DomainFusion: Generalizing To Unseen Domains with Latent Diffusion Models

Yuyang Huang[1], Yabo Chen[1], Yuchen Liu[1], Xiaopeng Zhang[2(✉)], Wenrui Dai[1(✉)], Hongkai Xiong[1], and Qi Tian[2]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Huawei Inc., Shenzhen, China
{huangyuyang, chenyabo, liuyuchen6666, daiwenrui, xionghongkai}@sjtu.edu.cn
zxphistory@gmail.com, tian.qi1@huawei.com

Abstract. Latent Diffusion Models (LDMs) are powerful and potential tools for facilitating generation-based methods for domain generalization. However, existing diffusion-based DG methods are restricted to offline augmentation using LDM and suffer from degraded performance and prohibitive computational costs. To address these challenges, we propose DomainFusion to simultaneously achieve knowledge extraction in the latent space and augmentation in the pixel space of the Latent Diffusion Model (LDM) for efficiently and sufficiently exploiting LDM. We develop a Latent Distillation module that distills gradient priors from LDM to guide the optimization of DG models. Moreover, we design an online lightweight augmentation method by decomposing candidate images into styles and contents for using LDM in a fast and online fashion. Experimental results demonstrate that DomainFusion outperforms diffusion-based methods by a large margin and achieves SOTA performance on existing DG benchmark datasets. Remarkably, DomainFusion can significantly reduce the number of generated images (e.g. by more than 97% on DomainNet) without finetuning LDM.

Keywords: Domain Generalization · Latent Diffusion Models · Latent Vision Knowledge · Data Augmentation

1 Introduction

Deep learning methods have demonstrated remarkable achievements under the independent and identically distributed (i.i.d.) assumption on the training and test datasets. Unfortunately, when generalized to out-of-distribution (OOD) data, this assumption becomes inadequate and causes diminished performance due to substantial domain shifts [32, 33, 43, 63]. To bridge domain shifts, domain generalization (DG) [36, 74] leverages multiple source domains to learn domain-invariant features for generalizing to unseen target domains.

Yuyang Huang and Yabo Chen—Equal contribution.
Correspondence to Xiaopeng Zhang and Wenrui Dai.

A notable portion of DG methods neglect the scarcity of cross-domain data and are limited in practical applications. Generation-based DG methods [62] are effective alternatives to address the issue. Data of new domains are generated to augment the source domain, and thereby assist the model in acquiring domain-invariant features. Latent Diffusion Models (LDMs) have been particularly demonstrated effective in generating high-quality images through stable and scalable denoising objectives [38, 46, 47, 51] and have been employed in DSI [68] and CDGA [17]. Despite utilizing the powerful LDM and introducing massive computational costs, DSI [68] and CDGA [17] fail to yield state-of-the-art (SOTA) performance. These diffusion-based DG methods are limited by insufficient and inefficient utilization of LDM, as summarized below.

First, as shown in Fig. 1, the inefficient utilization of LDM results in substantial computational costs. Specifically, CDGA [17] generates a massive volume of more than 5 million synthetic images for data augmentation. The offline generation approach along with the enormous generation scale substantially increases the training cost due to excessive computational expenses and generation time and significantly inflated dataset size. DSI [68] is prohibitive in computational costs since it employs an LDM for each source domain and requires fine-tuning each LDM individually. **Second**, augmenting the source domain does not sufficiently exploit the capability of LDM for benefiting DG. It is widely acknowledged that the latent space of LDM encapsulates valuable pre-trained vision knowledge for downstream perception tasks such as image segmentation [57] and object detection [8]. Considering the ability of LDM to transfer images across diverse domains without compromising the underlying semantics, we reasonably argue that the latent space of LDM learns rich knowledge about domain-invariant feature representation, which shows significant potential for DG tasks.

To address these limitations, we propose DomainFusion to simultaneously achieve knowledge extraction in the latent space of LDM and augmentation in the pixel space for DG, as shown in Fig. 1 and Fig. 3. The issues of inefficient and insufficient utilization of LDM are addressed by augmentation in the pixel space with online lightweight generation and knowledge extraction in the latent space of LDM with Latent Distillation. The proposed DomainFusion achieves SOTA performance with significantly reduced computational costs. DomainFusion is shown to outperform state-of-the-art methods in multiple benchmark datasets using multiple backbones and reduce the number of generated images. For example, the number of generated images (including candidates) reduces by more than 97% on DomainNet compared with CDGA [17].

Specifically, the online lightweight generation approach efficiently generates batches of samples that are more suitable for DG every few epochs in training. It substantially reduces computational costs by requiring only a shared pre-trained LDM for all datasets and domains without extra fine-tuning. Multiple candidate images are generated and decomposed into styles and contents such that a new sample is formed by selecting the most distinct style and adopting the most similar content. The Latent Distillation significantly enhances DG by distilling visual knowledge from the latent space of LDM and utilizing it to guide

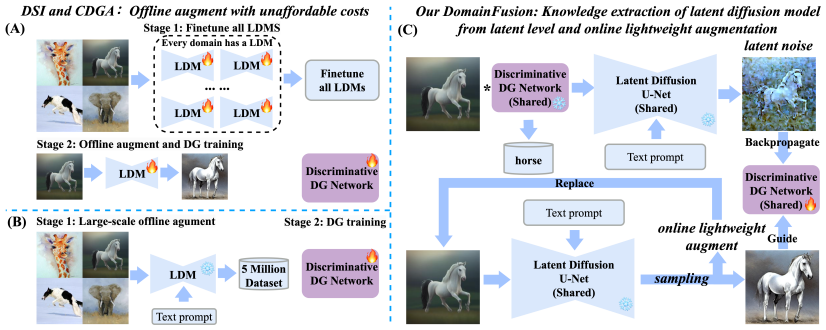

Fig. 1: Diffusion-based DG methods DSI [68] and CDGA [17] use LDM for offline augmentation only. They suffer from limitations of unaffordable costs and lagging behind SOTA methods. Specifically, (A) DSI [68] employs a separate LDM for each source domain and requires finetuning all LDMs before augmentation. (B) CDGA [17] generates more than 5 million synthetic images offline for augmentation, resulting in significant computational costs, prolonged generation time, and significantly increased training time of the DG model. (C) Proposed DomainFusion. To make full use of LDM's potential and improve performance, we believe that besides augmentation, the latent space of LDM contains valuable visual knowledge that can benefit DG. Therefore, we introduce Latent Distillation to extract knowledge from the latent space to guide the DG model. To reduce computational costs, we propose an online lightweight augmentation approach via a sampling strategy (illustrated in Fig. 4) that significantly reduces the number of generated samples (e.g. reduced the number of overall generated images by more than 97% on DomainNet) utilizing only a shared LDM without any finetuning.

