FDS: Feedback-guided Domain Synthesis with Multi-Source Conditional Diffusion Models for Domain Generalization

Mehrdad Noori* Milad Cheraghalikhani
Ali Bahri Gustavo A. Vargas Hakim David Osowiechi Moslem Yazdanpanah
Ismail Ben Ayed Christian Desrosiers

ÉTS Montreal, Canada
International Laboratory on Learning Systems (ILLS)

Abstract

*Domain Generalization techniques aim to enhance model robustness by simulating novel data distributions during training, typically through various augmentation or stylization strategies. However, these methods frequently suffer from limited control over the diversity of generated images and lack assurance that these images span distinct distributions. To address these challenges, we propose **FDS**, **F**eedback-guided **D**omain **S**ynthesis, a novel strategy that employs diffusion models to synthesize novel, pseudo-domains by training a single model on all source domains and performing domain mixing based on learned features. By incorporating images that pose classification challenges to models trained on original samples, alongside the original dataset, we ensure the generation of a training set that spans a broad distribution spectrum. Our comprehensive evaluations demonstrate that this methodology sets new benchmarks in domain generalization performance across a range of challenging datasets, effectively managing diverse types of domain shifts. The code can be found at:* https://github.com/Mehrdad-Noori/FDS

1. Introduction

Deep learning architectures, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have significantly advanced the field of computer vision, achieving state-of-art results in tasks like classification, semantic segmentation, and object detection. Despite these advancements, such models commonly operate under the simplistic assumption that training data (source domain) and post-deployment data (target domain) share identical distributions. This overlook of distributional shifts results in performance degradation when models are exposed to out-of-distribution (OOD) data [24, 47]. Domain adaptation (DA) [35, 51] and Test-Time Adaptation (or Training) [22, 43, 60] strategies have been developed to mitigate this issue by adjusting models trained on the source domain to accommodate a predefined target domain. Nonetheless, these strategies are constrained by their dependence on accessible target domain data for adaptation, a prerequisite that is not always feasible in real-world applications. Furthermore, adapting models to each novel target domain entails considerable computational overhead, presenting a practical challenge to their widespread implementation.

Domain generalization (DG) [6] aims to solve the issue of domain shift by training models using data from one or more source domains so that they perform well *out-of-the-box* on new, unseen domains. Recently, various techniques have been developed to tackle this problem [61, 69], including *domain aligning* [26, 33, 36], *meta-learning* [5, 31], *data augmentation* [52, 53, 59], *ensemble learning* [72], *self-supervised learning* [1, 9], and *regularization methods* [11, 28]. These strategies are designed to make models more adaptable and capable of handling data that they were not explicitly trained on, making them more useful in real-world situations where the exact nature of future data cannot be predicted. Among these techniques, a notable category focuses on synthesizing samples from different distributions to mimic target distributions. This is achieved through strategies like *image transformation* [53, 68], *style transfer* [7, 54], *learnable augmentation networks* [10, 70, 71], and *feature-level stylization* [14, 42, 72]. However, many of these methods face challenges in controlling the synthesis process, often resulting in limited diversity where primarily only textures are altered.

In this study, we introduce an innovative approach using diffusion model, named Feedback-guided Domain Synthesis (FDS), to address the challenge of domain generalization. Known for their exceptional ability to grasp intricate distributions and semantics, diffusion models excel at producing high-quality, realistic samples [16, 25, 48, 55]. We exploit this strength by training a single diffusion model that is conditioned on various *domains* and *classes* present in the training

*Correspondence to mehrdad.noori.1@ens.etsmtl.ca

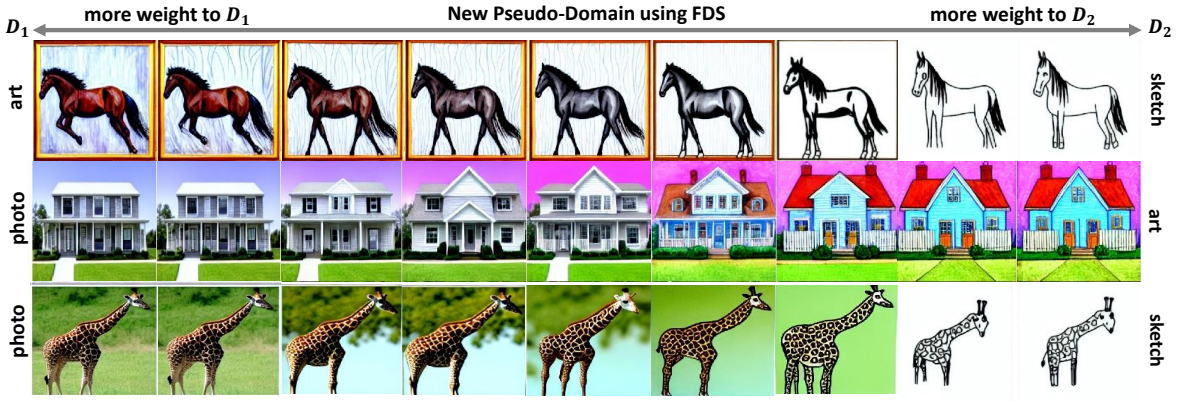

Figure 1. Generating new, pseudo-domains with FDS: Comprehensive distribution coverage from domain D_1 to D_2.

dataset, aiming to master the distribution of source domains.

As illustrated in Figure 1, through the process of domain interpolation and mixing during generation, we create images that appear to originate from novel, pseudo-domains. To ensure the development of a robust classifier, we initially select generated samples that are difficult for a model trained solely on the original source domains to classify. Subsequently, we train the model using a combination of these challenging images and the original dataset. In this manner, we make sure that our model is fed with images of the widest possible diversity, thereby significantly enhancing its ability to generalize across unseen domains. Our contributions can be summarized as follows:

- We introduce a novel DG approach that leverages diffusion models conditioned on multiple domains and classes to generate samples from novel, pseudo-domains through domain interpolation. This approach increases the diversity and realism of the generated images;
- We propose an innovative strategy for selecting images that pose a challenge to a classifier trained on original images, ensuring the diversity of the final sample set. By incorporating these challenging images with the original dataset, we ensure a comprehensive and diverse training set, significantly improving the model's generalization capabilities;
- We conduct extensive experiments across various benchmarks and perform different analyses to validate the effectiveness of our method. These experiments demonstrate FDS's ability to significantly improve the robustness and generalization of models across a wide range of unseen domains, achieving SOTA performance.

