k}, \mathbb{P}^{S,k}\right)$.

2. $d_{1/2}\left(\mathbb{P}^{\pi}_{\mathcal{Y}}, \mathbb{P}^{T}_{\mathcal{Y}}\right) \leq \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi}_j}{K} d_{1/2}\left(\mathbb{P}^{S,i}_{g}, \mathbb{P}^{S,j}_{g}\right) + \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi_j}}{K} d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{S,i}_{g}\right) + \mathcal{L}\left(\hat{h} \circ g, f^{T}, \mathbb{P}^{T}\right) + \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h} \circ g, f^{S,i}, \mathbb{P}^{S,i}\right)$.

Here we note that the general loss $\mathcal{L}$ is defined based on the Hellinger loss ℓ which is define as $\ell(\hat{f}(x), f(x)) = D_{1/2}(\hat{f}(x), f(x)) = 2\sum_{i=1}^{C}\left(\sqrt{\hat{f}(x,i)} - \sqrt{f(x,i)}\right)^2$ *(more discussion can be found in Appendix C).*

Compared to the trade-off bound in the work of Zhao et al. [49], our context is more general, concerning MSDA and DG problems with multiple source domains and multi-class probabilistic labeling functions, rather than single source DA with binary-class and deterministic setting. Therefore, the Hellinger distance is more universal, in the sense that it does not depend on the choice of classifier family $\mathcal{H}$ as in the case of $\mathcal{H}$-divergence in [49].

We base on the first inequality of Theorem 8 to analyze the trade-off of learning general DI representations. Because the term $\sum_{i=1}^{K} \pi_k \mathcal{L}\left(\hat{h} \circ g, f^{S,i}, \mathbb{P}^{S,i}\right)$ is controllable and tends to be small when enforcing learning general DI representations, if *two label marginal distributions* $\mathbb{P}^{\pi}_{\mathcal{Y}}$ and $\mathbb{P}^{T}_{\mathcal{Y}}$ are distant (i.e., $d_{1/2}\left(\mathbb{P}^{\pi}_{\mathcal{Y}}, \mathbb{P}^{T}_{\mathcal{Y}}\right)$ is high), the sum $d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{\pi}_{g}\right) + \mathcal{L}\left(\hat{h} \circ g, f^{T}, \mathbb{P}^{T}\right)$ tends to be high. For the *MSDA setting*, the data shift $d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{\pi}_{g}\right)$ can be trained to get smaller, meaning that $\mathcal{L}\left(\hat{h} \circ g, f^{T}, \mathbb{P}^{T}\right)$ tends to be high, hence the target performance is hurt. For the *DG setting*, if trained feature extractor g by chance helps to reduce the data shift $d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{\pi}_{g}\right)$, the target loss $\mathcal{L}\left(\hat{h} \circ g, f^{T}, \mathbb{P}^{T}\right)$ tends to be higher and the target performance is more hurt. In contrast, if for some target domains, the data shift $d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{\pi}_{g}\right)$ is high by some reasons, by linking to Inequality (1) of Theorem 1, the target general loss has a high upper-bound, hence is possibly high.

Based on the second inequality of Theorem 8, we observe that if *two label marginal distributions* $\mathbb{P}^{\pi}_{\mathcal{Y}}$ and $\mathbb{P}^{T}_{\mathcal{Y}}$ are distant while enforcing learning compressed DI representations (i.e., $\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi}_j}{K} d_{1/2}\left(\mathbb{P}^{S,j}_{g}, \mathbb{P}^{S,i}_{g}\right) + \sum_{i=1}^{K} \pi_i \mathcal{L}\left(\hat{h} \circ g, f^{S,i}, \mathbb{P}^{S,i}\right)$ is low), the sum $\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi_j}}{K} d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{S,i}_{g}\right) + \mathcal{L}\left(h \circ g, f^{T}, \mathbb{P}^{T}\right)$ is high. For the *MSDA setting*, the data shift $\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi}_j}{K} d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{S,i}_{g}\right)$ is trained to get smaller, meaning that $\mathcal{L}\left(\hat{h} \circ g, f^{T}, \mathbb{P}^{T}\right)$ tends to be high, hence hurting the target performance. For the *DG setting*, if the trained feature extractor g occasionally reduces $\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi}_j}{K} d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{S,i}_{g}\right)$ for some unseen target domain, it certainly increases the target loss $\mathcal{L}\left(h \circ g, f^{T}, \mathbb{P}^{T}\right)$, hence hurting the target performance. In contrast, if for some target domains, the data shift $\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\sqrt{\pi}_j}{K} d_{1/2}\left(\mathbb{P}^{T}_{g}, \mathbb{P}^{S,i}_{g}\right)$ is high by some reasons, by linking to Inequality (2) of Theorem 3, the target general loss has a high upper-bound, hence is possibly high.

