Beyond Loss Functions: Exploring Data-Centric Approaches with Diffusion Model for Domain Generalization

Anonymous authors
Paper under double-blind review

Abstract

There has been a huge effort to tackle the Domain Generalization (DG) problem with a focus on developing new loss functions. Inspired by the image generation capabilities of the diffusion models, we pose a pivotal question: Can diffusion models function as data augmentation tools to address DG from a data-centric perspective, rather than relying on the loss functions? Our findings reveal that trivial cross-domain data augmentation (CDGA) along with the vanilla ERM using readily available diffusion models without additional finetuning outperforms state-of-the-art (SOTA) training algorithms.

This paper delves into the exploration of why and how this rudimentary data generation can outperform complicated DG algorithms. The domain shift quantification results confirm that CDGA reduces the gap between domains. We show that in practice, CDGA along with ERM is equivalent to replacing the pointwise kernel estimates in ERM with new density estimates in the *vicinity of domain pairs* which can diminish the true data estimation error of ERM under domain shift scenario. Further, we empirically reveal connections between the loss landscape, adversarial robustness, and data generation, illustrating that CDGA reduces loss sharpness and improves robustness against adversarial shifts in data. These insights advocate for further investigation into the potential of data-centric approaches in DG.

1 Introduction

Out-of-distribution (OOD) generalization stands as a crucial capability for deep learning models in real-world scenarios. The prevalent setting for investigating OOD generalization is termed *domain generalization* (DG) Blanchard et al. (2011), involving multiple source domains to generalize to an unseen target domain. In DG problems, there is a shift between the training domains and the target domain which makes the models trained using Empirical Risk Minimization (ERM) Vapnik (1999b) struggle to maintain their performance in the target domain.

To enhance OOD generalization within the DG framework, researchers have proposed innovative loss functions—typically achieved by introducing regularizers to ERM—to facilitate the learning of domain-invariant mechanisms across domains. Nevertheless, none of these approaches consistently outperforms others across all datasets, as illustrated by results from the DomainBed benchmark (Gulrajani & Lopez-Paz, 2020). This observation suggests that a singular regularizer capable of capturing all invariances might not exist. We posit that the absence of such a universal regularizer arises from the diverse shifts present in each dataset, encompassing correlation shift, diversity shift, label shift, etc. Consequently, a rigid, data-independent regularizer may fall short in eliminating all types of spurious correlations and shifts. Additionally, the incorporation of sub-optimal regularizers can exacerbate optimization challenges, introducing excessive risk (Sener & Koltun, 2022), additional hyperparameters, and computational bottlenecks in ERM.

Rather than relying solely on traditional loss functions, recent advancements in generative foundation models open up new avenues for addressing the DG problem from a data-centric standpoint. Specifically, the capability of Denoising Diffusion Models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022) in generating high-fidelity synthetic images offers an innovative approach for advanced data augmentation, enabling the creation of domain-invariant images to enhance OOD generalization. To examine this hypothesis,

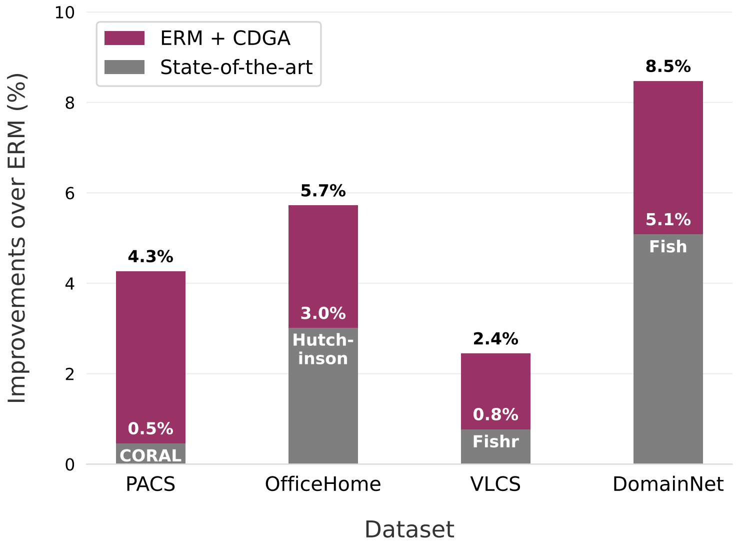


Figure 1: Improvements of CDGA + ERM and state-of-the-art (SOTA) DG training methods over ERM across four dataset using DomainBed benchmark. See Section 4 for experimental details. CDGA + ERM outperforms SOTA algorithms across all datasets, highlighting the absence of a singular algorithm achieving best results across datasets.

we employ a straightforward *Cross Domain Generative Augmentation* (CDGA) method. In CDGA, synthetic images are generated conditioned on images or text descriptions from all possible combinations of the training domains using a pre-trained latent diffusion model (LDM) (Rombach et al., 2022). In Figure 1, we show that applying vanilla ERM to combined generated and real images outperforms the previous state-of-the-art algorithms *across all datasets*.

This paper delves into an exploration of the reasons and mechanisms behind the superior performance of CDGA's simplistic data generation strategy compared to complex DG training algorithms. Our investigation involves quantifying and visualizing domain shifts across domains of generated synthetic images, validating that cross-domain data generation mitigates the gap between domains. From a theoretical standpoint, we illustrate that CDGA is equivalent to replacing pointwise kernel estimates in ERM with new density estimates in the proximity of *domain pairs*. This modification to ERM reduces the inherent data estimation error in the presence of domain shift, subsequently enhancing its out-of-distribution (OOD) performance. Furthermore, our empirical results establish connections between the loss landscape, adversarial robustness, and data generation, revealing that cross-domain data generation lessens loss sharpness and improves robustness against adversarial shifts in data. To the best of our knowledge, we are the first to utilize latent diffusion models as a data-centric approach for DG.