parameter space updates of the classification model. It establishes a connection between the discriminative and latent diffusion parameter spaces and successfully produces supervised signals for the DG network. The discriminative parameter space is optimized using gradient prior derived from the latent parameter space.

The main contributions of this paper are summarized as below.

- We propose DomainFusion that simultaneously achieves knowledge extraction in the latent space and augmentation in the pixel space of the Latent Diffusion Model (LDM) to efficiently and sufficiently exploit LDM for DG.
- We develop Latent Distillation (LD) that leverages gradient priors from the latent space of LDM to guide the optimization of DG models.
- We design an online lightweight augmentation method that generates image samples with distinct styles and similar contents for DG to reduce the scale of generated images with a shared latent diffusion model without finetuning.

2 Related Work

2.1 Domain Generalization

Most domain generalization (DG) methods operate under the assumption of having access to a sufficient amount of cross-domain data [34,35,62]. The main focus

of these methods is to eliminate domain-specific biases and retain invariant features across multiple source domains, including learning more generalized feature representations [26, 39, 48, 64, 69] and optimization-based methods [2, 5, 31, 71]. Despite their success, these methods are limited by the scarcity of real-world cross-domain data, which hinders their practical applicability [40]. The alternative strategy focuses on data augmentation to generate new domains and diverse samples [22, 29, 30, 61, 66, 73, 75]. Recent DG methods utilize latent diffusion models for data augmentation, namely CDGA [17] and DSI [68]. However, they are inefficient in utilizing latent diffusion models and fall short of state-of-the-art performance. CDGA [17] generates over 5 million new samples, while DSI [68] requires a separate latent diffusion model for each domain and necessitates finetuning them before offline generation. These methods incur substantial computational costs and prolonged generation times.

2.2 Diffusion Models for Perception Vision Tasks

Diffusion models have emerged as the state-of-the-art in image generation [38,46, 47,51]. Moreover, they have proven to be successful in various perception vision tasks, including image segmentation [57], object detection [8], monocular depth estimation [72], and semantic correspondence [70]. Significantly, a substantial amount of research has been dedicated to extracting valuable vision knowledge from diffusion models. We believe that this knowledge can also benefit domain generalization as latent diffusion models have exhibited superb generalization capacity to transfer images to various domains while maintaining semantic information. In line with our objective of leveraging the diffusion model for domain generalization in image classification tasks, we classify existing approaches into two distinct groups based on their methodologies of utilizing latent diffusion models. The first group focuses on extracting feature maps and cross-attention maps from the denoiser to train an extra decoder for downstream tasks [70, 72]. However, this approach often requires prior knowledge of image categories, which are used as conditional inputs into the denoising process. As a result, it is not suitable for high-level visual tasks like image classification. The second group is based on Score Distillation Sampling (SDS) [18, 24, 42, 65], which demonstrates good scalability. However, to the best of our knowledge, this approach is only applicable to generative models intended for generating diffusion-like images, which means the critical prerequisite is ensuring that the trained model shares the same image generation objectives with latent diffusion models. Hence, we seek to explore leveraging high-level semantic knowledge from LDM in a more natural manner to optimize discriminative models for DG tasks.

3 Method

3.1 Preliminary of Diffusion Models

Diffusion models are generative models employing a dual-phase strategy [19]. The forward phase, represented by $\{q_t\}_{t\in[0,T]}$, diffuses clean data $\mathbf{x}_0$ into noisy states

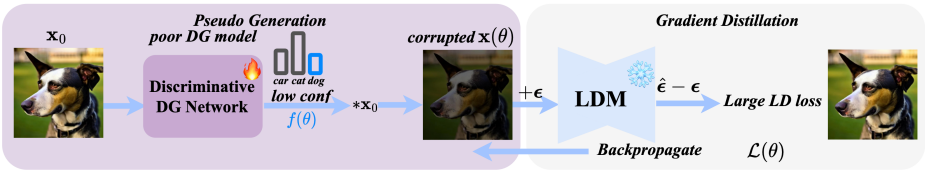

Fig. 2: Our Latent Distillation can be divided into two parts. In the pseudo generation part, the input image $\mathbf{x_0}$ is passed through the DG model θ, resulting in the confidence $f(\theta)$ for the corresponding class. The input image $\mathbf{x_0}$ is then multiplied by this confidence to obtain a corrupted image. In the gradient distillation part, noise $\boldsymbol{\epsilon}$ is added to the corrupted image, and then fed into the LDM. The LDM predicts the noise as $\hat{\boldsymbol{\epsilon}}$. We then compute the Latent Distillation (LD) loss based on $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$. Since the LD loss is dependent on θ, backpropagation can be performed to update the DG model θ. From an intuitive perspective, when the DG model poorly performs, it tends to generate low confidence, resulting in a corrupted image that becomes significantly darker. Consequently, LDM needs to reconstruct numerous details, leading to a large LD loss, which guides the update of the DG model. By repetition, the DG model gradually develops a robust understanding of various domains.

