Model-Based Domain Generalization

Alexander Robey **George J. Pappas** **Hamed Hassani**
Department of Electrical and Systems Engineering
University of Pennsylvania
{arobey1,pappasg,hassani}@seas.upenn.edu

Abstract

Despite remarkable success in a variety of applications, it is well-known that deep learning can fail catastrophically when presented with out-of-distribution data. Toward addressing this challenge, we consider the *domain generalization* problem, wherein predictors are trained using data drawn from a family of related training domains and then evaluated on a distinct and unseen test domain. We show that under a natural model of data generation and a concomitant invariance condition, the domain generalization problem is equivalent to an infinite-dimensional constrained statistical learning problem; this problem forms the basis of our approach, which we call Model-Based Domain Generalization. Due to the inherent challenges in solving constrained optimization problems in deep learning, we exploit nonconvex duality theory to develop unconstrained relaxations of this statistical problem with tight bounds on the duality gap. Based on this theoretical motivation, we propose a novel domain generalization algorithm with convergence guarantees. In our experiments, we report improvements of up to 30% over state-of-the-art domain generalization baselines on several benchmarks including ColoredMNIST, Camelyon17-WILDS, FMoW-WILDS, and PACS. Our code is publicly available at the following link: `https://github.com/arobey1/mbdg`.

1 Introduction

Despite well-documented success in numerous applications [1–4], the complex prediction rules learned by modern machine learning methods can fail catastrophically when presented with out-of-distribution (OOD) data [5–9]. Indeed, rapidly growing bodies of work conclusively show that state-of-the-art methods are vulnerable to distributional shifts arising from spurious correlations [10–12], adversarial attacks [13–17], sub-populations [18–21], and naturally-occurring variation [22–24]. This failure mode is particularly pernicious in *safety-critical applications*, wherein the shifts that arise in fields such as medical imaging [25–28], autonomous driving [29–31], and robotics [32–34] are known to lead to unsafe behavior. And while some progress has been made toward addressing these vulnerabilities, the inability of modern machine learning methods to generalize to OOD data is one of the most significant barriers to deployment in safety-critical applications [35, 36].

In the last decade, the *domain generalization* community has emerged in an effort to improve the OOD performance of machine learning methods [37–40]. In this field, predictors are trained on data drawn from a family of related training domains and then evaluated on a distinct and unseen test domain. Although a variety of approaches have been proposed for this setting [41, 42], it was recently shown that that no existing domain generalization algorithm can significantly outperform empirical risk minimization (ERM) [43] over the training domains when ERM is properly tuned and equipped with state-of-the-art architectures [44, 45] and data augmentation techniques [46]. Therefore, due to the prevalence of OOD data in safety critical applications, it is of the utmost importance that new algorithms be proposed which can improve the OOD performance of machine learning methods.

In this paper, we introduce a new framework for domain generalization which we call *Model-Based Domain Generalization* (MBDG). The key idea in our framework is to first learn transformations

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

that map data between domains and then to subsequently enforce invariance to these transformations. Under a general model of covariate shift and a novel notion of invariance to learned transformations, we use this framework to rigorously re-formulate the domain generalization problem as a semi-infinite constrained optimization problem. We then use this re-formulation to prove that a tight approximation of the domain generalization problem can be obtained by solving the empirical, parameterized dual for this semi-infinite problem. Finally, motivated by these theoretical insights, we propose a new algorithm for domain generalization; extensive experimental evidence shows that our algorithm advances the state-of-the-art on a range of benchmarks by up to thirty percentage points.

Contributions. Our contributions can be summarized as follows:

- We propose a new framework for domain generalization in which invariance is enforced to underlying transformations of data which capture inter-domain variation.
- Under a general model of covariate shift, we rigorously prove the equivalence of the domain generalization problem to a novel semi-infinite constrained statistical learning problem.
- We derive *data-dependent* duality gap bounds for the empirical parameterized dual of this semi-infinite problem, proving that tight approximations of the domain generalization problem can be obtained by solving this dual problem under the covariate shift assumption.
- We introduce a primal-dual style algorithm for domain generalization in which invariance is enforced over unsupervised generative models trained on data from the training domains.
- We empirically show that our algorithm significantly outperforms state-of-the-art baselines on several standard benchmarks, including `ColoredMNIST`, `Camelyon17-WILDS`, and `PACS`.

2 Related work

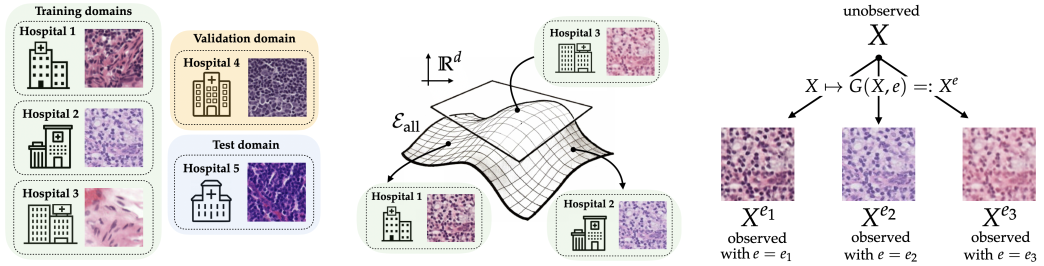
Domain generalization. The rapid acceleration of domain generalization research has led to an abundance of principled algorithms, many of which distill knowledge from an array of disparate fields toward resolving OOD failure modes [47–50]. Among such works, one prominent thrust has been to learn predictors which have internal feature representations that are consistent across domains [51–62]. This approach is also popular in the field of unsupervised domain adaptation [63–67], wherein it is assumed that unlabeled data from the test domain is available during training [68–70]. Also related are works that seek to learn a kernel-based embedding of each domain in an underlying feature space [71, 72], and those that employ Model-Agnostic Meta Learning [73] to adapt to unseen domains [42, 74–81]. Recently, another prominent direction has been to design weight-sharing [82–85] and instance re-weighting schemes [86–88]. Unlike any of these approaches, we explicitly enforce hard invariance-based constraints on the underlying statistical domain generalization problem.

Data augmentation. Another approach to improve OOD performance is to augment the available training data. Among such methods, perhaps the most common is to leverage various forms of data augmentation [89–96]. Recently, several approaches have used style-transfer techniques and image-to-image translation networks [97–104] to augment the training domains with artificially-generated data [105–112]. Alternatively, rather than generating new data, [113–115] all remove textural features in the data to encourage domain invariance. Unlike the majority of these works, we do not perform data augmentation directly on the training objective; rather, we derive a principled primal-dual style algorithm which enforces invariance constraints on data generated by unsupervised generative models.

