DGFamba: Learning Flow Factorized State Space for Visual Domain Generalization

Qi Bi[1*], Jingjun Yi[1*], Hao Zheng[1†], Haolan Zhan[2], Wei Ji[3], Yawen Huang[1], Yuexiang Li[4†]

[1]Jarvis Research Center, Tencent YouTu Lab, ShenZhen, China
[2]Faculty of Information Technology, Monash University, Melbourne, Australia
[3]School of Medicine, Yale University, New Haven, United States
[4]Faculty of Science and Technology, University of Macau, Macau
{q_bi, rsjingjuny}@whu.edu.cn, howzheng@tencent.com, yuexiang.li@ieee.org

Abstract

Domain generalization aims to learn a representation from the source domain, which can be generalized to arbitrary unseen target domains. A fundamental challenge for visual domain generalization is the domain gap caused by the dramatic style variation whereas the image content is stable. The realm of selective state space, exemplified by VMamba, demonstrates its global receptive field in representing the content. However, the way exploiting the domain-invariant property for selective state space is rarely explored. In this paper, we propose a novel Flow Factorized State Space model, dubbed as DGFamba, for visual domain generalization. To maintain domain consistency, we innovatively map the style-augmented and the original state embeddings by flow factorization. In this latent flow space, each state embedding from a certain style is specified by a latent probability path. By aligning these probability paths in the latent space, the state embeddings are able to represent the same content distribution regardless of the style differences. Extensive experiments conducted on various visual domain generalization settings show its state-of-the-art performance.

Introduction

In many real-world scenarios, the distributions between the source and target domains are not independently and identically distributed (i.i.d). Visual domain generalization handles the domain shift. It learns an image representation extracted from the source domain images, and aims to generalize to arbitrary unseen target domains (Perry, Von Kügelgen, and Schölkopf 2022; Fang et al. 2020; Hendrycks et al. 2021; Geirhos et al. 2021). Its key challenge lies in the domain gap caused by the dramatic style variation whereas the cross-domain image content is stable.

Visual domain generalization has been extensively studied in the past decade. A variety of advanced machine learning techniques (Sagawa et al. 2019; Krueger et al. 2021; Huang et al. 2020; Blanchard et al. 2021; Zhou et al. 2024; Nam et al. 2021) have been proposed to eliminate the impact of cross-domain styles. However, these methods rely heavily on the convolutional neural network (CNN) (He et al. 2016) as the image encoder, which has a limited local receptive field. Since the local receptive field is more sensitive to the style variation and less expressive to the global-wise image content, modern domain generalization methods (Sultana et al. 2022; Li et al. 2023) have shifted the image encoder from CNN to Vision Transformer (ViT) (Dosovitskiy et al. 2020), which is more capable to represent the image content owing to the self-attention mechanism.

More recently, selective state space model (SSM), exemplified by VMamba (Liu et al. 2024) and Vision Mamba (Zhu et al. 2024), has become the new paradigm for visual representation learning. SSM converts the image into patch sequences and exploits the visual information from recurrent modeling, which demonstrates a more global receptive field to represent the content. Such property provides a new feasible path for visual domain generalization, where a robust image content representation is critical. However, the selective scanning is implemented in a fixed way regardless of

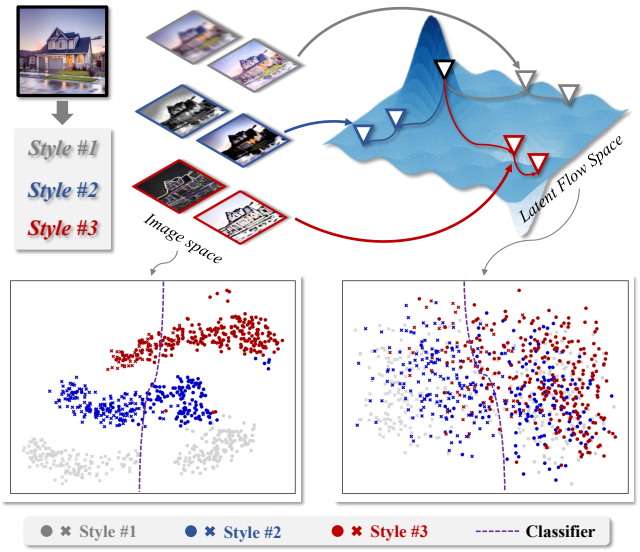

Figure 1: General idea of the proposed DGFamba. The latent flow space aligns the probability path between the pre- and post- style augmented state embeddings to enhance the robustness to the style change.

*These authors contributed equally.

