Representation Enhancement-Stabilization: Reducing Bias-Variance of Domain Generalization

Wei Huang[1], Yilei Shi[1], Zhitong Xiong[1], and Xiao Xiang Zhu[1,2]

1. Technical University of Munich, Arcisstraße 21, 80333 München, Germany
2. Munich Center for Machine Learning, 80333 Munich
{w2wei.huang, yilei.shi, zhitong.xiong, xiaoxiang.zhu}@tum.de

Abstract. Domain Generalization (DG) focuses on enhancing the generalization of deep learning models trained on multiple source domains to adapt to unseen target domains. This paper explores DG through the lens of bias-variance decomposition, uncovering that test errors in DG predominantly arise from cross-domain bias and variance. Inspired by this insight, we introduce a Representation Enhancement-Stabilization (RES) framework, comprising a Representation Enhancement (RE) module and a Representation Stabilization (RS) module. In RE, a novel set of feature frequency augmentation techniques is used to progressively reduce cross-domain bias during feature extraction. Furthermore, in RS, a novel Mutual Exponential Moving Average (MEMA) strategy is designed to stabilize model optimization for diminishing cross-domain variance during training. Collectively, the whole RES method can significantly enhance model generalization. We evaluate RES on five benchmark datasets and the results show that it outperforms multiple advanced DG methods. Our code will be available at `https://github.com/zhu-xlab/DG-RES`.

Keywords: Domain generalization · Frequency Domain · Data Augmentation · Bias-variance Decomposition

1 Introduction

In recent years, deep learning models have achieved remarkable advancements in computer vision tasks, primarily under the assumption that training and test data are independent and identically distributed. However, in real-world scenarios, this assumption often breaks down due to domain shift/bias, where the distribution of test data markedly differs from that of training data. Domain shift can significantly degrade the performance of deep models on unfamiliar, unseen target domains. For instance, self-driving models trained under daylight conditions might fail to perform effectively in nighttime environments.

To address these challenges, Domain Generalization (DG) has been proposed as a robust solution. DG focuses on training models on multiple diverse yet related source domains for robust performance on arbitrary unseen target domains. Numerous DG approaches have been developed, including adversarial training [23, 48], disentangled representation learning [37], meta-learning [7], model ensemble [2], flatness optimization [4], and domain data augmentation [8, 41, 47].

In this study we hightlight the cross-domain generalization of deep models from the perspective of bias-variance decomposition. Geman et al. [12] first used bias-variance to decompose the mean square error (MSE) of neural networks. Building on this, Yang et al. [43] developed the bias-variance decomposition for the cross-entropy (CE) loss in classification tasks, which was extended by Arpit et al. [2] to analyze the impact of variance in DG, as illustrated below:

$$\mathbb{E}_{x,y\sim D_s}[\mathrm{CE}(y, f(x; D_s))] = \underbrace{\mathbb{E}_{x,y\sim D_s}[\mathrm{CE}(y, \bar{f}(x))]}_{\mathrm{Bias}^2} + \underbrace{\mathbb{E}_{x,D_s}[\mathrm{KL}(\bar{f}(x), f(x; D_s))]}_{\mathrm{Variance}}, \tag{1}$$

where D_s represents source domains used for training. The formula decomposes the expected source CE loss into two components: bias and variance. Bias evaluates the discrepancy between model's average predictions and actual outcomes, while variance measures sensitivity to fluctuations during training.

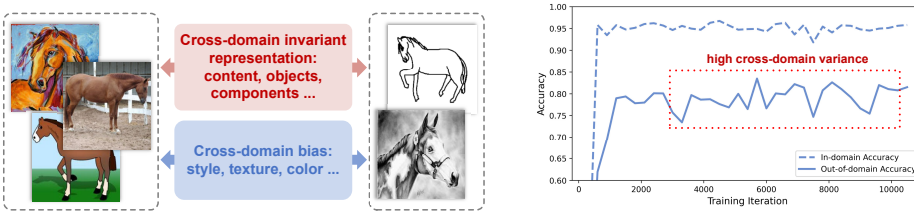
Motivated by them, we extend the bias-variance decomposition to DG to comprehensively analysis the cross-domain prediction errors as the following:

$$\begin{aligned}\mathbb{E}_{x,y\sim D_t}[\mathrm{CE}(y, f(x_t; D_t))] &= \mathbb{E}_{x,y\sim D_s}[\mathrm{CE}(y, f(x_s+\delta; D_s))] \\ &= \underbrace{\mathbb{E}_{x,y\sim D_s}[\mathrm{CE}(y, \bar{f}(x_s+\delta))]}_{\text{Cross-Domain Bias}^2} \\ &\quad + \underbrace{\mathbb{E}_{x\sim D_s}[\mathrm{KL}(\bar{f}(x_s+\delta), f(x_s+\delta; D_s))]}_{\text{Cross-Domain Variance}} \\ &\quad + \Delta_{domain}(\mathrm{Bias}^2, \mathrm{Variance}).\end{aligned} \tag{2}$$

This formula analyzes the expected CE loss when a model trained in source domains D_s, denoted as $f(x_s; D_s)$, is applied to new target domains D_t, with respect to domain shift δ between source and target images, x_s and x_t. The formula is decomposed into three components:

- (1) **Cross-Domain Bias**, as shown in the left of Fig. 1. It quantifies the discrepancy between the model's average prediction in the target domain and the actual labels. High cross-domain bias dramatically enlarges the divergence between the distribution learnt from source domains and the true underlying distribution of target domains. By progressively reducing bias in feature extraction during training, i.e., $f(x_s) \leftarrow f(x_s+\delta)$, there is an approximate bias error $\mathbb{E}_{x,y\sim D_s}[\mathrm{CE}(y, \bar{f}(x))] \leftarrow \mathbb{E}_{x,y\sim D_s}[\mathrm{CE}(y, \bar{f}(x_s+\delta))]$.
- (2) **Cross-Domain Variance**, as shown in the right of Fig. 1. It measures the variability and sensitivity in the model's predictions from source to target domains. The model trained on the source domain tends to generate unstable predictions when confronted with the target domain data, with elevated cross-domain variance. Based on cross-domain bias reduction in (1), further decreasing cross-domain variance results in an approximate variance error $\mathbb{E}_{x\sim D_s}[\mathrm{KL}(\bar{f}(x_s), f(x_s; D_s))] \leftarrow \mathbb{E}_{x\sim D_s}[\mathrm{KL}(\bar{f}(x_s+\delta), f(x_s+\delta; D_s))]$.
- (3) $\Delta_{domain}(\mathrm{Bias}^2, \mathrm{Variance})$. This term denotes the extra test error of the intertwined effects of the cross-domain bias and variance. The decrease of both bias and variance can contribute to its reduction.

Overall, diminishing cross-domain bias and variance can narrow the gap between Eq. (1) and Eq. (2), enabling models trained on source domains to effectively adapt to target domains. This analytical framework provides actionable insights for training models capable of generalizing across diverse data distributions.



(a) Challenge of cross-domain bias **(b)** Challenge of cross-domain variance

Fig. 1: Two challenges impeding DG from the view of bias-variance decomposition.

Based on this analytical framework, we proposed a joint Representation Enhancement-Stabilization (RES) method, which consists of a Representation Enhancement (RE) module and a Representation Stabilization (RS) module to alleviate cross-domain bias ans cross-domain variance, respectively. In RE, we propose a novel set of feature augmentation techniques within the frequency domain extracted through Fast Fourier Transform (FFT), including `random noise`, `random dropout`, and `mixup`. RE effectively broadens cross-domain representation space and thereby enhances robust representation learning, leading to a reduction in domain bias. Notably, the augmentations of RE target only the amplitude spectrum of high-dimensional features, with their phase spectrum remaining unchanged for maintaining domain-irrelevant semantics. In RS, we persent a novel Mutual Exponential Meaning Average (MEMA) model parameter optimization strategy. This approach facilitates a dynamic exchange and integration of parameters between teacher and student models, enhancing model stability in out-of-domain performance throughout training process.

We evaluate the proposed RES method on five DG benchmark datasets, including `PACS` [22], `VLCS` [10], `OfficeHome` [36], `TerraIncognita` [3], and `DomainNet` [30], and comparison experiment results indicate our RES outperforms multiple state-of-the-art DG methods.

Our contributions are summarized as follows: (1) We rethink the prediction error of DG from an extended bias-variance decomposition analytical framework. (2) Based on the bias-variance decomposition, we propose a RES method for DG, consisting of a Representation Enhancement (RE) module and a Representation Stabilization (RS) module, to respectively reduce cross-domain bias and variance. (3) Comparison experiments demonstrate that our RES method outperforms multiple state-of-the-art DG methods, verifying its effectiveness.

2 Related Work

In this section, we first give a brief review of the mainstreams DG methods. Then we delve into a detailed introduction to data augmentation methods and flatness-aware methods, which are relevant to our RES.

2.1 Domain Generalization

Various DG methods have been developed. One type of DG methods focus on distribution alignment among source domains to learn domain-invariant representations via all kinds of strategies, including domain alignment [23, 28, 46], adversarial training [23, 48], causal learning [25, 26], disentangled representation learning [31, 37], low-rank decomposition [33], self-supervised learning [1, 19], meta-learning [6, 7, 44], and data normalization [20, 34, 34].

2.2 Data Augmentation Methods

Our RE module aligns with another paradigm of data augmentation methods, which enhance model generalization to unseen domains by diversifying source domain data at two distinct levels.

(1) **Image level** [8, 41, 42, 47]. Yang et al. [42] and Zhou et al. [47] have used cross-domain image-to-image translation based on domain-adversarial learning, further applying these augmented images in model training. Xu et al. [41] introduced a novel Fourier-based image augmentation strategy for DG, based on the assumption that phase information in the frequency domain contains high-level semantics and is less affected by domain shift.

(2) **Feature level** [14, 24, 39]. Techniques such as the straightforward perturbation of high-dimensional feature embeddings with Gaussian noise by Li et al. [24], the use of Adaptive Instance Normalization (AdaIN) [17] for feature-level style randomization augmentation by Wang et al. [39], and the DomainDrop framework by Guo et al. [14] which drops high domain-activated channels to enhance feature channel robustness against domain shift, have been proposed.

2.3 Flatness-aware Methods

Flatness-aware methods aiming at avoiding unstable sharp optimization draws rising interest under the situation where a validation set has a different distribution from the test data, leading to a non-i.i.d. scenario. Cha et al. [4] hypothesized that in the non-i.i.d scenario the generalization gap between flat and sharp minima is more pronounced, and then proposed a Stochastic Weight Averaging Densely (SWAD) method aiming to locate flat minima and reduce domain gap. Moreover, Arpit et al. [2] argued that models trained on specific training domains often display erratic performance on test domains with distribution shifts. To mitigate this, they suggested the use of ensembling moving average (EoA) models, thereby decreasing uncertainty in test domains for enhanced domain generalization. Sharpness-Aware Minimization (SAM) [11] and an improved

Sharpness-Aware Gradient Matching (SGAM) [38] were further proposed to enhance model generalization by simultaneously minimizing the loss value and its sharpness, leading to more robust models.