Our primary contributions are as follows:

1. Demonstrating superior performance, our study reveals that combining CDGA with vanilla ERM outperforms state-of-the-art DG training algorithms. We validate this across four datasets using three model selection strategies.

2. Theoretical analysis illuminates the success of our data-centric approach, framing CDGA as a practical implementation of the Vicinal Risk Minimization principle (VRM) (Chapelle et al., 2000) within DG.

3. Employing various metrics such as transfer measure, diversity shift, and near-duplicate quantification, we empirically demonstrate that CDGA reduces distribution shift among domains, attributing to its superior performance.

4. Through empirical analysis, we calculate the loss landscape sharpness of CDGA during training with ERM, showcasing its reduced sharpness compared to vanilla ERM. Additionally, we demonstrate the robustness of CDGA + ERM against two adversarial shift attacks.
5. Our extensive ablation studies compare single-domain and cross-domain data generation, underscoring the necessity of cross-domain generation for substantial improvements. Furthermore, our ablation on the size of generated data highlights its potential as a method for mitigating class imbalance challenges.

2 Problem Settings and Related Work

We denote our prediction model as f and its parameters as $\boldsymbol{\theta}$. In DG, the goal is to learn a shared model from n source (train) population domains (environments) $\{\mathcal{E}_1, \ldots, \mathcal{E}_n\}$, to generalize to an unseen target domain $\mathcal{T}$. For a given domain $\mathcal{E}$, the classification loss is:

$$\mathcal{L}_{\mathcal{E}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{E}}[\ell(f(x;\theta), y)], \tag{1}$$

where each x and y are the input data point and its corresponding label and $\ell(f(x;\theta), y)$ is the cross entropy loss between $f(x;\theta)$ and y.

ERM The standard baseline for training deep learning models is ERM Vapnik (1999b), which minimizes the average of losses over the entire available training domains i.e., $\{S_1, \ldots, S_n\}$

$$\min_{\theta} \sum_i \frac{1}{|S_i|} \sum_{k=1}^{|S_i|} \ell(f(x_k^i;\theta), y_k^i) \tag{2}$$

where $(x_k^i, y_k^i) \sim \mathcal{E}_i$ is an i.i.d. sample from the training set S_i of domain i with $|S_i|$ samples, x_k^i is the k-th data point in S_i and y_k^i is its associated label. However, in the case of domain shift between domains, the pointwise kernel estimation of true data distribution proposed by ERM in Eq. 2, becomes less accurate which results in a lack of OOD generalization of ERM.

Improving the Loss Function: In the pursuit of improving OOD performance within ERM, diverse avenues have been explored. Notable efforts include robust optimization (Sagawa et al., 2020; Hu et al., 2018), invariant representation learning on the feature level (Sun & Saenko, 2016; Ganin et al., 2016; Li et al., 2018; Tzeng et al., 2014), and classifier head adjustments (Arjovsky et al., 2019). Additionally, advancements in loss functions, such as those discussed by Krueger et al. (2021), reveal their efficacy in enhancing domain generalization. Further insights stem from investigations into loss gradients/Hessians (Parascandolo et al., 2020; Shahtalebi et al., 2021; Koyama & Yamaguchi, 2020; Shi et al., 2021; Hemati et al., 2023), showcasing the multifaceted approaches taken to fortify ERM against domain shifts.

Data Augmentation for DG: Various strategies enhance OOD performance in Empirical Risk Minimization (ERM). Classic data augmentation, as explored by Gulrajani & Lopez-Paz (2020), demonstrates improved results under the DomainBed evaluation protocol. Ilse et al. (2021) introduces Select Data Augmentation, a method that identifies transformations detrimentally affecting validation accuracy. Mixup (Zhang et al., 2017) generates mixed data points and soft labels through linear convex combinations from two classes, while MixStyle (Zhou et al., 2021) combines per-sample feature statistics (mean and variance) across domains. Somavarapu et al. (2020) introduces a stylization transformation based on in-domain data. To the best of our knowledge, we are the first to utilize latent diffusion models as a data-centric approach for DG.

Denoising diffusion models and their applications. Recent advances in diffusion-based generative models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Zhang & Agrawala, 2023) demonstrate their capability to achieve SOTA image quality. Additionally, works such as Unclip (Ramesh et al., 2022) have successfully integrated Foundation models like CLIP (Radford et al., 2021) with stable diffusion, introducing new generative functionalities such as image-to-image, text-to-image, image variation, and image mixing to diffusion-based models. The application of diffusion models in representation learning has also been explored, as exemplified by StableRep proposed by Tian et al. (2023). In a setup akin to SimCLR (Chen et al., 2020), they demonstrated that synthetic images generated by stable diffusion models can enhance self-supervised learning.

3 Cross Domain Generative Augmentation

In this section, we provide a detailed description of CDGA. CDGA utilizes LDM to perform a transformation denoted by $\mathcal{M}(\cdot)$. This transformation takes two arguments: a data point in one domain and a guidance attribute in another domain from the same class. Formally,

$$\widetilde{x}_k^{i,j} = \mathcal{M}(x_k^i, \texttt{guide}^j), \tag{3}$$

where $\widetilde{x}_k^{i,j}$ is a synthetic image transformed from domains i and j, generated from the k-th sample in S_i. The attribute $\texttt{guide}^j$ serves as guidance towards another domain, S_j, within the same class.