via the process $q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}\right)$, with σ_t and $\alpha_t^2 = 1 - \sigma_t^2$ controlling noise intensity. Conversely, the reverse phase encapsulated in $\{p_t\}_{t\in[0,T]}$ employs a neural network-based MSE denoiser [55] $\hat{\boldsymbol{\epsilon}}_\phi\left(\boldsymbol{x}_t, t\right)$ to denoise from $p\left(\mathbf{x}_T\right) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ back to the original $\mathbf{x}_0$ following the transitions $p_\phi\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mathbf{x}_t - \hat{\boldsymbol{\epsilon}}_\phi\left(\mathbf{x}_t; t\right), \sigma_t^2 \mathbf{I}\right)$. $\hat{\boldsymbol{\epsilon}}_\phi\left(\boldsymbol{x}_t, t\right)$ is trained via MSE minimization with time-dependent weighting $w(t)$ by minimizing:

$$\mathcal{L}_{\text{Diff}}\left(\phi; \mathbf{x}\right) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)\left\|\hat{\boldsymbol{\epsilon}}_\phi\left(\alpha_t \mathbf{x}_0 + \alpha_t \boldsymbol{\epsilon}; t\right) - \boldsymbol{\epsilon}\right\|_2^2\right]. \tag{1}$$

3.2 Latent Distillation

Existing diffusion-based DG methods only utilize LDM for data augmentation, which does not fully exploit the potential of LDM. As LDM can transfer images to various domains while preserving semantic information, we believe that LDM acquires valuable vision knowledge about domain-invariant feature representations, which can be leveraged to benefit DG discrimination tasks. However, extracting and incorporating this visual knowledge into DG tasks pose challenges, because LDM is inherently designed for generation tasks, and is hard to be involved in discrimination tasks.

Therefore, we propose a novel method called Latent Distillation to leverage this latent knowledge for DG. Our primary objective is to establish a pathway for knowledge transfer from LDM to the DG model by designing a connection between these two tasks. To achieve this, we introduce a pseudo generation operation as a bridge connecting the latent space of LDM and the parameter space of the DG model. As depicted in Fig. 2, the pseudo generation part involves

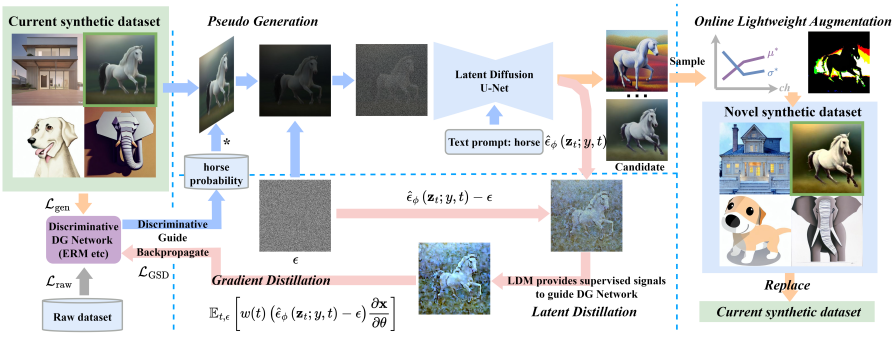

Fig. 3: In DomainFusion, we start with the original source dataset and the synthetic dataset generated later for supervised training. In Latent Distillation, each image $\mathbf{x}$ is fed into the DG Discriminative Network θ and weighted by the output confidence w.r.t. its label y, which corresponds to Pseudo Generation process in Fig. 2. The corrupted $\mathbf{x}$ is then added noise $\boldsymbol{\epsilon}$ and passed through the latent diffusion U-Net ϕ along with y as text prompt, resulting in predicted noise $\hat{\boldsymbol{\epsilon}}_\phi\left(\mathbf{x}_t; y; t\right)$. The discrepancy between the predicted and real noise $\left[\hat{\boldsymbol{\epsilon}}_\phi\left(\mathbf{x}_t; y; t\right) - \boldsymbol{\epsilon}\right]$ is utilized to obtain $\mathcal{L}_{\mathrm{LD}}$, which updates the parameter space of θ by backpropagation, corresponding to Gradient Distillation in Fig. 2. In Online Lightweight Augmentation, each image $\mathbf{x}$ in the current synthetic dataset generates N candidates via LDM. Through a sampling strategy, we decompose candidates into styles and contents, then select the most distinct style and the most similar content to sample one novel image, which then replaces $\mathbf{x}$ in the current synthetic dataset to form a novel synthetic dataset.

feeding the image into the DG model to obtain classification confidence for the corresponding class. The image is then multiplied by this confidence to generate a corrupted image. It is important to note that the purpose of the pseudo generation is not to obtain the final corrupted image, but rather to serve as a bridge for extracting loss from the latent space of the LDM, as illustrated in the gradient distillation part in Fig. 2. In this part, the corrupted image is first subjected to noise addition and then passed through the LDM. We utilize the LDM's prediction of the difference between the added noise and the predicted noise to calculate the Latent Distillation (LD) loss. Since the corrupted image is dependent on the parameters of the DG model, the loss derived from the LDM becomes a function of the DG model's parameters, which can be used to update the DG model. Besides, in Latent Distillation lower confidence scores yield higher loss, compelling the DG model to output higher confidence scores. The specific process is described in detail below.