3 Domain generalization

The domain generalization setting is characterized by a pair of random variables (X, Y) over instances $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and corresponding labels $y \in \mathcal{Y}$, where (X, Y) is jointly distributed according to an unknown probability distribution $\mathbb{P}(X, Y)$. Ultimately, the objective in this setting is to learn a predictor f such that $f(X) \approx Y$, meaning that f should be able to predict the labels y of corresponding instances x for each $(x, y) \sim \mathbb{P}(X, Y)$. However, unlike in standard supervised learning tasks, the domain generalization problem is complicated by the assumption that one cannot sample directly from $\mathbb{P}(X, Y)$. Rather, it is assumed that we can only measure (X, Y) under different environmental conditions, each of which corrupts or varies the data in a different way. For example, in medical imaging tasks, these environmental conditions might correspond to the imaging techniques and stain patterns used at different hospitals; this is illustrated in Figure 1a.

To formalize this notion of environmental variation, we assume that data is drawn from a set of *environments* or *domains* $\mathcal{E}_{\text{all}}$ (see Figure 1b). Concretely, each domain $e \in \mathcal{E}_{\text{all}}$ can be identified with



(a) In domain generalization, the data are drawn from a family of related domains. For example, in the `Camelyon17-WILDS` dataset [20], which contains images of cells, the domains correspond to different hospitals where these images were captured.

(b) Each data point in a domain generalization task is observed in a particular domain $e \in \mathcal{E}_{\text{all}}$. The set of all domains $\mathcal{E}_{\text{all}}$ can be thought of as an abstract space lying in $\mathbb{R}^p$. In `Camelyon17-WILDS`, this space $\mathcal{E}_{\text{all}}$ corresponds to the set of all possible hospitals.

(c) We assume that the variation from domain to domain is characterized by an underlying generative model $G(x, e)$, which transforms the unobserved random variable $X \mapsto G(X, e) := X^e$, where X^e represents X observed in any domain $e \in \mathcal{E}_{\text{all}}$.

Figure 1: An overview of the domain generalization problem setting used in this paper.

a pair of random variables (X^e, Y^e), which together denote the observation of the random variable pair (X, Y) in environment e. Given samples from a finite subset $\mathcal{E}_{\text{train}} \subsetneq \mathcal{E}_{\text{all}}$ of domains, the goal of the domain generalization problem is to learn a predictor f that generalizes across all possible environments, implying that $f(X) \approx Y$. This can be summarized as follows:

Problem 3.1 (Domain generalization)**.** Let $\mathcal{E}_{\text{train}} \subsetneq \mathcal{E}_{\text{all}}$ be a finite subset of training domains, and assume that for each $e \in \mathcal{E}_{\text{train}}$, we have access to a dataset $\mathcal{D}^e := \{(x_j^e, y_j^e)\}_{j=1}^{n_e}$ sampled i.i.d. from $\mathbb{P}(X^e, Y^e)$. Given a function class $\mathcal{F}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, our goal is to learn a predictor $f \in \mathcal{F}$ using the data from the datasets $\mathcal{D}^e$ that minimizes the worst-case risk over the entire family of domains $\mathcal{E}_{\text{all}}$. That is, we want to solve the following optimization problem:

$$\underset{f \in \mathcal{F}}{\text{minimize}} \; \max_{e \in \mathcal{E}_{\text{all}}} \; \mathbb{E}_{\mathbb{P}(X^e, Y^e)} \, \ell(f(X^e), Y^e). \tag{DG}$$

In essence, in Problem 3.1 we seek a predictor $f \in \mathcal{F}$ that generalizes from the finite set of training domains $\mathcal{E}_{\text{train}}$ to perform well on the set of all domains $\mathcal{E}_{\text{all}}$. However, note that while the inner maximization in (DG) is over the set of all training domains $\mathcal{E}_{\text{all}}$, by assumption we do not have access to data from any of the domains $e \in \mathcal{E}_{\text{all}} \backslash \mathcal{E}_{\text{train}}$, making this problem challenging to solve. Indeed, as generalizing to arbitrary test domains is impossible [116], further structure is often assumed on the topology of $\mathcal{E}_{\text{all}}$ and on the corresponding distributions $\mathbb{P}(X^e, Y^e)$.

Disentangling the sources of variation across environments. The difficulty of a particular domain generalization task can be characterized by the extent to which the distribution of data in the unseen test domains $\mathcal{E}_{\text{all}} \backslash \mathcal{E}_{\text{train}}$ resembles the distribution of data in the training domains $\mathcal{E}_{\text{train}}$. For instance, if the domains are assumed to be convex combinations of the training domains, as is often the case in multi-source domain generalization [117–119], Problem 3.1 can be seen as an instance of distributionally robust optimization [120]. More generally, in a similar spirit to [116], we identify two forms of variation across domains: *covariate shift* and *concept shift*. These shifts characterize the extent to which the marginal distributions over instances $\mathbb{P}(X^e)$ and the instance-conditional distributions $\mathbb{P}(Y^e | X^e)$ differ between domains. We capture these shifts in the following definition:

Definition 3.2 (Covariate shift & concept shift)**.** Problem 3.1 is said to experience **covariate shift** if environmental variation is due to differences between the set of marginal distributions over instances $\{\mathbb{P}(X^e)\}_{e \in \mathcal{E}_{\text{all}}}$. On the other hand, Problem 3.1 is said to experience **concept shift** if environmental variation is due to changes amongst the instance-conditional distributions $\{\mathbb{P}(Y^e | X^e)\}_{e \in \mathcal{E}_{\text{all}}}$.

The growing domain generalization literature encompasses a great deal of past work, wherein both of these shifts have been studied in various contexts [121–125]. Indeed, as this literature has grown, new benchmarks have been developed which span the gamut between covariate and concept shift [126]. However, a large-scale empirical study recently showed that no existing algorithm can significantly outperform ERM across these standard domain generalization benchmarks when ERM is carefully implemented [46]. As ERM is known to fail in the presence natural distribution shifts [127], this result highlights the critical need for new algorithms that can go beyond ERM toward solving Problem 3.1.

4 Model-based domain generalization

In what follows, we introduce a new framework for domain generalization that we call *Model-Based Domain Generalization* (MBDG). In particular, we prove that when Problem 3.1 is characterized solely by covariate shift, then under a natural invariance-based condition, Problem 3.1 is equivalent to an infinite-dimensional constrained statistical learning problem, which forms the basis of MBDG.

Formal assumptions for MBDG. In general, domain generalization tasks can be characterized by both covariate and concept shift. However, in this paper, we restrict the scope of our theoretical analysis to focus on problems in which inter-domain variation is due solely to covariate shift through an underlying model of data generation. Formally, we assume that the data in each domain $e \in \mathcal{E}_{\text{all}}$ is generated from the underlying random variable pair (X, Y) via an unknown function G.

Assumption 4.1. Let δ_e denote a Dirac distribution for $e \in \mathcal{E}_{\text{all}}$. We assume that there exists[1] a measurable function $G : \mathcal{X} \times \mathcal{E}_{\text{all}} \to \mathcal{X}$, which we refer to as a *domain transformation model*, that parameterizes the inter-domain covariate shift via $\mathbb{P}(X^e) =^d G\,\#\,(\mathbb{P}(X) \times \delta_e)\ \forall e \in \mathcal{E}_{\text{all}}$, where $\#$ denotes the push-forward measure and $=^d$ denotes equality in distribution.