In CDGA, each data point in domain S_i undergoes transformation to all n domains, including its own domain. This augmentation increases the number of samples for domain S_i from $|S_i|$ to $(b \times n + 1) \times |S_i|$, where n is the number of training domains, $|S_i|$ is the number of data points in S_i, and b is the generation batch size. Furthermore, we introduce CDGA*, where we assume access to a guidance attribute of the target domain. In this scenario, the size of domain S_i increases from $|S_i|$ to $(b \times n + 2) \times |S_i|$.

CDGA with Prompt Guidance (CDGA-PG): In CDGA-PG, given the k-th image in S_i, i.e., x_i^k, the guidance attribute $\texttt{guide}^j$ is a domain description text prompt that represents the same class in S_j. Having the image and the prompt guidance, we use the LDM to generate b synthetic images which we expect to interpolate domains i and j for the same class. For each image in S_i, we perform these image translations for all the training domains j, $\forall j \in \{1, ..., n\}$. We also consider the scenario where we can utilize the target domain description, i.e., $\texttt{guide}^{\mathcal{T}}$ as the guidance.

CDGA with Image Guidance (CDGA-IG): For scenarios where a text prompt description of domains is not available, CDGA-IG is used where the guidance is an image from S_j instead of a text description. More precisely, in CDGA-IG we attempt to mix two images from two different domains which is also known as the image mixer in the literature.

4 CDGA Outperforms SOTA

In this section, we compare CDGA + ERM with SOTA DG training methods, demonstrating its superior performance. We assess CDGA and CDGA* on for datasets, namely VLCS (Fang et al., 2013), PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019), using the DomainBed benchmark (Gulrajani & Lopez-Paz, 2020). This benchmark has gained popularity as a fair and standard evaluation platform for domain generalization algorithms. The evaluation process involves comparing DG algorithms across 20 hyperparameter choices and 3 trials, utilizing three distinct model selection techniques. To demonstrate CDGA's effectiveness, we present its evaluation results using the DomainBed benchmark in Tables 1-4. The tables follow a format of presenting the **first** and <u>second</u> results. For brevity, we report only the top five performing algorithms for each model selection, with full results available in the appendix. Examining Tables 1-4, CDGA* consistently achieves SOTA performance across all datasets and model selection techniques. Specifically, we applied prompt guidance for PACS, OfficeHome, and DomainNet, while using image guidance (i.e., image mixer) for VLCS. The code implementation for deploying CDGA-generated data within the DomainBed scheme is detailed in Appendix H.

5 CDGA with ERM Extends Vicinal Risk Minimization Principle to DG Setup

In this section, we attempt to provide a theoretical justification for the reasons behind the success of CDGA. Our theoretical justification relies on the Vicinal Risk Minimization principle (VRM) (Chapelle et al., 2000) initially introduced by Vapnik (1999a). The motivation behind VRM is to improve true data distribution estimation in ERM. First note that we can rewrite ERM (Vapnik, 1999a) loss in Eq. 2 as

$$\min_{\boldsymbol{\theta}} \sum_i \frac{1}{|S_i|} \sum_k^{|S_i|} \int \ell(f(x;\theta), y)\delta(x; x_k^i)\delta(y; y_k^i)dxdy,$$

Table 1: DomainBed benchmark for **training-domain validation set** model selection method.

Algorithm	**PACS**	**OfficeHome**	**DomainNet**	**Avg**
ERM	85.5 ± 0.2	66.5 ± 0.3	40.9 ± 1.8	64.3
CORAL	86.2 ± 0.3	68.7 ± 0.3	41.5 ± 0.1	65.5
SagNet	86.3 ± 0.2	68.1 ± 0.1	40.3 ± 0.1	64.9
Fish	85.5 ± 0.3	68.6 ± 0.4	42.7 ± 0.2	65.6
Fishr	85.5 ± 0.4	67.8 ± 0.1	41.7 ± 0.0	65.0
HGP	84.7 ± 0.0	68.2 ± 0.0	41.1 ± 0.0	64.7
ERM + CDGA-PG	88.5 ± 0.5	68.2 ± 0.6	43.7 ±0.1	66.6
ERM + CDGA-PG*	**89.5** ± 0.3	**70.8** ± 0.6	**44.8** ±0.0	**68.4**

Table 2: DomainBed benchmark for **leave-one-domain-out cross-validation** model selection.

Algorithm	**PACS**	**OfficeHome**	**DomainNet**	**Avg**
ERM	83.0 ± 0.7	65.7 ± 0.5	40.6 ± 0.2	63.1
CORAL	82.6 ± 0.5	68.5 ± 0.2	41.1 ± 0.1	64.1
SagNet	82.3 ± 0.1	67.6 ± 0.3	40.2 ± 0.2	63.4
MLDG	82.9 ± 1.7	66.1 ± 0.5	41.0 ± 0.2	63.3
HGP	82.2 ± 0.0	67.5 ± 0.0	41.1 ± 0.0	63.6
Hutchinson	84.8 ± 0.0	68.5 ± 0.0	41.4 ± 0.0	64.9
ERM + CDGA-PG	86.8 ± 0.4	68.7 ± 0.4	43.6 ±0.1	66.2
ERM + CDGA-PG*	**88.4** ± 0.5	**70.2** ± 0.4	**44.8** ±0.0	**67.8**

where $\delta(\boldsymbol{x}; x_k^i)$ is the Dirac delta distribution for the k-th data point in S_i. To unify this formulation with the loss on true distribution data we can rewrite Eq. 2 as

$$\min_{\theta} \mathbb{E}_{i\sim \text{Unif(n)}} \mathbb{E}_{(x,y)\sim \widehat{\mathcal{E}}_i} [\ell(f(x;\theta), y)], \tag{4}$$

where Unif(n) is the discrete uniform distribution over 0 to n and the sample distribution $\widehat{\mathcal{E}}_i$ is defined as

$$\widehat{\mathcal{E}}_i(x) = \frac{1}{|S_i|} \sum_{k}^{|S_i|} \delta(x; x_k^i)\delta(y; y_k^i). \tag{5}$$