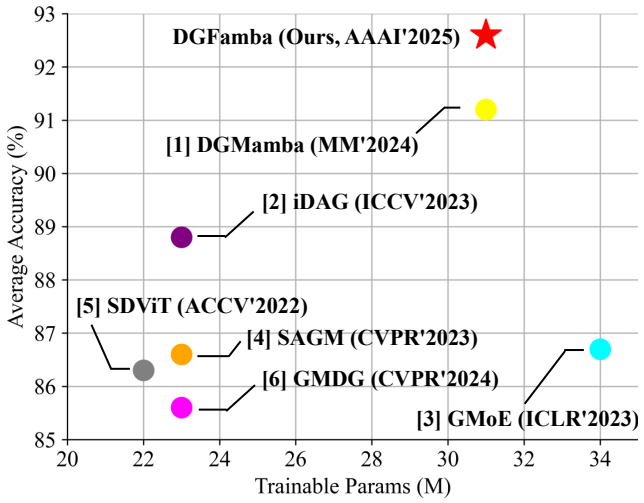

Figure 2: Performance on unseen domains (classification accuracy in %) v.s. trainable parameter number (in million, M). The proposed DGFamba outperforms the state-of-the-art methods.

the image features from different domains, which are usually distributed differently in the feature space. It may not necessarily learn a consistent state embedding before and after the style augmentation. Therefore, our question arises: *How to ensure that the selective state space is invariant to the cross-domain style shift?*

In this paper, we propose a flow factorized state space model, dubbed as DGFamba, for visual domain generalization. Its general idea is to maintain the global receptive field of SSM to represent the content while at the same time empower SSM with style-invariant property. Specifically, we introduce flow factorization (Song et al. 2023b,a), which maps the state embeddings between two styles by a probability path in the latent flow space (illustrated in Fig. 1). By aligning the probability paths between the pre- and post-style augmented state embeddings, the selective state space is able to represent the same content distribution regardless of the style differences.

The proposed DGFamba consists of three key components, namely, state style randomization, state flow encoding, and state flow constraint. Specifically, before the selective scanning and recurrent modeling, the state style randomization maximizes the style diversity of the state embedding. The style, parameterized by mean and standard deviation (Huang and Belongie 2017), is randomly sampled from a uniform distribution. Then, the state flow encoding component projects the pre- and post- style hallucinated state embeddings into the latent flow space, and factorizes their latent probability path. Finally, the state flow constraint aligns the latent probability path between the pre- and post- hallucinated state embeddings, so that the property of style invariant is achieved.

Our contributions can be summarized as follows.

- We conduct an initial exploration of harnessing SSM for visual domain generalization, and propose a flow factorized state space modeling method (DGFamba).
- We introduce the flow factorization to represent the pre- and post- style hallucinated state embedding, which theoretically warrants the style invariant property of SSM.
- Experiments demonstrate that the proposed DGFamba not only significantly outperforms existing CNN and ViT based methods, but also surpasses the contemporary DGMamba by up to 1.5% in top-1 accuracy.

Related Work

Mamba and Vision Mamba. Selective State Space Modeling (SSM), exemplified by Mamba and its variations (He et al. 2024; Li et al. 2024; Liu et al. 2024; Wang et al. 2024; Xiao et al. 2023; Bi et al. 2024a), is an emerging representation learning tool, which possesses global receptive fields with only linear complexity. In the field of computer vision, VMamba (Liu et al. 2024) and Vim (Zhou et al. 2021) are pioneering works that adapt SSM for visual representation learning.

Domain Generalization. Most prior works use CNN as their backbone. A variety of machine learning techniques, such as empirical risk minimization (Xu et al. 2021; Huang et al. 2020), domain alignment (Wang et al. 2023; Nam et al. 2021; Wang et al. 2022; Zhang et al. 2022), domain augmentation (Zhou et al. 2020, 2024), ensemble learning (Kim et al. 2021; Chen et al. 2022; Chu et al. 2022), frequency decoupling (Bi, You, and Gevers 2024a; Yi et al. 2024; Bi, You, and Gevers 2024b; Bi et al. 2024b), and meta learning (Dou et al. 2019; Du et al. 2020; Zhao et al. 2021), have been proposed. Vision Transformer has demonstrated its superiority in visual domain generalization (Noori et al. 2024; Zhang et al. 2022). Techniques such as mixture of experts (Li et al. 2023) and token-wise stylization (Noori et al. 2024) have been studied. More recently, (Long et al. 2024) made an earlier exploration to harness SSM for this task. However, the proposed DGMamba (Long et al. 2024) only focused on improving the hidden space and patch embedding. The style invariant property, which is crucial for visual domain generalization, remains unaddressed.

