

Towards Principled Disentanglement for Domain Generalization

Hanlin Zhang^{1,*} Yi-Fan Zhang^{2,*} Weiyang Liu^{3,4} Adrian Weller^{3,5} Bernhard Schölkopf⁴ Eric P. Xing^{1,6}

¹Carnegie Mellon University ²Chinese Academy of Science ³University of Cambridge

⁴Max Planck Institute for Intelligent Systems, Tübingen ⁵Alan Turing Institute ⁶MBZUAI *Equal Contribution

Abstract

A fundamental challenge for machine learning models is generalizing to out-of-distribution (OOD) data, in part due to spurious correlations. To tackle this challenge, we first formalize the OOD generalization problem as constrained optimization, called **Disentanglement-constrained Domain Generalization (DDG)**. We relax this non-trivial constrained optimization problem to a tractable form with finite-dimensional parameterization and empirical approximation. Then a theoretical analysis of the extent to which the above transformations deviates from the original problem is provided. Based on the transformation, we propose a primal-dual algorithm for joint representation disentanglement and domain generalization. In contrast to traditional approaches based on domain adversarial training and domain labels, DDG jointly learns semantic and variation encoders for disentanglement, enabling flexible manipulation and augmentation on training data. DDG aims to learn intrinsic representations of semantic concepts that are invariant to nuisance factors and generalizable across domains. Comprehensive experiments on popular benchmarks show that DDG can achieve competitive OOD performance and uncover interpretable salient structures within data.

1. Introduction

Learning representations that can reflect intrinsic class semantics and also render strong invariance to cross-domain variation is of great significance to robustness and generalization in deep learning. Despite being empirically effective on many visual recognition benchmarks [65], modern neural networks are still prone to learning shortcuts that stem from spurious correlations [24], resulting in poor out-of-distribution (OOD) generalization. To tackle this challenge, domain generalization (DG) has emerged as an increasingly important task where the goal is to learn invariant representations over source domains that are generalizable to distributions different from those seen during training [55].

In order to improve OOD generalization, efforts have

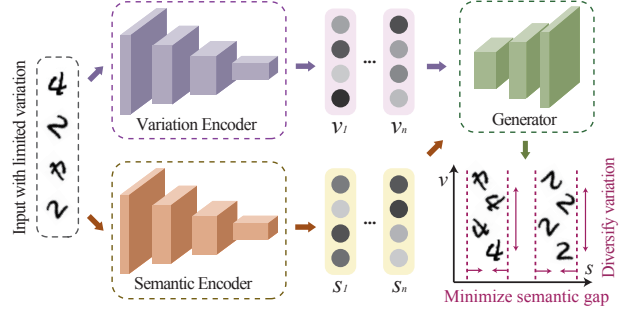


Figure 1. An illustration of DDG based on disentanglement of digit labels (semantics) and rotated angles (variation across domains). DDG seeks to minimize the semantic difference of the generated samples from the same class while diversifying the variation across source domains.

been made from a diverse set of directions, such as domain adaptation [5, 6, 23, 84], self-supervised learning [12, 29], causal inference [56, 61, 67, 81], invariant risk regularization [3, 42, 51], angular alignment regularization [18, 48, 49], distributionally robust optimization [7] and data augmentation [4, 77, 82]. However, having no access to target domain data poses great challenges. Two main lines of research seek to address this. First, training domain labels are assumed to be available in [3, 23, 47, 60, 66] such that the divergence to different domains can be minimized. However, these domain labels are often impractical or prohibitively expensive to obtain [28]. Moreover, it is non-trivial to minimize domain divergence with domain adversarial training which is notoriously hard to converge [63]. The second line of works try to model the cross-domain distribution shifts and capture the semantic invariance [37, 62, 78]. However, it has been found in [27] that this goal can be very difficult to achieve. What makes the problem even more challenging is the inconsistency of the evaluation protocol. Surprisingly, [27] shows that even the standard empirical risk minimization could outperform many recently proposed models under certain conditions. Motivated by these challenges, we aim to disentangle variations and semantics in a principled way, and then verify the effectiveness of our method under a consistent evaluation protocol [27].

A key desideratum for DG is to ensure the invariance of learned representations to all possible inter-class variations.

Therefore, our intuition is to first diversify the inter-class variation by modeling potential seen or unseen variations, and then minimize the discrepancy of the inter-class variation on a representation space where the target is to predict semantic labels. To this end, we first formalize distribution shifts and invariance based on disentanglement. Concretely, we formulate the disentanglement between class semantics and both intra- and inter-domain variations as constraints to the DG problem. Then we propose a novel framework called **Disentanglement-constrained Domain Generalization (DDG)**. An illustration of DDG is given in Fig. 1. When the semantic (i.e. the labels of digits) and variation factors (i.e. rotating angles $-60^\circ, 0^\circ, 60^\circ, 120^\circ$) are well disentangled, we expect that learned representations can be effectively constrained to be invariant to inter-class variation. In order to achieve such a non-trivial goal, we first derive a constrained optimization problem and then propose a principled algorithm based on primal-dual iterations to solve it. To understand how well the transformed solution approximates the solution to the original problem, we provide comprehensive theoretical guarantees for the parameterization gap and empirical gap. We also verify the empirical effectiveness of DDG by showing that it can consistently outperform current popular DG methods by a considerable margin.

As a useful side product, DDG simultaneously obtains an automated, domain-agnostic data augmentation network based on learned disentangled representations. This requires no usage of domain-specific knowledge or gradient estimation [4, 77]. The intuition why such a data augmentation network is useful comes from the fact that the learned variation encoder can well approximate some of the intrinsic intra- and inter-domain variations. It also serves as *feature removal* since more training examples augmented by specific variation factors lead to more invariant representations for those variations. Moreover, the increased diversity of source domain data improves the likelihood that an unseen distribution lies within the convex hull of source domains [2]. For example, in Fig. 1, the original dataset can be augmented via a learned manipulator by composing a diverse combination of semantic and variation factors. Such a disentanglement can be a good predictor for OOD generalization according to [21]. We highlight the following advantages of DDG:

- DDG adopts a principled constrained learning formulation based on disentanglement, yielding rigorous theoretical guarantees on the empirical duality gap.
- Our algorithm is conceptually simple yet effective. DDG promotes semantic invariance via a constrained optimization setup. This is done without the usage of adversarial training and domain labels. Moreover, there is no additional computational overhead for modeling variations.
- Our framework can be viewed as as a controllable and interpretable data generation paradigm for DG. Data manip-

ulation under domain transformation can be challenging under settings where domain-specific signals like image styles vary greatly across domains, constituting more complicated superficial variation factors. Yet DDG can uncover salient structure within data by imposing constraints on the semantic and variation factors.

- Comprehensive experiments are conducted under a consistent evaluation protocol to verify the effectiveness of DDG. We show that DDG is able to produce interpretable qualitative results and achieve competitive performance on a number of challenging DG benchmarks including RotatedMNIST, VLCS, PACS and WILDS.

2. Related Work

Domain Generalization. Domain/Out-of-distribution generalization [55] aims to learn representations that are invariant across domains so that the model can extrapolate well in unseen domains. Invariant Risk Minimization (IRM) [3], which extends [56], and its variants [1, 42, 51] are proposed to tackle this challenge. However, IRM entails challenging bi-level optimization and can fail catastrophically unless the test data are sufficiently similar to the training distribution [62]. DG via domain alignment [2, 18, 55] aims to minimize the difference between source domains for learning domain-invariant representations. The motivation is straightforward: features that are invariant to the source domain shift should also be robust to any unseen target domain shift. The main difference is that we propose to learn invariant representations by reconstructing images from various domains and class semantics to simulate variations and minimize domain divergence. PAC constrained learning [15, 16] is adopted for modeling cross-domain variations under domain transformation in MBDG [60]. We highlight several major differences between our approach and MBDG below: (1) DDG imposes weaker assumptions; (2) MBDG consumes additional domain labels, which are often hard to obtain in many applications, while DDG does not; (3) DDG enforces invariance constraints via parameterizing semantic and variation encoders, which does not belong to a model-based approach. In contrast, MBDG requires a pre-trained domain transformation model (e.g., CycleGAN) during training. Appendix D provides the detailed comparison to MBDG.

Disentangled Representation Learning. The goal of disentangled representation learning is to model distinct and explanatory factors of variation in the data [8, 68]. [21] shows that disentanglement is a good predictor for out-of-distribution (OOD) tasks. [64] proposes to disentangle the semantic latent variables and the domain latent variables for stronger generalization performance in domain adaptation. [54] shows that existing disentangled learning models are not sufficient to support compositional generalization and extrapolation while hypothesizing that the richness of the training domain matters more. However, previous works

[19, 31, 39] are limited to single-dimensional latent codes and developed with different purposes like generation and interpretability. Thus they are hard to scale well beyond toy datasets and adapt to complicated DG tasks [54]. In contrast, we harness the disentangled effects to learn invariant representations for realistic OOD generalization tasks.

Data Augmentation. The diversity of the training distribution is of great importance in improving DG performance [2, 27, 29, 75, 82]. Data augmentation is an effective way to increase data diversity [79] and it can therefore improve OOD generalization as well as robustness to spurious correlations [4, 38]. In particular, [14] devises an active learning scheme to learn causal manipulations on images, which enriches the dataset from observational data and improves generalization on both causal and predictive learning tasks. In contrast, DDG seeks to learn underlying causal features by approximating the data manipulation function. This is done without a task-specific metric to differentiate the augmented data and the oracle. Our work introduces a simple yet effective approach for augmenting training data, which reinforces the importance of data diversity in DG.

Fairness. Fairness research [22, 28, 52] aims to develop a model that performs well under group assignments according to some fairness criterion for addressing the underperformance in minority subgroups. Learning fair representations can be naturally translated to a constrained optimization problem [15, 16, 43]. There are also exchanging lessons between algorithmic fairness and domain generalization [20], showing that both fields are optimizing similar statistics for common goals. DDG well aligns with the formulation and goals of *fairness without demographics* [28] and has the potential to improve context-specific fairness without prior knowledge about domains or demographics.