To improve the true data estimation error of ERM, this is necessary to improve the sample distribution $\widehat{\mathcal{E}}_i$. To this end, the VRM principle suggests replacing the point-wise estimate in ERM i.e., $\delta(x; x_k^i)$ in eq. 5 with some better kernel estimate of the density in the vicinity of the data point k within S_i which we call $K(x; x_k^i)$. An example for $K(x; x_k^i)$ can be Gaussian kernel functions that act as smooth $\delta(x; x_k^i)$ functions. In this case, the VRM sample distribution represented by $\widetilde{\mathcal{E}}_i$ is written as

$$\widetilde{\mathcal{E}}_i(x) = \frac{1}{|S_i|} \sum_{k}^{|S_i|} K(x; x_k^i)\delta(y; y_k^i), \tag{6}$$

and the overall loss is obtained by replacing $\widehat{\mathcal{E}}_i(x)$ in Equation 4 by $\widetilde{\mathcal{E}}_i(x)$. In practice, VRM is implemented by employing data augmentation along with ERM. VRM suggests that data augmentation can improve ERM estimation error by adding additional synthetic examples from the vicinity distribution around each observation *within each domain* in the data. We argue in the DG setting, if we only employ classic data augmentation the estimation error of VRM is still high. This is because, in the DG setting, the estimation error of true data distribution by ERM is mainly caused by the distribution shift *between domains* which cannot be fully addressed by simple data augmentation techniques within each domain.

Extending the VRM (Chapelle et al., 2000) principle to the DG setup, we argue employing CDGA is equivalent to replacing the pointwise estimates in ERM with new density estimates in the *vicinity of domain pairs* so

Table 3: DomainBed benchmark **test-domain validation set (oracle)**model selection method.

Algorithm	**PACS**	**OfficeHome**	**DomainNet**	**Avg**
ERM	86.7 ± 0.3	66.4 ± 0.5	41.3 ± 0.1	64.8
Mixup	86.8 ± 0.3	68.0 ± 0.2	39.6 ± 0.1	64.8
MLDG	86.8 ± 0.4	66.6 ± 0.3	41.6 ± 0.1	65.0
CORAL	87.1 ± 0.5	68.4 ± 0.2	41.8 ± 0.1	65.8
SagNet	86.4 ± 0.4	67.5 ± 0.2	40.8 ± 0.2	64.9
Fish	85.8 ± 0.6	66.0 ± 2.9	43.4 ± 0.3	65.1
Fishr	86.9 ± 0.2	68.2 ± 0.2	41.8 ± 0.2	65.6
Hutchinson	86.3 ± 0.0	68.4 ± 0.0	41.9 ± 0.0	65.5
ERM + CDGA-PG	89.6 ± 0.3	68.8 ± 0.3	44.4 ±0.1	67.2
ERM + CDGA-PG*	**90.4** ± 0.3	**70.2** ± 0.2	**44.8** ±0.0	**68.5**

Table 4: DomainBed benchmark on **VLCS** dataset.

Method	Training domain	Leave-one -domain-out	Oracle
ERM	77.5 ± 0.4	77.2 ± 0.4	77.6 ± 0.3
CORAL	78.8 ± 0.6	78.7 ± 0.4	77.7 ± 0.2
SagNet	77.8 ± 0.5	77.5 ± 0.3	77.6 ± 0.1
Fishr	77.8 ± 0.1	78.2 ± 0.0	78.2 ± 0.2
HGP	77.6 ± 0.0	76.7 ± 0.0	77.3 ± 0.0
Hutchinson	76.8 ± 0.0	**79.3** ± 0.0	77.9 ± 0.0
ERM + CDGA-IG	**78.9** ± 0.3	77.9 ± 0.5	**79.5** ± 0.1

that the distribution shift between domains is further reduced. We believe this step can potentially reduce the estimation error induced by ERM under the domain shift scenario. To see the difference between classic augmentation and CDGA more clearly, here we provide the sample distribution that the CDGA along with the ERM brings to realization. To do this, first, we need to define the projection operator:

Definition 5.1 (**domain projection operator**). Given a metric d defined on $\mathrm{supp}(\mathcal{E}_i) \cup \mathrm{supp}(\mathcal{E}_j)$, the domain projection operator $P_{i \to j}(\cdot)$ that projects a data point x in $\mathcal{E}_i$ (i.e., $x \in \mathrm{supp}(\mathcal{E}_i)$) onto domain j is defined as

$$P_{i \to j}(x) = \underset{z \,\in\, \mathrm{supp}(\mathcal{E}_j)}{\mathrm{argmin}} \; d(z, x).$$

The projection operator gives the mathematical definition for what we mean by "finding the closest point on another domain." For example, if x is a cat image of painting style ($\mathcal{E}_i$), then $P_{i \to j}(x)$ would mean the most similar image to x in the sketch domain $\mathcal{E}_j$. The projection operator relies on the metric d, which measures the similarity of two samples. This similarity can be implemented, e.g., by computing the L_2 distance in some latent feature embedding space.