2. Related Works

2.1. Domain Generalization (DG)

Domain Generalization, a concept first introduced by Blanchard et al. in 2011 [6], has seen growing interest in the field of computer vision. This interest has spurred the development of a broad spectrum of methods aimed at enabling models to generalize across unseen domains. These include approaches based on *domain alignment* like moment matching [44], discriminant analysis [26] and domain-adversarial learning [34], *meta-learning* approaches [5, 31] which solve a bi-level optimization problem where the model is fine-tuned on meta-source domains to minimize the error on a meta-target domain, *ensemble learning* techniques [38, 72] improving robustness to OOD data by training multiple models tailored to specific domains, *self-supervised learning* methods fostering domain-agnostic representations through the pre-training of models on unsupervised tasks [9, 20, 62] or via contrastive learning [29], approaches leveraging *disentangled representation learning* [13, 30] segregating domain-specific features from those common across all domains, and *regularization* methods which build on the empirical risk minimization (ERM) framework [21], incorporating additional objectives such as distillation [56, 63], stochastic weight averaging [11], or distributionally robust optimization (DRO) [49] to promote generalization.

A key strategy in DG, *data augmentation* aims to enhance model robustness against domain shifts encountered during deployment. This objective is pursued through a variety of techniques, such as learnable augmentation, off-the-shelf style transfer, and augmentation at the feature level. Learnable augmentation models utilizes networks to create images from training data, ensuring their distribution diverges from that of the source domains [10, 70, 71]. Meanwhile, off-the-shelf style transfer based methods seek to transform the appearance of images from one domain to another or modify their stylistic elements, often through Adaptive Instance Normalization (AdaIN) [27, 54]. Unlike most augmentation approaches that modify pixel values, some propose altering features directly, a technique inspired by the finding that CNN features encapsulate style information [37, 72] in their statistics.

2.2. Diffusion Models

Diffusion models have recently surpassed Generative Adversarial Networks (GANs) as the leading technique for image synthesis. Innovations like denoising diffusion proba-

bilistic models (DDPMs) [25] and denoising diffusion implicit models (DDIMs) [55] have significantly sped up the image generation process. Rombach et al. [48] introduced latent diffusion models (LDMs), also known as stable diffusion models, which enhance both training and inference efficiency while facilitating text-to-image and image-to-image conversions. Extensions of these models, such as stable diffusion XL (SDXL) [45] and ControlNet [67], have been developed to further guide the generation process with additional inputs, such as depth or semantic information, allowing for more controlled and versatile image creation. Recent studies have shown that augmenting datasets using diffusion models can improve performance in general vision tasks [4, 17].

Despite the capabilities of diffusion models in generating images, their application in domain generalization has been minimally explored. Miao et al. [39] introduced DomainDiff, a method that boosts OOD generalization by training a Word-to-Image Mapping (WIM) with diffusion models to generate additional synthetic data. Yue et al. [66] proposed a method called DSI to address OOD prediction by transforming testing samples back to the training distribution using multiple diffusion models, each trained on a single source distribution. While effective, DSI requires multiple models during prediction, making it impractical for real-time deployment. In contrast, our method only uses diffusion during training to synthesize pseudo-domains and select more challenging images, keeping the computational complexity the same during testing while achieving significantly better performance.

In another study, CDGA [23] uses diffusion models to generate synthetic images that fill the gap between different domain pairs through a simple interpolation. While effective, CDGA relies on generating a large number of synthetic images (e.g., 5M for PACS) and employs naive interpolation. Additionally, CDGA generates samples with additional text descriptions that are not provided in standard DG benchmarks. In contrast, our method employs novel and more efficient interpolation techniques and a unique filtering mechanism that selects challenging images. We demonstrate that both of these innovations are crucial for improving generalization.

3. Theoretical Motivation

Our classifier, represented by f with parameters θ, aims to craft a unified model from n source domains $\{\mathcal{D}^1, \ldots, \mathcal{D}^n\}$ that adapts to a novel target domain $\mathcal{D}^{\mathcal{T}}$. Within any domain $\mathcal{D}$, we measure classification loss by

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

where x and y denote the input and its corresponding label, respectively, and $\ell(f(x;\theta), y)$ is the cross entropy loss in this work.

Empirical Risk Minimization (ERM) [50] forms the foundation for training our models, aiming to reduce the mean loss across training domains:

$$\min_{\theta} \sum_{i} \frac{1}{|\mathcal{D}^i|} \sum_{k=1}^{|\mathcal{D}^i|} \ell(f(x_i^k;\theta), y_i^k) \tag{2}$$

Here, $|\mathcal{D}^i|$ counts the number of samples in domain i, with x_i^k as the k-th sample and y_i^k its label. However, the accuracy of ERM-trained models drops when data shifts occur across domains due to inadequate OOD generalization. To enhance ERM, Chapelle et al. introduced Vicinal Risk Minimization (VRM) [12], which substitutes point-wise estimates with density estimation in the vicinity distribution around every observation within each domain. Practical implementation often involves data augmentation to introduce synthetic samples from these density estimates. Traditional augmentation processes a single data point x_i^k from domain i to yield $\tilde{x}_i^k = g(x_i^k)$, where $g(\cdot)$ denotes a basic transformation and $\tilde{x}_i^k$ is the modified data point.

Despite the potential of VRM to boost performance on OOD samples, it does not completely bridge domain gaps. The root cause lies in the inability of ERM methods to anticipate shifts in data distribution, which simple augmentation within domains does not address. Hence, domain generalization requires more robust transformations to extend the model's applicability beyond training domains. Muller et al. [40] have demonstrated that a wider range of transformations outperforms standard methods. However, Aminbeidokhti et al. [2] advise that aggressive augmentations could distort the essential characteristics of images, pointing to the necessity of a mechanism to filter out extreme alterations. The emergence of diffusion models, proficient in reproducing diverse data distributions, facilitates such advanced sampling approaches. Our goal, therefore, is to generate images that span across domain gaps, lessen the variability between data distributions, and introduce a system to exclude trivial or excessive modifications.