3. Disentanglement-constrained Optimization for Domain Generalization

Notations. We consider a classification problem from feature space $\mathcal{X} \in \mathbb{R}^d$ to label space $\mathcal{Y} \in \{0, 1\}$ where $(X, Y) \sim \mathbb{P}(X, Y)$. The infinite-dimensional functional space and finite-dimensional hypothesis space are denoted as \mathcal{F} and $\mathcal{H} \subseteq \mathbb{R}^p$, respectively. The parameterized latent spaces for semantic and variation factors are denoted as \mathcal{S} and \mathcal{V} , respectively. $\tilde{\mathbf{x}}$ denotes a different sample with \mathbf{x} from the training distribution \mathbb{P} . $d(\cdot, \cdot)$ denotes a distance metric over $\mathcal{X} \times \mathcal{X}$.

Problem setting. Suppose we observe a dataset denoted as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathcal{D}_{data}$ where (\mathbf{x}_i, y_i) is a realization of random vector (X, Y) with support $(\mathcal{X} \times \mathcal{Y})$. We consider a set of domains $\{d_i\}_{i=1}^{n_d} \subset \mathcal{D}$ of size n_d , where each domain corresponds to a distinct data distribution D^{d_i} over some input and label space. The set of domains \mathcal{D} is partitioned into multiple training domains $\mathcal{D}_S \subset \mathcal{D}$ and a test domain $D_U \in \mathcal{D}$ which is inaccessible during training.

Our formulation generally follows the prior work of PAC constrained learning [15, 16, 58, 60], but we use a more flexible parameterization and derive a new algorithm to solve the resulting constrained optimization problem. More specifically, we emphasize that DDG, motivated by an analysis on the multi-source domain adaptation upper bound (Appendix B), requires no domain labels and pre-trained domain transformation models during training. DDG can also be trained in an end-to-end manner, yielding a more flexible and potentially better solution.

3.1. Formulation

The basic idea of DDG is to learn disentangled representations by imposing invariant constraints in the semantic space \mathcal{S} and variation space \mathcal{V} . Such a disentanglement can also be applicable for augmenting the training data so that the learned representations can be more invariant to both inter- and intra-domain variations. To formalize this, we begin by introducing some necessary definitions and assumptions.

Definition 1 (*Invariance based on disentanglement*). Given a decoder $D: \mathcal{S} \times \mathcal{V} \rightarrow \mathcal{X}$, a semantic featurizer f_s is invariant if for all domains $d_i \in \mathcal{D}$ and a variation featurizer f_v , $\mathbf{x} = D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))$ holds almost surely when $\mathbf{x}, \tilde{\mathbf{x}} \sim P(X)$.

The property enforces the invariance of the original input \mathbf{x} and the one $D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))$ that reconstructs jointly from semantic and variation latent spaces when semantic factors remain constant while variation factors vary.

Assumption 1 (*Domain shift based on disentanglement*). Denote $f_s(\mathbf{x}; \theta)$ as the semantic factor of input \mathbf{x} and $f_v(\tilde{\mathbf{x}}; \phi)$ as the variation factor of any other one $\tilde{\mathbf{x}}$. Similar to the covariate shift assumption [71], we assume the domain/distribution shift stems from the variation of the marginal distribution $P(X)$ and the following invariance condition holds with the proposed f_s, f_v and D

$$P(Y = y | X = \mathbf{x}) = P(Y = y | X = D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))). \quad (1)$$

This assumption shows that the prediction depends only on the semantic factor $f_s(\mathbf{x}; \theta)$ regardless of the variation one $f_v(\tilde{\mathbf{x}}; \phi)$. It also subsumes as a special case the domain shifts based on domain labels, i.e., $P(Y = y | X = \mathbf{x}) = P(Y^d = y | X^d = G(X = \mathbf{x}, d))$ given a domain transformation model [59, 60], since our variation factors includes both inter- and intra-domain variations.

Note that the above assumptions follows the formulation in [60]. To elaborate this, we introduce the notion of invariance based on a decoder D taking as input the disentanglement results $f_s(\mathbf{x}; \theta)$ and $f_v(\tilde{\mathbf{x}}; \phi)$. In practice, D can be parameterized as a pre-trained model or a trainable component D_ψ updated in the primal step as in our implementation.

Assumption 2 (Regularity conditions) The loss function ℓ and distance metric d are convex, non-negative, B -bounded. ℓ is a L_ℓ -Lipschitz function, the distance metric is also a L_d -Lipschitz function.

Assumption 3 (Feasibility) There exist semantic and variation featurizers $f_s, f_v \in \mathcal{F}$ such that $\mathcal{L}_{con}(f_s, f_v) < \gamma - \max\{L_\ell \epsilon_s, L_d \epsilon_g\} = \gamma - m$ with ϵ -parameterization.

Definition 2 (Domain generalization problem). Similar to prior works [60, 66, 72, 78], we formulate domain generalization as a minimax optimization problem, optimizing the worst-domain risk over the entire family of domains \mathcal{D}

$$\min_{f_s \in \mathcal{F}} \max_{d \in \mathcal{D}} \mathbb{E}_{\mathbb{P}(X, Y)} \ell(f_s(D(X, d)), Y). \quad (2)$$

The above formulation (2) requires the availability of domain labels and is hard to optimize. However, the domain labels are expensive or even impossible to obtain in part due to privacy and fairness issues [28]. Therefore, under the disentanglement-based invariance and domain shift assumptions, we constrain the model to be invariant with respect to variation factors, then the problem is converted to an inequality-constrained optimization problem:

Definition 3 (Constrained domain generalization problem) Given a fixed margin $\gamma > 0$, with Assumption 3 and enforcing the invariance on the semantic featurizer f_s , we transform the vanilla formulation Eq. (2) to the following inequality-constrained optimization

$$\begin{aligned} \mathcal{P}^* &\triangleq \min_{f_s \in \mathcal{F}} \mathcal{L}(f_s) \triangleq \mathbb{E}_{\mathbb{P}(X, Y)} \ell(f_s(X), Y), \\ \text{s.t. } d(\mathbf{x}, D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))) &\leq \gamma, \text{ a.e. } \mathbf{x}, \tilde{\mathbf{x}} \sim \mathbb{P}(X). \end{aligned} \quad (3)$$

One intriguing property of Eq. (3) is that learning with inequality constraints¹ does not produce additional sample complexity overhead under some regularity conditions on the loss function ℓ [15]. However, it is difficult to satisfy the strictness and provide theoretical guarantees for learning in practical cases. In the following section, with the parameterization and saddle-point condition, we can relax the invariant constraint and obtain a version that is amenable to a provable PAC learning framework.

3.2. Parameterization

We first discuss how to parameterize the learnable components in DDG. The DG problem (Eq. (3)) yields an infinite dimensional optimization. A *de facto* way to enable tractable optimization is using finite-dimensional parameterization of \mathcal{F} like neural networks [33] or reproducing kernel Hilbert spaces (RKHS) [9]. To further discuss the parameterization gap, we formalize the approximation power of such parameterization by the following definition of ϵ -parameterization.

¹We simply enforce one constraint based on prior knowledge about semantic invariance and leave other design choices as future work.

Definition 4 (ϵ -parameterization) Let $\mathcal{H} \subseteq \mathbb{R}^p$ be a finite-dimensional parameter space. For $\epsilon > 0$, a function $h : \mathcal{H} \times \mathcal{X} \rightarrow \mathcal{Y}$ is an ϵ -parameterization of \mathcal{F} if for each $f_s, f_v \in \mathcal{F}$, there exist parameters $\theta, \phi \in \mathcal{H}$ such that

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim P(X)} \|h_s(\mathbf{x}; \theta) - f_s(\mathbf{x})\|_\infty &\leq \epsilon_s, \\ \mathbb{E}_{\mathbf{x} \sim P(X)} \|h_v(\mathbf{x}; \phi) - f_v(\mathbf{x})\|_\infty &\leq \epsilon_v, \\ \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim P(X)} \|D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi)) - D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}}))\|_2 &\leq \epsilon_g. \end{aligned}$$

With the help of ϵ -parameterization, tractable optimization can be performed over finite-dimensional parameterized space. Note that a regularity condition about D is introduced to allow DDG to faithfully reconstruct inputs under finite-dimensional parameterization. With the above formulation and to provide guarantees for the DG problem, we consider a corresponding saddle-point problem as follows:

$$\mathcal{D}_\epsilon^*(\gamma) \triangleq \max_{\lambda} \min_{\theta, \phi \in \mathcal{H}} \mathcal{L}(\theta) + \lambda \mathcal{L}_{con}(\theta, \phi), \quad (4)$$

where the constraint-related risk is defined as

$$\mathcal{L}_{con}(\theta, \phi) = \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim P(X)} [d(\mathbf{x}, D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi))) - \gamma].$$

The challenge for the parameterized problem (4) is the unaccessability of the ground truth data distribution $\mathbb{P}(X, Y)$. To address it, we resort to a corresponding empirical dual problem using finite n empirical training samples:

$$\begin{aligned} \mathcal{D}_{\epsilon, n}^*(\gamma) &\triangleq \max_{\lambda} \min_{\theta, \phi \in \mathcal{H}} L(\theta, \phi, \gamma) \triangleq \hat{\mathcal{L}}(\theta) + \lambda \hat{\mathcal{L}}_{con}(\theta, \phi) \\ &= \max_{\lambda} \min_{\theta, \phi \in \mathcal{H}} \sum_{i=1}^n \ell(f_s(\mathbf{x}_i), y_i) + \\ &\quad \lambda \sum_{i=1}^n \sum_{j \neq i}^n [d(\mathbf{x}_i, D(h_s(\mathbf{x}_i; \theta), h_v(\mathbf{x}_j; \phi))) - \gamma], \end{aligned} \quad (5)$$

which gives us the final optimization objective for DDG. Compared to the previous optimization problems, this is much easier and more tractable to solve.

3.3. Algorithm

Motivated by the above analysis, we use a primal-dual algorithm for efficient optimization [15, 17, 58, 60]. The algorithm alternates between optimizing θ (and/or ϕ) via minimizing the empirical Lagrangian with fixed dual variable λ and updating the dual variable according to the minimizer:

$$\begin{aligned} \theta^{(t+1)} &\leftarrow \arg \min_{\theta} L(\theta^{(t)}, \phi^{(t)}, \gamma) + \rho, \\ \phi^{(t+1)} &\leftarrow \arg \min_{\phi} L(\theta^{(t)}, \phi^{(t)}, \gamma) + \rho, \\ \lambda^{(t+1)} &\leftarrow \max \left\{ \left[\lambda^{(t)} + \eta_2 \left(\hat{\mathcal{L}}_{con} - \gamma \right) \right], 0 \right\}, \end{aligned} \quad (6)$$

where the η_2 denotes the learning rate of the dual step.