Informally, this assumption specifies that there should exist a function G that relates the random variables X and X^e via $X \mapsto G(X, e) = X^e$. In past work, this particular setting in which the instances X^e measured in an environment e are related to the underlying random variable X has been referred to as *domain shift* [128, §1.8]. In our medical imaging example, the domain shift captured by a domain transformation model would characterize the mapping from the underlying distribution $\mathbb{P}(X)$ over different cells to the distribution $\mathbb{P}(X^e)$ of images of these cells observed at a particular hospital; this is illustrated in Figure 1c, wherein inter-domain variation is due to varying colors and stain patterns encountered at different hospitals. On the other hand, in this example example, the label $y \sim Y$ describing whether a given cell contains a cancerous tumor should not depend on the lighting and stain patterns used at different hospitals. In this sense, while in other applications, e.g. the datasets introduced in [10], the instance-conditional distributions can vary across domains, in this paper we assume that inter-domain variation is *solely* characterized by the domain shift due to G.

Assumption 4.2 (Domain shift)**.** We assume that inter-domain variation is solely characterized by domain shift in the marginal distributions $\mathbb{P}(X^e)$, as described in Assumption 4.1. As a consequence, we assume that the instance-conditional distributions $\mathbb{P}(Y^e|X^e)$ are stable across domains, meaning that Y^e and Y are equivalent in distribution and that for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, it holds that

$$\mathbb{P}(Y = y|X = x) = \mathbb{P}(Y^e = y|X^e = G(x, e)) \quad \forall e \in \mathcal{E}_{\text{all}}. \tag{1}$$

Pulling back Problem 3.1. The structure imposed on Problem 3.1 by Assumptions 4.1 and 4.2 provides a concrete way of parameterizing large families of distributional shifts in domain generalization problems. Indeed, the utility of these assumptions is that when taken together, they provide the basis for pulling-back Problem 3.1 onto the underlying distribution $\mathbb{P}(X, Y)$ via the domain transformation model G. This insight is captured in the following proposition:

Proposition 4.3. Under Assumptions 4.1 and 4.2, Problem 3.1 is equivalent to

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \max_{e \in \mathcal{E}_{\text{all}}} \mathbb{E}_{\mathbb{P}(X,Y)}\, \ell(f(G(X, e)), Y). \tag{2}$$

The proof of this fact is a straightforward consequence of the decomposition $\mathbb{P}(X^e, Y^e) = \mathbb{P}(Y^e|X^e) \cdot \mathbb{P}(X^e)$ in conjunction with Assumptions 4.1 and 4.2 (see Appendix C.2). Note that this result allows us to implicitly absorb each of the domain distributions $\mathbb{P}(X^e, Y^e)$ into the domain transformation model. Thus, the outer expectation in (2) is defined over the underlying distribution $\mathbb{P}(X, Y)$. On the other hand, just as in (DG), this problem is still a challenging statistical min-max problem. To this end, we next introduce a new notion of invariance with respect to domain transformation models, which allows us to reformulate the problem in (2) as a semi-infinite constrained optimization problem.

[1] Crucially, although we assume the existence of a domain transformation model G, we emphasize that for many problems, it may be impossible to obtain or derive a simple analytic expression for G. This topic will be discussed at length in Section 6 and in Appendix G.

A new notion of model-based invariance. Common to much of the domain generalization literature is the idea that predictors should be invariant to inter-domain changes. For instance, in [10] the authors seek to learn an *equipredictive representation* $\Phi : \mathcal{X} \to \mathcal{Z}$ [129], i.e. an intermediate representation that satisfies $\mathbb{P}(Y^{e_1}|\Phi(X^{e_1})) = \mathbb{P}(Y^{e_2}|\Phi(X^{e_2})) \quad \forall e_1, e_2 \in \mathcal{E}_{\text{all}}$. Despite compelling theoretical motivation for this approach, it has been shown that current algorithms which seek equipredictive representations do not significantly improve over ERM [130–133]. With this in mind and given the additional structure introduced in Assumptions 4.1 and 4.2, we introduce a new definition of invariance with respect to the variation captured by the underlying domain transformation model G.

Definition 4.4 (G-invariance)**.** Given a domain transformation model G, we say a classifier f is **G-invariant** if it holds for all $e \in \mathcal{E}_{\text{all}}$ that $f(x) = f(G(x, e))$ almost surely when $x \sim \mathbb{P}(X)$.

Concretely, this definition says that a predictor f is G-invariant if environmental changes under $G(x, e)$ cannot change the prediction returned by f. Intuitively, this notion of invariance couples with the definition of domain shift, in the sense that we expect that a prediction should return the same prediction for any realization of data under G. Thus, whereas equipredictive representations are designed to enforce invariance of in an intermediate representation space $\mathcal{Z}$, Definition 4.4 is designed to enforce invariance directly on the predictions made by f. In this way, in the setting of Figure 1, G-invariance would imply that the predictor f would return the same label for a given cluster of cells regardless of the hospital at which these cells were imaged.

Formulating the MBDG optimization problem. The G-invariance property described in the previous section is the key toward reformulating the min-max problem in (2). Indeed, the following proposition follows from Assumptions 4.1 and 4.2 and from the definition of G-invariance.

Proposition 4.5. Under Assumptions 4.1 and 4.2, if we restrict the domain $\mathcal{F}$ of Problem 3.1 to the set of G-invariant predictors, then Problem 3.1 is equivalent to the following constrained problem:

$$\begin{aligned} P^\star \triangleq \underset{f \in \mathcal{F}}{\text{minimize}} \quad & R(f) \triangleq \mathbb{E}_{\mathbb{P}(X,Y)}\, \ell(f(X), Y) \\ \text{subject to} \quad & f(x) = f(G(x, e)) \quad \text{a.e. } x \sim \mathbb{P}(X) \;\; \forall e \in \mathcal{E}_{\text{all}}. \end{aligned} \tag{MBDG}$$

Here a.e. stands for "almost everywhere" and $R(f)$ is the statistical risk of a predictor f with respect to the underlying random variable pair (X, Y). Note that unlike (2), (MBDG) is not a composite optimization problem, meaning that the inner maximization has been eliminated. In essence, the proof of Proposition 4.6 relies on the fact that G-invariance implies that predictions should not change across domains (see Appendix C.2). The optimization problem in (MBDG) forms the basis of our Model-Based Domain Generalization framework. To explicitly contrast this problem to Problem 3.1, we introduce the following concrete problem formulation for Model-Based Domain Generalization.

Problem 4.6 (Model-Based Domain Generalization)**.** As in Problem 3.1, let $\mathcal{E}_{\text{train}} \subsetneq \mathcal{E}_{\text{all}}$ be a finite subset of training domains and assume that we have access to datasets $\mathcal{D}^e \; \forall e \in \mathcal{E}_{\text{train}}$. Then under Assumptions 4.1 and 4.2, the goal of Model-Based Domain Generalization is to use the data from the training datasets to solve the semi-infinite constrained optimization problem in (MBDG).