4. Method

To advance the generalization ability of our classifier, we aim to utilize generated image samples that traverse domain gaps yet retain class semantics. This objective is realized through a three-step methodology at the core of our FDS approach. **Step 1:** we begin by training an image generator that masters the class-specific distributions of all source domains, enabling the synthesis of class-consistent samples within those domains. **Step 2:** we then introduce a strategy for generating inter-domain images to span the domain gaps effectively. **Step 3:** last, we filter overly simplistic samples from synthetic inter-domain images and train the final model on this refined set alongside original domain images for enhanced OOD generalization. The process is illustrated in Figure 2.

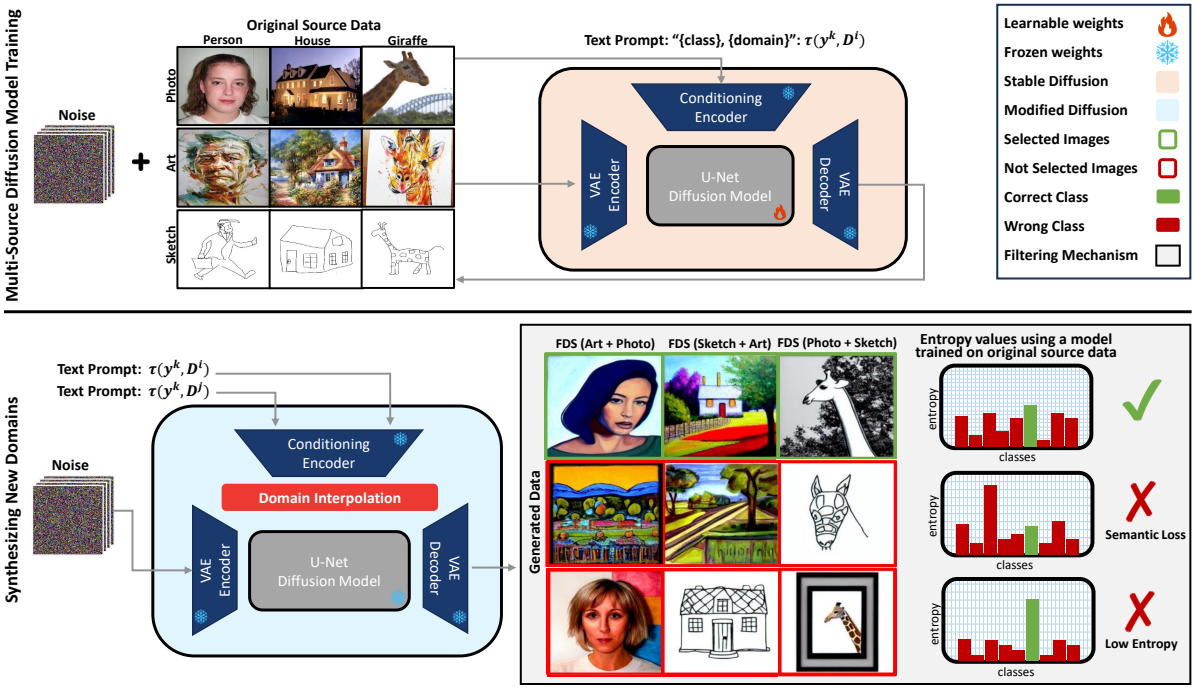

Figure 2. Overview of the proposed architecture for FDS. (top) Multi-source training of diffusion model conditioned on class and domain of the training images. (bottom) Generating novel pseudo-domain using the proposed interpolation and filtering mechanism of FDS.

4.1. Image Generator

Diffusion Models are designed to approximate the data distribution $p(x)$ by reversing a predefined Markov Chain of T steps, effectively denoising a sample in stages. These stages are modeled as a sequence of denoising autoencoder applications $e_\theta(x_t, t)$, for $t = 1 \dots T$, which gradually restore their input x_t. This iterative restoration is formalized by the following objective:

$$\mathcal{L}_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),\,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \quad (3)$$

where t is drawn uniformly from $\{1, \dots, T\}$. Utilizing perceptual compression models, encoded by $\mathcal{E}$ and decoded by $\tilde{\mathcal{E}}$, we access a latent space that filters out non-essential high-frequency details. By placing the diffusion process in this compressed space, our objective is then reformulated as

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),\,t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right]. \quad (4)$$

Diffusion models can also handle conditional distributions $p(z|c)$. To enable this conditioning, building upon the work of Rombach et al. [48], we utilize a condition encoder τ_θ that maps c onto an intermediate representational space $\tau(c) \in \mathbb{R}^{M\times d_\tau}$. Following Stable Diffusion, we employ the CLIP-tokenizer and implement τ as a transformer to infer a latent code. This representation is subsequently integrated into the UNet using a cross-attention mechanism, culminating in our final enhanced objective:

$$\mathcal{L}^{\text{cond}}_{LDM} = \mathbb{E}_{\mathcal{E}(x),\,c,\,\epsilon\sim\mathcal{N}(0,1),\,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(c))\|^2\right], \quad (5)$$

This approach enables a single diffusion model to understand and generate images across different domains for each class, using text-based conditions (prompts). By dynamically adjusting conditions, the model efficiently learns varied representations without needing multiple models for each domain. Thus, we employ a textual template $c(x)$, denoting "`[y], [D]`" where `y` is the class label and `D` is the domain name of the input x, and proceed to train our diffusion model using this template across all images from the source domains. Upon completing training, we can create a new sample for class y^k, belonging to the set $\{y^1, \dots, y^m\}$, within domain $\mathcal{D}^i$. This is achieved by decoding a denoised representation after t timesteps, $\tilde{x}_t^{i,k} = \tilde{\mathcal{E}}(\Phi_t(\mathcal{D}^i, y^k))$, where $\Phi_t(\mathcal{D}^i, y^k)$ is the denoised representation that originates from random Gaussian noise conditioned on $\mathcal{D}^i$ and y^k.