The framework of Latent Distillation is elaborated on in Fig. 3. Given the DG network θ to be trained and an image $\mathbf{x}_0$ with its class label y_0 (note that we employ y throughout to represent both the numeric class label within the DG network θ and the textual class label within the latent diffusion model ϕ), our initial step involves forwarding $\mathbf{x}_0$ through θ for image classification, and

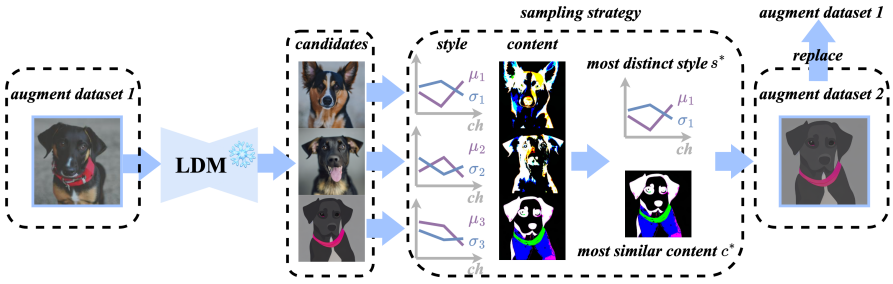

Fig. 4: In online lightweight augmentation, every T epochs, each image from the current augmentation dataset is sequentially passed through the LDM to generate N candidates. A final new image is obtained based on a sampling strategy. The new augment dataset composed of these selected images then replaces the current augment dataset. Specifically, for the sampling strategy, the styles and content of each candidate are computed. Then the style s^* that is most distinct from the input image's style and the content c^* that is most similar to the input image's content are selected. These style and content components are then combined to form the final new image.

compute the element-wise product of $\mathbf{x}_0$ with the confidence score corresponding to class y_0 to obtain a θ-related image $\mathbf{x}$. We denote this pseudo generation process as $\mathbf{x} = g(\theta) = p_\theta(y \mid \mathbf{x}_0)\,\delta(y_0)\,\mathbf{x}_0$. Subsequently, we feed $\mathbf{x}$ into the denoising process of ϕ, and we denote the loss generated from denoising as $\mathcal{L}_{\text{LD}}$. Note that $\mathcal{L}_{\text{LD}}$ is mathematically equivalent to $\mathcal{L}_{\text{Diff}}$ in Eq. 1, which yields:

$$\mathcal{L}_{\text{LD}}(\phi; \mathbf{x} = g(\theta)) = \mathcal{L}_{\text{Diff}}(\phi; \mathbf{x}) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)\,\|\hat{\boldsymbol{\epsilon}}_\phi(\alpha_t \mathbf{x} + \alpha_t \boldsymbol{\epsilon}; t) - \boldsymbol{\epsilon}\|_2^2\right]. \quad (2)$$

We compute the gradient of $\mathcal{L}_{\text{LD}}$ w.r.t. θ while omitting the U-Net Jacobian term following the SDS setting [42]:

$$\nabla_\theta \mathcal{L}_{\text{LD}}(\phi, \mathbf{x} = g(\theta)) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)\,(\hat{\boldsymbol{\epsilon}}_\phi(\mathbf{z}_t; y, t) - \boldsymbol{\epsilon})\,\frac{\partial \mathbf{x}}{\partial \theta}\right]. \quad (3)$$

$\nabla_\theta \mathcal{L}_{\text{LD}}$ is then used to update θ through backward propagation, as is shown in Fig. 2 and Fig. 3. Through this approach, we establish a pathway for gradient propagation from the latent diffusion model to the DG model.

Latent Distillation Provides Supervised Signals. We now delve into an explanation of why LD provides supervised signals. Given an image $\mathbf{x}_0$, LD corrupts the image using the confidence score predicted by the DG network θ. In our algorithm, we employ a corruption method by confidence multiplies RGB values, which is a kind of contrast reduction as it narrows the range between the max and min pixel intensities. This corruption method is simple yet effective because it reduces image details, as low-contrast images lack detail features [52], which is also mathematically supported by image quality metrics like Laplace Gradient and Average Gradient that assess intelligibility and details. Therefore, when the DG network θ exhibits poorer semantic understanding, it assigns a

lower confidence score, leading to more pronounced corruption to $\mathbf{x}_0$. As a result, larger $\mathcal{L}_{\text{LD}}$ is observed when the corrupted image is fed into the LDM. We substantiate this claim with experiments in Sec. 4.4, along with more discussion about the effectiveness of the corruption method we employed.

3.3 Online lightweight Augmentation

Existing diffusion-based DG method CDGA [17] and DSI [68] employ offline generation for data augmentation. However, they are limited by substantial computational costs and unendurable generation time, with CDGA [17] generating over 5 million synthetic samples and DSI [68] employing multiple LDMs and requiring finetuning them individually. These methods are highly inefficient and impractical when applied to large-scale scenarios. Thus, we propose an online lightweight augmentation to address this.

As shown in Fig. 4, starting from the first training epoch of the DG model, each sample involved in the training process is passed through the LDM to generate N candidate samples. Through a sampling strategy, we ultimately select one sample from the N candidates. Consequently, in the first training epoch, each image undergoes a one-to-one generation process to be replaced by a generated new image, resulting in the formation of a new synthetic dataset. This synthetic dataset is updated every T epochs, wherein each image from the current synthetic dataset is input into the LDM to obtain a new synthetic dataset. This new synthetic dataset is then used as input for the LDM in the subsequent augmentation epoch, repeating the aforementioned process iteratively.

Sampling Strategy. Data augmentation proves effective for DG by generating novel samples with distribution shifts compared to the source domain data. This allows the DG model to encounter diverse images with varying distributions during training, preventing the model from being influenced by domain-specific features and instead facilitating the learning of domain-invariant features. Given our aim to achieve lightweight augmentation, it is important to consider how to more effectively benefit DG using only a modest number of new samples.