Problem 4.6 offers a principled perspective on Problem 3.1 when data varies WRT an underlying domain transformation model. However, just as solving the min-max problem of Problem 3.1 is known to be difficult, the problem in (MBDG) is also challenging to solve for several reasons:

(C1) The G-invariance constraint in (MBDG) is strict and thus challenging to enforce.
(C2) Problem 4.6 is a constrained problem over an infinite-dimensional functional space $\mathcal{F}$.
(C3) We do have access to the set of all domains $\mathcal{E}_{\text{all}}$ or to the underlying distribution $\mathbb{P}(X, Y)$.
(C4) We also generally do not have access to the underlying domain transformation model G.

In the ensuing sections, we explicitly address each of these challenges toward developing a tractable method for approximately solving Problem 4.6 with guarantees on optimality. In particular, we discuss challenges (C1), (C2), and (C3) in Section 5. We then discuss (C4) in Section F.

5 Data-dependent duality gap for MBDG

In this section, we offer a principled analysis of Problem 4.6. In particular, we first address (C1) by introducing a tight relaxation of the G-invariance constraint. Next, to resolve the fundamental difficulty involved in solving constrained statistical problems highlighted in (C2), we formulate the parameterized dual problem, which is unconstrained and thus more suitable for learning with deep neural networks. Finally, to address (C3), we introduce an empirical version of the parameterized dual problem and explicitly characterize the data-dependent duality gap between this problem and Problem 4.6. At a high level, this analysis results in an *unconstrained* optimization problem which is guaranteed to produce a solution that is close to the solution of Problem 3.1 (see Theorem 5.3). In this section, we have chosen to present our results somewhat informally by deferring preliminary results, regularity assumptions, and proofs to the appendices.

Addressing (C1) by relaxing the G-invariance constraint. One of the most fundamental challenges in solving Problem 4.6 is the difficulty of enforcing the G-invariance equality constraint. To alleviate some of this difficulty, we introduce the following relaxation of Problem 4.6:

$$P^\star(\gamma) \triangleq \underset{f\in\mathcal{F}}{\text{minimize}} \;\; R(f) \quad \text{s.t.} \quad \mathcal{L}^e(f) \triangleq \mathbb{E}_{\mathbb{P}(X)}\, d\big(f(X), f(G(X,e))\big) \leq \gamma \quad \forall e \in \mathcal{E}_{\text{all}} \tag{3}$$

where $\gamma > 0$ is a fixed margin the controls the extent to which we enforce G-invariance and $d : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_{\geq 0}$ is a distance metric over the space of probability distributions on $\mathcal{Y}$. While at first glance this problem may appear to be a significant relaxation of the MBDG optimization problem in (MBDG), when $\gamma = 0$ and under mild conditions on d, the two problems are equivalent in the sense that $P^\star(0) = P^\star$ (see Proposition B.1). Indeed, we note that the conditions we require on d are not restrictive, and include the KL-divergence and more generally the family of f-divergences. Moreover, when the margin γ is strictly larger than zero, under the assumption that the perturbation function $P^\star(\gamma)$ is L-Lipschitz continuous, we show in Remark B.2 that $|P^\star - P^\star(\gamma)| \leq L\gamma$, meaning that the gap between the problems is relatively small when γ is chosen to be small. In particular, when strong duality holds for (MBDG), this Lipschitz constant L is equal to the L^1 norm of the optimal dual variable for (MBDG) (see Remark B.4).

Addressing (C2) by formulating the parameterized dual problem. As written, the relaxation in (3) is an infinite-dimensional constrained optimization problem over a functional space $\mathcal{F}$ (e.g. L^2 or the space of continuous functions). Optimization in this infinite-dimensional function space is not tractable, and thus we follow the standard convention by leveraging a finite-dimensional parameterization of $\mathcal{F}$, such as the class of deep neural networks [134, 135]. The approximation power of such a parameterization can be captured in the following definition:

Definition 5.1 (ϵ-parameterization)**.** Let $\mathcal{H} \subseteq \mathbb{R}^p$ be a finite-dimensional parameter space. For $\epsilon > 0$, a function $\varphi : \mathcal{H} \times \mathcal{X} \to \mathcal{Y}$ is said to be an **ϵ-parameterization** of $\mathcal{F}$ if it holds that for each $f \in \mathcal{F}$, there exists a parameter $\theta \in \mathcal{H}$ such that $\mathbb{E}_{\mathbb{P}(X)} \left\|\varphi(\theta, x) - f(x)\right\|_\infty \leq \epsilon$.

The benefit of using such a parameterization is that optimization is generally more tractable in the parameterized space $\mathcal{A}_\epsilon := \{\varphi(\theta, \cdot) : \theta \in \mathcal{H}\} \subseteq \mathcal{F}$. However, typical parameterizations often lead to nonconvex problems, wherein methods such as SGD cannot guarantee constraint satisfaction. And while several heuristic algorithms have been designed to enforce constraints over common parametric classes [136–141], these approaches cannot provide guarantees on the underlying statistical problem of interest [142]. Thus, to provide guarantees on the underlying statistical problem in Problem 4.6, given an ϵ-parameterization φ of $\mathcal{F}$, we consider the following saddle-point problem:

$$D_\epsilon^\star(\gamma) \triangleq \underset{\lambda\in\mathcal{P}(\mathcal{E}_{\text{all}})}{\text{maximize}} \; \min_{\theta\in\mathcal{H}} \; R(\theta) + \int_{\mathcal{E}_{\text{all}}} \left[\mathcal{L}^e(\theta) - \gamma\right] \mathrm{d}\lambda(e). \tag{4}$$

where $\mathcal{P}(\mathcal{E}_{\text{all}})$ is the space of normalized probability distributions over $\mathcal{E}_{\text{all}}$ and $\lambda \in \mathcal{P}(\mathcal{E}_{\text{all}})$ is the (semi-infinite) dual variable. Here we have slightly abused notation to write $R(\theta) = R(\varphi(\theta, \cdot))$ and $\mathcal{L}^e(\theta) = \mathcal{L}^e(\varphi(\theta, \cdot))$. One can think of (4) as the dual problem to (3) solved over the parametric space $\mathcal{A}_\epsilon$. Notice that unlike Problem 4.6, the problem in (4) is *unconstrained*, making it much more amenable for optimization over the class of deep neural networks. Moreover, under mild conditions, the optimality gap between (3) and (4) can be explicitly bounded as follows:

Proposition 5.2 (Parameterization gap)**.** Let $\gamma > 0$ be given. Under mild regularity assumptions (see Assumption C.1 in Appendix C.3) on ℓ and d, there exists a small universal constant k such that

$$P^\star(\gamma) \leq D_\epsilon^\star(\gamma) \leq P^\star(\gamma) + \epsilon k \left(1 + \left\|\lambda_{\text{pert}}^\star\right\|_{L^1}\right), \tag{5}$$

where $\lambda_{\text{pert}}^\star$ is the optimal dual variable for a perturbed version of (3) in which the constraints are tightened to hold with margin $\gamma - k\epsilon$.