4.2. Domain Mixing

We propose two mixing strategies to synthesize images from new, pseudo distributions, based on *noise-level interpolation* and *condition-level interpolation*.

4.2.1 Noise Level Interpolation

Consider our dataset comprising n source domains $\{\mathcal{D}^1, \dots, \mathcal{D}^n\}$ and target classes y^k from the set $\{y^1, \dots, y^m\}$. Utilizing a trained image generator that initiates with random Gaussian noise, we denoise this input over t steps to produce a synthetic, denoised representation $z_t^{i,k} = \Phi_t(\mathcal{D}^i, y^k)$ indicative of domain $\mathcal{D}^i$ and class y^k. To

synthesize a sample that merges the characteristics of domains $\mathcal{D}^i$ and $\mathcal{D}^j$ for class y^k, we employ a single diffusion model, conditioned dynamically to capture the essence of both domains. This process aims to generate a sample that embodies the transitional features between these domains, effectively bridging the domain gap.

The model begins its process from the same initial random Gaussian noise, adapting its denoising trajectory under two distinct conditions, $\tau(\mathcal{D}^i, y^k)$ and $\tau(\mathcal{D}^j, y^k)$, up to a specific timestep T. This dual-conditioned approach ensures that the evolving representation up to T incorporates influences from both domains, guided by the respective conditional inputs.

From timestep T onwards, until the final representation is formed, the model blends the outputs from these dual paths at each step. Specifically, for each step $t > T$, we form a mixed representation $\tilde{z}_t$ as:

$$\tilde{z}_t \;=\; \alpha\, \Phi_t(\mathcal{D}^i, y^k) \;+\; (1-\alpha)\, \Phi_t(\mathcal{D}^j, y^k), \qquad (6)$$

where α is a predefined mixing coefficient that dictates the blend of domain characteristics in the output. This combined representation $\tilde{z}_t$ is then used as the basis for the model's next denoising step, integrating features from both $\mathcal{D}^i$ and $\mathcal{D}^j$ for class y^k. Through this iterative mixing, the model ensures a gradual and cohesive fusion of domain-specific attributes, leading to a synthesized sample that seamlessly spans the gap between the domains for class y^k.

4.2.2 Condition Level Interpolation

We also propose Condition Level Interpolation to generate images that effectively bridge domain gaps. This technique relies on manipulating the conditions fed into our diffusion model to guide the synthesis of new samples. Specifically, for a target class y^k and two distinct domains, $\mathcal{D}^i$ and $\mathcal{D}^j$, we employ our encoder $\tau(c)$ to create separate condition representations for each domain-class pair: $(y^k, \mathcal{D}^i)$ and $(y^k, \mathcal{D}^j)$.

The core of this strategy involves blending these condition representations using a mixing coefficient α, leading to a unified condition:

$$c_{\text{mixed}} \;=\; \alpha\, \tau(c_{y^k}, \mathcal{D}^i) \;+\; (1-\alpha)\, \tau(c_{y^k}, \mathcal{D}^j). \qquad (7)$$

This mixed condition c_{mixed} then orchestrates the generation process from the initial step, ensuring that the diffusion model is consistently influenced by attributes from both domains. By initiating this conditioned blending from the beginning of the diffusion process, we ensure a harmonious integration of domain characteristics throughout the generation of the synthetic image.

4.3. Filtering Mechanism

Through our mixing strategies, we create synthetic samples $\tilde{x}^k_{i,j}$ that not only synthesize class y^k traits but also blend features from domains $\mathcal{D}^i$ and $\mathcal{D}^j$, thereby aiming to bridge the domain gaps. This approach generates a synthetic dataset comprising $\tilde{N}$ samples for each combination of class index k and domain index pair (i, j), structured as $S^k_{i,j} = \{\tilde{x}^{k,(r)}_{i,j} \,|\, 1 \le i < j \le n,\ 1 \le k \le m,\ 1 \le r \le \tilde{N}\}$, ensuring diversity via distinct random Gaussian noise initiation for each sample.

The utility of these synthetic samples in improving model generalization varies, prompting an entropy-based evaluation to identify those with the greatest potential. High entropy scores, indicating prediction uncertainty by a classifier $h(x)$ trained on the original dataset, suggest that such samples may come from previously unseen distributions. This characteristic posits these high-entropy samples as prime candidates for training, hypothesized to challenge the classifier significantly and aid in covering the domain gaps. Further refining this selection, we only include samples correctly predicted as their target class by $h(x)$, ensuring the exclusion of samples that have lost semantic integrity during the diffusion process.

Let $\mathcal{C}^k_{i,j}$ be the subset of samples in $S^k_{i,j}$ which are correctly classified by $h(x)$:

$$\mathcal{C}^k_{i,j} \;=\; \{\tilde{x}^{k,(r)}_{i,j} \,|\, h(\tilde{x}^{k,(r)}_{i,j}) = y^k\}. \qquad (8)$$

We choose from $\mathcal{C}^k_{i,j}$ the N_L samples with highest entropy (the entropy of a k-class discrete probability distribution p is given by $\mathcal{H}(p) = -\sum_k p_k \log p_k$) to form a set of selected samples $\tilde{\mathcal{D}}^k_{i,j}$. Last, we combine the synthetic samples, created for each class and domain pair, to the original dataset $\mathcal{O}$ to obtain the final augmented training set

$$\mathcal{A} \;=\; \mathcal{O} \,\cup\, \left\{\tilde{\mathcal{D}}^k_{i,j} \,|\, 1 \le i < j \le n,\ 1 \le k \le m\right\} \qquad (9)$$

Training the final classifier on $\mathcal{A}$ not only enriches the dataset but also ensures robust model generalization across diverse domain landscapes.

5. Experimental Setup

Datasets. Following [21], we compare our proposed approach to the current state-of-art using three challenging datasets–**12 individual target domains**–with different characteristics: `PACS` [30], `VLCS` [18], and `OfficeHome` [58]. The PACS dataset has a total of 9,991 photos divided into four distinct domains, $d \in$ {Art, Cartoon, Photo, Sketch}, and seven distinct classes. The second dataset, `VLCS`, comprises 10,729 photos from four separate domains, $d \in$ {Caltech101, LabelMe , SUN09, VOC2007}, and five different classes. The third dataset, `OfficeHome`, includes a total of 15,588 photos taken from four domains, $d \in$ {Art, Clipart, Product, Real}, and 65 classes.