The primal-dual iteration has clear advantages over stochastic gradient descent in solving constrained optimization problems. Specifically, it avoids introducing extra balancing hyperparameters. Moreover, it provides convergence

Algorithm 1: DDG: Disentanglement-constrained Optimization for Domain Generalization

Input: $\mathcal{D}_S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, batch size B , primal and dual learning rate η_1, η_2 , Adam hyperparameters β_1, β_2 , initial coefficients λ , margin γ

Initial: Parameters of DDG (i.e. parameters θ, ϕ and ψ for semantic encoder h_s , variation encoder h_v and decoder D .)

```

repeat
  for  $i, j = 1, \dots, B, i \neq j$  do
     $\mathcal{L}_{\text{con}}^i = \max \left\{ \|\mathbf{x}_i - D(h_s(\mathbf{x}_i) \oplus h_v(\mathbf{x}_j))\|_{l_1} - \gamma, 0 \right\}$ 
     $\mathcal{L}_{\text{ERM}}^i = \ell(h_s(\mathbf{x}_i), y_i)$ 
     $\mathcal{L}_i = \mathcal{L}_{\text{ERM}}^i + \lambda \mathcal{L}_{\text{con}}^i$ 

    if Data Augmentation then
       $\mathbf{x}^* = D(h_s(\mathbf{x}_i) \oplus h_v(\mathbf{x}_j))$ 
       $\mathcal{L}_{\text{ERM}}^* = \ell(h_s(\mathbf{x}^*), y_i)$ 
       $\mathcal{L}_i = \mathcal{L}_i + \mathcal{L}_{\text{ERM}}^*$ 
    end
  end
  Primal step
   $\theta \leftarrow \text{Adam} \left( \frac{1}{B} \sum_{i=1}^B \mathcal{L}_i, \theta, \eta_1, \beta_1, \beta_2 \right)$ 
   $\phi \leftarrow \text{Adam} \left( \frac{1}{B} \sum_{i=1}^B \mathcal{L}_{\text{con}}^i, \phi, \eta_1, \beta_1, \beta_2 \right)$ 
  if Training  $D$  then
     $\psi \leftarrow \text{Adam} \left( \frac{1}{B} \sum_{i=1}^B \mathcal{L}_{\text{con}}^i, \psi, \eta_1, \beta_1, \beta_2 \right)$ 
  end
  Dual step
   $\lambda \leftarrow \max \left\{ \left[ \lambda + \eta_2 \left( \frac{1}{B} \sum_{i=1}^B \mathcal{L}_{\text{con}}^i - \gamma \right) \right], 0 \right\}$ 
until  $\theta$  is converged or  $\mathcal{D}_S = \emptyset$ ;

```

guarantees once we have sufficient iterations and a sufficiently small step size. We refer readers to [17, 60] for more in-depth and complete discussions of related conditions and convergence bounds.

One intriguing property of disentanglement is that it can be applicable for augmenting the training data. Based on this, DDG approximates a manipulator function by learning “hard” data points from fictitious target distributions for promoting invariance and improving generalization.

We give the detailed procedure of our DDG learning algorithm in Algorithm 1, where ℓ denotes the cross entropy loss and \oplus denotes the concatenation in the implementation. We also use l_1 norm as the distance metric d in the experiments.

3.4. Theoretical Insights and Guarantees

In this subsection, we provide a comprehensive analysis on the statistical guarantees of our solution (Eq. (5)). In order

to derive the bound on the final empirical duality gap, we start by proving two lemmas on the parameterization gap and empirical gap. Specifically, they elaborate the corresponding approximation gaps of two transformations (i.e. Eq. (4) and (5)) in above sections.

We first discuss the gap between the finite-dimensional model parameterization (e.g. neural networks) and the model over infinite functional space \mathcal{F} .

Lemma 1 (*Parameterization gap*) With Assumption 2 about ℓ and d , the gap between optimum of a statistical problem and its finite dimensional, deterministic version $\mathcal{D}_\epsilon^*(\gamma) - \mathcal{P}^*$ can be bounded as

$$0 \leq \mathcal{D}_\epsilon^*(\gamma) - \mathcal{P}^* \leq (1 + |\lambda_p^*|) \max \{L_\ell \epsilon_s, L_d \epsilon_g\}, \quad (7)$$

where λ_p^* is the dual variable with a tighter constraint $\gamma - \max \{L_\ell \epsilon_s, L_d \epsilon_g\}$ in Eq. (3).

The upper bound indicates that the parameterization gap is dominated by both the semantic function parameterization and the reconstruction-based transformation on perturbed inputs, which makes intuitive senses and also emphasizes the important role of disentanglement.

Then we compare Eq. (7) to the parameterization gap of MBDG which is shown as follows:

$$0 \leq \mathcal{D}_\epsilon^*(\gamma) - \mathcal{P}^* \leq (1 + |\lambda_p^*|) \max \{L_\ell, L_d\} \epsilon_s, \quad (8)$$

and the parameterization gap in [16]:

$$0 \leq \mathcal{D}_\epsilon^*(\gamma) - \mathcal{P}^* \leq (1 + |\lambda_p^*|) L_\ell \epsilon_s. \quad (9)$$

From the comparison, we notice that our formulation and analysis are closely connected to Proposition 1 (with $m = 1$) in [16]. We shall also see that in a perfect case where representations are well disentangled, i.e., $\epsilon_g \rightarrow 0$, our bound will become $(1 + |\lambda_p^*|) L_\ell \epsilon_s$. We note that this bound is strictly tighter than that in Eq. (8).

In practice, we approximate the expectation by its empirical average. By the classical VC-dimension bound, the following bound on the empirical gap holds:

Lemma 2 (*Empirical gap*) Denote d_{VC} as the VC-dimension of the hypothesis class \mathcal{H}_θ . Assume that ℓ and d obey the regularity condition in Assumption 2. Then given n samples, with probability $1 - \delta$, we can upper bound the deviation $|\mathcal{D}_\epsilon^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)|$ with

$$|\mathcal{D}_\epsilon^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| \leq 2B \sqrt{\frac{1}{n} \left[1 + \log \left(\frac{4(2n)^{d_{VC}}}{\delta} \right) \right]}. \quad (10)$$

With the above heavy lifting, we start deriving the empirical duality gap, which is our ultimate goal of the theoretical analysis. The empirical duality gap includes the above two components. Combining the above bounds on two gaps, we can bound the deviation between \mathcal{P}^* and $\mathcal{D}_{\epsilon,n}^*(\gamma)$ (i.e., $|\mathcal{P}^* - \mathcal{D}_{\epsilon,n}^*(\gamma)|$) under some mild conditions.

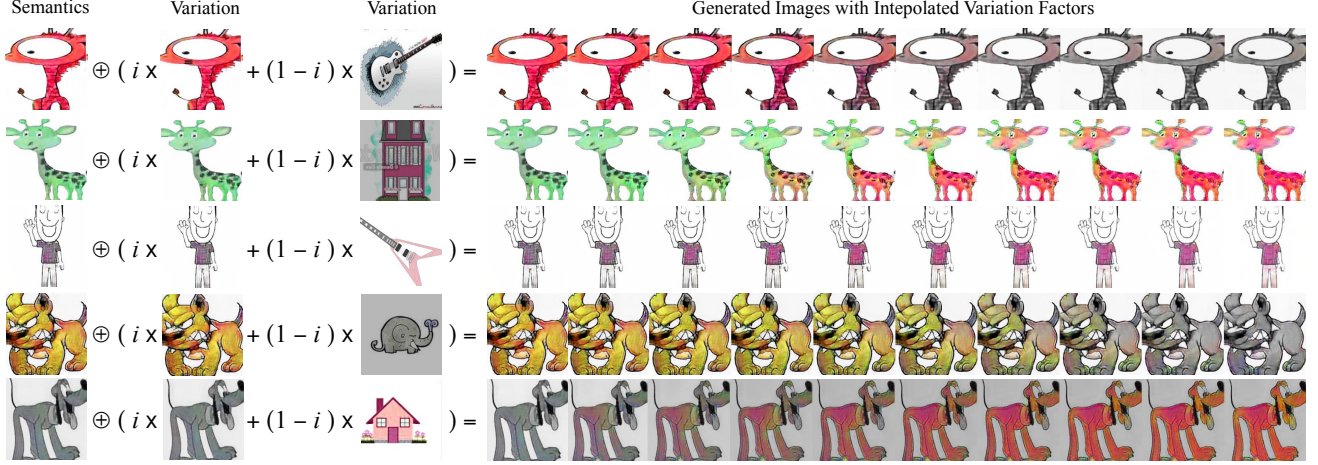


Figure 2. Interpolation disentanglement results. Different proportions of variation factors are mixed to generate the image by varying $i \in \{1.0, 0.9, \dots, 0.1\}$.

Theorem 1 (*Empirical duality gap*) When Assumption 2 holds, by denoting $\max \{L_{\ell \in s}, L_{d \in g}\}$ as m , we have

$$|p^* - \mathcal{D}_{\epsilon, n}^*(\gamma)| \leq L_P \gamma + (1 + |\lambda|)m + \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right). \quad (11)$$

The final bound tells us that the quality of the empirical, dual approximation of the primal problem is determined by the sample size, the hardness of learning problem and the richness of parameterization. The proof can be easily shown using triangle inequality as in Appendix A.3. As suggested by Theorem 1, we can improve the performance of our algorithm by using neural networks with larger capacity or training our model with more data.

4. Experiments

Datasets. We consider the following four datasets: Rotated MNIST [25], PACS [45], VLCS [74] and WILDS [41] to evaluate DDG against previous methods. We include the visualization of datasets in Appendix C.2.

Rotated MNIST [25] consists of 10,000 digits in MNIST with different rotated angles d such that each domain is determined by the degree $d \in \{0, 15, 30, 45, 60, 75\}$.

PACS [45] includes 9,991 images with 7 classes $y \in \{\text{dog, elephant, giraffe, guitar, horse, house, person}\}$ from 4 domains $d \in \{\text{art, cartoons, photos, sketches}\}$.