Flow Factorization (Song et al. 2023b,a), as an emerging representation learning tool, holds a unique position to understand both disentangled and equivalent representations. Inspired by the general idea that the representation distribution is encouraged to be factorial without substantially affecting the quality (Kim and Mnih 2018), the probability of each transformation is modeled as a flow by the gradient field of some learned potentials following fluid mechanical dynamic optimal transport. However, *to the best of our knowledge*, flow factorization has so far rarely been explored in the context of domain generalization, especially for the style invariant properties.

Methodology

Fig. 3 gives an overview of the proposed DGFamba, which consists of three key components, namely, state style randomization (SSR), State Flow Encoding (SFE), and State Flow Constraint (SFC), respectively.

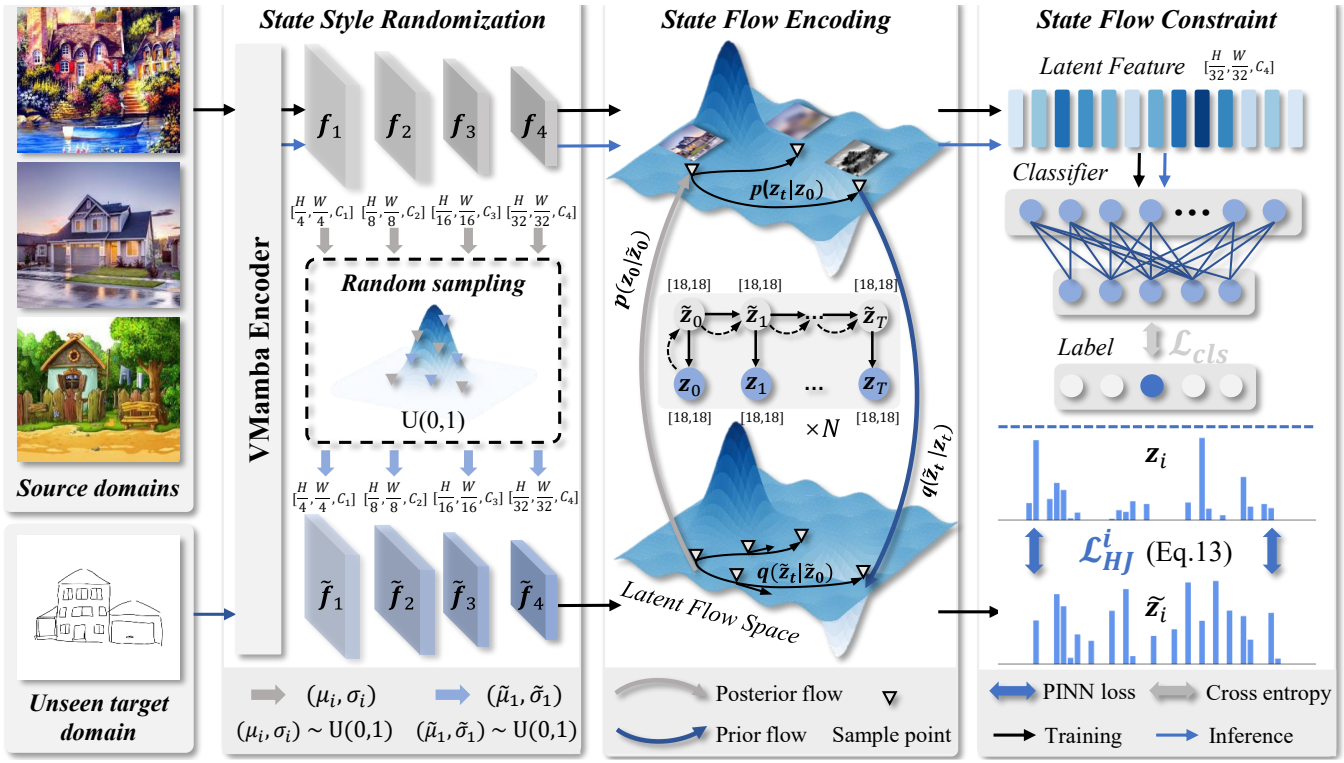

Figure 3: The proposed flow factorized state space model, dubbed as DGFamba, consists of three key components, namely, State Style Randomization, State Flow Encoding, and State Flow Constraint.

State Style Randomization

Backbone & State Embedding. We use the VMamba (Liu et al. 2024) as the backbone, assuming it consists of N sequential layers from four feature blocks, denoted as L_1, L_2, $\cdots$, L_N. For each layer, the output state embedding is denoted as $\boldsymbol{f}_i$.