Implementation Details. To ensure a fair comparison, we adopt the DomainBed framework [21], a comprehensive benchmark that encompasses prominent domain generalization (DG) methodologies under a uniform evaluation protocol. Following this framework, we employ a leave-one-out

	Method	Aug.	PACS	VLCS	Office	Avg.
Standard Methods	ERM (*baseline*) [21]	✗	85.5 ±0.2	77.5 ±0.4	66.5 ±0.3	76.5
	ERM (*reproduced*)	✗	84.3 ±1.1	76.2 ±1.1	64.6 ±1.1	75.0
	IRM [3]	✗	83.5 ±0.8	78.5 ±0.5	64.3 ±2.2	75.4
	GroupDRO [49]	✗	84.4 ±0.8	76.7 ±0.6	66.0 ±0.7	75.7
	Mixup [64]	✓	84.6 ±0.6	77.4 ±0.6	68.1 ±0.3	76.7
	CORAL [57]	✗	86.2 ±0.3	78.8 ±0.6	68.7 ±0.3	77.9
	MMD [32]	✗	84.6 ±0.5	77.5 ±0.9	66.3 ±0.1	76.1
	DANN [19]	✗	83.6 ±0.4	78.6 ±0.4	65.9 ±0.6	76.0
	SagNet [41]	✓	86.3 ±0.2	77.8 ±0.5	68.1 ±0.1	77.4
	RSC [28]	✓	85.2 ±0.9	77.1 ±0.5	65.5 ±0.9	75.9
	Mixstyle [72]	✓	85.2 ±0.3	77.9 ±0.5	60.4 ±0.3	74.5
	mDSDI [8]	✗	86.2 ±0.2	79.0 ±0.3	69.2 ±0.4	78.1
	SelfReg [29]	✓	85.6 ±0.4	77.8 ±0.9	67.9 ±0.7	77.1
	DCAug [2]	✓	86.1 ±0.7	78.6 ±0.4	68.3 ±0.4	77.7
	DomainDiff [39]	✓	85.6 ±0.6	–	63.7 ±0.6	–
	DSI [66]	✓	86.9 ±1.4	–	–	–
	CDGA [23]	✓	88.5 ±0.5	79.6 ±0.3	68.2 ±0.6	78.8
	ERM + FDS (ours)	✓	**88.8** ±0.1	**79.8** ±0.5	**71.1** ±0.1	**79.9**
WA Methods	SWAD (*baseline*) [11]	✗	88.1 ±0.1	79.1 ±0.1	70.6 ±0.2	79.3
	SWAD (*reproduced*)	✗	88.1 ±0.4	78.9 ±0.5	70.3 ±0.4	79.1
	SelfReg SWA [29]	✓	86.5 ±0.3	77.5 ±0.0	69.4 ±0.2	77.8
	DNA [15]	✗	88.4 ±0.1	79.0 ±0.1	71.2 ±0.1	79.5
	DIWA [46]	✓	88.8 ±0.4	79.1 ±0.2	71.0 ±0.1	79.6
	TeachDCAug [2]	✓	88.4 ±0.2	78.8 ±0.4	70.4 ±0.2	79.2
	SWAD + FDS (ours)	✓	**90.5** ±0.3	**79.7** ±0.5	**73.5** ±0.4	**81.3**

Table 1. Leave-one-out accuracy (%) results on the PACS, VLCS, and OfficeHome benchmarks. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and second-best results are highlighted.

strategy for DG dataset assessment where one domain serves as the test set while the others form the training set. A subset of the training data, constituting 20%, is designated as the validation set[1]. The aggregate result for each dataset represents the mean accuracy derived from varying the test domain. To ensure reliability, experiments are replicated three times, each with a unique seed. Moreover, to rigorously test our method, we try both ERM and SWAD classifiers with FDS. The former is considered a baseline in standard training methods, while the latter serves as a baseline for weight averaging (WA) methods. SWAD is essentially ERM but with weight averaging applied during training using multiple steps based on the validation set. To have a fair comparison with other methods, for the ERM baseline, we strictly follow the hyperparameter tuning proposed in [11]. For SWAD, as in the original work, we did not tune any parameters and used the default values. For image synthesis, the original Stable Diffusion framework [48] is utilized with DDIM$=50$ steps. The PACS and VLCS datasets prompt the generation of $N = 32,000$ samples per class, whereas OfficeHome, with its 65 classes, necessitates $N = 16,000$ samples. Image generation spans an interpolation range of $\alpha \in [0.3, 0.7]$ and a Noise Level Interpolation range of $\mathcal{T} \in [20, 45]$. This diversified parameter selection, rather than optimizing hyperparameters per dataset, acknowledges each domain's unique shift. Our filtering mechanism then identifies the most informative images from the generated pool. The impact of dataset size, N_L, is studied further in Section 6.2.

[1] The model with peak accuracy on this validation set is selected for evaluation on the test domain, providing unseen domain accuracy.

Module	Target Domains				
	Art	Cartoon	Photo	Sketch	Avg.
Baseline (SWAD [11])	89.49 ±0.2	83.65 ±0.4	97.25 ±0.2	82.06 ±1.0	88.11 ±0.4
+ Basic Gen.	89.87 ±0.1	85.59 ±0.6	97.50 ±0.3	83.07 ±0.4	89.01 ±0.4
+ Interpolation	91.38 ±0.2	85.20 ±0.6	97.73 ±0.1	84.27 ±0.9	89.65 ±0.4
+ Filtering	**91.80** ±0.3	**86.03** ±0.8	**98.05** ±0.2	**86.11** ±0.1	**90.50** ±0.3

Table 2. Comparative analysis of FDS component effects on accuracy (%) across PACS dataset domains. "**Basic Gen.**" refers to generation without interpolation or filtering.