Algorithm 1 Model-Based Domain Generalization (MBDG)

```
Hyperparameters: Primal step size $\eta_p > 0$, dual step size $\eta_d \geq 0$, margin $\gamma > 0$
repeat
    for minibatch $\{(x_j, y_j)\}_{j=1}^m$ in training dataset do
        $\tilde{x}_j \leftarrow$ GenerateImage($x_j$) $\forall j \in [m]$            ▷ Generate model-based images
        $\text{distReg}(\theta) \leftarrow (1/m)\sum_{j=1}^m d(\varphi(\theta, x_j), \varphi(\theta, \tilde{x}_j))$    ▷ Calculate distance regularizer
        $\text{loss}(\theta) \leftarrow (1/m)\sum_{j=1}^m \ell(x_j, y_j; \varphi(\theta, \cdot))$        ▷ Calculate classification loss
        $\theta \leftarrow \theta - \eta_p \nabla_\theta[\,\text{loss}(\theta) + \lambda \cdot \text{distReg}(\theta)\,]$            ▷ Primal step for $\theta$
        $\lambda \leftarrow [\lambda + \eta_d\,(\text{distReg}(\theta) - \gamma)]_+$                 ▷ Dual step for $\lambda$
    end for
until convergence

procedure GenerateImage(x)
    Sample $e \sim \mathcal{N}(0, I)$                                     ▷ $e$ is a latent code for MUNIT
    return $G(x, e)$                                            ▷ Return image produced by MUNIT
end procedure
```

In this way, solving the parameterized dual problem in (4) provides a solution that can be used to recover a close approximation of the primal problem in (3). To see this, observe that Prop. 5.2 implies that $|D^\star_\epsilon(\gamma) - P^\star(\gamma)| \leq \epsilon k(1 + ||\lambda^\star_{\text{pert}}||_{L^1})$. This tells us that the gap between $P^\star(\gamma)$ and $D^\star_\epsilon(\gamma)$ is small when we use a tight ϵ-parameterization of $\mathcal{F}$.

Addressing (C3) by bounding the empirical duality gap. The parameterized dual problem in (4) gives us a principled way to address Problem 4.6 in the context of deep learning. However, complicating matters is the fact that we do not have access to the full distribution $\mathbb{P}(X, Y)$ or to data from any of the domains in $\mathcal{E}_{\text{all}}\backslash\mathcal{E}_{\text{train}}$. In practice, it is ubiquitous to solve optimization problems such as (4) over a finite sample of N data points drawn from $\mathbb{P}(X, Y)$[2]. More specifically, given $\{(x_j, y_j)\}_{j=1}^N$ drawn i.i.d. according to (X, Y), we consider the empirical counterpart of (4):

$$D^\star_{\epsilon,N,\mathcal{E}_{\text{train}}}(\gamma) \triangleq \underset{\lambda(e)\geq 0,\; e\in\mathcal{E}_{\text{train}}}{\text{maximize}} \; \min_{\theta\in\mathcal{H}} \; \hat{\Lambda}(\theta, \lambda) \triangleq \hat{R}(\theta) + \frac{1}{|\mathcal{E}_{\text{train}}|} \sum_{e\in\mathcal{E}_{\text{train}}} \left[\hat{\mathcal{L}}^e(\theta) - \gamma\right] \lambda(e) \tag{6}$$

where $\hat{R}(\theta)$ and $\hat{\mathcal{L}}^e(\theta)$ are the empirical counterparts of $R(f)$ and $\mathcal{L}(f)$. Notably, the duality gap between the solution to (6) and (MBDG) can be explicitly bounded as follows.

Theorem 5.3 (Data-dependent duality gap)**.** Let $\epsilon > 0$ be given, and let φ be an ϵ-parameterization of $\mathcal{F}$. Under mild regularity assumptions on ℓ and d and assuming that $\mathcal{A}_\epsilon$ has finite VC-dimension, with probability $1 - \delta$ over the N samples from $\mathbb{P}(X, Y)$ we have that

$$|P^\star - D^\star_{\epsilon,N,\mathcal{E}_{\text{train}}}(\gamma)| \leq L\gamma + \epsilon k\left(1 + \left\|\lambda^\star_{\text{pert}}\right\|_{L^1}\right) + \mathcal{O}\left(\sqrt{\log(N)/N}\right) \tag{7}$$

where L is the Lipschitz constant of $P^\star(\gamma)$ and k and $\lambda^\star_{\text{pert}}$ are as defined in Proposition 5.2.

The key message to take away from Theorem 5.3 is that given samples from $\mathbb{P}(X, Y)$, the duality gap incurred by solving the empirical problem in (6) is small when (a) the G-invariance margin γ is small, (b) the parametric space $\mathcal{A}_\epsilon$ is a close approximation of $\mathcal{F}$, and (c) we have access to sufficiently many samples. Thus, assuming that Assumptions 4.1 and 4.2 hold, the solution to Problem 3.1 is closely-approximated by the solution to the empirical, parameterized dual problem in (6).

6 MBDG: A principled algorithm for domain generalization

Motivated by these theoretical insights, we now introduce a new domain generalization algorithm which is widely applicable to problems with or without covariate shift. Our algorithm consists of two steps. First, we learn an approximation of the underlying domain transformation model $G(x, e)$ using the data from the training domains $\mathcal{E}_{\text{train}}$. Next, we leverage G toward solving the unconstrained dual optimization problem in (6) via a primal-dual iteration.

[2] Indeed, in practice we do not have access to any samples from $\mathbb{P}(X, Y)$. In Section 6, we argue that the N samples from $\mathbb{P}(X, Y)$ can be replaced by the $\sum_{e\in\mathcal{E}_{\text{train}}} n_e$ samples drawn from the training datasets $\mathcal{D}^e$.

Learning domain transformation models from data. Regarding challenge (C4), critical to our approach is having access to the domain transformation model G. For the vast majority of settings, the underlying function $G(x, e)$ is not known a priori and cannot be represented by a simple expression. For example, obtaining a closed-form expression for a model that captures the variation in coloration, brightness, and contrast in the dataset shown in Figure 1 would be very challenging. While in general it is impossible to learn the true underlying domain transformation model when one only has access to data from the training domains, we argue that a realistic *approximation* of the underlying model can be learned from this data. To this end, to learn a domain transformation model, we train multimodal image-to-image translation networks on the training data. These networks are designed to transform samples from one dataset so that they resemble a diverse collection of images from another dataset. In particular, in each of the experiments in Section 7, we use the MUNIT architecture introduced in [102] to parameterize learned domain transformation models. As shown in Figure 5 and in Appendix G, models trained using the MUNIT architecture learn accurate and diverse transformations of the training data, which often generalize to generate images from new domains.