Inspired by [9], synthetic data with similar content and distinct style can benefit DG more effectively. Therefore, we adopt a sampling strategy to get one sample from N candidates, as shown in Fig. 4. Given an input image $\mathbf{x}_0$ and N generated candidate samples $\{\mathbf{x}_i\}_{i\in[1,N]}$, we decompose them into content $\{c_i\}_{i\in[0,N]}$ and style $\{s_i\}_{i\in[1,N]}$ following [9], where the content $c_i = (\mathbf{x}_i - \mu_i)/\sigma_i$, s_i has two components μ_i and σ_i, with $\mu_i = [\mu_i^R, \mu_i^G, \mu_i^B]$ and $\sigma_i = [\sigma_i^R, \sigma_i^G, \sigma_i^B]$:

$$\mu^{ch} = \sum_{h=1}^{H}\sum_{w=1}^{W} \frac{\mathbf{x}_{hw}^{ch}}{(HW)}, \sigma^{ch} = \left[\sum_{h=1}^{H}\sum_{w=1}^{W} \frac{(\mathbf{x}_{hw}^{ch} - \mu^{ch})^2}{(HW)} + \epsilon\right]^{\frac{1}{2}}, ch = R, G, B, \quad (4)$$

To select the most distinct style $s^* = [\mu^*, \sigma^*]$, we apply the KL divergence to measure their distance w.r.t. s_0 by :

$$s^* = \arg\max_{s_i} KL(\mu_i, \mu_0) + KL(\sigma_i, \sigma_0). \quad (5)$$

Additionally, for selecting the most similar content, we employ both the cosine similarity and the θ classification confidence by :

$$c^* = \arg\max_{c_i} \lambda cos(f_\theta(c_i), f_\theta(c_0)) + (1-\lambda) p_\theta(y|c_i), \tag{6}$$

where f_θ represents the feature map extracted by θ and λ represents a predetermined constant. The feature maps extracted via DG network θ serve to assess the low-level and high-level semantic similarity between the generated candidates' contents and that of the input image. Furthermore, the classification confidence is employed to ascertain whether the generated candidates' contents have preserved the correct class-level semantic information. Subsequently, we utilize AdaIN style transfer [20] to sample the ultimate new sample $\mathbf{x}^*$ by $\mathbf{x}^* = \sigma^* c^* + \mu^*$.

Comparison with Offline Augmentation of DSI and CDGA. Our method improves the efficiency of using LDM for augmentation in several aspects. First, we employ a sampling strategy to generate samples that are better suited for DG. Given that LDM has encountered diverse data from various domains, we believe it can generate cross-domain data with a wide range of styles. Through the sampling strategy, we select the most distinct styles and transfer the most similar content, enabling the generated samples to efficiently expand the source domain. Consequently, our augmentation method is lightweight. Taking the DomainNet dataset [41] as an example, we only need to generate 120k new samples including candidates, whereas CDGA [17] requires 4 million new samples. In the same scenario, specifically considering the generation aspect, our overall generation scale, including candidates, is over 97% smaller than that of CDGA. Second, due to the significantly reduced scale of our generated data, we employ an online generation pattern to train the DG model end-to-end, further enhancing time utilization efficiency. Third, by utilizing the sampling strategy, we eliminate the need for multiple LDMs and finetuning, thus avoiding the lengthy finetuning process and substantial computational costs associated with DSI [68].

3.4 Loss Extraction at Both Latent and Pixel Levels

The overall training architecture of DomainFusion is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{raw}} + \lambda_2 \mathcal{L}_{\text{gen}} + \lambda_3 \mathcal{L}_{\text{LD}}, \tag{7}$$

where $\mathcal{L}_{\text{raw}}$ and $\mathcal{L}_{\text{gen}}$ denote the cross-entropy loss in the source dataset and the synthesized dataset, and λ_1, λ_2, and λ_3 are predetermined hyper-parameters.

4 Experiments

4.1 Experimental Settings

Settings and Datasets. Following DomainBed [15], we conduct a series of experiments on five prominent real-world benchmark datasets: PACS [27], VLCS [12], OfficeHome [60], TerraIncognita [3], and DomainNet [41]. To ensure a fair and

Table 1: Comparison with DG methods. The DG accuracy on five domain generalization benchmarks is presented with the best results highlighted in bold. The results denoted by † correspond to the results from DomainBed [15]. Results of other DG methods including Fish [53], SelfReg [23], mDSDI [5], MIRO [6], Fishr [45], DSI [68], CDGA [17] are from corresponding paper.

Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
Using ResNet-50 backbone: Non-generation method						
ERM† [59]	85.5	77.5	66.5	46.1	40.9	63.3
MLDG† [28]	84.9	77.2	66.8	47.7	41.2	63.6
CORAL† [56]	86.2	78.8	68.7	47.6	41.5	64.5
MMD† [58]	84.7	77.5	66.3	42.2	23.4	58.8
DANN† [13]	83.6	78.6	65.9	46.7	38.3	62.6
MTL† [4]	84.6	77.2	66.4	45.6	40.6	62.9
SagNet† [37]	86.3	77.8	68.1	48.6	40.3	64.2
RSC† [21]	85.2	77.1	65.5	46.6	38.9	62.7
Fish [53]	85.5	77.8	68.6	45.1	42.7	63.9
SelfReg [23]	85.6	77.8	67.9	47.0	42.8	64.2
mDSDI [5]	86.2	79.0	69.2	48.1	42.8	65.1
MIRO [6]	85.4	79.0	70.5	50.4	44.3	65.9
Fishr [45]	85.5	77.8	68.6	47.4	41.7	64.2
Using ResNet-50 backbone: Non-diffusion-based generation method						
GroupDRO† [50]	84.4	76.7	66.0	43.2	33.3	60.7
Mixup† [67]	84.6	77.4	68.1	47.9	39.2	63.4
Mixstyle‡ [75]	85.2	77.9	60.4	44.0	34.0	60.3
Using ResNet-50 backbone: Diffusion-based generation method						
DSI [68]	78.3	-	67.3	-	-	-
CDGA [17]	88.5	78.9	68.2	-	43.1	-
Ours	**90.0**	**79.2**	**72.4**	**51.1**	**44.6**	**67.5**
Using RegNetY-16GF backbone with SWAG pre-training						
ERM [59]	89.6	78.6	71.9	51.4	48.5	68.0
MIRO [6]	**97.4**	79.9	80.4	58.9	53.8	74.1
Ours	96.6	**80.0**	**83.4**	**60.6**	**55.9**	**75.3**

consistent comparison, we follow DomainBed's training and evaluation protocol. We provide full details in the supplementary material.