3 RES-based Domain Generalization

3.1 Notations and Overview

Assuming there are K distinct yet semantic-related source domains represented as $D_s = \{D_s^1, D_s^2, \ldots, D_s^K\}$, each originating from different distributions and comprising N_k image-label pairs, denoted as $D_s^k = \{(x_i, y_i)\}_{i=1}^{N_k}$. In this study, the model training is independent of domain labels, so there is no domain subscript in (x, y). The primary objective of DG is to leverage these source domains to train a model, parameterized by θ, that exhibits strong generalization capabilities and can effectively adapt to unseen target domains. In RES, there are two models sharing the same architecture, a student model M^S parameterized with θ^S and a teacher model M^T parameterized with θ^T. M^S is used for model training and M^T is utilized for model validation and test.

The proposed RES-based DG framework is shown in Fig. 2 with its training process summarized in Algorithm 1. RES consists of: (1) a RE module that augments amplitude spectrum of feature maps while maintaining the phase spectrum, to expand cross-domain feature representation; and (2) a RS module that mutually fuses the parameters of the teacher and student models to each other.

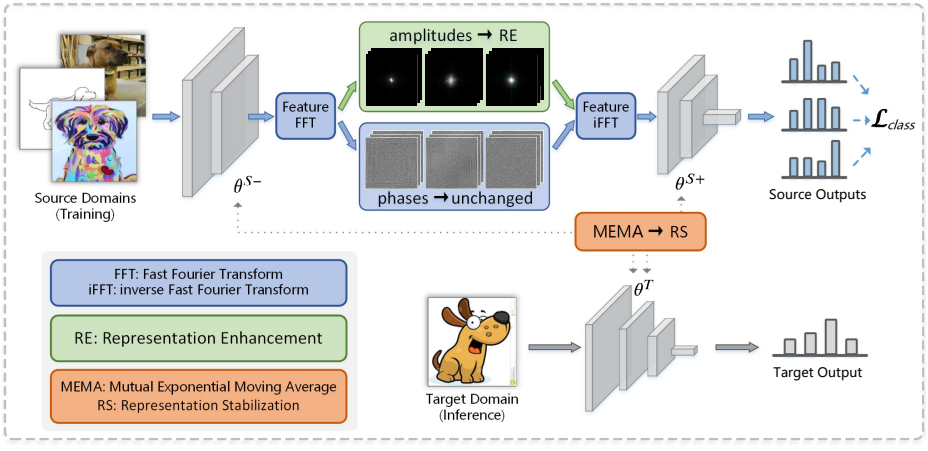


Fig. 2: Workflow of the proposed RES-based domain generalization framework, which consists of Representation Enhancement (RE) and Representation Stabilization (RS). RE aims to enhance feature representation to reduce cross-domain bias via a set of augmentations performing on the amplitude spectrum of frequency domain, while RS aims to further stabilize cross-domain performance to reduce cross-domain variance through a novel MEMA optimization strategy between student and teacher models.

3.2 Feature Frequency Augmentation-Based RE

Our RE modules augments source data at the feature level. During training, the student model M^S is split into two parts: M^{S-} which includes all layers up to a middle layer l for feature map extraction, and M^{S+} which consists of the remaining layers for the remaining feature extraction and classification.

Decomposing Phase and Amplitude From Features. Given an image $x \in \mathcal{R}^{3\times\bar{H}\times\bar{W}}$ where $\bar{H}\times\bar{W}$ represents image spatial size, a corresponding multi-layer feature map $f \in \mathcal{R}^{C\times H\times W}$ can be extracted by M^{S-}, where C signifies the number of channels, and H, W denote the height and width of the feature maps respectively. It is formulated as

$$f = M^{S-}(x). \tag{3}$$

To delve into the frequency properties of these feature maps, we employ a 2D FFT across the spatial dimensions of each channel, ensuring that every layer of f undergoes this transformative process. This approach allows us to comprehensively analyze the frequency content within each individual feature channel. For channel c, its FFT formulation is as follows:

$$F^c(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f^c(h,w)\cdot e^{-j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}, \tag{4}$$

where $f^c(h,w)$ represents the pixel value at the spatial location (h,w) within the range of $[H-1, W-1]$ in channel c, and $F^c(u,v)$ denotes the complex frequency spectrum at the frequency coordinates (u,v) of channel c. The term j is the imaginary unit.

Then we compute a multi-layer phase spectrum denoted as $\mathcal{P}$ and a multi-layer amplitude spectrum represented as $\mathcal{A}$, from the complex frequency spectrum F, as

$$\mathcal{P}(F^c(u,v)) = \arctan 2(\mathrm{Imag}(F^c(u,v)), \mathrm{Real}(F^c(u,v))), \tag{5}$$

$$\mathcal{A}(F^c(u,v)) = \sqrt{\mathrm{Real}(F^c(u,v))^2 + \mathrm{Imag}(F^c(u,v))^2}. \tag{6}$$

Here, $\mathrm{Real}(F^c(u,v))$ and $\mathrm{Imag}(F^c(u,v))$ indicate the real and imaginary parts of the channel-c frequency spectrum F^c, respectively. $\mathcal{P}$ and $\mathcal{A}$ have the same size of $C\times H\times W$ as f. Drawing inspiration from [15, 29, 32, 41], there is a common assumption that the phase component of frequency spectrum maintains the high-level semantics of the original data while the amplitude component contains its low-level statistics. In this study, we extend it from the image level to feature level, where we regard feature maps as high-dimension signals with the same characteristic of phase-amplitude decomposition as images.