VLCS [74] is composed of 10,729 images, 5 classes $y \in \{\text{bird, car, chair, dog, person}\}$ from domains $d \in \{\text{Caltech101, LabelMe, SUN09, VOC2007}\}$.

Camelyon17-WILDS [11, 41] is about tumor detection in tissues. This dataset is composed of 455,954 images from 5 different medical centers/domains in total, which defines a binary classification problem about whether the patch image contains a tumor or not.

Baselines. We compare our model with ERM [76], IRM

[3], GDRO [66], Mixup [80], MLDG [44], CORAL [73], MMD [46], DANN [23], CDANN [47], AugMix [30].

All the baselines in DG tasks are implemented using the codebase of Domainbed [27]. We adapt AugMix for ablation studies using the official implementations as indicated in [30]. The two-sample classifier, implemented as a RBF kernel SVM using Scikit-learn, is used for calculating the generalization error for \mathcal{A} -distance.

Hyperparameter search. Following the experimental settings in [27], we conduct a random search of 20 trials over the hyperparameter distribution for each algorithm and test domain. Specifically, we split the data from each domain into 80% and 20% proportion, where the larger split is used for training and evaluation, and the smaller one is for selecting hyperparameters. We repeat the entire experiment twice using different seeds to reduce the randomness. Finally we report the mean over these repetitions as well as their estimated standard error.

Model selection. The model selection in domain generalization is intrinsically a learning problem, and we use test-domain validation, one of the three selection methods in [27]. This strategy is an oracle-selection one since we choose the model maximizing the accuracy on a validation set that follows the same distribution of the test domain.

Model architectures. Following [27], we use as encoders ConvNet for RotatedMNIST (detailed in Appendix D.1 in [27]) and ResNet-50 for the remaining datasets.

Motivated by the observation that GAN is able to improve image quality for evaluating the disentanglement effects in the latent spaces [57, 69, 70], we use adversarial training [26] on real samples \mathbf{x} against fake ones $D(h_s(\mathbf{x}) \oplus h_v(\tilde{\mathbf{x}}))$ to attain high-quality images \mathbf{x}' :

$$\mathcal{L}_{GAN} = \log \text{Disc}(\mathbf{x}) + \log(1 - \text{Disc}(\mathbf{x}')). \quad (12)$$

In practice, we can train the generator using adversarial train-

Domain	RotatedMNIST							Camelyon17-WILDS					
	0°	15°	30°	45°	60°	75°	Avg	d ₁	d ₂	d ₃	d ₄	d ₅	Avg
ERM [76]	96.0 ± 0.2	98.8 ± 0.1	98.8 ± 0.1	99.0 ± 0.0	99.0 ± 0.0	96.8 ± 0.1	98.1	96.8 ± 0.3	94.9 ± 0.2	95.9 ± 0.2	95.8 ± 0.2	94.8 ± 0.3	95.6
IRM [3]	96.0 ± 0.2	98.9 ± 0.0	99.0 ± 0.0	98.8 ± 0.1	98.9 ± 0.1	95.7 ± 0.3	97.9	95.0 ± 0.7	92.0 ± 0.2	95.2 ± 0.3	94.3 ± 0.1	93.3 ± 0.6	94.0
GDRO [66]	96.2 ± 0.1	98.9 ± 0.0	99.0 ± 0.1	98.7 ± 0.1	99.1 ± 0.0	96.8 ± 0.1	98.1	96.5 ± 0.1	95.0 ± 0.3	95.9 ± 0.9	96.0 ± 0.1	95.7 ± 0.4	95.8
MIXUP [80]	95.8 ± 0.3	98.9 ± 0.1	99.0 ± 0.1	99.0 ± 0.1	98.9 ± 0.1	96.5 ± 0.1	98.0	96.2 ± 0.0	94.3 ± 0.1	95.7 ± 0.4	96.7 ± 0.0	95.1 ± 0.1	95.6
MLDG [44]	96.2 ± 0.1	99.0 ± 0.0	99.0 ± 0.1	98.9 ± 0.1	99.0 ± 0.1	96.1 ± 0.2	98.0	97.0 ± 0.1	95.0 ± 0.3	96.6 ± 0.5	96.0 ± 0.2	96.1 ± 0.3	96.1
CORAL [73]	96.4 ± 0.1	99.0 ± 0.0	99.0 ± 0.1	99.0 ± 0.0	98.9 ± 0.1	96.8 ± 0.2	98.2	96.5 ± 0.2	95.2 ± 0.1	96.9 ± 0.1	96.8 ± 0.3	94.8 ± 0.3	96.0
MMD [46]	95.7 ± 0.4	98.8 ± 0.0	98.9 ± 0.1	98.8 ± 0.1	99.0 ± 0.0	96.3 ± 0.2	97.9	96.3 ± 0.1	94.9 ± 0.1	96.8 ± 0.1	96.5 ± 0.2	93.3 ± 0.1	95.6
DANN [23]	96.0 ± 0.1	98.8 ± 0.1	98.6 ± 0.1	98.7 ± 0.1	98.8 ± 0.1	96.4 ± 0.1	97.9	93.9 ± 0.3	89.6 ± 1.0	94.5 ± 0.1	93.9 ± 0.5	92.0 ± 0.2	92.8
CDANN [47]	95.8 ± 0.2	98.8 ± 0.0	98.9 ± 0.0	98.6 ± 0.1	98.8 ± 0.1	96.1 ± 0.2	97.8	94.3 ± 0.1	91.7 ± 0.7	95.0 ± 0.1	94.7 ± 0.2	92.9 ± 0.5	93.7
DDG	96.6 ± 0.1	99.0 ± 0.1	99.0 ± 0.2	99.1 ± 0.2	99.1 ± 0.2	97.4 ± 0.4	98.4	97.4 ± 0.2	95.4 ± 0.2	96.8 ± 0.1	96.8 ± 0.1	96.6 ± 0.3	96.6
DDG w/ AUG	96.7 ± 0.4	99.0 ± 0.3	99.1 ± 0.2	99.2 ± 0.2	99.0 ± 0.3	97.4 ± 0.3	98.4	97.7 ± 0.4	96.6 ± 0.2	96.9 ± 0.1	97.2 ± 0.3	96.9 ± 0.1	96.9

Domain	PACS					VLCS				
	A	C	P	S	Avg	C	L	S	V	Avg
ERM [76]	87.8 ± 0.4	82.8 ± 0.5	97.6 ± 0.4	80.4 ± 0.6	87.2	97.7 ± 0.3	65.2 ± 0.4	73.2 ± 0.7	75.2 ± 0.4	77.8
IRM [3]	85.7 ± 1.0	79.3 ± 1.1	97.6 ± 0.4	75.9 ± 1.0	84.6	97.6 ± 0.5	64.7 ± 1.1	69.7 ± 0.5	76.6 ± 0.7	77.2
GDRO [66]	88.2 ± 0.7	82.4 ± 0.8	97.7 ± 0.2	80.6 ± 0.9	87.2	97.8 ± 0.0	66.4 ± 0.5	68.7 ± 1.2	76.8 ± 1.0	77.4
MIXUP [80]	87.4 ± 1.0	80.7 ± 1.0	97.9 ± 0.2	79.7 ± 1.0	86.4	98.3 ± 0.3	66.7 ± 0.5	73.3 ± 1.1	76.3 ± 0.8	78.7
MLDG [44]	87.1 ± 0.9	81.3 ± 1.5	97.6 ± 0.4	81.2 ± 1.0	86.8	98.4 ± 0.2	65.9 ± 0.5	70.7 ± 0.8	76.1 ± 0.6	77.8
CORAL [73]	87.4 ± 0.6	82.2 ± 0.3	97.6 ± 0.1	80.2 ± 0.4	86.9	98.1 ± 0.1	67.1 ± 0.8	70.1 ± 0.6	75.8 ± 0.5	77.8
MMD [46]	87.6 ± 1.2	83.0 ± 0.4	97.8 ± 0.1	80.1 ± 1.0	87.1	98.1 ± 0.3	66.2 ± 0.2	70.5 ± 1.0	77.2 ± 0.6	78.0
DANN [23]	86.4 ± 1.4	80.6 ± 1.0	97.7 ± 0.2	77.1 ± 1.3	85.5	95.3 ± 1.8	61.3 ± 1.8	74.3 ± 1.0	79.7 ± 0.9	77.7
CDANN [47]	87.0 ± 1.2	80.8 ± 0.9	97.4 ± 0.5	77.6 ± 0.1	85.7	98.9 ± 0.3	68.8 ± 0.6	73.7 ± 0.6	79.3 ± 0.6	80.2
DDG	88.9 ± 0.6	85.0 ± 1.9	97.2 ± 1.2	84.3 ± 0.7	88.9	99.1 ± 0.6	66.5 ± 0.3	73.3 ± 0.6	80.9 ± 0.6	80.0
DDG w/ AUG	89.0 ± 0.3	86.3 ± 0.3	97.0 ± 0.5	84.8 ± 1.1	89.3	99.4 ± 0.2	68.9 ± 2.3	73.4 ± 1.1	81.2 ± 0.3	80.7

Table 1. Domain generalization accuracies (%) on RotatedMNIST, PACS, VLCS and WILDS.

ing. In this stage, we employ an adversarial objective \mathcal{L}_{GAN} and an additional cycle consistency constraint. With a slight abuse of notation, we denote $h_s(\cdot)$ and $h_v(\cdot)$ for $h_s(\cdot; \theta)$ and $h_v(\cdot; \phi)$, respectively. The detail of the cycle consistency constraint is that: we encode \mathbf{x} and $\tilde{\mathbf{x}}$ into the latent space as $h_s(\mathbf{x})$, $h_s(\tilde{\mathbf{x}})$, $h_v(\mathbf{x})$, $h_v(\tilde{\mathbf{x}})$. We then swap their variation factors and generate $\mathbf{x}_{x \rightarrow \tilde{x}} = D(h_s(\mathbf{x}) \oplus h_v(\tilde{\mathbf{x}}))$, $\mathbf{x}_{\tilde{x} \rightarrow x} = D(h_s(\tilde{\mathbf{x}}) \oplus h_v(\mathbf{x}))$. Again, the generated images will be encoded, and their variation factors will be swapped and used to generate $\mathbf{x}_{x \rightarrow \tilde{x} \rightarrow x}$ and $\mathbf{x}_{\tilde{x} \rightarrow x \rightarrow \tilde{x}}$. Finally, the cycle consistency constraint for \mathbf{x} is implemented by the proposed reconstruction loss (also similarly for $\tilde{\mathbf{x}}$):

$$\begin{aligned} \mathcal{L}_{cyc} &= d(\mathbf{x}_{x \rightarrow \tilde{x} \rightarrow x}, \mathbf{x}) \\ &= d(D(h_s(D(h_s(\mathbf{x})) \oplus h_v(\tilde{\mathbf{x}})))) \oplus \\ &\quad h_v(D(h_v(\tilde{\mathbf{x}}) \oplus h_v(\mathbf{x}))))), \mathbf{x}) \end{aligned} \quad (13)$$

In most experiments, the generator is a simple autoencoder, which converts the concatenation of $[h_s(\mathbf{x}), h_v(\mathbf{x})]$ to \mathbf{x}' . For qualitative evaluation and data augmentation experiments, the main idea of our generator follows [35, 85]. The decoder uses four MLPs to produce a set of AdaIN [34] parameters from the semantic factor. The variation factor is then processed by four residual blocks and four convolutional layers with these AdaIN parameters. Finally, the processed latent vector is decoded to the image space by upsampling and convolutional layers. The discriminator D follows the popular multi-scale PatchGAN [36] on three input image scales: 14×14 , 28×28 and 56×56 . The gradient punishment [53] is also applied when updating D .