State Style Representing. The state embedding $\boldsymbol{f}_i$ from a certain layer L_i ($i = 1, 2, \cdots, N$) is supposed to generalize well to unseen target domains, where the dramatically different style variation leads to the distribution shift. To realize this objective, the first step is to quantify and represent the style. Following the common definition of style (Huang and Belongie 2017), the channel-wise mean $\boldsymbol{\mu}_i$ and standard deviation $\boldsymbol{\sigma}_i$ is used to quantize the style of $\boldsymbol{f}_i$, given by

$$\boldsymbol{\mu}_i = \frac{1}{C}\sum_{c=1}^{C} \boldsymbol{f}_i^c, \quad \boldsymbol{\sigma}_i = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{f}_i^c - \boldsymbol{\mu}_i)^2}. \tag{1}$$

where C denotes the channel size. After normalization, the per-channel mean and standard deviation value ranges from 0 to 1, denoted as $\mu_{i,m} \in \mathbb{R}^{[0,1]}$ and $\sigma_{i,m} \in \mathbb{R}^{[0,1]}$.

Style Randomization. The model can learn the styles only from the source domain. To enrich the style diversity, we hallucinate the state embedding $\boldsymbol{f}_i$ and generate the style augmented state embedding $\widetilde{\boldsymbol{f}}_i$ with a random style. Specifically, we randomize the hallucinated styles $[\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\sigma}}_i]^{\mathrm{T}}$ from the entire style space $\mathcal{S} \subset \mathbb{R}^{[0,1]\times[0,1]}$:

$$\widetilde{\mu}_{i,m} \sim [0,1], \quad \widetilde{\sigma}_{i,m} \sim [0,1]. \tag{2}$$

State Embedding Stylization. The randomized styles $[\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\sigma}}_i]^{\mathrm{T}}$ are injected into the original state embedding by AdaIN (Huang and Belongie 2017), computed as

$$\widetilde{\boldsymbol{f}}_i = \widetilde{\boldsymbol{\sigma}}_i \cdot \frac{\boldsymbol{f}_i - \boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i} + \widetilde{\boldsymbol{\mu}}_i. \tag{3}$$

State Flow Encoding

After enriching the state style, how to learn a state embedding invariant to pre- and post- hallucinated styles is the key bottleneck. To address this, we introduce flow factorization (Song et al. 2023b,a), a recent representation disentanglement tool, to model the state embedding probability from each style as a flow by the gradient field in the latent space. The mechanical dynamic optimal transport of the factorized flows allows the state embedding to be invariant to the styles.

Generating Flow Embeddings. Both $\boldsymbol{f}_i$ and $\widetilde{\boldsymbol{f}}_i$ are mapped to a latent embedding $\boldsymbol{z}_i$ and $\widetilde{\boldsymbol{z}}_i$ in the latent flow space by a Variational Auto-Encoder (VAE). The architecture of VAE, for simplicity, directly follows (Song et al. 2023a).

Prior State Flow Factorization. The prior flow maps the transformation from the original state embedding $\boldsymbol{z}_i$ to the hallucinated state embedding $\widetilde{\boldsymbol{z}}_i$. Assume there are a total of T steps in the factorization, the prior state flow can be factorized to T terms, given by

$$p(\boldsymbol{z}_i, \widetilde{\boldsymbol{z}}_i) = p(\widetilde{\boldsymbol{z}}_{i,0})p(\boldsymbol{z}_{i,0}|\widetilde{\boldsymbol{z}}_{i,0}) \prod_{t=1}^{T} p(\widetilde{\boldsymbol{z}}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t-1})p(\boldsymbol{z}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t}). \tag{4}$$

Prior State Flow Evolution. The prior flow evolves from the original state embedding to the hallucinated state embedding, which allows the probability density of the flow to be defined by the factorization. The conditional update $p(\widetilde{\boldsymbol{z}}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t-1})$ is computed under a continuity equation form, given by $\partial_t p(\widetilde{z}_i) = -\nabla \cdot (p(\widetilde{z}_i)\nabla\psi(\widetilde{z}_i))$. Here, $\nabla\psi(\widetilde{z}_i)$ denotes the induced velocity field, which is adverted by the potential function ψ on the probability density $p(\widetilde{z}_i)$. As the discrete particle evolution on the density is modeled as $\widetilde{\boldsymbol{z}}_{i,t} = f(\widetilde{\boldsymbol{z}}_{i,t-1}) = \widetilde{\boldsymbol{z}}_{i,t-1} + \nabla_{\widetilde{\boldsymbol{z}}}\psi(\widetilde{\boldsymbol{z}}_{i,t-1})$, the conditional update of the prior flow evolution is computed as

$$p(\widetilde{\boldsymbol{z}}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t-1}) = p(\widetilde{\boldsymbol{z}}_{i,t-1})|\frac{df(\widetilde{\boldsymbol{z}}_{i,t-1})}{d\widetilde{\boldsymbol{z}}_{i,t}}|^{-1}. \quad (5)$$

The diffusion equation $\psi = -D\log p(\widetilde{\boldsymbol{z}}_{i,t})$, which represents the random trajectories with a minimum of informative prior, allows the prior flow to be evolved as

$$\partial_t p(\widetilde{\boldsymbol{z}}_{i,t}) = -\nabla \cdot (p(\widetilde{\boldsymbol{z}}_{i,t})\nabla\psi) = D\nabla^2 p(\widetilde{\boldsymbol{z}}_{i,t}), \quad (6)$$

where D is a constant coefficient.