Frequency Augmentations on Features. To avoid damaging the semantics of features, augmentations are only implemented on the amplitude $\mathcal{A}$ to obtain an augmented $\mathcal{A}'$ while the phase $\mathcal{P}$ keeps unchanged, as

$$\mathcal{A}' = \texttt{Aug}(\mathcal{A}). \tag{7}$$

Given a batch of images $\{x_1, x_2, \ldots, x_B\}$ where B is batch size, a corresponding batch of amplitudes $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_B\}$ can be obtained by Eqs. (3), (4), and (6). During each training iteration, we randomly select and apply one augmentation strategy from a pre-defined augmentation list, including `none`, `random noise`, `random dropout`, and `mixup`. Here `none` represents no augmentation, and the details of the rest three strategies are as follows:

- **Random noise**, denoted as RE^{ns}. Given a multi-layer amplitude $\mathcal{A}_i \in \mathcal{R}^{C\times H\times W}$, we add Gaussian noise $G \in \mathcal{R}^{1\times H\times W}$ to it. The elements of G adhere to a Gaussian distribution characterized by mean μ and standard deviation σ, i.e., $G \sim \mathcal{N}(\mu, \sigma^2)$. Leveraging the broadcasting mechanism, this noise addition is equally conducted to each layer of $\mathcal{A}_i$, as:

$$\mathcal{A}'_i = \mathcal{A}_i * (I + G) = \mathcal{A}_i * (I + \mathcal{N}(\mu, \sigma^2)), \tag{8}$$

  where I represents the Identity Matrix. This approach guarantees that the same frequency components throughout all channels of the feature map are enhanced with the same noise intensity.
- **Random dropout**, denoted as RE^{dp}. We randomly generate a binary dropout mask, $M \in \mathcal{R}^{1\times H\times W}$, where each element is set either to 0 with probability p or to 1 with a reversal probability $1 - p$. The dimension of $\mathcal{M}$ is also expand to $C \times H \times W$ by broadcasting. The dropout-augmented amplitude $\mathcal{A}'$ is calculated by:

$$\mathcal{A}'_i = \mathcal{A} \cdot M. \tag{9}$$

  This operation introduces sparsity of amplitudes, aiming at reducing their overfitting for better robustness. It is worth mentioning that the element of M at the location $[0, 0]$ always equals 1, which represents the mean value of the feature map, to avoid over-masking on features.
- **Mixup**, denoted as RE^{mix}. Unlike `random noise` and `random dropout` which operate on single amplitudes extracted from individual samples, `mixup` engages in a cross-sample amplitude interaction by linear interpolation on different amplitudes extracted from different samples. Specifically, for a given amplitude spectrum $\mathcal{A}_i$, `mixup` randomly selects another amplitude $\mathcal{A}_j$ from the set $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_B\}$, and then blends $\mathcal{A}_i$ with $\mathcal{A}_j$, resulting in a new, hybrid amplitude spectrum $\mathcal{A}'_i$ as follows:

$$\mathcal{A}'_i = \alpha\mathcal{A}_i + (1 - \alpha)\mathcal{A}_j, \tag{10}$$

  where α is a uniform random coefficient between [0,1]. This procedure enriches the amplitude characteristics by promoting a more diverse fusion of feature styles across domains. It boosts the model's generalization by augmenting intricate cross-domain amplitude patterns effectively.

Reconstructing Features. For each pair of phase and augmented amplitude components, $\mathcal{P}$ and $\mathcal{A}'$, an augmented multi-layer feature map, f', can be reconstructed via inverse Fast Fourier Transform (iFFT) as

$$f'^{c}(h,w) = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\mathcal{A}'(F^c(u,v)) \cdot e^{j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)} \cdot e^{j\mathcal{P}(F^c(u,v))} \tag{11}$$

Here, $e^{j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}$ is a complex exponential term representing the frequency component in the inverse Fourier transform, and $e^{j\mathcal{P}(F^c(u,v))}$ reintroduces the phase information into the transform.

Finally, f' is used for further feature extraction and class classification by the remaining layers of the student model, M^{S+}, as

$$p = M^{S+}(f'). \tag{12}$$

3.3 MEMA-based RS

As shown in Fig. 2, there are two models, one student model M^S with parameters θ^S and one teacher model M^T with parameters θ^T. Their parameters undergo mutual updates to provide stabilization to each other, referred to as RE^T and RE^S for the teacher and student models, respectively. Their processes are formulated as

$$\begin{aligned} \theta^T &\leftarrow \tau^T\theta^T + (1-\tau^T)\theta^S, \quad \tau^T \in [0,1), \\ \theta^S &\leftarrow (1-\tau^S)\theta^S + \tau^S\theta^T, \quad \tau^S \in [0,1). \end{aligned} \tag{13}$$

Here, only the student model M^S is actively trained, while the teacher model M^T is responsible solely for parameter updates. The student model M^S is the primary learner, actively absorbing new knowledge and adapting to training data. In contrast, the teacher model M^T acts as a robust and stable reference using moving average updating strategy, providing more robust and consistent parameters by reducing the variance of training batches [2].

Both the values of τ^T and τ^S are small and they play a mutual role in this process: (1) Stability from the teacher model. With a small τ^T, the teacher model's parameters change slowly, ensuring that they remains stable and consistent.(2) Gradual integration for the student model. A small τ^S implies that the student model integrates the teacher's parameters slowly and steadily. This allows the student model to gradually assimilate the stable characteristics of the teacher model. To simplify the hyperparameter tuning process in experiments, τ^T is set to a small constant value of 1e-3, while τ^S is variable for different datasets.