See Appendix C for full details of all experimental set-

tings including datasets statistics and visualization, baselines and its implementation, hyperparameter search and model selection protocols. See Appendix D for many more results.

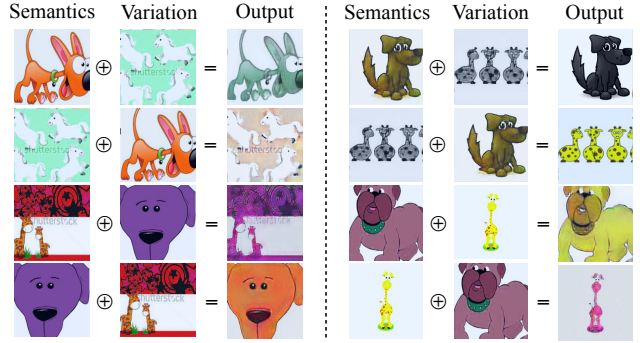


Figure 3. Qualitative disentanglement results. Swapping semantic and variation codes enables controllable generation for qualitative evaluation.

4.1. Qualitative Studies

We showcase some of the reconstructed images training with GAN in Fig. 2 and Fig. 3 (see appendix for many more similar results). The results show that the representations are disentangled with respect to various variation factors like background, color etc, which supports diverse manipulations for enriching training datasets. Moreover, DDG captures proper semantics over data, which allows diverse manipulations on variation factors like object colors and backgrounds without changing object semantics. For example, DDG changes the color of a dog in the first panel of Fig. 3 but retains the same color of its nose and eyes. Such

disentanglement enables flexible and controllable generation by manipulating semantic and variation factors via swapping (Fig. 3) or interpolation (Fig. 2).

Interpolation details. In order to better understand the learned semantic and variation representations, we further perform a linear interpolation experiment between two variation factors and generate the corresponding images as shown in Fig. 2. We denote the variation code of the first and second column as $h_v(\tilde{\mathbf{x}}), h_v(\mathbf{x})$, respectively. The images from 4 – 13 column are generated by $D(h_s(\mathbf{x}) \oplus (i \times h_v(\mathbf{x}) + (1 - i) \times h_v(\tilde{\mathbf{x}})))$ where $i \in \{1.0, 0.9, \dots, 0.1\}$. These interpolation results verify the smoothness and continuity in the variation space, and also show that our model is able to generalize in the embedding space instead of simply memorizing existing visual information. As a complementary study, we also generate images by linearly interpolating between two semantic factors while keeping the variation factors intact. We provide additional qualitative results in Appendix D.

4.2. Numerical Results

Comprehensive experiments show that DDG consistently outperforms all the baselines by a considerable margin. From Table 1, we observe that DDG achieves better DG results both in most single domain and on average. In particular, the performance gain is greater in the **worst-case** scenario like the C and S domain in PACS. This is particularly important since average performance is not an effective indicator of OOD generalization, and bad worst-case performance is tightly connected to issues like disparity amplification [28]. The performance gain of DDG is larger under the variation-rich dataset PACS. This makes intuitive senses because DDG is able to better capture inter-domain variations for improving OOD generalization.

4.3. Empirical Analyses and Ablations

Effect of data augmentation.

We first evaluate the effect of data augmentation by comparing our learned data augmentation network with a heuristic-based augmentation method AugMix [30]. Fig. 4 shows that the constraints optimization brings great performance gain over vanilla ERM and a data augmentation heuristics Augmix, especially the worst-case (i.e. the S domain in PACS) performance. The effectiveness of the data augmentation procedure in DDG is well connected to many empirical evidences in [29, 75, 82] and also validates the hypothesis in [54] that the richness of training domain data is crucial for extrapolation.

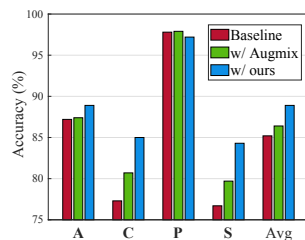


Figure 4. Data augmentation on PACS with different target domain.

Convergence analysis.

We investigate the training dynamics of DDG and several baselines over WILDS, where the target domain is d_5 . The learning curves in Fig. 5 show that domain adversarial training methods like DANN, CDANN are unstable and hard to converge due to the adversarial training nature. IRM has a similar pattern yet more stable. Thanks to the primal-dual algorithm, DDG can converge much better than the above methods, bear a resemblance to the ERM counterpart.

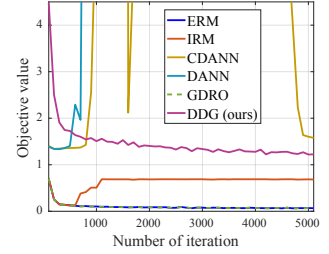


Figure 5. Convergence comparison.

Evaluation of domain divergence.

We use the \mathcal{A} -distance to measure domain discrepancy [5]. This can be approximated as $d_{\mathcal{A}} = 2(1 - 2\sigma)$, where σ is the error of a two-sample classifier distinguishing features of samples from source and target domains [50]. Fig. 6 shows that DDG can learn more invariant features to *minimize the divergence* between source and target domains than IRM and ERM.

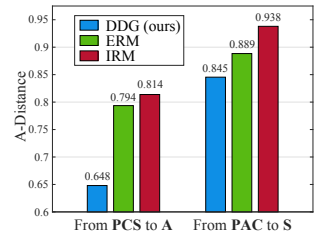


Figure 6. \mathcal{A} -distance on learned features for different generalization tasks.

Qualitative comparison with AugMix. Both Fig. 3 and Fig. 10b in the appendix show that DDG produces samples with diverse styles. In contrast, Fig. 8 in the appendix shows that it is much more difficult for heuristic-based methods such as AugMix to generate samples with diverse styles for training. The qualitative results validate the effectiveness of DDG as an automatic data augmentation method.

5. Concluding Remarks

We propose a novel disentangled learning framework for domain generalization, with both theoretical analyses and practical algorithmic implementation. By separating semantic and variation representations into different subspaces while enforcing invariance constraints, DDG yields superior OOD performance with improved empirical convergence and also yields interpretable and controllable generative results. In this work, we only consider the disentangled effects between semantic and variation factors since it is hard to provide known generative variation factors that manifest the distribution shifts precisely. It remains an open problem to improve the disentanglement between different variation factors with limited supervision and evaluate the treatment effects of data augmentation in a controllable manner [83].

Acknowledgement. WL is supported by a Cambridge-Tübingen Fellowship, DeepMind and the Leverhulme Trust via CFI. AW acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

References

- [1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *ICML*, 2020. 2
- [2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv:1911.00804*, 2019. 2, 3
- [3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019. 1, 2, 6, 7
- [4] Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. In *AAAI*, 2021. 1, 2, 3
- [5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 2010. 1, 8
- [6] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, 2006. 1, 15
- [7] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009. 1
- [8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013. 2
- [9] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 4
- [10] Stephen Boyd, Stephen P Boyd, and Lieven Vandenbergh. *Convex optimization*. Cambridge university press, 2004. 13
- [11] Péter Bárdi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermesen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandeveld, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2019. 6
- [12] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 1, 15
- [13] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *ACL*, 2020. 15
- [14] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *UAI*, 2015. 3
- [15] Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. In *NeurIPS*, 2020. 2, 3, 4
- [16] Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. The empirical duality gap of constrained statistical learning. In *ICASSP*, 2020. 2, 3, 5, 13
- [17] Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *arXiv:2103.05134*, 2021. 4, 5
- [18] Beidi Chen, Weiyang Liu, Zhiding Yu, Jan Kautz, Anshumali Shrivastava, Animesh Garg, and Animashree Anandkumar. Angular visual hardness. In *ICML*, 2020. 1, 2
- [19] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and DavidK Duvenaud. Isolating sources of disentanglement in variational autoencoders. *NIPS*, 2018. 3
- [20] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, 2021. 3
- [21] Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. In *ICLR*, 2021. 2, 15
- [22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, 2012. 3
- [23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 1, 6, 7
- [24] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1
- [25] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 6
- [26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 6
- [27] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 1, 3, 6, 19
- [28] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, 2018. 1, 3, 4, 8
- [29] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedziec, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *ACL*, 2020. 1, 3, 8
- [30] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. 6, 8, 18
- [31] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 3
- [32] Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. In *ICLR*, 2020. 15
- [33] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators.