Posterior State Factorization. In contrast to the prior flow, the posterior flow maps the approximation from the latent embedding of the hallucinated state $\widetilde{\boldsymbol{z}}_i$ to the latent embedding of the original state $\boldsymbol{z}_i$. The posterior flow can be factorize as

$$q(\widetilde{\boldsymbol{z}}_i|\boldsymbol{z}_i) = q(\widetilde{\boldsymbol{z}}_{i,0}|\boldsymbol{z}_{i,0})\prod_{t=1}^{T} q(\widetilde{\boldsymbol{z}}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t-1}). \quad (7)$$

Posterior State Evolution. Same as the prior flow evolution of the token features, the continuity equation is used to model the posterior flow evolution. Specifically, given the particle evolution function $\widetilde{\boldsymbol{z}}_{i,t} = g(\widetilde{\boldsymbol{z}}_{i,t-1}) = \widetilde{\boldsymbol{z}}_{i,t-1} + \nabla_{\widetilde{\boldsymbol{z}}}u$, it is mathematically computed as

$$q(\widetilde{\boldsymbol{z}}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t-1}) = q(\widetilde{\boldsymbol{z}}_{i,t-1})|\frac{dg(\widetilde{\boldsymbol{z}}_{i,t-1})}{d\widetilde{\boldsymbol{z}}_{i,t-1}}|^{-1}. \quad (8)$$

After discretizing the above equation and implementing the logarithm operation, Eq. 8 can be mathematically reformulated as

$$\log q(\widetilde{\boldsymbol{z}}_{i,t}|\widetilde{\boldsymbol{z}}_{i,t-1}) = \log q(\widetilde{\boldsymbol{z}}_{i,t-1}) - \log|1 + \nabla^2_{\widetilde{\boldsymbol{z}}}u|. \quad (9)$$

State Flow Constraint

After modeling the original state embedding and the hallucinated state embedding by the prior and posterior flow, it is necessary to constrain them in the latent flow space, so that both state embeddings are enforced to learn the same representation despite style variance. In flow factorization (Song et al. 2023b,a), this constraint is realized by the latent posterior with the optimal transport path.

Definition 1. Benamou-Brenier Formula. *Given two probability measures μ_0 and μ_1, their $L2$ Wasserstein distance is defined by*

$$W_2(\mu_0,\mu_1)^2 = \min_{\rho,\nu}\{\int\int \frac{1}{2}\rho(x,t)|\nu(x,t)|^2 dxdt\}. \quad (10)$$

where the density ρ and the velocity ν satisfy:

$$\frac{d\rho(x,t)}{dt} = -\nabla\cdot(\nu(x,t)\rho(x,t)), \nu(x,t) = \nabla u(x,t). \quad (11)$$

Specifically, when ∇u satisfies certain partial differential equation (PDE), the probability density evolution between the original state embedding and the hallucinated state embedding can be minimized by the $L2$ Wasserstein distance. The generalized Hamilton-Jacobi (HJ) equation (i.e., $\partial_t u + 1/2||\nabla u||^2 \leq 0$) determines the optimality condition of the velocity. Consequently, the posterior flow of the state embedding is supposed to satisfy the HJ equation with an external driving force, given by

$$\frac{\partial}{\partial t}u(\widetilde{\boldsymbol{z}}_i,t) + \frac{1}{2}||\nabla_{\widetilde{\boldsymbol{z}}_i}u(\widetilde{\boldsymbol{z}}_i,t)||^2 = f(\widetilde{\boldsymbol{z}}_i,t) \quad s.t. f(\widetilde{\boldsymbol{z}}_i,t) \leq 0. \quad (12)$$

To realize the negative constraint of this external force $f(\widetilde{\boldsymbol{z}}_i,t)$, a MLP is used for parameterization, given by $f(\widetilde{\boldsymbol{z}}_i,t) = -\mathrm{MLP}([\widetilde{\boldsymbol{z}}_i;t])^2$. For simplicity, the MLP we use for $f(\widetilde{\boldsymbol{z}}_i,t)$ shares the same architecture as $u(\widetilde{\boldsymbol{z}}_i,t)$ does. The above PDE constraint is realized by a physics-informed neural network loss (Raissi, Perdikaris, and Karniadakis 2019), given by

$$\begin{aligned}\mathcal{L}^i_{HJ} = \frac{1}{T}\sum_{t=1}^{T}(\frac{\partial}{\partial t}u(\widetilde{\boldsymbol{z}}_i,t) + \frac{1}{2}||\nabla_{\widetilde{\boldsymbol{z}}_i}u(\widetilde{\boldsymbol{z}}_i,t)||^2 \\ - f(\widetilde{\boldsymbol{z}}_i,t))^2 + ||\nabla u(\widetilde{\boldsymbol{z}}_0,0)||^2,\end{aligned} \quad (13)$$

where the first term allows the flow to be constrained by the HJ equation, and the second term matches the initial condition. This constraint allows the posterior flow from the original state embedding and the hallucinated state embedding to be optimally aligned, so that the impact caused by the style variation is eliminated.