In summary, the student model acts as an active learner, continually evolving and assimilating new information, whereas the teacher model functions as a reliable guide, providing a stable parameter reference to shield the student model from the effects of cross-domain variance in training batches.

Algorithm 1 Training Process of our RES-based DG

Student model M^S with initialized parameters θ^S, teacher model M^T with initialized parameters θ^T, all training sample pairs $\{(x_i, y_i)\}_i^N$, and batch size B
for *iter* $= 1$ to N_{iters} **do**
  Sample Batch Data: $\{(x_i, y_i)\}_i^B$
  Extract Feature: $f \leftarrow \mathcal{M}^{S-}(x)$
  Decompose Phase and Amplitude: $\mathcal{P}, \mathcal{A} \leftarrow FFT(f)$
  Augmentation Selection: Aug $\leftarrow$ [`random noise`, `random dropout`, `mixup`, `none`]
  Augment Amplitude: $\mathcal{A}' \leftarrow \text{Aug}(\mathcal{A})$
  Reconstruct feature: $f' = iFFT(\mathcal{P}, \mathcal{A}')$
  Compute prediction: $p \leftarrow \mathcal{M}^{S+}(f')$
  Compute Loss: $\mathcal{E}(\theta^S) \leftarrow \ell(p, y)$
  Update Student Model with Loss: $\theta^S \leftarrow \theta^S - \eta \nabla_\theta L(\theta^S)$
  Integrate Student Model to Teacher Model: $\theta^T = \tau^T \theta_T + (1 - \tau^T)\theta_S$
  Stabilize Student Model with Teacher Model: $\theta^S = (1 - \tau^S)\theta^S + \tau^S \theta^T$
end for

3.4 Supervision Loss

During each training iteration with a batch size of B, the empirical risk minimization (ERM) loss function is used to optimize the student model M^S across all the training domains as

$$\mathcal{E}(\theta^S) = \frac{1}{B}\sum_{i=1}^{B} \ell(p_i, y_i) = \frac{1}{B}\sum_{i=1}^{B} \ell(M^S(x_i; \theta^S), y_i), \tag{14}$$

where the detailed progress of $M^S(x_i; \theta^S)$ come from Eqs. (3)-(7) and (12) and ℓ represents the CE loss. Here $\mathcal{E}(\theta^S)$ represents the basic ERM loss.

4 Experiments

This section outlines benchmark datasets and experimental settings, conducts an ablation study on RES, compares it with advanced DG methods, examines the impact of τ^S, and quantitatively assesses RES's effects in reducing cross-domain bias and variance and improving accuracy across domains. In Supplementary Material, we compare the effects of data augmentation from two perspectives: image-level vs. feature-level frequency augmentation, and normal feature augmentation vs. feature frequency augmentation. This comparison demonstrates the superiority of our feature frequency augmentation.

4.1 Benchmark Datasets

We evaluate the proposed RES method on five DG datasets: (1) `PACS` [22], which contains 9,991 images in 7 object categories across 4 diverse domains; (2) `VLCS` [10], which comprises 10,729 examples from 5 categories, collected from 4

domains; (3) `OfficeHome` [36], which encompasses around 15,500 samples in 65 categories from four domains. (4) `TerraIncognita` [3], which contains 24,788 images within 10 categories from four domains; and (5) `DomainNet` [30], a vast dataset with 586,575 images across 345 categories, spanning 6 domains.

4.2 Experimental Settings

Basic settings. We utilize ResNet-50 [16], pretrained on ImageNet-1K [9], as the backbone. Our codes are constructed on Domainbed's framework [13], allocating 80% of in-domain data for training and the remaining 20% for validation, with test performed on an unseen domain. `PACS`, `VLCS`, and `OfficeHome` are trained for 10K iterations and validated every 300 iterations, `TerraIncognita` is trained for 3K iterations and validated every 300 iterations, and `DomainNet` undergoes 400K iterations of training and 1K iterations of validation. Adam is used as the optimizer, with a batch size of 128 and a learning rate of 5e-5.

Settings of RES. In every training iteration, we apply the frequency augmentation of RE randomly to one of the four potential middle-layer positions within ResNet-50. These positions are strategically located at the ends of the 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} residual blocks, ensuring diverse augmentation effects across different layers of the model. There are some hyperparameters in RES summarized in Table 1, where τ^S is variable and tuned based on the validation performance as discussed in Section 4.4. All the experiments of RES are conducted for three times and the mean values and standard deviations are reported.

Table 1: Hyperparamters of RES.

RE^{ns}	RE^{dp}	RE^{mix}	$\text{RS}^{S\to T}$	$\text{RS}^{T\to S}$
$G \sim \mathcal{N}(0.75, 0.75^2)$	$p = 0.5$	$\alpha \sim U(0,1)$	$\tau^T = 0.001$	τ^S is variable

4.3 Ablation Study

The ablation study of RES's modules, including sub-strategies of RE and RS, on PACS are provided in Table 2 and Fig. 3. RE employs `random noise` (RE^{ns}), `random dropout` (RE^{dp}), and `mixup` (RE^{mix}) techniques for feature augmentation, and RS employs MEMA containing representation stabilization for both student models (RS^s) and teacher models (RS^t).