Primal-dual iteration. Given a learned approximation $G(x, e)$ of the underlying domain transformation model, the next step in our procedure is to use a primal-dual iteration [143] toward solving (6) using the datasets $\mathcal{D}^e$. We note that while our theory calls for data drawn from $\mathbb{P}(X, Y)$, the G-invariance condition implies that when (6) is feasible, $\varphi(\theta, x) \approx \varphi(\theta, x^e)$ when $x \sim \mathbb{P}(X)$, $x^e \sim \mathbb{P}^e(X)$, and $x^e = G(x, e)$. Therefore, the data from $\cup_{e\in\mathcal{E}_{\text{train}}}\mathcal{D}^e$ is a useful proxy for data drawn from $\mathbb{P}(X, Y)$. As the outer maximization in (6) is a linear program in λ, the primal-dual iteration can be characterized by alternating between the following steps:

$$\theta^{(t+1)} \in \rho\text{-}\operatorname*{argmin}_{\theta\in\mathcal{H}} \hat{\Lambda}(\theta, \lambda^{(t)}) \qquad (8) \qquad \lambda^{(t+1)}(e) \leftarrow \left[\lambda^{(t)}(e) + \eta\left(\hat{\mathcal{L}}^e(\theta) - \gamma\right)\right]_+ \qquad (9)$$

Here $[\cdot]_+ = \max\{0, \cdot\}$, η is the dual step size, and $\rho\text{-}\operatorname{argmin}$ denotes a solution that is ρ-close to being a minimizer, i.e. we should that have $\hat{\Lambda}(\theta^{(t+1)}, \lambda^{(t)}) \leq \min_{\theta\in\mathcal{H}} \hat{\Lambda}(\theta, \lambda^{(t)}) + \rho$. We call (8) the primal step, and we call (9) the dual step. Furthermore, it can be shown that if this iteration is run for sufficiently many steps and with small enough step size, the iteration convergences with high probability to a solution which closely approximates the solution to Problem 4.6.

Theorem 6.1 (Primal-dual convergence)**.** Assuming that ℓ and d are $[0, B]$-bounded, $\mathcal{H}$ has finite VC-dimension, and under mild regularity conditions on (6), the primal-dual pair $(\theta^{(T)}, \lambda^{(T)})$ obtained after running the alternating primal-dual iteration in (8) and (9) for T steps with step size η, where

$$T = \left\lceil \frac{1}{2\eta\kappa} \right\rceil + 1 \qquad \text{and} \qquad \eta \leq \frac{2\kappa}{|\mathcal{E}_{\text{train}}|B^2} \qquad (10)$$

satisfies $|P^\star - \hat{\Lambda}(\theta^{(T)}, \mu^{(T)})| \leq K(\rho, \kappa, \epsilon) + \mathcal{O}(\sqrt{\log(N)/N})$. Here κ is a constant that captures the regularity of (6) (see Appendix C.6) and $K(\rho, \kappa, \epsilon)$ is a small constant depending on ρ, κ, and ϵ.

This means that by solving the empirical dual problem for sufficiently many steps, we can reach a solution that is close to solving the Model-Based Domain Generalization problem in Problem 4.6.

Implementation of MBDG. In practice, because (a) it may not be tractable to find a ρ-minimizer over $\mathcal{H}$ at each iteration and (b) there may be a large number of domains in $\mathcal{E}_{\text{train}}$, we propose two modifications of the primal-dual iteration in which we replace (8) with a stochastic gradient step and we use only one dual variable for all of the domains; we call this algorithm MBDG (see Algorithm 1). We provide results in Appendix E where one dual variable is used per training domain.

7 Experiments

We now evaluate the performance of MBDG on a range of standard domain generalization benchmarks. In the main text, we present results on `ColoredMNIST`, `Camelyon17-WILDS`, `FMoW-WILDS`, and `PACS`; we defer results on `VLCS` to the supplemental. For `ColoredMNIST`, `PACS`, and `VLCS`, we use the DomainBed package [46], facilitating comparison to a range of baselines. Model selection for each of these datasets was performed using hold-one-out cross-validation. For `Camelyon17-WILDS` and `FMoW-WILDS`, we use the repository provided with the WILDS dataset suite, and we perform model-selection using the out-of-distribution validation set provided in the WILDS repository. Further details concerning hyperparameter tuning and model selection are deferred to Appendix E.

Table 1: **ColoredMNIST.** We report accuracies for ColoredMNIST. Model-selection was performed via cross-validation.

Algorithm	**+90%**	**+80%**	**-90%**	**Avg**
ERM	50.0 ± 0.2	50.1 ± 0.2	10.0 ± 0.0	36.7
IRM	46.7 ± 2.4	51.2 ± 0.3	23.1 ± 10.7	40.3
GroupDRO	50.1 ± 0.5	50.0 ± 0.5	10.2 ± 0.1	36.8
Mixup	36.6 ± 10.9	53.4 ± 5.9	10.2 ± 0.1	33.4
MLDG	50.1 ± 0.6	50.1 ± 0.3	10.0 ± 0.1	36.7
CORAL	49.5 ± 0.0	59.5 ± 8.2	10.2 ± 0.1	39.7
MMD	50.3 ± 0.2	50.0 ± 0.4	9.9 ± 0.2	36.8
DANN	49.9 ± 0.1	62.1 ± 7.0	10.0 ± 0.1	40.7
CDANN	63.2 ± 10.1	44.4 ± 4.5	9.9 ± 0.2	39.1
MTL	44.3 ± 4.9	50.7 ± 0.0	10.1 ± 0.1	35.0
SagNet	49.9 ± 0.4	49.7 ± 0.3	10.0 ± 0.1	36.5
ARM	50.0 ± 0.3	50.1 ± 0.3	10.2 ± 0.0	36.8
VREx	50.2 ± 0.4	50.5 ± 0.5	10.1 ± 0.0	36.9
RSC	49.6 ± 0.3	49.7 ± 0.4	10.1 ± 0.0	36.5
MBDA	72.0 ± 0.1	50.7 ± 0.1	22.5 ± 0.0	48.3
MBDG-DA	72.7 ± 0.2	71.4 ± 0.1	33.2 ± 0.1	59.0
MBDG-Reg	73.3 ± 0.0	73.7 ± 0.0	27.2 ± 0.1	58.1
MBDG	**73.7** ± **0.1**	**68.4** ± **0.0**	**63.5** ± **0.0**	**68.5**

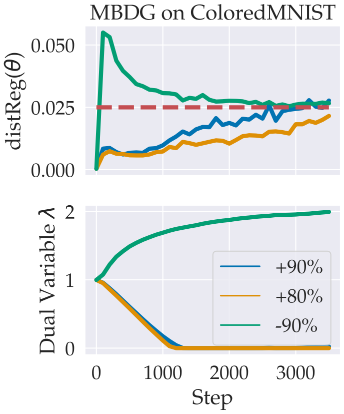


Figure 2: **Tracking dual variables.** We show the values of $\text{distReg}(\theta)$ and the dual variables λ for the each MBDG models in Table 1. The margin $\gamma = 0.025$ is shown in red.

7.1 ColoredMNIST

We first consider the ColoredMNIST dataset [10], which is a standard domain generalization benchmark created by colorizing subsets of the MNIST dataset [144]. This dataset contains three domains, each of which is characterized by a different level of correlation between the label and digit color. As shown in Table 1, the MBDG algorithm improves over each baseline by nearly 30%. To understand the reasons behind this improvement, we consider three ablation studies on ColoredMNIST.