Implementation Details. For the LDM, we employ the stable diffusion v1-4 model. The batch size is set to 16 and we employ the Adam optimizer [25] and cosine learning rate schedule. For algorithm-specific parameters, the candidate number $N = 3$, the interval of generation epochs $T = 5$. λ_1, λ_2, λ_3 are 1, 0.5, and 0.5. λ is 0.4 for c^*.We provide full details in the supplementary material.

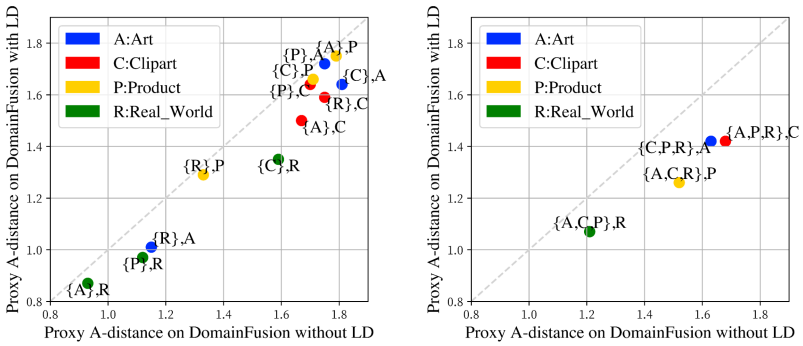

Fig. 5: Proxy A-distance (PAD) on OfficeHome. x-axis: PAD computed upon DomainFusion without LD; y-axis: PAD computed upon DomainFusion with LD. We employ the DG model to extract features across diverse domains to train a linear domain classifier, and PAD is proportional to its classification accuracy. A superior DG model yields a lower PAD, indicating its ability to extract domain-invariant features. DomainFusion with LD demonstrates lower PAD compared to its non-LD version in both cases: (a) PAD between the single source domain and the target domain. For example, {C}, A denotes measuring PAD between one source domain Clipart and the target domain Art. (b) PAD between all source domains and the target domain.

4.2 Main Results

Comparison with Domain Generalization Methods. We compare DomainFusion with baseline methods and recent DG algorithms and present results in Table 1. In the first section, we evaluated DomainFusion using the ResNet-50 [16] architecture as the backbone. The experimental results demonstrate that DomainFusion outperforms the current state-of-the-art methods in all benchmark datasets, yielding accuracy improvements of +1.5pp, +0.2pp, +1.9pp, +0.7pp, and +0.3pp in each dataset, where 'pp' is short for 'percentage point'.

In the second section of Table 1, we employ RegNet-Y-16GF [44] as the backbone and utilize the SWAG [54] method to obtain a pre-trained model on the ImageNet [49] dataset, aiming to investigate the maximum performance potential of the DomainFusion algorithm. The experimental results convincingly demonstrate a significant performance improvement exhibited by the DomainFusion algorithm compared to ERM across all datasets. Moreover, our proposed approach outperforms the current SOTA algorithm, MIRO [6], in all datasets except PACS, with performance gains of +0.1pp, +3pp, +1.7pp, and +2.1pp in VLCS, OfficeHome, TerraInc, and DomainNet, respectively. The effectiveness of our algorithm has been substantiated through a wide range of experiments.

Comparison with Non-diffusion-based Generation Methods. Experimental results demonstrate that non-diffusion-based generation methods exhibit notable shortcomings, as they typically exhibit subpar performance and often

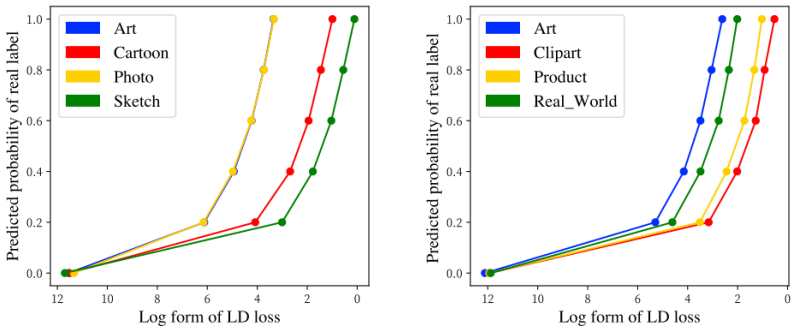

Fig. 6: Predicted probability and LD curve. x-axis: Log form of $\mathcal{L}_{\text{LD}}$; y-axis: Predicted probability of real label by DG network. (a) on PACS and (b) on OfficeHome.

Table 2: Effects of Different Components in DomainFusion.

$\mathcal{L}_{\text{raw}}$	$\mathcal{L}_{\text{gen}}$	$\mathcal{L}_{\text{LD}}$	Art	Clipart	Product	Real	Avg.
✓	✗	✗	69.3	61.3	81.6	82.5	73.7
✓	✗	✓	77.3	66.2	84.3	85.4	78.3
✓	✓	✗	73.6	71.2	80.7	88.7	78.6
✓	✓	✓	**81.2**	**73.9**	**88.5**	**90.1**	**83.4**

struggle to surpass certain baseline algorithms. Compared with non-diffusion-based generation methods, DomainFusion demonstrates a significant improvement by +4.8pp, +1.3pp, +4.3pp, +3.2pp, and +5.2pp in the 5 benchmark datasets respectively.

Comparison with Diffusion-based Generation Methods. Both DSI [68] and CDGA [17] require substantial computational resources, yet their performance falls short of SOTA methods, making them cost-ineffective. DomainFusion surpasses the reported results of other Diffusion-based DG methods in PACS, VLCS, OfficeHome, and DomainNet.

4.3 Ablation Study

We conduct ablation study on OfficeHome based on RegNet-Y-16GF.