Define the smoothed sample distribution of $\widetilde{\mathcal{E}}_i$ projected to domain j by $\widetilde{\mathcal{E}}_{i,j}$:

$$\widetilde{\mathcal{E}}_{i,j}(x) = \frac{1}{|S_i|} \sum_{k}^{|S_i|} K(x; P_{i \to j}(x_k^i))\delta(y; y_k^i). \tag{7}$$

where i and j are domain counters and k is the data point index, $K(x; P_j(x_k^i))$ represents the vicinity of k-th data point in $\mathcal{E}_i$ projected onto $\mathcal{E}_j$ which is label invariant. Similar to VRM, the expression for the final loss function can be obtained by replacing $\widehat{\mathcal{E}}_i(x)$ with $\frac{1}{n}\sum_j \widetilde{\mathcal{E}}_{i,j}(x)$ in equation 4.

In practice, thanks to image manipulation of diffusion models, such projection to other domains and sampling from the vicinity of projected data points has become feasible. To this end, in CDGA, we employ LDM,

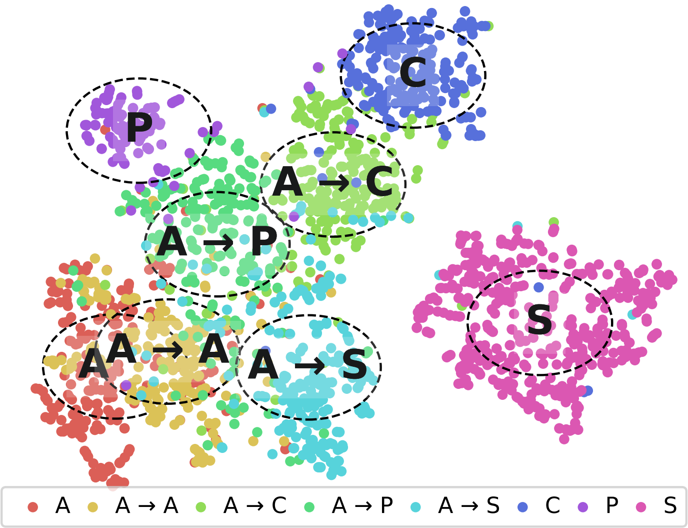


Figure 2: The *t*-SNE plot of features extracted from the original PACS dataset and generated images by CDGA from A domain. This figure shows that CDGA can fill the gap between the original domains. Check Section 6 for details.

denoted by $\mathcal{M}(\cdot)$ which takes two arguments, one is a data point in a domain and the second argument is a guidance attribute in another domain from the same class i.e.,

$$\widetilde{x}_k^{i,j} = \mathcal{M}(x_k^i, \texttt{guide}^j) \sim K(\cdot; P_{i\to j}(x_k^i)).$$

Effectively, we are drawing an augmented sample near the projected sample in the target domain,[1] and the kernel estimation relies on the LDM we choose. This is equivalent to drawing samples $(\widetilde{x}_k^{i,j}, y_k^i)$ from $\widetilde{\mathcal{E}}_{i,j}$ and estimating the smoothed empirical loss using the Monte Carlo approach:

$$\min_{\theta} \mathbb{E}_{i\sim \mathrm{Unif(n)}} \mathbb{E}_{j\sim \mathrm{Unif(n)}} \mathbb{E}_{(x,y)\sim \widetilde{\mathcal{E}}_{i,j}} [\ell(f(x;\theta), y)], \tag{8}$$

Since our kernel relies on the LDM we use, the estimation error comparison between equation 8 and equation 5 seems infeasible. More theoretical exploration along this direction is left as future work.

6 CDGA Reduces Domain Shift

In this section, we empirically confirm that CDGA reduces domain shift. To validate the efficacy of CDGA in mitigating domain shift, we employ five domain shift quantification techniques from the literature on the PACS dataset. Specifically, we utilize t-SNE visualization of feature embeddings, near-duplicate analysis (Oquab et al., 2023), transferability (Zhang et al., 2021; Hemati et al., 2023), and diversity shift metrics (Ye et al., 2022) to quantify the shift between domains.

6.1 Domain shift Visualization

To visualize domain shifts in CDGA-based data for the class "dog" across all domains (P, A, C, and S), we generate synthetic images for A → A, A → P, A → C, and A → S. Subsequently, we utilize the pretrained CLIP ViT-B/32 image encoder (Radford et al., 2021) to extract features from both real and synthetic images. These features are then projected onto a two-dimensional space using t-SNE and presented in Figure 2. Notably, the cross-domain synthetic images effectively interpolate between different domains, addressing the

---

[1] We implicitly assume that $P_{i\to j}(x)$ is a singleton for any x. Otherwise, we can extend the definition of $\widetilde{\mathcal{E}}_{i,j}$ by averaging over all "closest" projections.