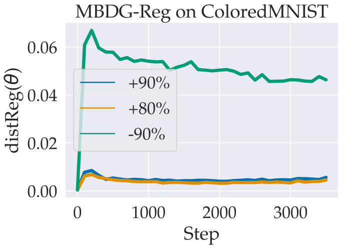


Figure 3: **Regularized MBDG.** We show the regularization value for each domain in ColoredMNIST for a fixed dual variable $\lambda = 1.0$.

Tracking the dual variables. For the three MBDG classifiers in Table 1, we plot the regularization term $\text{distReg}(\theta)$ and the corresponding dual variable at each training step in Figure 2. Observe that for the +90% and +80% domains, the dual variables decay to zero, as the constraint is satisfied early on in training. On the other hand, the constraint for the -90% domain is not satisfied early on in training, and in response, the dual variable increases, gradually forcing constraint satisfaction. As we shown in the next subsection, without the dual update step, the constraints may never be satisfied (see Figure 3).

Regularization vs. dual ascent. A common trick for encouraging constraint satisfaction is to introduce soft constraints by adding a regularizer multiplied by a fixed multiplier to the objective. While this approach yields a related problem to (6) (see Appendix B.4) where the dual variables are fixed, there are few formal guarantees for this approach. Moreover, we show in Table 1 that when the dual variable is fixed during training (MBDG-Reg in Table 1), the performance drops significantly vis-a-vis MBDG. Notice that relative to Figure 2, the value of $\text{distReg}(\theta)$ is much larger than the margin, meaning that the constraint is not being satisfied.

Ablation on data augmentation. To study the efficacy of the MBDG algorithm, we consider two natural alternatives MBDG: (1) ERM with data augmentation through the learned model $G(x, e)$ (MBDA); and (2) MBDG with data augmentation through $G(x, e)$ on the training objective (MBDG-DA). As shown at the bottom of Table 1, while these variants significantly outperform the baselines, they not perform nearly as well as MBDG. Thus, while data augmentation can in some cases improve performance, the primal-dual iteration is a much more effective tool for enforcing invariance.

7.2 Camelyon17-WILDS and FMoW-WILDS

We next consider the Camelyon17-WILDS and FMoW-WILDS datasets from the WILDS family of domain generalization baselines [20]. Table 2 shows that on Camelyon17-WILDS, MBDG improves by more than 20 percentage points over the state-of-the-art baselines.

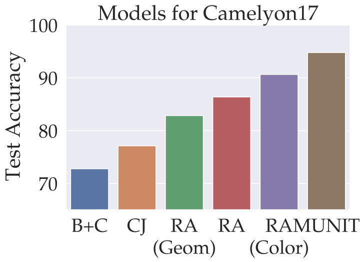


Figure 4: **Known vs. learned models $\mathbf{G(x,e)}$.** We compare the performance of MBDG for known models (first five columns) against a model that was trained with the data from the training domains using MUNIT.

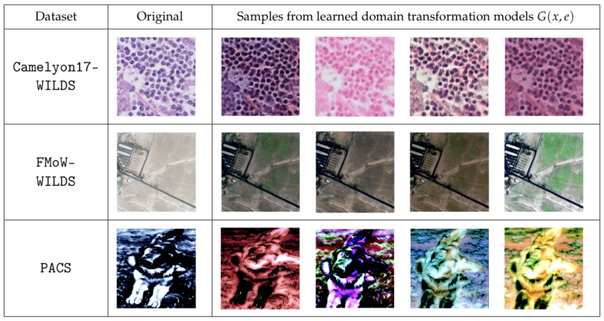


Figure 5: **Samples from learned models $\mathbf{G(x,e)}$.** We show samples from domain transformation models learned from the training data via the MUNIT architecture for `Camelyon17-WILDS`, `FMOW-WILDS`, and `PACS`.

Table 2: **WILDS datasets.** We report accuracies for `Camelyon17` and `FMoW`. For both datasets, we use the out-of-distribution validation set provided in the WILDS repository to perform model selection.

Algorithm	Camelyon17-WILDS	FMoW-WILDS
ERM	73.3 ± 9.9	51.3 (0.4)
IRM	60.9 ± 15.3	51.1 (0.4)
ARM	62.1 ± 6.4	47.9 (0.3)
CORAL	59.2 ± 15.1	49.6 (0.5)
MBDG	**94.8** ± **0.4**	**52.3** ± **0.5**

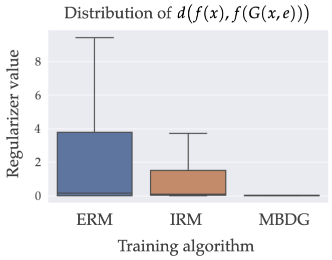


Figure 6: **Measuing invariance.** We measure the invariance of ERM, IRM, and MBDG to images generated by a model G learned for `Camelyon17-WILDS`.

Measuring G-invariance of trained classifiers. In Section 4, we restricted our attention predictors satisfying the G-invariance condition. To test whether our algorithm successfully enforces G-invariance when G is learned from data, we measure the distribution of $\text{distReg}(\theta)$ over all of the instances from the training domains of `Camelyon17-WILDS` for ERM, IRM, and MBDG. In Figure 6, observe that whereas MBDG is quite robust to changes under G, ERM and IRM are not nearly as robust. This property is key to the ability of MBDG to learn invariant representations across domains.

Ablation on learning models vs. data augmentation. Rather than learning G from data, a heuristic alternative is to replace the GENERATEIMAGE procedure in Algorithm 1 with standard data augmentation transformations. In Figure 4, we investigate this approach with five different forms of data augmentation: B+C (brightness and contrast), CJ (color jitter), and three variants of RandAugment [145] (RA, RA-Geom, and RA-Color). More details concerning these data augmentation schemes are given in Appendix E. The bars in Figure 4 show that although these schemes offer strong performance in our MBDG framework, the learned model trained using MUNIT performs best.

7.3 PACS

In this subsection, we highlight selected results for the `PACS` dataset. Due to spatial limitations, we report the top-performing baselines in the main text, and defer the full set of results to Appendix E. Notably, MBDG beats the current SOTA by nearly 2% when averaged over the four domains. Of note in Table 3 is the result on the "sketch" (S) subset, wherein MBDG improves by more than 10% over the baselines.

Table 3: **PACS.** We report classification accuracies for `PACS`. Model-selection was performed via cross-validation.

Algorithm	A	C	P	S	Avg
ERM	83.2 ± 1.3	76.8 ± 1.7	**97.2** ± **0.3**	74.8 ± 1.3	83.0
MTL	**85.6** ± 1.5	78.9 ± 0.6	97.1 ± 0.3	73.1 ± 2.7	83.7
RSC	83.7 ± 1.7	**82.9** ± **1.1**	95.6 ± 0.7	68.1 ± 1.5	82.6
MBDG	80.6 ± 1.1	79.3 ± 0.2	97.0 ± 0.4	**85.2** ± **0.2**	**85.6**

8 Acknowledgements and disclosure of funding

The authors would like to thank Samuel Sokota for his helpful feedback.

The research of Alexander Robey and Hamed Hassani is supported by NSF Grants 1837253, 1943064, 1934876, AFOSR Grant FA9550-20-1-0111, and DCIST-CRA. Alexander Robey and George J. Pappas are supported by the NSF-Simons Program on the Mathematics of Deep Learning, ARL CRA DCIST W911NF-17-2-0181, and NSF 2038873.