Effects of Different Components. Table 2 presents the impact of different components of DomainFusion on DG performance. It is worth noting that using only $\mathcal{L}_{\text{raw}}$ is equivalent to the ERM method. To ensure fairness, we also searched for the parameters of the ERM method, resulting in improved ERM performance compared to that in Table 1. As shown in Table 2, we first incorporate LD on top of ERM and observe that the introduction of LD leads to a 4.6% improvement in DG performance. Then, we incorporate $\mathcal{L}_{\text{gen}}$ on top of ERM to evaluate the

effectiveness of the generation part. It is observed that $\mathcal{L}_{\text{gen}}$ improves the average accuracy by 4.9% by generating a more diverse set of samples to augment the source domain, resulting in a significant improvement in DG performance. However, using $\mathcal{L}_{\text{gen}}$ alone still exhibits a considerable performance gap compared to state-of-the-art methods. To address this discrepancy, $\mathcal{L}_{\text{LD}}$ bridges this gap by further enhancing the accuracy by 4.8% compared to use $\mathcal{L}_{\text{gen}}$ solely.

Effects of Synthetic Dataset Updating in Augmentation. We ablate on the effect of synthetic dataset updating, where we only generate a fixed synthetic dataset on the first epoch, detailed in the supplementary material.

Effects of the Sampling Strategy. We also provide an ablation study on the effect of the sampling strategy. Please refer to the supplementary material.

Effects of the Candidate Number. Moreover, we provide an ablation study on the impact of the candidate number, detailed in the supplementary material.

4.4 Effectiveness Analysis of Latent Distillation

Proxy A-distance (PAD) Analysis of Latent Distillation. We compute the Proxy A-distance (PAD) [10] to verify the effectiveness of Latent Distillation. PAD requires extracting image features separately from the source and target domains, labeling them as 1 and 0, and subsequently training a classifier to discriminate between these two domains. Given a test error of ε, PAD is defined as $2(1-2\varepsilon)$. A superior DG algorithm yields a lower PAD, indicating its ability to extract domain-invariant features. Consistent with prior studies [1, 7, 10, 14], we employ DomainFusion with/without LD to extract image features from source and target domains, labeled as 1 and 0, and train a linear SVM for classification. As is shown in Fig. 5, we first quantify PAD between a single source domain and the target domain, demonstrating that incorporating LD consistently yields lower PAD values. Then we measure the PAD between all source domains and the target domain, revealing a larger margin between the two versions, thus validating the effectiveness of LD.

Probability-$\mathcal{L}_{\text{LD}}$ Curve Analysis of Latent Distillation. According to Sec. 3.2, LD can provide effective training signals to supervise DG model. We substantiate this claim with experiments. As is shown in Fig. 6, on both PACS and OfficeHome, as $\mathcal{L}_{\text{LD}}$ decreases, the predicted probability of the real label by the DG network increases significantly. Consequently, by minimizing $\mathcal{L}_{\text{LD}}$, we effectively impose a constraint on the DG network θ to assign a higher confidence score to the real label, thereby providing supervised signal to guide θ.

Contrast Reduction Is Simple Yet Effective. Firstly, contrast reduction is one of the typical corruptions according to [11]. Secondly, it diminishes image intricacies, since images with diminished contrast have fewer distinguishing features [52], which is further substantiated mathematically by metrics such as the Laplacian Gradient and Mean Gradient, which gauge clarity and the presence of fine details. Thirdly, it is more effective than other corruptions like blurring, noise, and random masking, detailed in the supplementary material.

Fig. 7: Visualization of generated samples and LD noise. The left section is online lightweight augmentation samples and the right section is corresponding LD noise.

4.5 Visualization

Visualization of Generated Samples. Fig. 7 showcases the visualization results of the synthetic dataset generated by online lightweight augmentation at various epochs, with each row representing the evolution of a specific image. In terms of visual effects, it is apparent that as the synthesized dataset is updated, the image sequences retain a certain degree of content similarity and also introduce new styles, serving as evidence of the effectiveness of our method.
Visualization of LD Noise. Fig. 7 illustrates the visualization of the LD noise for all images. We first calculate the difference between predicted denoising latent and latent with noise, and then use the stable diffusion decoder to decode this difference. It is evident that these noise patterns can also effectively capture the high-level semantic information in the images while reducing the influence of irrelevant domain-specific elements, such as the background. This finding demonstrates the strong generalization capability of the latent diffusion model, as it can extract transferable feature representations, which contribute to optimizing the DG semantic understanding network in our LD, further confirming the effectiveness of the LD method.

5 Conclusion

In this paper, we propose a novel framework that utilizes the latent diffusion model (LDM) in both the latent space and pixel space for domain generalization (DG). In latent space, we propose Latent Distillation (LD) that extracts transferable knowledge as gradient priors from the LDM to optimize the DG model. In pixel space, we propose an online lightweight augmentation method that significantly reduces the number of generated images and computational costs compared with previous diffusion-based DG methods. Experimental results demonstrate that our method achieves state-of-the-art performance on DG classification.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62125109, Grant 61931023, Grant 61932022, Grant 62371288, Grant 62320106003, Grant 62301299, Grant T2122024, Grant 62120106007.

References

1. Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M.: Domain-adversarial neural networks. arXiv preprint arXiv:1412.4446 (2014)
2. Arpit, D., Wang, H., Zhou, Y., Xiong, C.: Ensemble of averages: Improving model selection and boosting performance in domain generalization. Advances in Neural Information Processing Systems **35**, 8265–8277 (2022)
3. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
4. Blanchard, G., Deshmukh, A.A., Dogan, U., Lee, G., Scott, C.: Domain generalization by marginal transfer learning. Journal of machine learning research **22**(2), 1–55 (2021)
5. Bui, M.H., Tran, T., Tran, A., Phung, D.: Exploiting domain-specific features to enhance domain generalization. Advances in Neural Information Processing Systems **34**, 21189–21201 (2021)
6. Cha, J., Lee, K., Park, S., Chun, S.: Domain generalization by mutual-information regularization with pre-trained models. In: European Conference on Computer Vision. pp. 440–457. Springer (2022)
7. Chen, M., Xu, Z., Weinberger, K., Sha, F.: Marginalized denoising autoencoders for domain adaptation. arXiv preprint arXiv:1206.4683 (2012)
8. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. arXiv preprint arXiv:2211.09788 (2022)
9. Dai, R., Zhang, Y., Fang, Z., Han, B., Tian, X.: Moderately distributional exploration for domain generalization. arXiv preprint arXiv:2304.13976 (2023)
10. Ding, Y., Wang, L., Liang, B., Liang, S., Wang, Y., Chen, F.: Domain generalization by learning and removing domain-specific features. Advances in Neural Information Processing Systems **35**, 24226–24239 (2022)
11. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 eighth international conference on quality of multimedia experience (QoMEX). pp. 1–6. IEEE (2016)
12. Fang, C., Xu, Y., Rockmore, D.N.: 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)
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
14. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 513–520 (2011)
15. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