6. Results

We first compare the performance of our FDS approach against SOTA DG methods across three benchmarks. We then present a detailed analysis investigating several key aspects of our approach, including the effectiveness of each proposed component, a comparison of different mixing strategies, its regularization capabilities, domain diversity visualization and quantification, the impact of data size, the efficacy of our filtering mechanism, and stability analysis during training. Additional analysis, visualizations, and detailed tables can be found in the Supplementary Material.

6.1. Comparison with the State-of-the-art

In Table 1, we compare our approach with recent methods for domain generalization as outlined in the DomainBed framework [21]. Our method, when added to the baseline ERM classifier, shows an impressive improved accuracy of 4.5% on the PACS dataset using the ResNet-50 model. On the VLCS dataset, our method sees a 3.6% increase in accuracy over the ERM. For the OfficeHome dataset, our method outperforms the ERM baseline by 6.5%. To test the strength of our method, we also applied it to SWAD, which is a the baseline for weight averaging (WA) domain generalization methods. Here, our method improves the performance by 2.4%, 0.8%, and 3.2% on the PACS, VLCS, and OfficeHome datasets, respectively.

Furthermore, when comparing with previous SOTA methods, our method outperforms CDGA [23] by 0.3% on PACS, 0.2% on VLCS, and 2.9% on OfficeHome. In the context of weight averaging methods, our approach surpasses DIWA [46], which trains multiple independent models, by 1.7% on PACS, 0.6% on VLCS, and 2.5% on OfficeHome, setting a new benchmark for domain generalization. Please refer to the Supplementary Material for the full, detailed results for each dataset and its domains.

6.2. Further Analysis

For this section, we use the SWAD baseline to analyze the performance of our proposed method FDS, due to its stable performance as a WA method. Final settings are applied to the ERM baseline. All analyses in this section use the PACS dataset and the SWAD baseline unless otherwise stated.

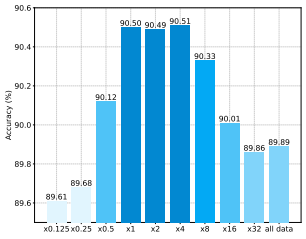

Figure 3. Impact of varying scales of sample size N_L relative to the average number of images per class on PACS dataset.

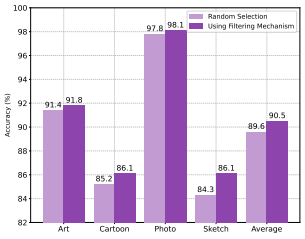

Figure 4. Impact of using Random Selection vs. Proposed Filtering Strategy of FDS on PACS accuracy (%).

Strategy	Target Domains				
	Art	Cartoon	Photo	Sketch	Avg.
Baseline (SWAD [11])	89.49 ±0.2	83.65 ±0.4	97.25 ±0.2	82.06 ±1.0	88.11 ±0.4
Noise Level Interpol.	**91.95 ±0.3**	83.23 ±0.1	97.56 ±0.1	85.85 ±0.3	89.65 ±0.2
Condition Level Interpol.	91.80 ±0.3	**86.03 ±0.8**	**98.05 ±0.2**	**86.11 ±0.1**	**90.50 ±0.3**
Both	91.13 ±0.4	82.98 ±0.2	97.90 ±0.2	85.62 ±0.8	89.41 ±0.4

Table 3. Impact of different interpolation strategies of FDS on PACS accuracy (%).

Ablation Study on Different Components. To evaluate the impact of each component on generalization, we systematically introduce each module and observe the enhancements. As detailed in Table 2, first incorporating domain-specific synthetic samples, generated without interpolation or filtering (Basic Gen.), yields a 0.9% gain over the baseline. This increment validates our diffusion model's proficiency in capturing and replicating the class-specific distributions within domains, thus refining OOD performance through enhanced density estimation near original samples. Subsequently, integrating images from pseudo-novel domains via our interpolation strategy leads to a further 0.5% enhancement, underscoring the mechanism's effectiveness in connecting distinct domains. Lastly, applying our entropy-based filtering to eliminate overly simplistic images and images with diminished semantic relevance results in a significant 2.4% improvement against the SWAD model, a robust benchmark. These outcomes collectively underscore the efficacy of our approach in improving OOD generalization.

Mixing Strategies. Our ablation study, presented in Table 3, reveals that both Noise Level Interpolation and Condition Level Interpolation significantly outperform the baseline, yet their effectiveness varies with the dataset's attributes and the nature of the domain shift. Noise Level Interpolation is optimal for minimal domain shifts, focusing on adjusting the noise aspect to bridge domain gaps. However, it falls short in scenarios with substantial domain differences, such as the transition to Cartoon or Sketch, where the source and target domains diverge significantly. In these cases, Condition Level Interpolation proves more advantageous, offering a robust mechanism for navigating complex domain shifts by manipulating higher-level semantic representations. When applying both methods simultaneously, the performance did not improve compared to using Condition Level Interpolation alone, possibly due to the increased complexity. Nonetheless, it still performed better than the baseline. Based on these results, we use Condition Level Interpolation as our final interpolation method.

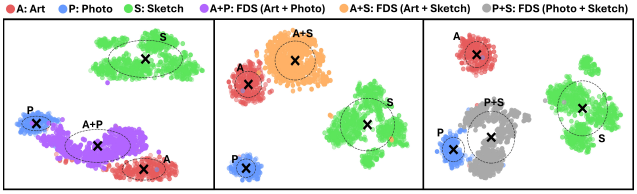

Figure 5. t-SNE plots showcasing the original "giraffe" class samples for the "Art", "Photo" and "Sketch" source domains of the PACS dataset, as well as the data generated with FDS.

Method	PACS	VLCS	Office	Avg.
Baseline (SWAD [11])	88.11 ±0.4	78.87 ±0.5	70.34 ±0.4	79.11
Filtered Based on Entropy	90.50 ±0.4	79.33 ±0.8	71.97 ±0.3	80.60
+ Reject Semantic Loss	**90.50 ±0.3**	**79.73 ±0.5**	**73.51 ±0.5**	**81.25**

Table 4. Impact of filtering strategy components of FDS on accuracy (%), using three benchmarks.