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]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
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]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All of the external code used for this project is open-sourced and freely-available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
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]

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. *arXiv preprint arXiv:1709.01889*, 2017.

[3] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[5] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[6] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. *arXiv preprint arXiv:2007.08558*, 2020.

[7] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[9] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

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

[11] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

[12] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

[13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[16] Eric Wong and J Zico Kolter. Provable Defenses Against Adversarial Examples Via the Convex Outer Adversarial Polytope. *arXiv preprint arXiv:1711.00851*, 2017.

[17] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

[18] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

[19] Nimit S Sohoni, Jared A Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *arXiv preprint arXiv:2011.12945*, 2020.

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[21] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[22] Alexander Robey, Hamed Hassani, and George J Pappas. Model-based robust deep learning. *arXiv preprint arXiv:2005.10247*, 2020.

[23] Eric Wong and J Zico Kolter. Learning perturbation sets for robust machine learning. *arXiv preprint arXiv:2007.08450*, 2020.

[24] Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotham, Timothy Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1211–1220, 2020.

[25] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[26] Li Yao, Jordan Prosky, Ben Covington, and Kevin Lyman. A strong baseline for domain adaptation and generalization in medical imaging. *arXiv preprint arXiv:1904.01638*, 2019.

[27] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *arXiv preprint arXiv:2009.12829*, 2020.

[28] Vishnu M Bashyam, Jimit Doshi, Guray Erus, Dhivya Srinivasan, Ahmed Abdulkadir, Mohamad Habes, Yong Fan, Colin L Masters, Paul Maruff, Chuanjun Zhuo, et al. Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging. *arXiv preprint arXiv:2010.05355*, 2020.

[29] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[30] Luona Yang, Xiaodan Liang, Tairui Wang, and Eric Xing. Real-to-virtual domain unification for end-to-end autonomous driving. In *Proceedings of the European conference on computer vision (ECCV)*, pages 530–545, 2018.

[31] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.

[32] Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning. *arXiv e-prints*, pages arXiv–2004, 2020.

[33] Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. *arXiv preprint arXiv:2006.01096*, 2020.

[34] Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[36] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[37] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24:2178–2186, 2011.

[38] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[39] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

[40] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *arXiv preprint arXiv:2007.02454*, 2, 2020.

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

[42] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[43] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[44] 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.

[45] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[46] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[47] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[48] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.

[49] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[50] Alexis Bellot and Mihaela van der Schaar. Accounting for unobserved confounding in domain generalization. *arXiv preprint arXiv:2007.10653*, 2020.

[51] 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.

[52] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H. Falk, and Ioannis Mitliagkas. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 2019.

[53] 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*, pages 5400–5409, 2018.

[54] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[55] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

[56] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020.

[57] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020.

[58] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 315–331. Springer, 2019.

[59] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020.

[60] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, pages 7728–7738. PMLR, 2020.

[61] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[62] 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)*, pages 624–639, 2018.

[63] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[64] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[65] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[66] 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.

[67] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*, pages 567–583. Springer, 2020.

[68] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[69] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[70] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[71] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. *arXiv preprint arXiv:2103.15796*, 2021.

[72] Aniket Anand Deshmukh, Yunwen Lei, Srinagesh Sharma, Urun Dogan, James W Cutler, and Clayton Scott. A generalization error bound for multi-class domain generalization. *arXiv preprint arXiv:1905.10392*, 2019.

[73] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[74] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31:998–1008, 2018.

[75] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *arXiv preprint arXiv:1910.13580*, 2019.

[76] 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*, pages 1446–1455, 2019.

[77] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. *arXiv preprint arXiv:2104.03620*, 2021.

[78] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2019.

[79] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020.

[80] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[81] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020.

[82] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Robust place categorization with deep domain generalization. *IEEE Robotics and Automation Letters*, 3(3):2093–2100, 2018.

[83] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357. IEEE, 2018.

[84] 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.

[85] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017.

[86] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[87] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[88] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

[89] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[90] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[91] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. Invariance reduces variance: Understanding data augmentation in deep learning and beyond. *arXiv preprint arXiv:1907.10905*, 2019.

[92] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

[93] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

[94] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020.

[95] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[96] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[97] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[98] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[99] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[100] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[101] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[102] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[103] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

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

[105] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020.

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

[107] Simon Vandenhende, Bert De Brabandere, Davy Neven, and Luc Van Gool. A three-player gan: generating hard samples to improve classification networks. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[108] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[109] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE, 2019.

[110] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019.

[111] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[112] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. *arXiv preprint arXiv:2104.05833*, 2021.

[113] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

[114] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *arXiv preprint arXiv:1910.11645*, 2019.

[115] Nader Asadi, Amir M Sarfi, Mehrdad Hosseinzadeh, Zahra Karimpour, and Mahdi Eftekhari. Towards shape biased unsupervised representation learning for domain generalization. *arXiv preprint arXiv:1909.08245*, 2019.

[116] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[117] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 87–97, 2016.

[118] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11749–11756, 2020.

[119] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2774–2783, 2015.

[120] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

[121] 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):151–175, 2010.

[122] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

[123] J Andrew Bagnell. Robust supervised learning. In *AAAI*, pages 714–719, 2005.

[124] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[125] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[126] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

[127] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[128] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

[129] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

[130] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[131] Pritish Kamath, Akilesh Tangella, Danica J Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? *arXiv preprint arXiv:2101.01134*, 2021.

[132] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

[133] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020.

[134] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[135] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[136] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

[137] Steven Chen, Kelsey Saulnier, Nikolay Atanasov, Daniel D Lee, Vijay Kumar, George J Pappas, and Manfred Morari. Approximating explicit model predictive control using constrained neural networks. In *2018 Annual American control conference (ACC)*, pages 1520–1527. IEEE, 2018.

[138] Thomas Frerix, Matthias Nießner, and Daniel Cremers. Homogeneous linear inequality constraints for neural network activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 748–749, 2020.

[139] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[140] Sathya N Ravi, Tuan Dinh, Vishnu Lokhande, and Vikas Singh. Constrained deep learning using conditional gradient and applications in computer vision. *arXiv preprint arXiv:1803.06453*, 2018.

[141] Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. *arXiv preprint arXiv:2104.12225*, 2021.

[142] Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. The empirical duality gap of constrained statistical learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8374–8378. IEEE, 2020.

[143] Dimitri P Bertsekas and Athena Scientific. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

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

[145] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[146] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[147] Richard F Bass. *Real analysis for graduate students*. Createspace Ind Pub, 2013.

[148] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[149] Miguel A Goberna and MA López. Recent contributions to linear semi-infinite optimization. *4OR*, 15(3):221–264, 2017.

[150] Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[151] Elias M Stein and Rami Shakarchi. *Functional analysis: introduction to further topics in analysis*, volume 4. Princeton University Press, 2011.

[152] Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *arXiv preprint arXiv:2103.05134*, 2021.

[153] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[154] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[155] 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.

[156] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 783–790, 2018.

[157] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.