17. Hemati, S., et al.: Cross domain generative augmentation: Domain generalization with latent diffusion models. arXiv:2312.05387 (2023)
18. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. arXiv preprint arXiv:2304.07090 (2023)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
21. Huang, Z., et al.: Self-challenging improves cross-domain generalization. In: 16th European Conference on Computer Vision. pp. 124–140 (2020)
22. Jackson, P.T., Abarghouei, A.A., Bonner, S., Breckon, T.P., Obara, B.: Style augmentation: data augmentation via style randomization. In: CVPR workshops. vol. 6, pp. 10–11 (2019)
23. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9619–9628 (2021)
24. Kim, S., Lee, K., Choi, J.S., Jeong, J., Sohn, K., Shin, J.: Collaborative score distillation for consistent visual synthesis. arXiv preprint arXiv:2307.04787 (2023)
25. Kinga, D., Adam, J.B., et al.: A method for stochastic optimization. In: International conference on learning representations (ICLR). vol. 5, p. 6. San Diego, California; (2015)
26. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning. pp. 5815–5826. PMLR (2021)
27. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5542–5550 (2017)
28. Li, H., e.a.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5400–5409 (2018)
29. Li, P., Li, D., Li, W., Gong, S., Fu, Y., Hospedales, T.M.: A simple feature augmentation for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8886–8895 (2021)
30. Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., Duan, L.Y.: Uncertainty modeling for out-of-distribution generalization. arXiv preprint arXiv:2202.03958 (2022)
31. Li, Z., Ren, K., Jiang, X., Shen, Y., Zhang, H., Li, D.: Simple: Specialized model-sample matching for domain generalization. In: The Eleventh International Conference on Learning Representations (2022)
32. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In: European Conference on Computer Vision. pp. 511–528. Springer (2022)
33. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
34. Liu, Y., Chen, Y., Gou, M., Huang, C.T., Wang, Y., Dai, W., Xiong, H.: Towards unsupervised domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20654–20664 (2023)

35. Liu, Y., Wang, Y., Chen, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Promoting semantic connectivity: Dual nearest neighbors contrastive learning for unsupervised domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3510–3519 (2023)
36. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International conference on machine learning. pp. 10–18. PMLR (2013)
37. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8690–8699 (2021)
38. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
39. Niu, Z., Yuan, J., Ma, X., Xu, Y., Liu, J., Chen, Y.W., Tong, R., Lin, L.: Knowledge distillation-based domain-invariant representation learning for domain generalization. IEEE Transactions on Multimedia (2023)
40. Palakkadavath, R., Nguyen-Tang, T., Gupta, S., Venkatesh, S.: Improving domain generalization with interpolation robustness. In: NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications (2022)
41. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019)
42. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
43. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. Mit Press (2008)
44. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10428–10436 (2020)
45. Rame, A., Dancette, C., Cord, M.: Fishr: Invariant gradient variances for out-of-distribution generalization. In: International Conference on Machine Learning. pp. 18347–18377. PMLR (2022)
46. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
48. Rosenfeld, E., Ravikumar, P., Risteski, A.: The risks of invariant risk minimization. arXiv preprint arXiv:2010.05761 (2020)
49. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
50. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
51. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

52. Saleem, A., Beghdadi, A., Boashash, B.: Image fusion-based contrast enhancement. EURASIP Journal on Image and Video Processing **2012**, 1–17 (2012)
53. Shi, Y., Seely, J., Torr, P.H., Siddharth, N., Hannun, A., Usunier, N., Synnaeve, G.: Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 (2021)
54. Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., Van Der Maaten, L.: Revisiting weakly supervised pre-training of visual perception models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 804–814 (2022)
55. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
56. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016)
57. Tan, H., Wu, S., Pi, J.: Semantic diffusion network for semantic segmentation. Advances in Neural Information Processing Systems **35**, 8702–8716 (2022)
58. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
59. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
60. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5018–5027 (2017)
61. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. Advances in neural information processing systems **31** (2018)
62. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P.: Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering (2022)
63. Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., Li, T.: Contrastive regression for domain adaptation on gaze estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19376–19385 (2022)
64. Wang, Y., Li, H., Chau, L.P., Kot, A.C.: Variational disentanglement for domain generalization. arXiv preprint arXiv:2109.05826 (2021)
65. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
66. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14383–14392 (2021)
67. Yan, S., Song, H., Li, N., Zou, L., Ren, L.: Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677 (2020)
68. Yu, R., Liu, S., Yang, X., Wang, X.: Distribution shift inversion for out-of-distribution prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3592–3602 (2023)
69. Zhang, H., Zhang, Y.F., Liu, W., Weller, A., Schölkopf, B., Xing, E.P.: Towards principled disentanglement for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8024–8034 (2022)

70. Zhang, J., Herrmann, C., Hur, J., Cabrera, L.P., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. arXiv preprint arXiv:2305.15347 (2023)
71. Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., Finn, C.: Adaptive risk minimization: Learning to adapt to domain shift. Advances in Neural Information Processing Systems **34**, 23664–23678 (2021)
72. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. arXiv preprint arXiv:2303.02153 (2023)
73. Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., Sebe, N.: Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6277–6286 (2021)
74. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
75. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)