Impact of Sample Size. Next, we assess the impact of the selected data size, denoted as N_L, on final performance. We consider the PACS dataset for this analysis, where the average number of images per class is 570. We systematically explore varying scales relative to this average class size, aiming to discern the optimal dataset size for enhancing OOD generalization. Our findings, detailed in Figure 3, reveal an initial improvement in OOD generalization with increased data size. However, excessively enlarging the dataset size begins to diminish the benefits of our filtering mechanism, as it incorporates a broader array of samples, including those that are overly simplistic and not conducive to model improvement.

Filtering Mechanism. One might consider that the enhanced accuracy observed with our filtering mechanism might be the result of constraining the sample size for a balanced final training set. To investigate this hypothesis, Figure 4 contrasts the outcome of selecting N_L samples at random from all generated images per class against employing our entropy-based filtering strategy. The consistent improvement in out-of-domain (OOD) generalization across all domains, facilitated by our filtering approach, underscores its effectiveness.

Additionally, to examine the impact of excluding samples that deviate semantically which will be identified through misclassification by a classifier trained solely on the original dataset, we contrasted the accuracy between selections purely based on entropy and those refined this way in Table 4. This comparison highlights the significance and efficiency of our filtering strategy in preserving semantic integrity.

Domain Diversity Visualization. To illustrate the effectiveness of our method, we present t-SNE plots of original PACS dataset samples and those generated by FDS in Figure 5. The t-SNE plots show distinct clusters for original source domains (here Art, Photo, Sketch) and highlight how the gen-

Data	Target Domains				
	Art	Cartoon	Photo	Sketch	Avg.
Original PACS	0.39	0.87	0.54	0.99	0.70
Basic Gen.	0.54	0.72	0.45	1.01	0.68
FDS	0.44	0.58	0.49	0.92	**0.61**

Table 5. Domain diversity metric [65] between source domains and the target domain of the PACS dataset. "**Basic Gen.**" refers to generation without interpolation or filtering.

erated samples clearly bridge the gaps between these clusters, thus enhancing domain diversity. This expanded diversity is crucial for improving the generalization capabilities of models, as it ensures a broader spectrum of data distributions in the training set. By providing a continuous representation of the domain space, our method facilitates smoother transitions and better prepares models to handle unseen domains, ultimately contributing to more robust performance in real-world applications. Please check the supplementary materials (Figure 8) for further details.

Domain Diversity Quantification. We further validate the effectiveness of our method using a domain diversity metric based on the methodology proposed in [65]. This metric quantifies the diversity shift between the source domains and the target domain, providing valuable insight into how the newly generated domains using our method can improve generalization on unseen domains. Table 5 presents the domain diversity metric for the original PACS dataset, comparing it with samples generated using the diffusion model in its basic form (no interpolation and no filter) and our FDS method (including interpolation and filtering). We observe that our FDS method results in a lower diversity metric compared to using the original PACS dataset or the samples generated with basic generation, suggesting that our method more effectively promotes generalization to new, unseen domains.

In-Domain Regularization Effect. In this section, we study the impact of incorporating images generated by our method on the network's accuracy, when testing on the same domains as those used for training. To provide a more comprehensive analysis, we adopted an 80/10/10 split for the source domains[2]. We selected the best model on the validation set and reported its performance on the held-out in-domain test set. As indicated in Table 6, our FDS approach surpasses the baseline accuracy in this setup for both the ERM and SWAD backbones. Additionally, we compared our method to two alternative strategies: *i*) duplicating and augmenting original samples (Dup. Aug.), and *ii*) generating synthetic in-domain images (same number of images as FDS) without domain interpolation (Basic Gen.). While these strategies also improve the performance, FDS achieves the highest accuracy due to its ability to generate more diverse and challenging pseudo-domains, rather than simply increasing dataset size. This confirms the value of our method not only in OOD conditions but also in standard in-domain validation, aligning with

[2] This setting was only used for in-domain experiments, while the rest of the paper follows the standard 80/20 setting as used in DomainBed.

Method	Source Domains				
	C, S, P	A, P, S	A, C, S	A, C, P	Avg.
ERM	98.05 ±0.5	96.06 ±0.9	95.66 ±0.3	96.83 ±0.6	96.65
ERM + Dup. Aug.	98.01 ±0.2	97.07 ±0.3	96.46 ±0.2	97.28 ±0.3	97.21
ERM + Basic Gen.	96.96 ±0.4	96.74 ±0.2	96.23 ±0.2	97.37 ±0.5	96.83
ERM + FDS (ours)	97.77 ±0.4	97.53 ±0.5	96.71 ±0.1	98.02 ±0.4	**97.51**
SWAD	98.48 ±0.3	97.87 ±0.5	97.61 ±0.2	98.44 ±0.0	98.10
SWAD + Dup. Aug.	98.35 ±0.2	97.86 ±0.3	97.38 ±0.1	98.16 ±0.1	97.94
SWAD + Basic Gen.	98.42 ±0.4	98.23 ±0.3	97.87 ±0.5	98.64 ±0.2	98.29
SWAD + FDS (ours)	98.70 ±0.3	98.25 ±0.5	97.80 ±0.4	98.61 ±0.0	**98.34**

Table 6. Impact of FDS on in-domain PACS accuracy (%). 'A', 'C', 'P', and 'S' refer to 'Art', 'Cartoon', 'Photo', and 'Sketch'.

Method	Target Domains				
	Art	Cartoon	Photo	Sketch	Avg.
ERM	86.94 ±0.6	80.21 ±0.7	96.61 ±0.4	74.45 ±2.9	84.30
ERM + Dup. Aug.	85.34 ±0.9	80.99 ±0.9	94.98 ±0.8	76.88 ±2.0	84.55
ERM + Basic Gen.	87.21 ±0.3	80.90 ±1.9	95.71 ±0.4	80.31 ±2.3	86.03
ERM + FDS (ours)	90.69 ±0.9	84.19 ±0.6	97.21 ±0.1	82.99 ±0.4	**88.77**
SWAD	89.49 ±0.2	83.65 ±0.4	97.25 ±0.2	82.06 ±1.0	88.11
SWAD + Dup. Aug.	89.55 ±0.3	83.00 ±1.7	97.65 ±0.2	81.86 ±1.1	88.02
SWAD + Basic Gen.	89.87 ±0.1	85.59 ±0.6	97.50 ±0.3	83.07 ±0.4	89.01
SWAD + FDS (ours)	91.80 ±0.3	86.03 ±0.8	98.05 ±0.2	86.11 ±0.1	**90.50**

Table 7. Leave-one-out accuracy (%) of FDS compared to other augmentation strategies.

the principles of Vicinal Risk Minimization (VRM) [12]. Hence, our method may also be viewed as a regularization strategy, suitable for a broad spectrum of applications. For completeness, we also report the out-domain performance of FDS and these strategies in Table 7. As can be seen, the beneficial impact of FDS on OOD generalization is not simply the result of increasing the dataset size.