3 Experiment

To investigate the trade off of learning DI representations, we conduct experiments on the *colored MNIST* dataset (CC BY-SA 3.0) [2, 22]. In particular, the task is to predict binary label Y of colored input images X generated from binary digit feature Z_d and color feature Z_c. We refer readers to Appendix D for more information of this dataset.

We conduct 7 source domains by setting color-label correlation $\mathbb{P}(Z_c = 1|Y = 1) = \mathbb{P}(Z_c = 0|Y = 0) = \theta^{S,i}$ where $\theta^{s,i} \sim Uni\left([0.6, 1]\right)$ for $i = 1, ..., 12$, while two target domains are created with $\theta^{T,i} \in \{0.05, 0.7\}$ for $i = 1, 2$. Here we note that colored images in the target domain with

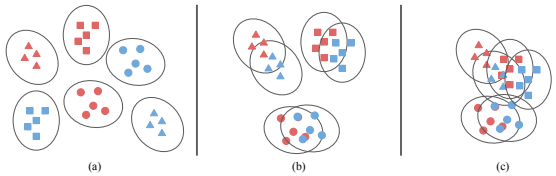


Figure 1: Phase transition of representations with increased λ. (a) General DI representations. (b) Just-right compressed DI representations. (c) Overly-compressed DI representations.

$\theta^{T,2} = 0.7$ are *more similar* to those in the source domains, while colored images in the target domain with $\theta^{T,1} = 0.05$ are *less similar*.

We wish to study the characteristics and trade-off between two kinds of DI representations when predicting on various target domains. Specifically, we apply adversarial learning [16] similar to [14], in which a min-max game is played between domain discriminator $\hat{h}^d$ trying to distinguish the source domain given representation, while the feature extractor (generator) g tries to fool the domain discriminator. Simultaneously, a classifier is used to classify label based on the representation. Let $\mathcal{L}_{gen}$ and be the label classification and domain discrimination losses, the training objective becomes:

$$\min_{g} \left(\min_{\hat{h}} \mathcal{L}_{gen} + \lambda \max_{\hat{h}^d} \mathcal{L}_{disc} \right),$$

where the source compression strength $\lambda > 0$ controls the compression extent of learned representation. More specifically, general DI representation is obtained with $\lambda = 0$, while larger λ leads to more compressed DI representation. Finally, our implementation is based on DomainBed [17] repository, and all training details are included in Appendix D.

3.1 Trade-off of Two Kinds of Domain-Invariant Representations

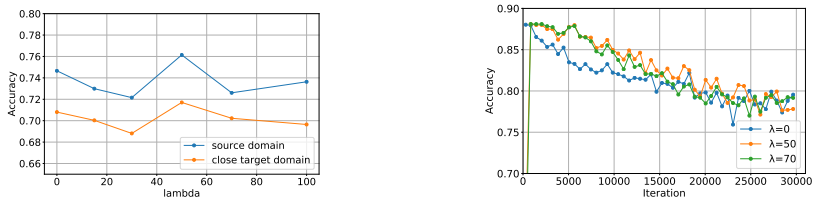
We consider target domain which is similar to source domain in this experiment. To govern the trade-off between two kinds of DI representations, we gradually increase the source compression strength λ with noting that $\lambda = 0$ means only focusing on learning general DI representations. According to our theory and as shown in Figure 1, *learning only general DI representations* ($\lambda = 0$) maximizes the cross-domain class divergence by encouraging the classes of source domains more separate arbitrarily, while by increasing λ, we enforce compressing the latent representations of source domains together. Therefore, for an appropriate λ, the class separation from general DI representations and the source compression from compressed DI representations (i.e., *just-right compressed DI representations*), are balanced as in the case (b) of Figure 1, while for overly high λ, source compression from compressed DI representations dominates and compromises the class separation from general DI representations (i.e., *overly-compressed DI representations*).