- Neural networks*, 1989. 4
- [34] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 7
 - [35] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 7
 - [36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 7
 - [37] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *AISTATS*, 2021. 1
 - [38] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2019. 3
 - [39] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018. 3
 - [40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 17
 - [41] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 6
 - [42] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, 2021. 1, 2
 - [43] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NeurIPS*, 2017. 3
 - [44] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 6, 7
 - [45] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 6
 - [46] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 6, 7
 - [47] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 1, 6, 7
 - [48] Weiyang Liu, Rongmei Lin, Zhen Liu, James M Rehg, Liam Paull, Li Xiong, Le Song, and Adrian Weller. Orthogonal over-parameterized training. In *CVPR*, 2021. 1
 - [49] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical learning. In *NIPS*, 2017. 1
 - [50] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 8
 - [51] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv:2102.12353*, 2021. 1, 2
 - [52] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018. 3
 - [53] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 7
 - [54] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *ICLR*, 2021. 2, 3, 8, 15
 - [55] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 1, 2
 - [56] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2016. 1, 2
 - [57] Antoine Plummerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *ICLR*, 2019. 6
 - [58] Alexander Robey, Luiz Chamon, George Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial robustness with semi-infinite constrained learning. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 4
 - [59] Alexander Robey, Hamed Hassani, and George J. Pappas. Model-based robust deep learning: Generalizing to natural, out-of-distribution data, 2020. 3
 - [60] Alexander Robey, George J. Pappas, and Hamed Hassani. Model-based domain generalization. In *NeurIPS*, 2021. 1, 2, 3, 4, 5, 13, 19
 - [61] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *JMLR*, 2018. 1
 - [62] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *ICLR*, 2021. 1, 2
 - [63] Kevin Roth, Aurélien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NIPS*, 2017. 1
 - [64] Ruichu, Cai, Zijian, Pengfei, Wei, Jie, Qiao, Kun, Zhang, and Zhifeng and. Learning disentangled semantic representation for domain adaptation. In *IJCAI*, 2019. 2, 15
 - [65] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1
 - [66] 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. *ICLR*, 2020. 1, 4, 6, 7
 - [67] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *ICML*, 2012. 1
 - [68] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. 2
 - [69] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 6
 - [70] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 6
 - [71] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 2000. 3
 - [72] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial

- training. In *ICLR*, 2018. 4
- [73] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 6, 7
 - [74] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 6
 - [75] Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *TACL*, 2020. 3, 8
 - [76] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 6, 7
 - [77] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *NIPS*, 2018. 1, 2
 - [78] Haohan Wang, Zeyi Huang, Hanlin Zhang, and Eric Xing. Toward learning human-aligned cross-domain robust models by countering misaligned features. *arXiv:2111.03740*, 2021. 1, 4
 - [79] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020. 3
 - [80] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*, 2020. 6, 7
 - [81] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *ICML*, 2013. 1
 - [82] 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:1906.03347*, 2019. 1, 3, 8
 - [83] Yi-Fan Zhang, Hanlin Zhang, Zachary C Lipton, Li Erran Li, and Eric P Xing. Can transformers be strong treatment effect estimators? *arXiv:2202.01336*, 2022. 8
 - [84] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *NIPS*, 2018. 1
 - [85] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 7

Appendix

A. Proofs

A.1. Proof for parameterization gap

Proof A.1 Recall the feasibility assumption 3 ensures the strong duality such that the primal and dual optimal objectives are equal, which means for $\forall \tilde{\lambda} \in \mathbb{R}^+$, $\tilde{f}_s, \tilde{f}_v \in \mathcal{H}$, the following saddle point condition holds

$$\begin{aligned} \mathcal{R}(f_s^*, f_v^*, \tilde{\lambda}) &\leq \max_{\lambda} \min_{f_s, f_v} \mathcal{R}(f_s, f_v, \lambda) = D^* = P^* \\ &= \min_{f_s, f_v} \max_{\lambda} \mathcal{R}(f_s, f_v, \lambda) \leq \mathcal{R}(\tilde{f}_s, \tilde{f}_v, \lambda^*) \end{aligned} \quad (14)$$

Recall the definition of \mathcal{D}_ϵ^* and by the inclusion relation $\mathcal{H}_\theta \subseteq \mathcal{F}$, we can derive the lower bound as

$$\begin{aligned} \mathcal{D}_\epsilon^*(\gamma) &\triangleq \max_{\lambda} \min_{\theta, \phi \in \mathcal{H}} \mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{con}}(\theta, \phi) \\ &\geq \min_{\theta, \phi \in \mathcal{H}} \mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{con}}(\theta, \phi), \forall \lambda \in \mathbb{R}^+ \\ &\geq \min_{f_s, f_v \in \mathcal{F}} \mathcal{L}(f_s) + \lambda \mathcal{L}_{\text{con}}(f_s, f_v) = \mathcal{P}^* \end{aligned} \quad (15)$$

For the upper bound, we add and subtract $\mathcal{R} \equiv \mathcal{R}(f_s, f_v, \lambda) = \min_{f_s \in \mathcal{F}} \mathcal{L}(f_s) + \min_{f_s, f_v \in \mathcal{F}} \lambda \mathcal{L}_{\text{con}}(f_s, f_v)$ from $\mathcal{D}_\epsilon^*(\gamma)$ to get

$$\begin{aligned} \mathcal{D}_\epsilon^*(\gamma) &= \max_{\lambda} \min_{\theta, f_s, f_v} \mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{con}}(\theta, \phi) + \mathcal{R} - \mathcal{R} \\ &= \max_{\lambda} \min_{\theta, f_s, f_v} \mathcal{R} + [\mathcal{L}(\theta) - \mathcal{L}(f_s)] + \lambda [\mathcal{L}_{\text{con}}(\theta, \phi) - \mathcal{L}_{\text{con}}(f_s, f_v)] \end{aligned} \quad (16)$$

With a slight abuse of notation, \mathbb{E} is short for $\mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}(X, Y), \tilde{\mathbf{x}} \sim \mathbb{P}(X)}$. Then we consider the combination of the second and third term as

$$\begin{aligned} &[1, \lambda] \begin{bmatrix} \mathcal{L}(\theta) - \mathcal{L}(f_s) \\ \mathcal{L}_{\text{con}}(\theta, \phi) - \mathcal{L}_{\text{con}}(f_s, f_v) \end{bmatrix} \\ &\leq (1 + \|\lambda\|_1) \left\| \begin{bmatrix} \mathcal{L}(\theta) - \mathcal{L}(f_s) \\ \mathcal{L}_{\text{con}}(\theta, \phi) - \mathcal{L}_{\text{con}}(f_s, f_v) \end{bmatrix} \right\|_\infty \\ &= (1 + |\lambda|) \max \{ \mathcal{L}(\theta) - \mathcal{L}(f_s), \mathcal{L}_{\text{con}}(\theta, \phi) - \mathcal{L}_{\text{con}}(f_s, f_v) \} \\ &= (1 + |\lambda|) \max \{ \mathbb{E}[\ell(h_s(\mathbf{x}; \theta); y) - \ell(f_s(\mathbf{x}); y)], \mathbb{E}[d(\mathbf{x}, D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi))) - d(\mathbf{x}, D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}})))] \} \\ &\leq (1 + |\lambda|) \mathbb{E} \{ \max \{ [\ell(h_s(\mathbf{x}; \theta); y) - \ell(f_s(\mathbf{x}); y)], \mathbb{E}[d(\mathbf{x}, D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi))) - d(\mathbf{x}, D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}})))] \} \} \\ &\leq (1 + |\lambda|) \mathbb{E} \{ \max \{ L_\ell |h_s(\mathbf{x}; \theta) - f_s(\mathbf{x})|, L_d |D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi)) - D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}}))| \} \} \\ &\leq (1 + |\lambda|) \mathbb{E} \max \{ L_\ell \epsilon_s, L_d \epsilon_g \} \\ &= (1 + |\lambda|) \max \{ L_\ell \epsilon_s, L_d \epsilon_g \} \end{aligned} \quad (17)$$

where the first inequality is using Hölder's inequality (18) when $p = 1, q = \infty$ and the second one is by the convexity of max-norm and Jensen's inequality. The third inequality is applying L_ℓ and L_d lipschitzness on ℓ and d . The fourth one is due to ϵ_s and ϵ_g parameterization on f_s and D .

$$\sum_{k=1}^n |x_k y_k| \leq \left(\sum_{k=1}^n |x_k|^p \right)^{1/p} \left(\sum_{k=1}^n |y_k|^q \right)^{1/q} \quad (18)$$

Then Eq. (16) becomes

$$\mathcal{D}_\epsilon^*(\gamma) \leq \max_{\lambda} \min_{f_s, f_v} \mathcal{R} + (1 + |\lambda|) \max \{ L_\ell \epsilon_s, L_d \epsilon_g \} \triangleq D_p^* \quad (19)$$

In order to use strong duality to bound \mathcal{D}_p^* , we need to construct primal problem whose optimum is \mathcal{P}_p^* from \mathcal{D}_p^* . Therefore, the remaining proof is to show \mathcal{D}_p^* is the dual problem to a constraint statistical learning problem with perturbation function as $m = \max \{L_{\ell \in s}, L_{d \in g}\}$. Specifically, when $\lambda > 0$, \mathcal{D}_p^* can be expanded and rearranged as

$$\begin{aligned}\mathcal{D}_p^* &= \max_{\lambda} \min_{f_s, f_v} \mathcal{L}(f_s) + \lambda \mathcal{L}_{\text{con}}(f_s, f_v) + (1 + |\lambda|)m \\ &= \max_{\lambda} \min_{f_s, f_v} \ell(f_s(\mathbf{x}; \theta); y) + m + \lambda[d(\mathbf{x}, D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))) - \gamma + m]\end{aligned}\quad (20)$$

By the feasibility assumption 3, Eq. (20) can be considered as the dual problem of the following one:

$$\begin{aligned}\mathcal{P}_p^*(\gamma) &\triangleq \min_{f_s \in \mathcal{F}} \mathcal{L}(f_s) + m \\ \text{s.t.} \quad &d(\mathbf{x}, D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))) \leq \gamma - m\end{aligned}\quad (21)$$

Denote f_s^* and f_v^* as the primal solutions to \mathcal{P}_p^* , λ^* and λ_p^* as the dual variable that give solutions \mathcal{D}_p^* , \mathcal{P}_p^* . With the regularity assumption 2, the saddle point condition shows for $\forall \tilde{f}_s, \tilde{f}_v \in \mathcal{H}$ and $\lambda, \lambda_p^* \in \mathbb{R}^+$, it holds that

$$\begin{aligned}\mathcal{R}(f_{s,p}^*, f_{v,p}^*, \tilde{\lambda}) + (1 + |\tilde{\lambda}|) &\leq \max_{\lambda} \min_{f_s, f_v} \mathcal{R}(f_s, f_v, \lambda) + (1 + |\lambda|)m = \mathcal{D}_p^* = \mathcal{P}_p^* \\ &= \min_{f_s, f_v} \max_{\lambda} \mathcal{R}(f_s, f_v, \lambda) + (1 + |\lambda|)m \\ &\leq \mathcal{R}(\tilde{f}_s, \tilde{f}_v, \lambda_p^*) + (1 + |\lambda_p^*|)m\end{aligned}\quad (22)$$

Recall that Eq. (22) holds for $\forall \tilde{f}_s, \tilde{f}_v \in \mathcal{H}$. Let $\tilde{f}_s = f_s^*, \tilde{f}_v = f_v^*$, then we use it to upper bound \mathcal{D}_ϵ^* as

$$\mathcal{D}_\epsilon^*(\gamma) \leq \mathcal{P}_p^* \leq \mathcal{R}(f_s^*, f_v^*, \lambda^*) + (1 + |\lambda_p^*|)m = \mathcal{P}^*(\gamma) + (1 + |\lambda_p^*|)m\quad (23)$$

The last step is due to Eq. (14).