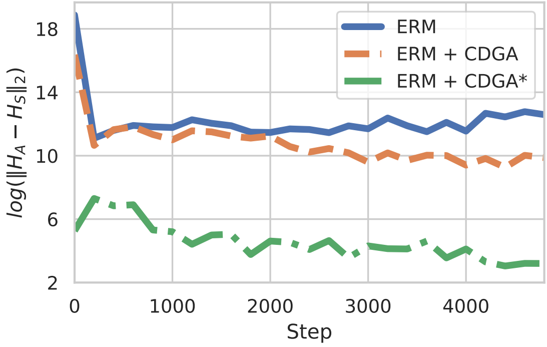
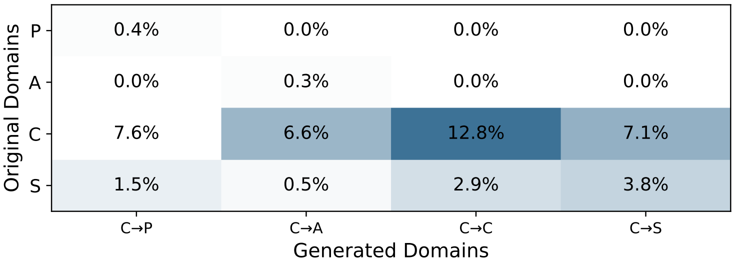


Figure 3: Classifier head Hessian difference between domains P & S during training.



Original Domains	C→P	C→A	C→C	C→S
P	0.4%	0.0%	0.0%	0.0%
A	0.0%	0.3%	0.0%	0.0%
C	7.6%	6.6%	12.8%	7.1%
S	1.5%	0.5%	2.9%	3.8%

Figure 4: Heat map of the percentage of near-duplicates of each original domain in the generated domains. This table shows that using target-domain description results in more near-duplicate images.

desired distribution shift. In Figure 2, examining domains A (in red) and S (in pink) reveals a significant distribution shift in their two-dimensional representations, despite all images belonging to the dog class. However, A → S synthetic images seamlessly bridge the gap between A and S representations. Refer to Figure 14 in the appendix for t-SNE plots of other classes.

6.2 Transferability Measurement

Transferability is another recent approach proposed by Zhang et al. (2021) to quantify the domain shift. The transferability measure is an upper bound of the difference between the source and the target domain excess risks. Subsequently, Hemati et al. (2023) showed that an upper bound for transferability measure is $\frac{1}{2}\delta^2\|\boldsymbol{H}_{\mathcal{T}} - \boldsymbol{H}_{\mathcal{S}}\|_2 + o(\delta^2)$ where δ is a constant, $\boldsymbol{H}_{\mathcal{T}}$ and $\boldsymbol{H}_{\mathcal{S}}$ are target and source classifier head's Hessians. Following these findings, to quantify the dynamics of domain shift, we monitor classifier heads Hessian distances between all possible domain pairs through the training steps for ERM and CDGA, where we set domains P, A, and C as train (source) and domain S as target. Figure 3 shows the difference between the classifier head's Hessians given data from domains A and S during the steps. We see that CDGA and CDGA* lead to lower classifier head's Hessian difference and subsequently smaller transferability and domain shift compared to ERM. This pattern consistently exists for other domain pairs where the full results are presented in Figure 13 in the appendix.

6.3 Diversity Shift

Ye et al. (2022) proposed a numerical method to measure diversity shift which is equivalent to total variation (Zhang et al., 2021) to quantify domain shift. Diversity shift is usually due to the novel domain-specific features in the data. We employ the proposed algorithm by Ye et al. (2022) to quantify and compare diversity shift between training domains and the target domain in a leave-one domain out scheme for PACS real data, CDGA-PACA, and CDGA*-PACS datasets. Figure 6 (left) shows both CDGA and CDGA* reduce the diversity shift between training domains and the target domain.

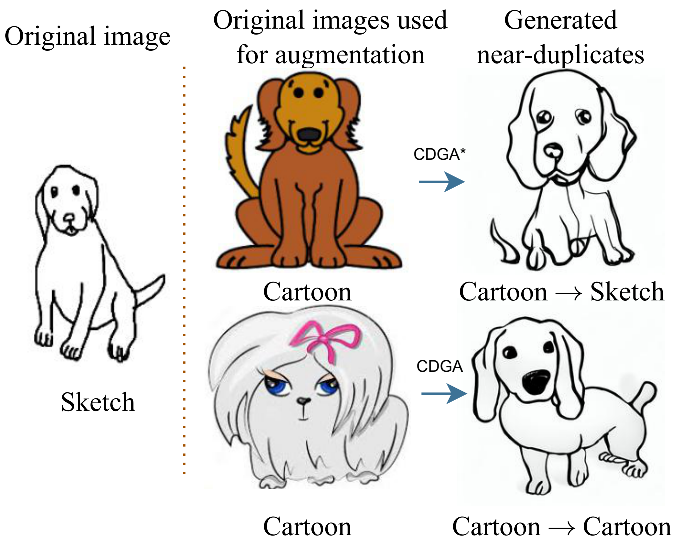


Figure 5: Examples of near-duplicates (right-most column) found for the dog image in Sketch domain (left-most column) that are generated using CDGA from the original images (middle column).