After applying all three RE strategies, there is significant performance improvement from ERM's 85.5 to (ERM+RE)'s 88.8 as shown in Table 2. with their training trajectory shown in Fig. 3. The results verify the effect of RE in reducing cross-domain bias. Additionally, the application of RS strategies reduces their standard deviations from ERM's 1.1 to (ERM+RS)'s 0.5, demonstrating their

effectiveness in reducing cross-domain variance. When RE and RS strategies are applied in conjunction, as shown in the last highlighted row of the table, the model achieves optimal test performance, with an average accuracy increase to 90.0 with the smallest standard deviation of 0.3, significantly surpassing ERM.

Through this ablation study, the complementary effects of RE and RS strategies in enhancing model generalization across unseen domains are clearly demonstrated, underscoring the significance of reducing cross-domain bias-variance.

Table 2: Ablation study of the proposed RES on PACS.

ERM	RE^{ns}	RE^{dp}	RE^{mix}	RS^t	RS^s	PACS				Avg
						Art	Clipart	Painting	Sketch	
✓						84.4±0.6	81.1±0.7	96.3±0.5	80.2±2.5	85.5±1.1
✓	✓					88.8±1.2	82.2±0.5	97.0±0.6	81.2±0.6	87.3±0.7
✓		✓				87.6±0.3	82.7±1.0	96.8±0.8	82.1±1.6	87.3±0.9
✓			✓			89.4±0.2	81.4±1.1	96.6±0.4	85.1±0.9	88.5±0.7
✓	✓	✓	✓			89.0±0.9	83.0±0.4	97.0±0.4	85.1±0.7	88.8±0.6
✓				✔		87.9±0.1	83.1±0.2	96.0±0.4	82.2±1.2	87.3±0.5
✓				✔	✔	90.9±0.4	82.0±0.5	97.9±0.3	82.2±0.8	88.2±0.5
✓	✓	✓	✓	✔	✔	**91.6**±0.4	**84.2**±0.3	**98.1**±0.1	**86.1**±0.4	**90.0**±0.3

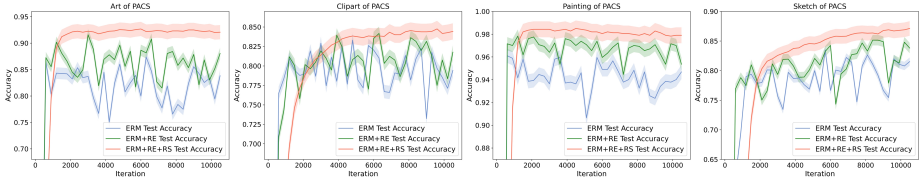


Fig. 3: Training convergence of RES's modules on PACS.

4.4 Values of τ^S in MEMA

In RES, the MEMA's τ^S is the sole hyperparameter requiring adjustment across various benchmarks. We explored five distinct values: [0, 1e-3, 5e-3, 1e-2, 5e-2] for each benchmark, with results shown in Fig. 4. The results indicate that the optimal τ^S value varies by dataset, attributed to differences in their convergence speed. Subsequent comparison experiments will report out-of-domain accuracies for τ^S corresponding to the highest in-domain accuracies observed.

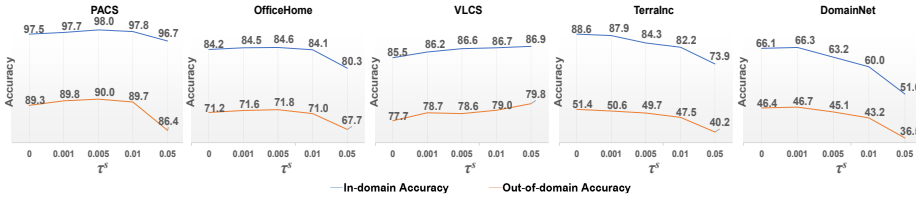


Fig. 4: In-domain (validation) and out-of-domain (test) accuracies vary with τ^S.

4.5 In Comparison With Advanced DG Methods

To objectively assess the effectiveness of the proposed RES, we conducted comparative analyses against several state-of-the-art DG methods. This comparison spans the baseline ERM, various data augmentation strategies (e.g., Mixup [40], CCFP [21], DomainDrop [14]), flatness-aware methods (such as SWAD [4], SMA [2], SAGM [38], FAD [45]), and some other DG approaches (including FISH [35], RSC [18], DAC-SC [20], GVRT [27], MIRO [5]). These evaluations were performed across five benchmarks utilizing the ResNet-50 backbone, with results detailed in Table 3. The domain-wise performance of the datasets of our RES are provided in Supplementary Material. Notably, RES incorporates both data augmentation and flatness-aware strategies, positioning it uniquely to tackle the dual challenges of cross-domain bias and variance simultaneously.

Table 3: Performance comparisons between some SOTA methods and our RES.

Method	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg
ERM [ICLR'20] [13]	85.7±0.5	77.4±0.3	67.5±0.5	47.2±0.4	41.2±0.2	63.8
RSC [ECCV'20] [18]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
Mixup [AAAI'20] [40]	84.3±0.5	77.7±0.4	69.0±0.1	48.9±0.8	39.6±0.1	63.9
FISH [ICLR'21] [35]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
GVRT [ECCV'22] [27]	85.1±0.3	79.0±0.2	70.1±0.1	48.0±0.2	44.1±0.1	65.2
MIRO [ECCV'22] [5]	85.4±0.4	79.0±0.0	70.5±0.4	50.4±1.1	44.3±0.2	65.9
SMA [NeurIPS'22] [2]	87.5±0.2	78.2±0.2	70.6±0.1	50.3±0.5	46.0±0.1	66.5
SWAD [NeurIPS'21] [4]	88.1±0.1	79.1±0.1	70.6±0.2	50.0±0.3	46.5±0.1	66.9
CCFP [ICCV'23] [21]	86.6±0.2	78.9±0.3	68.9±0.1	48.6±0.4	41.2±0.1	64.8
DAC-SC [CVPR'23] [20]	87.5±0.1	78.7±0.3	70.3±0.2	46.5±0.3	44.9±0.1	65.6
FAD [ICCV'23] [45]	88.2±0.2	78.9±0.8	69.2±0.5	45.7±1.0	44.4±0.1	65.3
SAGM [CVPR'23] [38]	86.6±0.2	**80.0±0.3**	70.1±0.2	48.8±0.9	45.0±0.2	66.1
DomainDrop [ICCV'23] [14]	87.9±0.3	79.8±0.3	68.7±0.1	**51.5±0.4**	44.4±0.5	66.5
Our RES	**90.0±0.3**	79.8±0.2	**71.8±0.3**	51.4±0.6	**46.7±0.2**	**67.9**