A.2. Proof for empirical gap

Proof A.2 Similar to [16, 60], by KKT conditions and complementary slackness conditions [10] shows

$$\begin{aligned}\lambda_\epsilon^* (\mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim \mathbb{P}(X)} [d(\mathbf{x}, D(f_s(\mathbf{x}; \theta_\epsilon^*), f_v(\tilde{\mathbf{x}}; \phi_\epsilon^*))) - \gamma] &= 0 \\ \lambda_{\epsilon,n}^* (\sum_{i=1}^n \sum_{j \neq i}^n d(\mathbf{x}_i, D(f_s(\mathbf{x}_i; \theta_\epsilon^*), f_v(\mathbf{x}_j; \phi_\epsilon^*))) - \gamma) &= 0\end{aligned}\quad (24)$$

where $(\theta_\epsilon^*, \phi_\epsilon^*, \lambda_\epsilon^*)$ and $(\theta_{\epsilon,n}^*, \phi_{\epsilon,n}^*, \lambda_{\epsilon,n}^*)$ are primal-dual pairs for achieving the optimum $\mathcal{D}_\epsilon^*(\gamma)$ and $\mathcal{D}_{\epsilon,n}^*(\gamma)$. Eq. (24) implies the constraint-related terms in the objectives to be zero, then consider the remaining term as

$$\begin{aligned}\mathcal{D}_\epsilon^*(\gamma) &= \mathbb{E}[\ell(f_s(\mathbf{x}; \theta); y)] \triangleq \mathcal{M}(\theta_\epsilon^*) \\ \mathcal{D}_{\epsilon,n}^*(\gamma) &= \sum_{i=1}^n \ell(f_s(\mathbf{x}_i), y_i) \triangleq \hat{\mathcal{M}}(\theta_{\epsilon,n}^*)\end{aligned}\quad (25)$$

Thus the empirical gap reduces to

$$|\mathcal{D}_\epsilon^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| = |\mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*)|\quad (26)$$

Using the fact that θ_ϵ^* and $\theta_{\epsilon,n}^*$ are optimal for $\mathcal{M}(\theta_\epsilon^*)$ and $\hat{\mathcal{M}}(\theta_{\epsilon,n}^*)$, the following holds

$$\mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_\epsilon^*) \leq \mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*) \leq \mathcal{M}(\theta_{\epsilon,n}^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*)\quad (27)$$

Therefore, using the above lower and upper bound, we can bound Eq. (26) as

$$|\mathcal{D}_\epsilon^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| \leq \max\{|\mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_\epsilon^*)|, |\mathcal{M}(\theta_{\epsilon,n}^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*)|\}\quad (28)$$

Then we resort to the classical VC-dimension bounds for the above two terms in Eq. (28) as follows

$$|\mathcal{M}(\theta) - \hat{\mathcal{M}}(\theta)| \leq 2B \sqrt{\frac{1}{n} \left[1 + \log \left(\frac{4(2n)^{d_{VC}}}{\delta} \right) \right]} \quad (29)$$

holds with probability $1 - \delta$ when d_{VC} is the VC dimension for all θ .

Combing Eq. (28) and (29) completes the proof.

A.3. Proof for empirical duality gap

Proof A.3 Simply combining the results in the above lemmas, i.e. parameterization gap and empirical gap, via applying the triangle inequality completes the proof.

$$\begin{aligned} |\mathcal{P}^* - \mathcal{D}_{\epsilon,n}^*(\gamma)| &= |\mathcal{P}^* + \mathcal{D}_{\epsilon}^*(\gamma) - \mathcal{D}_{\epsilon}^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| \\ &\leq |\mathcal{P}^* - \mathcal{D}_{\epsilon}^*(\gamma)| + |\mathcal{D}_{\epsilon}^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| \\ &\leq (1 + |\lambda|)m + 2B \sqrt{\frac{1}{n} \left[1 + \log \left(\frac{4(2n)^{d_{VC}}}{\delta} \right) \right]} \end{aligned}$$

B. Domain Generalization by Learning on Fictitious Distributions

This section gives a justification for disentanglement from a different perspective by connecting the dots with classical domain adaptation. Specifically, we construct a fictitious distribution to extend it to the DG setting and decompose the target learning objective into empirical learning errors, domain divergence and source domain data diversity. Moreover, we show that learning disentangled representations gives a tighter risk upper bound.

With a slight abuse of notation, let \mathcal{H} be a hypothesis space and denote $\tilde{\mathcal{D}}$ as the induced distribution over feature space \mathcal{Z} for every distribution \mathcal{D} over the raw space. Define \mathcal{D}_S^i as the source distribution over \mathcal{X} , which enables a mixture construction of source domains as $\mathcal{D}_S^\alpha = \sum_{i=1}^{N_S} \alpha_i \mathcal{D}_S^i(\cdot)$. Denote a fictitious distribution $\mathcal{D}_U^\alpha = \sum_{i=1}^{N_S} \alpha_i^* \mathcal{D}_S^i(\cdot)$ as the convex combination of source domains which is the closest to \mathcal{D}_U , where $\alpha_1^*, \dots, \alpha_{N_S}^* = \arg \min_{\alpha_1, \dots, \alpha_{N_S}} d_{\mathcal{H}}(\mathcal{D}_U, \sum_{i=1}^{N_S} \alpha_i \mathcal{D}_S^i(\cdot))$. The fictitious distribution induces a feature space distribution $\tilde{\mathcal{D}}_U^\alpha = \sum_{i=1}^{N_S} \alpha_i^* \tilde{\mathcal{D}}_S^i(\cdot)$. The following inequality holds for the risk $\epsilon_U(h)$ on any unseen target domain \mathcal{D}_U .

$$\epsilon_U(h) \leq \underbrace{\lambda_\alpha + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h)}_{\text{① Empirical}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha)}_{\text{② Divergence}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha)}_{\text{③ Diversity}} \quad (30)$$

where λ_α is the risk of the optimal hypothesis on the mixture source domain \mathcal{D}_S^α and \mathcal{D}_U .

We define the symmetric difference hypothesis space as $\mathcal{H}\Delta\mathcal{H} = \{h(\mathbf{x}) \oplus h'(\mathbf{x}) : h, h' \in \mathcal{H}\}$, where \oplus is the XOR operator. Applying [6], we have

$$\begin{aligned} \epsilon_U(h) &\leq \lambda_U + \Pr_{\mathcal{D}_U}[\mathcal{Z}_h \Delta \mathcal{Z}_h^*] \\ &\leq \lambda_U + \Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \Delta \mathcal{Z}_h^*] + |\Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \Delta \mathcal{Z}_h^*] - \Pr_{\mathcal{D}_U}[\mathcal{Z}_h \Delta \mathcal{Z}_h^*]| \\ &\leq \lambda_U + \Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \Delta \mathcal{Z}_h^*] + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_S^\alpha) \\ &\leq \lambda_U + \Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \Delta \mathcal{Z}_h^*] + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha) \\ &\leq \lambda_\alpha + \underbrace{\sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h)}_{\text{① Empirical}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha)}_{\text{② Divergence}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha)}_{\text{③ Diversity}} \end{aligned} \quad (31)$$

The fourth inequality holds because of the triangle inequality. We provide the explanation for our bound in the Eq. (30) and Eq. (31). The second term is the empirical loss for the convex combination of all source domains. The third term corresponds to “To what extent can the convex combination of the source domain approximate the target domain”. The minimization of the third term requires diverse data or strong data augmentation, such that the unseen distribution lies within the convex combination of source domains. For the fourth term, the following equation holds for any two distributions $\mathcal{D}'_U, \mathcal{D}''_U$, which are the convex combinations of source domains [12]

$$d_{\mathcal{H}}[\mathcal{D}'_U, \mathcal{D}''_U] \leq \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \alpha_l \alpha_k d_{\mathcal{H}}[\mathcal{D}_{S,l}, \mathcal{D}_{S,k}] \quad (32)$$

Such an upper bound will be minimized when $d_{\mathcal{H}}[\mathcal{D}_{S,l}, \mathcal{D}_{S,k}] = 0, \forall l, k \in \{1, \dots, N_S\}$. Namely projecting the source domain data into a feature space, where the source domain labels are hard to distinguish.