Figure 2a shows the source validation and target accuracies when increasing λ (i.e., encouraging the source compression). It can be observed that both source validation accuracy and target accuracy have the same pattern: increasing when setting appropriate λ for *just-right compressed DI representations* and compromising when setting overly high values λ for *overly-compressed DI representations*. Figure 2b shows in detail the variation of the source validation accuracy for each specific λ. In practice, we should encourage learning two kinds of DI representations simultaneously by finding an appropriate trade-off to balance them for working out just-right compressed DI representations.

3.2 Generalization Capacity on Various Target Domains

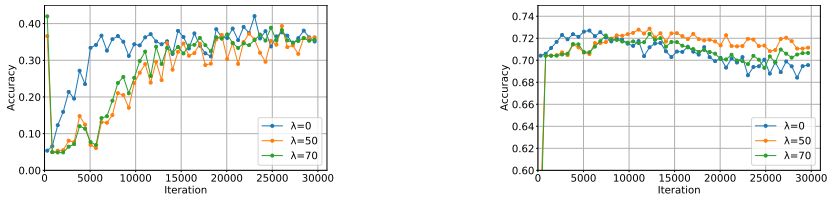
In this experiment, we inspect the generalization capacity of the models trained on the source domains with different compression strengths λ to different target domains (i.e., close and far ones). As shown in Figure 3, the target accuracy is lower for the far target domain than for the close one regardless of the source compression strengths λ. This observation aligns with our upper-bounds developed in Theorems 1 and 3 in which higher data shift increases the upper-bounds of the general target loss, hence more likely hurting the target performance.

Another interesting observation is that for the far target domain, the target accuracy for $\lambda = 0$ peaks. This can be partly explained as the general DI representations help to increase the span of source latent representations for more chance to match a far target domain (cf. Section 2.3.2). Specifically,



(a) Accuracy vs λ (b) Accuracy vs iteration

Figure 2: (**2a**) Source validation accuracy and target accuracy for close target domain (see Section 3.2) over compression strength. (**2b**) Validation accuracy over training step for different values of λ.



(a) Far target domain with $\theta = 0.05$. (b) Close target domain with $\theta = 0.7$.

Figure 3: Accuracy of models trained with different λ on two test domains. For the far domain with $\theta = 0.05$, the model corresponding with general DI representations performs better than that with more compact DI representations. For close domain with $\theta = 0.7$, the models with compression generalize significantly better than those with general DI representations.

the source-target data shift $d_{1/2}\left(\mathbb{P}_g^T, \mathbb{P}_g^\pi\right)$ could be small and the upper bound in Theorem 1 is lower. In contrast, for the close target domain, the compressed DI representations help to really improve the target performance, while over-compression degrades the target performance, similar to result on source domains. This is because the target domain is naturally close to the source domains, the source and target latent representations are already mixed up, but by encouraging the compressed DI representations, we aim to learning more elegant representations for improving target performance.

4 Related Works

Domain adaptation (DA) has been intensively studied from both theoretical [3, 9, 6, 49, 32, 39] and empirical [40, 48, 14, 7, 47] perspectives. Notably, the pioneering work [3] and other theoretical works [39, 48, 49, 21, 29, 6] have investigated DA in various aspects which lay foundation for the success of practical works [14, 43, 40, 48, 14, 7, 47]. *Multiple-source domain adaptation (MSDA)* extends DA by gathering training set from multiple source domains [8, 28, 20, 38, 39, 12, 50, 46, 36]. Different from DA with abundant theoretical works, theoretical study in MSDA is significantly limited. Notably, the works in [30, 31] assume target domain being a mixture of source domains to learn a mixture of expert classifiers. The work in [18] develops both theory and algorithm to learn such combination for different losses. *Domain generalization (DG)* [17, 24, 4, 19] is the most challenging setting among the three due to the unavailability of target data. The studies in [4, 10] use kernel method to find feature map which minimizes expected target loss. Moreover, those in [26, 27, 33, 15, 45, 11] learn compressed DI representations by minimizing different types of domain dissimilarity, while [2] learns a latent space on which representation distributions do not need to be aligned, but share a common optimal hypothesis. Moreover, there are other works which try to strike a balance between two types of representation spaces [5].

5 Conclusion

In this paper, we derive theoretical bounds for target loss in MSDA and DG settings which characterize two types of representation: general DI representation for learning invariant classifier which works on all source domains and compressed DI representation motivated from reducing inter-domain representation discrepancy. We further characterize the properties of these two representations, and develop a lower bound on the target loss which governs the trade-off between learning them. Finally, we conduct experiments on Colored MNIST dataset to illustrate our theoretical claims.