7. Conclusion

This work presented FDS, a domain generalization (DG) technique that leverages diffusion models for domain mixing, generating a diverse set of images to bridge the domain gap between source domains distribution. We also proposed an entropy-based filtering strategy that enriches the pseudo-novel generated set with images that test the limits of classifiers trained on original data, thereby boosting generalization. Our extensive experiments across multiple benchmarks demonstrate that our method not only surpasses existing DG techniques but also sets new records for accuracy. Our analysis indicate that our approach contributes to more stable training processes when confronted with domain shifts and serves effectively as a regularization method in in-domain contexts. Notably, our technique consistently enhances performance across diverse scenarios, from realistic photos to sketches. While we exploited our trained diffusion model for covering the domain gap, more sophisticated techniques could be considered. For instance, future work could investigate the idea of generating pseudo-novel distributions via extrapolation in the domain space, in addition to interpolation.

References

[1] Isabela Albuquerque, Nikhil Naik, Junnan Li, Nitish Keskar, and Richard Socher. Improving out-of-distribution generalization via multi-task self-supervised pretraining. *arXiv preprint arXiv:2003.13525*, 2020. 1

[2] Masih Aminbeidokhti, Fidel A Guerrero Pena, Heitor Rapela Medeiros, Thomas Dubail, Eric Granger, and Marco Pedersoli. Domain generalization by rejecting extreme augmentations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2215–2225, 2024. 3, 6

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 6

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 3

[5] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 1, 2

[6] 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. 1, 2

[7] Francesco Cappio Borlino, Antonio D'Innocente, and Tatiana Tommasi. Rethinking domain generalization baselines. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9227–9233. IEEE, 2021. 1

[8] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021. 6

[9] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 1, 2

[10] Fabio Maria Carlucci, Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Hallucinating agnostic images to generalize across domains. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3227–3234. IEEE, 2019. 1, 2

[11] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 6, 7

[12] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000. 3, 8

[13] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 301–318. Springer, 2020. 2

[14] Milad Cheraghalikhani, Mehrdad Noori, David Osowiechi, Gustavo A Vargas Hakim, Ismail Ben Ayed, and Christian Desrosiers. Structure-aware feature stylization for domain generalization. *Computer Vision and Image Understanding*, 244:104016, 2024. 1

[15] Xu Chu, Yujie Jin, Wenwu Zhu, Yasha Wang, Xin Wang, Shanghang Zhang, and Hong Mei. DNA: Domain generalization with diversified neural averaging. In *International Conference on Machine Learning*, pages 4010–4034. PMLR, 2022. 6

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[17] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, 36:79024–79034, 2023. 3

[18] 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*, pages 1657–1664, 2013. 5

[19] 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. 6

[20] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 2

[21] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2021. 2, 5, 6

[22] Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Ismail Ben Ayed, and Christian Desrosiers. Clust3: Information invariant test-time training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6136–6145, 2023. 1

[23] Sobhan Hemati, Mahdi Beitollahi, Amir Hossein Estiri, Bassel Al Omari, Xi Chen, and Guojun Zhang. Cross domain generative augmentation: Domain generalization with latent diffusion models. *arXiv preprint arXiv:2312.05387*, 2023. 3, 6

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[26] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020. 1, 2

[27] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[28] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In

European Conference on Computer Vision, pages 124–140. Springer, 2020. 1, 6

[29] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021. 2, 6

[30] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2, 5

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[32] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 6

[33] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129, 2020. 1

[34] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 2

[35] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 1

[36] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 1

[37] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 466–483. Springer, 2020. 2

[38] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1353–1357. IEEE, 2018. 2

[39] Qiaowei Miao, Junkun Yuan, Shengyu Zhang, Fei Wu, and Kun Kuang. Domaindiff: Boost out-of-distribution generalization with synthetic data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5640–5644. IEEE, 2024. 3, 6

[40] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 774–782, 2021. 3

[41] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8690–8699, June 2021. 6

[42] Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Gustavo A Vargas Hakim, David Osowiechi, Ismail Ben Ayed, and Christian Desrosiers. Tfs-vit: Token-level feature stylization for domain generalization. *Pattern Recognition*, 149:110213, 2024. 1

[43] David Osowiechi, Gustavo A Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani, Ismail Ben Ayed, and Christian Desrosiers. Tttflow: Unsupervised test-time training with normalizing flow. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2126–2134, 2023. 1

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

[45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[46] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022. 6

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1

[48] 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*, pages 10684–10695, 2022. 1, 3, 4, 6

[49] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2, 6

[50] Stephan R Sain. The nature of statistical learning theory, 1996. 3

[51] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1

[52] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. 1

[53] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020. 1

[54] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*, 2020. 1, 2

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 3

[56] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. *arXiv preprint arXiv:2207.12392*, 2022. 2

[57] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 6

[58] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 5

[59] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5334–5344, 2018. 1

[60] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1

[61] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[62] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020. 2

[63] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2595–2604, 2021. 2

[64] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 6

[65] 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*, pages 7947–7958, 2022. 8

[66] Runpeng Yu, Songhua Liu, Xingyi Yang, and Xinchao Wang. Distribution shift inversion for out-of-distribution prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3592–3602, 2023. 3, 6

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[68] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020. 1

[69] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[70] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 1, 2

[71] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 561–578. Springer, 2020. 1, 2

[72] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 1, 2, 6