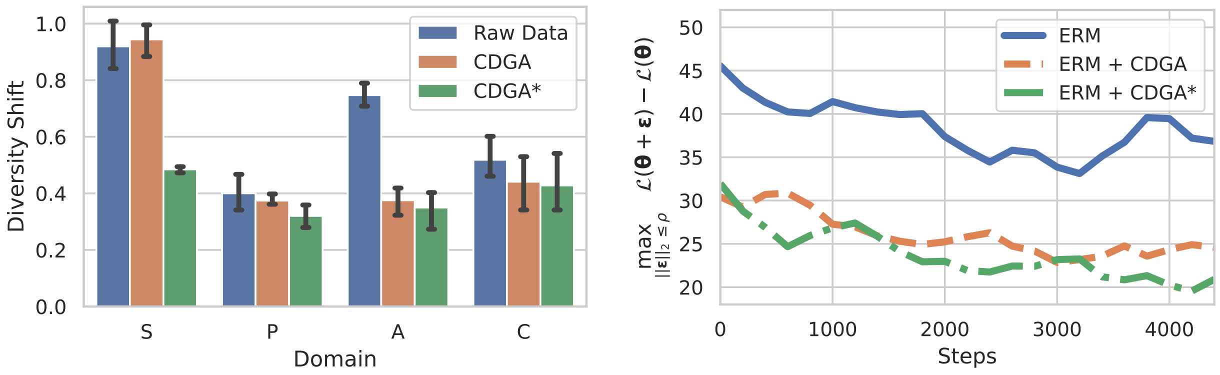


Figure 6: (left) Diversity shift as a measure of iid-ness of PACS (raw data) against CDGA, and CDGA* augmented datasets. Each column is the target domain and the rest of the domains are training domains. CDGA and CDGA* reduce the diversity shift. (right) Sharpness of loss landscape through training for ERM, CDGA, and CDGA* on PACS dataset.

6.4 Near-duplicate Analysis

We employ near-duplicate image detection on images generated using CDGA to quantify the similarity between the generated and original images in each domain. Following the self-supervised image retrieval technique outlined in (Oquab et al., 2023), we utilize the pretrained CLIP ViT-B/32 image encoder (Radford et al., 2021) to extract embeddings and calculate cosine similarity between original and generated images. For each original image, if at least one image in a generated domain exhibits a cosine similarity above 0.95, we categorize the original domain as having a near-duplicate. Figure 4 provides a summarized view of this experiment for the case of generated images from domain C, while the complete results are available in Figure 11 in the appendix. In Figure 4, we report, for each original domain, the percentage of near-duplicates relative to the original domain size. Clearly, generating synthetic images within the manifold between training domains allows us to obtain examples that are near-duplicates of the target domain. Figure 5 showcases some of the near-duplicates identified for a sample image in the S domain using this technique. Additional examples can be found in Figure 12 in the appendix.

7 Loss Landscape Analysis: CDGA Leads to a Flatter Local Minimum

Recently, there has been a growing interest in finding the relation between loss landscape structure and generalization ability of deep neural networks (Keskar et al., 2016; Foret et al., 2020). More precisely, according to the theorem stated by Foret et al. (2020) generalization error of deep neural networks is upper bounded by the sharpness of their loss landscapes:

$$\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\theta + \epsilon) - \mathcal{L}(\theta), \tag{9}$$

where ρ is the perturbation size. In other words, neural networks with flatter local minima have less generalization error. Inspired by this finding, we raise the following question. Given better OOD generalization of CDGA and CDGA*, do these methods lead to flatter local minima compared with ERM? To answer this question we calculate the sharpness using Eq. 9 and monitor it through the training of the model on the PACS dataset. The result of this experiment is reported in Figure 6 (right). Interestingly, CDGA and CDGA* lead to less sharp local minima, which explains their superior generalization ability.

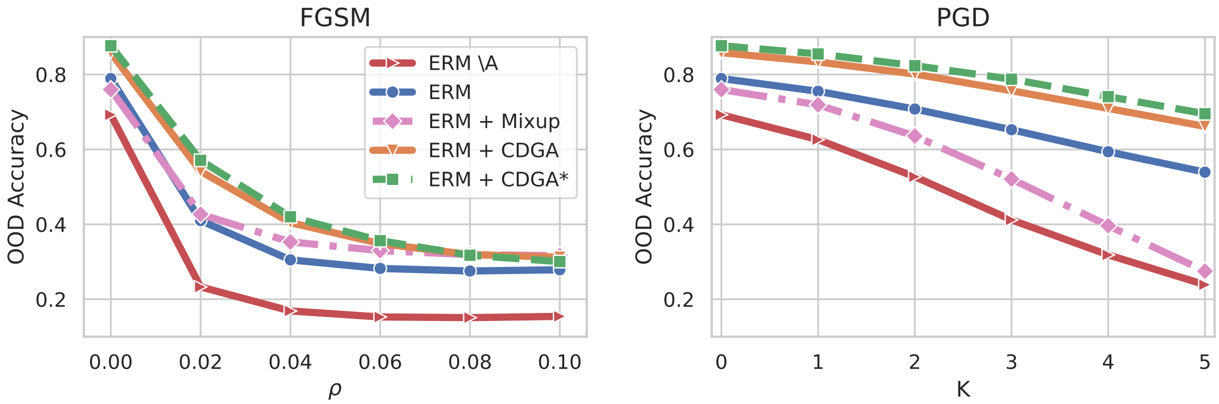


Figure 7: OOD accuracy of models trained using CDGA, CDGA*, classic augmentation, and MixUp after FGSM adversarial shift with varying perturbation size ρ (left Figure) and PGD adversarial shift with varying number of iterations K (right Figure). Models trained with CDGA and CDGA* are more robust against adversarial shifts. ERM \A means without data augmentation.

8 CDGA Improves Adversarial Robustness

The robustness of deep learning models to adversarial attacks is a notion of their generalization ability (Shorten & Khoshgoftaar, 2019). Here, we study the robustness of models trained with CDGA and CDGA* against adversarial shifts to the data. To this end, we evaluate their performance on the target domain after shifting the target domain using the fast gradient sign method (FGSM) (Goodfellow et al., 2015), and the more sophisticated projected gradient descent (PGD) adversarial attacks (Madry et al., 2018). We design an experiment using the PACS dataset where the training domains are P, A, and C and the target domain is S. After training the same model using different data augmentation techniques including MixUp and the classic data augmentation, we perturb the target domain using FSGM and PGD and record the OOD accuracy for different strength of attacks. Results are presented in Figure 7, where ρ controls the FGSM perturbation size, and K controls the number of iterations in PGD. Clearly, models trained with CDGA and CDGA* are more robust against adversarial shifts on the target domain compared with the classic data augmentation and Mixup.