The above ③ Diversity term is also supported by the evidence that compositional generalization and extrapolation can be improved if the training domain data are rich enough [13, 32, 54]. To this end, one can obviously simulate data points with predetermined data augmentation methods such as rotating, cropping, Gaussian blur, color jitter, etc. However, their developments require prior knowledge and domain-specific expertise like translation-invariance on images, which is likely to fail in the unseen domain due to distribution shifts. It motivates learning disentangled representations that are transferable across various domains [21]. Thus we discuss the benefits of disentanglement on the domain generalization gap in the following section. Assume that the semantic and the variation factors are disentangled in the latent space \mathcal{S} and \mathcal{V} , then the errors [64] on the disentangled source and target domain with a hypothesis h are

$$\epsilon_{S,i}(h) = \epsilon_{S,i}^s(h) + \epsilon_{S,i}^v(h), \epsilon_U(h) = \epsilon_U^s(h) + \epsilon_U^v(h) \quad (33)$$

Given $h^* = \arg \min_{h \in \mathcal{H}} (\epsilon_{S,i}^s(h), \epsilon_{S,i}^v(h)) \forall i \in \{1, \dots, N_S\}$, since $\epsilon_U(h) = \epsilon_U^s(h) + \epsilon_U^v(h)$, combining Eq. (30) and we have

$$\epsilon_U^s(h) + \epsilon_U^v(h) \leq \lambda_\alpha + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha) \quad (34)$$

where $\lambda_\alpha = \epsilon_U(h^*) + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h^*) = \epsilon_U^s(h^*) + \epsilon_U^v(h^*) + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}^s(h^*) + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}^v(h^*)$. Then the upper bound for the unseen domain can be further derived as follow,

$$\epsilon_U^s(h) \leq \lambda_\alpha + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha) - \epsilon_U^v(h) \quad (35)$$

Combining Eq. (30) and Eq. (33), we have

$$\epsilon_U^s(h) \leq \lambda_\alpha + \underbrace{\sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h)}_{\textcircled{1} \text{ Empirical}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha)}_{\textcircled{2} \text{ Divergence}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha)}_{\textcircled{3} \text{ Diversity}} - \underbrace{\epsilon_U^v(h)}_{\textcircled{4}} \quad (36)$$

where $\textcircled{1}$ denotes the empirical loss over every source domain, $\textcircled{2}$ means divergence minimization among source domains and $\textcircled{3}$ encourages the diversity and coverage of the mixture of source domains. The above result shows the benefits of disentanglement over representation spaces. Thus we propose to use two separate encoders representing semantic and variation subspaces respectively in the following section. Such a formulation combined with the disentanglement term $\textcircled{4}$ in Eq. (36) implies that we should perform ERM over the semantic space only. The above analysis essentially justifies the design of DDG from a classical domain adaptation perspective: optimizing empirical risk over semantic space while promoting diversity and divergence by taking disentanglement as a constraint.

C. Experimental Settings

C.1. Other Training Details

We optimize all models using Adam [40]. For all the detailed hyperparameter settings, please refer to our code which is publicly available at <https://github.com/hlzhang109/DDG>.

C.2. Dataset Statistics and Visualization

We show some images of the datasets in Fig. 7 to give an intuitive comparison among these image datasets. One can observe that these images have a diverse set of styles, making it very challenging to transfer knowledge from one to another.

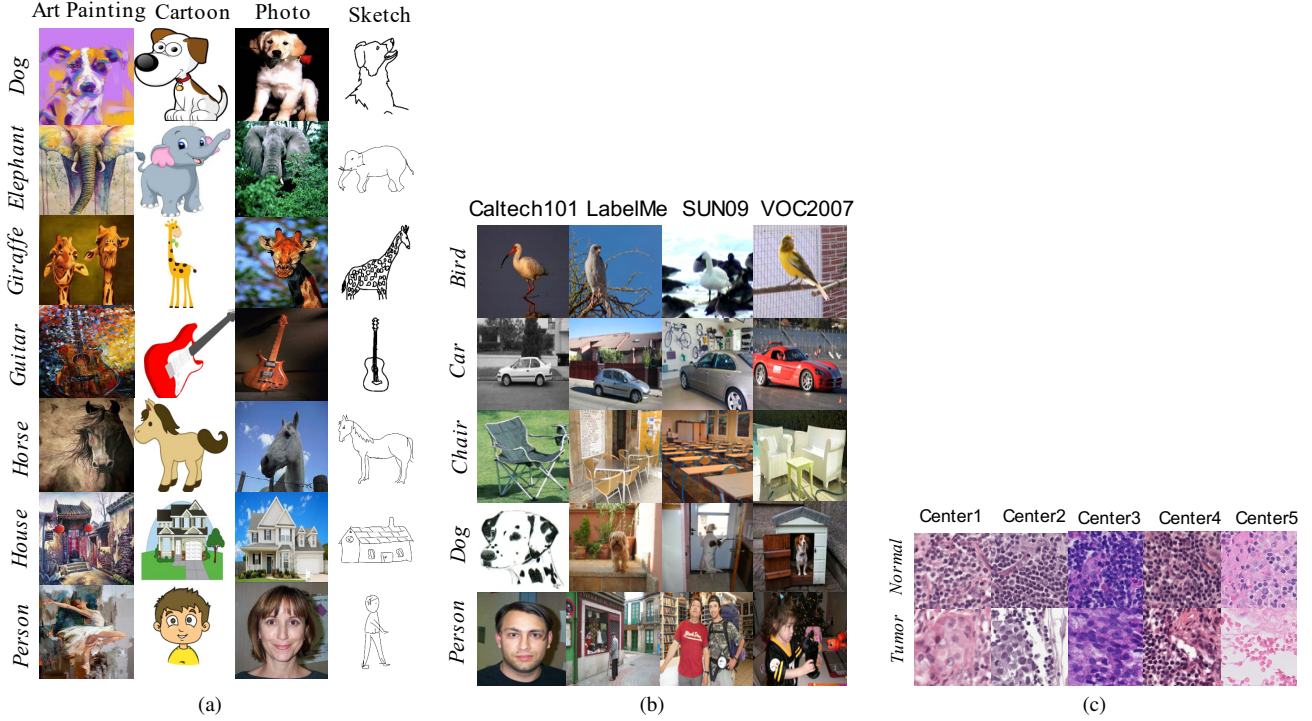


Figure 7. **Samples of DG datasets.** The training data (a) PACS and (b) VLCS (c) WILDS are shown. PACS has four domains *art* (A), *cartoons* (C), *photos* (P), *sketches* (S). VLCS contains four domains *Caltech101* (C), *LabelMe* (L), *SUN09* (S), *VOC2007* (V). The WILDS dataset includes data from five different medical centers as domains.

D. Additional Experimental Results

Qualitative comparison with AugMix. Looking into the Fig. 8, it is harder for the heuristic-based method AugMix to generate diverse samples via interpolation for training compared with DDG as shown in Fig. 3 and Fig. 10b.

More qualitative results via interpolation. Fig. 9 showcases the results of combining the semantic code of one image and the mixture of two variation codes. Results show that the model can generate samples with intermediate variation states.

More qualitative results via swapping variation and semantic factors. We showcase the qualitative results of swapping variation and semantic factors with PACS in Fig. 10b, MNIST in Fig. 10a, and WILDS in Fig. 10c. The results demonstrates the strong disentangled capability of DDG. Some interesting observations are DDG learns both intra (e.g. thickness) and inter (e.g. rotated angle) domain variations over RotatedMNIST. DDG also maintains semantic information like the color of distinct features across variation-rich data like PACS.

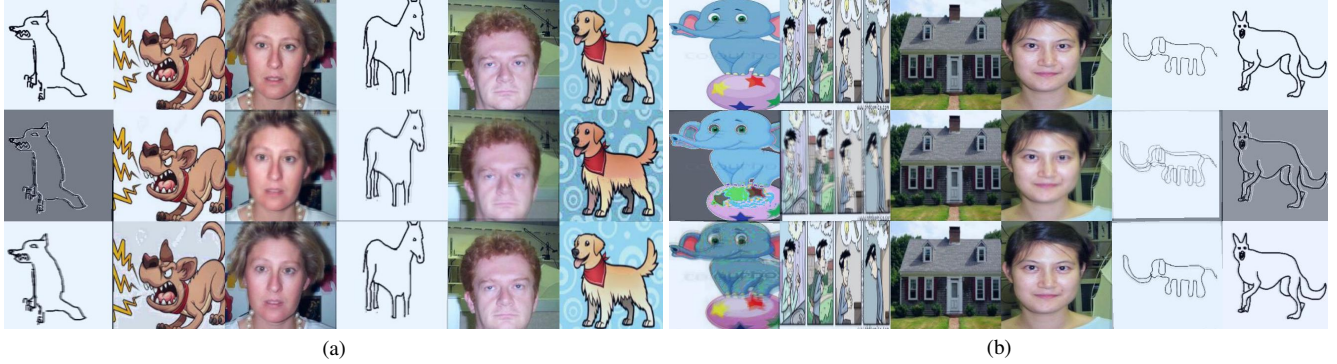


Figure 8. The augmented samples from AugMix [30]. The second and third rows are generated by applying AugMix to the first row.



Figure 9. Interpolation via mixing results on PACS.



Figure 10. **Qualitative disentanglement results on RotatedMNIST, PACS and Wilds.** In every panel, the training data in the first row manifests the semantic factors.

Numerical comparison with MBDG. We quantitatively compare DDG with MBDG [60] as in Table 2. For a consistent comparison, we run the source code of authors under a test domain validation protocol [27] on PACS with the results as:

	A	C	P	S	Avg
MBDG	82.0 \pm 0.0	78.4 \pm 0.01	93.9 \pm 0.0	85.0 \pm 0.0	85.8
MBDG _{Reg}	84.9 \pm 0.0	84.9 \pm 0.0	93.9 \pm 0.0	85.6 \pm 0.0	87.3
DDG	88.9 \pm 0.6	85.0 \pm 1.9	97.2 \pm 1.2	84.3 \pm 0.7	88.9

Table 2. Numerical results for comparing DDG and MBDG.

Though adopting similar PACCL frameworks and primal dual algorithms, DDG can consistently outperform MBDG and its variant except over domain S. The primary reason behind the clear performance gain of DDG can be that DDG is better at capturing variations within data via random sampling without relying on domain labels. Specifically, parameterizing h_s, h_v, D

and constrain them based on disentanglement makes the model more robust to both inter- and intra-domain nuisance factors compared to MBDG that only use a pretrained generator to simulate inter-domain variations. Moreover, the performance gain is also partly due to the three major differences between our approach and MBDG we highlight in the Related Work: First, our upper bound of the parameterization gap is tighter under mild conditions, whereas MBDG requires unrealistic assumptions on the distance metric, i.e., $d(\cdot, \cdot)$ satisfies Lipschitz-like inequality on both arguments, which is stronger than our normal L_d Lipschitzness assumption; second, MBDG consumes additional domain labels, which are prohibitively expensive or even infeasible to obtain in safety-critical applications or those containing sensitive demographics; third, DDG enforces invariance constraints via parameterizing semantic and variation encoders, which does not belong to a model-based approach. In contrast, MBDG requires a pre-trained domain transformation model (e.g., CycleGAN) during training, which may result in sub-optimal solutions and parameter inefficiency, while DDG is more flexible by treating this as a design choice.