Implementation Details

The proposed DGFamba uses VMamba (Liu et al. 2024) as the backbone. The initial weights of VMamba have been pre-trained on ImageNet (Deng et al. 2009). The image encoder consists of four blocks, with a number of 2, 2, 4 and 2 VMamba layers. The proposed three key steps are integrated into each of these VMamba layers. The total loss $\mathcal{L}$ is a linear combination between the classification loss $\mathcal{L}_{cls}$ and the HJ loss $\mathcal{L}_{HJ}$ defined in Eq. 13, given by $\mathcal{L} = \mathcal{L}_{cls} + \sum_{i=1}^{N}\mathcal{L}^i_{HJ}$.

For fair evaluation with DGMamba (Long et al. 2024), the configuration settings keep the same. The training terminates after 10000 iterations, with a batch size of 16 per source domain. The AdamW optimizer is used for optimization, with a momentum value of 0.9 and an initial learning rate of 3×10^{-4}. In addition, the cosine decay learning rate scheduler is adapted.

Experiments

Datasets & Evaluation Protocols

Datasets. Our experiments are conducted on four visual domain generalization datasets. Specifically, **PACS** (Li et al. 2017) consists of 9,991 samples from four domains. **VLCS** (Fang, Xu, and Rockmore 2013) consists of a total number of 10,729 samples from four domains. **Office-Home** (Venkateswara et al. 2017) consists of 15,588 samples from four different domains. **TerraIncognita** (Beery,

Method	Venue	Params.	Target domain				Avg.(↑)
			Art	Cartoon	Photo	Sketch	
ResNet-50 Based:							
GroupDRO	ICLR 2019	23M	83.5	79.1	96.7	78.3	84.4
VREx	ICML 2021	23M	86.0	79.1	96.9	77.7	84.9
RSC	ECCV 2020	23M	85.4	79.7	97.6	78.2	85.2
MTL	JMLR 2021	23M	87.5	77.1	96.4	77.3	84.6
Mixstyle	ICLR 2021	23M	86.8	79.0	96.6	78.5	85.2
SagNet	CVPR 2021	23M	87.4	80.7	97.1	80.0	86.3
ARM	NeurIPS 2021	23M	86.8	76.8	97.4	79.3	85.1
SWAD	NeurIPS 2021	23M	89.3	83.4	97.3	82.5	88.1
PCL	CVPR 2022	23M	90.2	83.9	98.1	82.6	88.7
SAGM	CVPR 2023	23M	87.4	80.2	98.0	80.8	86.6
iDAG	ICCV 2023	23M	90.8	83.7	98.0	82.7	88.8
GMDG	CVPR 2024	23M	84.7	81.7	97.5	80.5	85.6
DeiT-S Based:							
SDViT	ACCV 2022	22M	87.6	82.4	98.0	77.2	86.3
GMoE	ICLR 2023	34M	89.4	83.9	99.1	74.5	86.7
VMamba Based:							
DGMamba	MM 2024	31M	91.3	87.0	99.0	87.3	91.2
DGFamba	AAAI 2025	31M	**92.6**	**89.4**	**99.7**	**88.8**	**92.6**

Table 1: Performance comparison between the proposed DGFamba and existing methods on PACS dataset. M: in million.

Method	Target domain				Avg.(↑)
	C	L	S	P	
ResNet-50 Based:					
GroupDRO	97.3	63.4	69.5	76.7	76.7
VREx	98.4	64.4	74.1	76.2	78.3
RSC	97.9	62.5	72.3	75.6	77.1
MTL	97.8	64.3	71.5	75.3	77.2
Mixstyle	98.6	64.5	72.6	75.7	77.9
SagNet	97.9	64.5	71.4	77.5	77.8
ARM	98.7	63.6	71.3	76.7	77.6
SWAD	98.8	63.3	75.3	79.2	79.1
PCL	99.0	63.6	73.8	75.6	78.0
SAGM	99.0	65.2	75.1	80.7	80.0
iDAG	98.1	62.7	69.9	77.1	76.9
GMDG	98.3	65.9	73.4	79.3	79.2
DeiT-S Based:					
SDViT	96.8	64.2	76.2	78.5	78.9
GMoE	96.9	63.2	72.3	79.5	78.0
VMamba Based:					
DGMamba	98.9	64.3	79.2	80.8	80.8
DGFamba	**99.5**	**66.2**	**80.9**	**82.0**	**82.2**

Table 2: Performance comparison between the proposed DGFamba and existing state-of-the-art methods on VLCS dataset. C: Caltech; L: LabelMe; S: SUN; P: PASCAL.