References

[1] Isabela Albuquerque, Joao Monteiro, Mohammad Darvishi, Tiago H. Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching, 2020.

[2] Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1â2):151â175, May 2010.

[4] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[5] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. *arXiv preprint arXiv:2008.12839*, 2020.

[6] Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 308–323, 2011. Copyright: Copyright 2011 Elsevier B.V., All rights reserved.; 22nd International Conference on Algorithmic Learning Theory, ALT 2011 ; Conference date: 05-10-2011 Through 07-10-2011.

[7] Corinna Cortes, Mehryar Mohri, and Andres Munoz Medina. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

[8] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(57):1757–1774, 2008.

[9] Shai Ben David, Tyler Lu, Teresa Luu, and David Pal. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136. PMLR, 13–15 May 2010.

[10] Aniket Anand Deshmukh, Yunwen Lei, Srinagesh Sharma, Urun Dogan, James W. Cutler, and Clayton Scott. A generalization error bound for multi-class domain generalization, 2019.

[11] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6450–6461, 2019.

[12] 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, 2012.

[13] John C. Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence, and surrogate risk, 2017.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[15] 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*, pages 2551–2559, 2015.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *CoRR*, abs/2007.01434, 2020.

[18] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[19] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[20] Amaury Habrard Marc Sebban Ievgen Redko. On the analysis of adaptability in multi-source domain adaptation. *Mach. Learn.*, 108:1635â1652, 2019.

[21] F. D. Johansson, D. Sontag, and R. Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of Machine Learning Research*, volume 89, pages 527–536, 2019.

[22] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. 2018.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[25] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[26] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[28] Yishay Mansour, Mehryar Mohri, Jae Ro, Ananda Theertha Suresh, and Ke Wu. A theory of multiple-source adaptation with limited target labeled data. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2332–2340. PMLR, 13–15 Apr 2021.

[29] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. volume 22, 06 2009.

[30] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.

[31] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the renyi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 367â374, Arlington, Virginia, USA, 2009. AUAI Press.

[32] Yishay Mansour and Mariano Schain. Robust domain adaptation. *Annals of Mathematics and Artificial Intelligence*, 71 Issue 4:365â380, 2014.

[33] Krikamol Muandet, David Balduzzi, and Bernhard SchÃ¶lkopf. Domain generalization via invariant feature representation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[34] XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[35] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1406–1415. IEEE Computer Society, 2019.

[36] 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/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[37] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7728–7738. PMLR, 13–18 Jul 2020.

[38] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 849–858. PMLR, 16–18 Apr 2019.

[39] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport, 2017.

[40] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16282–16292. Curran Associates, Inc., 2020.

[41] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[42] Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve, 2021.

[43] 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 (CVPR)*, July 2017.

[44] Vladimir Vapnik and Vlamimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[45] Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019.

[46] 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 (CVPR)*, June 2018.

[47] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training, 2020.

[48] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7404–7413. PMLR, 09–15 Jun 2019.

[49] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7523–7532. PMLR, 09–15 Jun 2019.

[50] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[51] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16096–16107. Curran Associates, Inc., 2020.

[52] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12975–12983, Apr. 2020.

[53] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, pages 561–578. Springer, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our main contributions are in section 2. Moreover, in the introduction, we explicitly map each contribution to its section.
   (b) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] .
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is theoretical and concerns a general theme, i.e., representation learning in domain generalization and multi-source domain adaptation. Therefore, negative social impact is not well-linked to.
   (d) Did you describe the limitations of your work? [No]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We include assumptions in all theoretical statements. Please refer to our theorems and lemma in both main paper and Appendix A, B, C.
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide proofs to all theoretical results in the Appendix (A, B, C), while their motivations and implications are discussed under Section 2 of main paper.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our code can be found in the URL included in Appendix D.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? The full details can be provided with the code, but the important details should be in the main paper. [Yes] We briefly mention the data and training process in the main paper. More details can be found in the Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We would like to demonstrate only the tendency which supports the theory.
   (d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This information is included in the Appendix D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use ColoredMNIST dataset and PACS dataset, and have appropriately cited the authors.

(b) Did you mention the license of the assets? [Yes] We include license of MNIST dataset briefly when mentioning it under Section 3. For PACS dataset, we could not find its license, although it is a public and widely used dataset.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] We do not use any other datasets besides those already mentioned.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We could not find these information.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We could not find these information.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]