The table showcases the RES algorithm's performance across five benchmarks, with "Avg" representing the average performance across them. This provides a comprehensive view of each method's generalizability and effectiveness in DG. The proposed RES demonstrates superior performance, particularly high-

lighted by its leading average score, indicating its robustness and adaptability across a diverse set of domains. The comparative analysis confirms the superiority and effectiveness of the RES in mitigating cross-domain bias and variance.

4.6 Effects of RES on Domain Generalization

This subsection quantitatively describes the advantages of RES on DG from three perspectives, including reducing cross-domain bias, reducing cross-domain variance, and improving both in-domain and out-of-domain performance.

Reducing Cross-domain Bias. First, we measure the cosine similarity between the mean feature embeddings obtained from the global average pooling layer of ResNet-50 across both source and target datasets within PACS and OfficeHome. This assessment encompasses comparisons between the baseline ERM and our RES, as illustrated in Fig. 5. Our findings indicate that RES significantly enhances the cosine similarity between source and target data compared to ERM, demonstrating its capability in diminishing cross-domain bias.

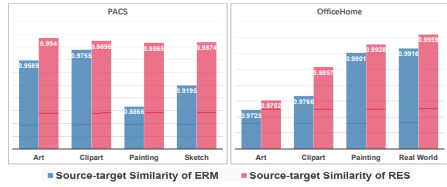


Fig. 5: Cosine similarity between mean features of source domains and target domain. The higher the similarity is, the lower the cross-domain bias is.

Reducing Cross-domain Variance. Secondly, we present the training convergence trajectories alongside the corresponding cumulative instability for four DG scenarios within the PACS dataset, as illustrated in Fig. 6. It is observable that, in comparison to the baseline method ERM represented by the blue color, the proposed RES depicted in red markedly reduces the fluctuations in both in-domain and out-of-domain accuracies with almost no instability increase after the initial increase stage during training, which spans approximately the first 2000 iterations. Benefiting form the stable and narrower performance gap between in-domain (validation) and out-of-domain (test) sets, the best models chosen based on validation sets demonstrate consistent performance on test sets.

Improving Both In-domain and Out-of-domain Performance. As shown in Table 4, the comparison between ERM and RES on PACS illustrates significant enhancements in both in-domain and out-of-domain accuracies achieved by RES. While ERM delivers reliable in-domain performance, it falls short in adapting to unseen domains. In contrast, RES not only boosts in-domain accuracy from ERM's 96.5 to its 98.0 but also lifts the out-of-domain accuracy

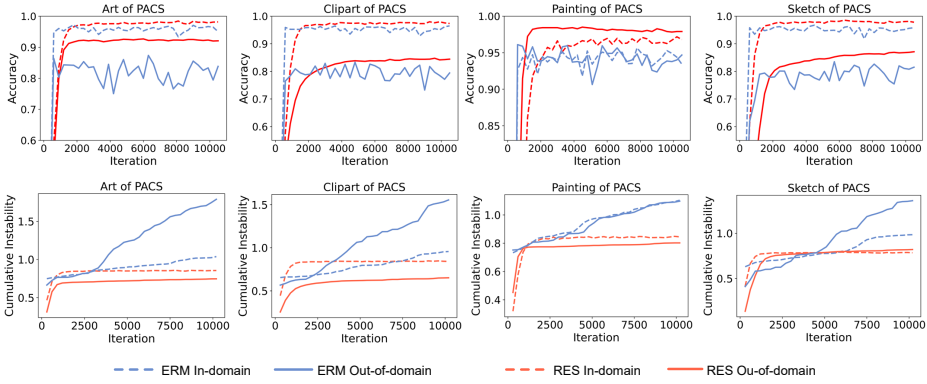


Fig. 6: Training stability of ERM and RES on PACS, evaluated by the absolute first derivative of accuracy every 300 iterations ($|f'(Acc)| = |Acc_t - Acc_{t-1}|$), capturing accuracy change rate. Accumulative instability aggregates $|f'(Acc)|$ up to iteration t, with lower values indicating more stable performance and less variance.

from 85.5±1.1 to 90.0±0.3. This robust improvement underscores RES's robust capability to refine model performance across various distributions.

Table 4: In-domain and out-of-domain accuracies of PACS between ERM and RES.