Method	Target domain				Avg. (↑)
	A	C	P	R	
ResNet-50 Based:					
GroupDRO	60.4	52.7	75.0	76.0	66.0
VREx	60.7	53.0	75.3	76.6	66.4
RSC	60.7	51.4	74.8	75.1	65.5
MTL	61.5	52.4	74.9	76.8	66.4
Mixstyle	51.1	53.2	68.2	69.2	60.4
SagNet	63.4	54.8	75.8	78.3	68.1
ARM	58.9	51.0	74.1	75.2	64.8
SWAD	66.1	57.7	78.4	80.2	70.6
PCL	67.3	59.9	78.7	80.7	71.6
SAGM	65.4	57.0	78.0	80.0	70.1
iDAG	68.2	57.9	79.7	81.4	71.8
GMDG	68.9	56.2	79.9	82.0	70.7
DeiT-S Based:					
SDViT	68.3	56.3	79.5	81.8	71.5
GMoE	69.3	58.0	79.8	82.6	72.4
VMamba Based:					
DGMamba	76.2	61.8	83.9	86.1	77.0
DGFamba	**77.4**	**63.7**	**85.6**	**87.3**	**78.5**

Table 3: Performance comparison between the proposed DGFamba and existing state-of-the-art methods on Office-Home. A: Art; C: Clipart; P: Product; R: Real.

Van Horn, and Perona 2018) consists of 24,330 samples from four different domains.

Evaluation Protocols. Following the evaluation protocols of existing methods (Gulrajani and Lopez-Paz 2020; Cha et al. 2021), experiments are conducted under the leave-one-domain-out protocol, where only one domain is used as the unseen target domain and the rest domains are used as the source domains for training. The classification accuracy (in percentage, %) is used as the evaluation metric.

Comparison with State-of-the-art

Existing visual domain generalization methods are compared. The first category is CNN based methods, including GroupDRO (Sagawa et al. 2019), VREx (Krueger et al. 2021), RSC (Huang et al. 2020), MTL (Blanchard et al. 2021), Mixstyle (Zhou et al. 2024), SagNet (Nam et al. 2021), ARM (Blanchard et al. 2021), SWAD (Cha et al. 2021), PCL (Yao et al. 2022), SAGM (Wang et al. 2023),

iDAG (Huang et al. 2023), and GMDG (Tan, Yang, and Huang 2024). The second category is ViT based methods, including SDViT (Sultana et al. 2022) and GMoE (Li et al. 2023). The third category is the contemporary VMamba based method, namely, DGMamba (Long et al. 2024).

Results on PACS are reported in Table 1. DGFamba shows the best performance on all the four experimental settings, yielding an accuracy of 92.6%, 89.4%, 99.7% and 88.8% on A, C, P and S unseen target domain, respectively. The average accuracy achieves 92.6%, outperforming the second-best DGMamba by 1.4%. Specifically, the accuracy improvement on C and S unseen target domains is 2.4% and 1.5%, respectively. It also significantly outperforms existing CNN and ViT based methods by an improvement about 6%, while at the same time has less parameter number.

Results on VLCS are reported in Table 2. DGFamba outperforms all the compared methods under all the four experiment settings. It outperforms the second-best, DGMamba, by 1.4% average accuracy. Notably, the accuracy improvement on unseen L and S domains is 1.9% and 1.7%, respectively. These outcomes indicate that the proposed DGFamba is more stable and more robust when generalized on unseen target domains. On the other hand, DGFamba outperforms the best CNN based method SAGM by an average accuracy of 2.2%, and outperforms the best ViT based method SDViT by an average accuracy of 3.3%.

Results on OfficeHome are reported in Table 3. DGFamba shows state-of-the-art performance on all the four unseen target domains, yielding an average accuracy of 78.5%. It significantly outperforms the second-best DGMamba. The accuracy improvement on the A, C, P and R unseen target domain is 1.2%, 1.9%, 1.7% and 1.2%, respectively. DGMamba outperforms the best CNN based method PCL by an average accuracy of 6.9%, and surpasses the best ViT based method GMoE by an average accuracy of 6.1%.