9 Ablation Studies

9.1 Single Domain Generative Augmentation (SDGA) vs CDGA

To show the advantage of employing cross-domain data to mitigate domain shift, We also explore the SDGA method, where, unlike CDGA, the image from S_i is augmented only from the guidance of the same domain i.e.,

$\mathcal{M}(\texttt{guide}^i)$, where guidance can either be an image (SDGA-IG) or a prompt (SDGA-PG). In SDGA-PG, for each image in S_i, i.e., x_k^i we create prompt guidance $\texttt{guide}^i$ that can contain label and/or domain information from S_j and feed $\texttt{guide}^i$ to the LDM. In SDGA-IG, for each image from S_i, i.e., x_k^i we construct guidance that contains both x_k^i and label information. To compare variations of CDGA and SDGA, we evaluate them on the PACS dataset using the DomainBed benchmark with twenty different hyperparameters and one trial. The results of this experiment are presented in Table 5. Clearly, CDGA consistently outperforms all variations of SDGA.

Table 5: OOD accuracy of models trained with variations of CDGA and SDGA on PACS dataset using Domainbed benchmark.

Method	Training domain	Leave-one -domain-out	Oracle
ERM	85.5 ± 0.2	83.0 ± 0.7	86.7 ± 0.3
+ SDGA-PG (label)	86.1 ± 0.5	83.7 ± 1.0	87.3 ± 0.5
+ SDGA-PG (label+domain)	85.9 ± 1.0	84.6 ± 0.8	87.5 ± 1.1
+ SDGA-IG (label)	87.5 ± 0.6	86.5 ± 1.1	89.5 ± 0.3
+ CDGA-PG (canny edge)	86.8 ± 0.9	79.2 ± 3.6	87.8 ± 0.3
+ CDGA-PG	88.5 ± 0.5	86.8 ± 0.4	89.6 ± 0.3
+ CDGA*-PG	**89.5** ± 0.3	**88.4** ± 0.5	**90.4** ± 0.3

9.2 Scaling Law of Data Size

In this experiment, we aim to measure the effect of generation batch size b which determines the final augmented dataset size on the OOD generalization of models trained with CDGA. To this end, on the PACS dataset, using DomainBed benchmark with 20 choices of hyperparameters and 1 trial, we apply CDGA-PG with b equal to 1, 2, 3, and 4. We conduct this experiment for two models: ResNet-18 model pretrained on ImageNet and ResNet-18 model with random initialization. As we see from Table 6, for both pretrained and random initializations, increasing b, i.e., increasing the data size, further improves OOD generalization. However, for the pretrained model, the performance gets saturated which can be due to the capacity of the model.

Table 6: Effect of generation batch size b on CDGA for PACS dataset with different initialization.

Initialization	b	Training domain	Leave-one -domain-out	Oracle
Random	1	66.1	61.6	65.0
	2	70.2	69.2	70.3
	3	71.8	71.8	74.5
	4	**73.9**	**73.2**	**74.6**
Pre-trained	1	87.0	86.4	89.0
	2	87.4	86.2	89.0
	3	**88.4**	**89.1**	88.9
	4	**88.4**	88.2	**89.2**

10 Conclusions

In this paper, we showed that a simple cross domain generative augmentation (i.e., CDGA) alongside ERM surpasses SOTA DG algorithms in the standard DomainBed benchmark. We explore the reasons behind the superior performance of this basic data generation technique compared to complex DG algorithms.

Theoretically, we establish that CDGA with ERM extends the VRM principle to the DG setup. Empirically, employing various distribution shift quantification methods, we observe a significant reduction in distribution shift between training domains after applying CDGA. Furthermore, we conduct comprehensive ablation studies, particularly focusing on adversarial robustness, loss landscape analysis, and data scaling laws. Notably, the use of CDGA enhances adversarial robustness and reduces the sharpness of the loss landscape, both contributing to improved model generalization. Our work provides a novel data-centric point of view for domain generalization, in the era when AI Generated Content (AIGC) becomes more and more popular.

References

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

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.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6572`.

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

Sobhan Hemati, Guojun Zhang, Amir Estiri, and Xi Chen. Understanding hessian alignment for domain generalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, pp. 4555–4562. PMLR, 2021.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

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). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

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*, pp. 5542–5550, 2017.

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*, pp. 5400–5409, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

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

Justin Pinkney. Image mixer. `https://huggingface.co/lambdalabs/image-mixer`, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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. In *International Conference on Learning Representations*, 2020.

Ozan Sener and Vladlen Koltun. Domain generalization without excess empirical risk. *Advances in Neural Information Processing Systems*, 35:13380–13391, 2022.

Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

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.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), July 2019. doi: 10.1186/s40537-019-0197-0. URL `https://doi.org/10.1186/s40537-019-0197-0`.

Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999a.

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

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7947–7958, 2022.

Guojun Zhang, Han Zhao, Yaoliang Yu, and Pascal Poupart. Quantifying and improving transferability in domain generalization. *Advances in Neural Information Processing Systems*, 34:10957–10970, 2021.

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

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.