Split	Method	PACS				Avg
		Art	Clipart	Painting	Sketch	
In-domain (Validation Set)	ERM	97.2±0.0	97.0±0.4	95.4±0.2	96.3±0.4	96.5±0.3
	Our RES	**98.4±0.2**	**98.0±0.0**	**97.4±0.3**	**98.2±0.3**	**98.0±0.2**
Out-of-domain (Test Set)	ERM	84.4±0.6	81.1±0.7	96.3±0.5	80.2±2.5	85.5±1.1
	Our RES	**91.6±0.4**	**84.2±0.3**	**98.1±0.1**	**86.1±0.4**	**90.0±0.3**

5 Conclusion

In this study, we explored DG through bias-variance decomposition, an analytical framework that offers deeper insights into the challenges and solutions for cross-domain generalization in deep learning models. Our approach quantified the detrimental effects of cross-domain bias and variance on model performance and therefore introduced the Representation Enhancement-Stabilization (RES) method as a potent solution. The RES method effectively minimizes the cross-domain bias and variance, enhancing model adaptability and generalization across unseen target domains. Empirical evaluations on five DG benchmarks reveals that the proposed RES markedly outperforms multiple advanced DG methods in improving model generalization.

Acknowledgements

The work was supported by the German Research Foundation (DFG GZ: ZH 498/18-1; Project number: 519016653). The work of Z. Xiong was supported by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) based on a resolution of the German Bundestag (grant number: 67KI32002B; Acronym: *EKAPEx*). The work of X. Zhu was also supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001), by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C) and by the Munich Center for Machine Learning.

References

1. Albuquerque, I., Naik, N., Li, J., Keskar, N., Socher, R.: Improving out-of-distribution generalization via multi-task self-supervised pretraining. arXiv preprint arXiv:2003.13525 (2020)
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. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems **34**, 22405–22418 (2021)
5. 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)
6. Chen, C., Li, J., Han, X., Liu, X., Yu, Y.: Compound domain generalization via meta-knowledge encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7119–7129 (2022)
7. Chen, J., Gao, Z., Wu, X., Luo, J.: Meta-causal learning for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7683–7692 (2023)
8. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. 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)
11. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
12. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural computation **4**(1), 1–58 (1992)
13. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (2020)
14. Guo, J., Qi, L., Shi, Y.: Domaindrop: Suppressing domain-sensitive channels for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19114–19124 (2023)
15. Hansen, B.C., Hess, R.F.: Structural sparseness and spatial phase alignment in natural scenes. JOSA A **24**(7), 1873–1885 (2007)
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. 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)

18. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 124–140. Springer (2020)
19. 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)
20. Lee, S., Bae, J., Kim, H.Y.: Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11776–11785 (2023)
21. Li, C., Zhang, D., Huang, W., Zhang, J.: Cross contrasting feature perturbation for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1327–1337 (2023)
22. 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)
23. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5400–5409 (2018)
24. 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)
25. Lv, F., Liang, J., Li, S., Zang, B., Liu, C.H., Wang, Z., Liu, D.: Causality inspired representation learning for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8046–8056 (2022)
26. Mahajan, D., Tople, S., Sharma, A.: Domain generalization using causal matching. In: International Conference on Machine Learning. pp. 7313–7324. PMLR (2021)
27. Min, S., Park, N., Kim, S., Park, S., Kim, J.: Grounding visual representations with texts for domain generalization. In: European Conference on Computer Vision. pp. 37–53. Springer (2022)
28. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International conference on machine learning. pp. 10–18. PMLR (2013)
29. Oppenheim, A., Lim, J., Kopec, G., Pohlig, S.: Phase in speech and pictures. In: ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 4, pp. 632–637. IEEE (1979)
30. 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)
31. Peng, X., Huang, Z., Sun, X., Saenko, K.: Domain agnostic learning with disentangled representations. In: International Conference on Machine Learning. pp. 5102–5112. PMLR (2019)
32. Piotrowski, L.N., Campbell, F.W.: A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. Perception **11**(3), 337–346 (1982)
33. Piratla, V., Netrapalli, P., Sarawagi, S.: Efficient domain generalization via common-specific low-rank decomposition. In: International Conference on Machine Learning. pp. 7728–7738. PMLR (2020)

34. Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., Han, B.: Learning to optimize domain specific normalization for domain generalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 68–83. Springer (2020)
35. 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)
36. 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)
37. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6678–6687 (2020)
38. Wang, P., Zhang, Z., Lei, Z., Zhang, L.: Sharpness-aware gradient matching for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3769–3778 (2023)
39. Wang, Y., Qi, L., Shi, Y., Gao, Y.: Feature-based style randomization for domain generalization. IEEE Transactions on Circuits and Systems for Video Technology **32**(8), 5495–5509 (2022)
40. Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., Zhang, W.: Adversarial domain adaptation with domain mixup. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 6502–6509 (2020)
41. 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)
42. Yang, F.E., Cheng, Y.C., Shiau, Z.Y., Wang, Y.C.F.: Adversarial teacher-student representation learning for domain generalization. Advances in Neural Information Processing Systems **34**, 19448–19460 (2021)
43. Yang, Z., Yu, Y., You, C., Steinhardt, J., Ma, Y.: Rethinking bias-variance trade-off for generalization of neural networks. In: International Conference on Machine Learning. pp. 10767–10777. PMLR (2020)
44. Zhang, J., Qi, L., Shi, Y., Gao, Y.: Mvdg: A unified multi-view framework for domain generalization. In: European Conference on Computer Vision. pp. 161–177. Springer (2022)
45. Zhang, X., Xu, R., Yu, H., Dong, Y., Tian, P., Cui, P.: Flatness-aware minimization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5189–5202 (2023)
46. Zhao, S., Gong, M., Liu, T., Fu, H., Tao, D.: Domain generalization via entropy regularization. Advances in Neural Information Processing Systems **33**, 16096–16107 (2020)
47. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13025–13032 (2020)
48. Zhu, W., Lu, L., Xiao, J., Han, M., Luo, J., Harrison, A.P.: Localized adversarial domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7108–7118 (2022)