Results on TerraIncognita. Table 4 compares the performance. Same as the above three experiments, the proposed DGFamba outperforms all existing methods on all the four unseen target domains, yielding an average accuracy of 56.1%. Compared with the second-best, the accuracy improvement on the L100, L38, L43 and L46 unseen target domain is 1.2%, 1.5%, 2.0% and 1.1%, respectively. Compared with the best-performed CNN based method PCL, the average accuracy improvement is 4.0%, Compared with the best-performed ViT based method GMoE, the average accuracy improvement is 10.5%.

Method	Target domain				Avg. (↑)
	L100	L38	L43	L46	
ResNet-50 Based:					
GroupDRO	41.2	38.6	56.7	36.4	43.2
VREx	48.2	41.7	56.8	38.7	46.4
RSC	50.2	39.2	56.3	40.8	46.6
MTL	49.3	39.6	55.6	37.8	45.6
Mixstyle	54.3	34.1	55.9	31.7	44.0
SagNet	53.0	43.0	57.9	40.4	48.6
ARM	49.3	38.3	55.8	38.7	45.5
SWAD	55.4	44.9	59.7	39.9	50.0
PCL	58.7	46.3	60.0	43.6	52.1
SAGM	54.8	41.4	57.7	41.3	48.8
iDAG	58.7	35.1	57.5	33.0	46.1
GMDG	59.8	45.3	57.1	38.2	50.1
DeiT-S Based:					
SDViT	55.9	31.7	52.2	37.4	44.3
GMoE	59.2	34.0	50.7	38.5	45.6
VMamba Based:					
DGMamba	62.7	48.3	61.1	46.4	54.6
DGFamba	**63.9**	**49.8**	**63.1**	**47.5**	**56.1**

Table 4: Performance comparison between the proposed DGFamba and existing state-of-the-art methods on TerraIncognita. The best results are marked in **bold**.

Component			Target domain				Avg. ↑
SSR	SFE	SFC	Art	Cartoon	Photo	Sketch	
✗	✗	✗	88.2	86.2	98.4	84.9	89.4
✓	✗	✗	89.4	87.1	98.7	86.2	90.4
✓	✓	✗	90.7	88.2	99.1	87.9	91.5
✓	✓	✓	**92.6**	**89.4**	**99.7**	**88.8**	**92.6**

Table 5: Ablation studies on each component in the proposed DGFamba. VMamba (Liu et al. 2024) as baseline. Experiments are conducted on the PACS dataset.

Ablation Studies

On Each Component. On top of the VMamba backbone, the proposed DGFamba consists of three key components, namely, State Style Randomization (SSR), State Flow Encoding (SFE), and State Flow Constraint (SFC), respectively. When there is no SFC component, the feature representation processed by SSR or SFE is processed by a MLP to finish the feature propagation. Table 5 inspects the performance of each individual component. SSR mainly focuses on enriching the style diversity. Naively using it functions as a type of feature augmentation, which leads to an average accuracy improvement of 1.0%. Similarly, SFE helps further condense the feature embedding from both the pre- and post- style randomized samples in the latent flow space. It also leads an average accuracy improvement of 1.1%. Finally, SFC constrains the feature distribution between the pre- and post- style randomized samples in the latent flow space, which contributes to the most significant improvement. Especially, the accuracy improvement on the A, C, P and S unseen target domain is 1.9%, 1.2%, 0.6% and 0.9%.

On Each Feature Block. As the proposed DGFamba implements the flow factorization in every layer, it is also necessary to inspect the contribution to generalization from different layers. To this end, Table 6 ablates the block-wise contribution, where $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$ and $\mathbf{F}_4$ denote the implementation on the first, second, third and fourth VMamba block.

It can be seen that, implementing the proposed learning scheme shows a more predominant performance improvement on unseen target domains. Specifically, implementing on the first VMamba block ($\mathbf{F}_1$) can lead to an improvement of 1.5%, 1.3%, 0.5% and 1.2% on the Art, Cartoon, Photo and Sketch unseen target domain, respectively. This may be explained that the shallower features are usually more sensitive to the shift of color, shape and etc, which are typical factors of the cross-domain style. In contrast, the deeper fea-

Feature Block				Target domain			
$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$	Art	Cartoon	Photo	Sketch
✗	✗	✗	✗	88.2	86.2	98.4	84.9
✓	✗	✗	✗	89.7	87.5	98.9	86.1
✓	✓	✗	✗	91.0	88.6	99.3	87.3
✓	✓	✓	✗	91.9	89.0	99.5	88.2
✓	✓	✓	✓	**92.6**	**89.4**	**99.7**	**88.8**

Table 6: Ablation studies on each component in the proposed DGFamba. VMamba (Liu et al. 2024) as baseline. Experiments are conducted on the PACS dataset.

T	Art	Cartoon	Photo	Sketch	avg.
2	91.5	88.6	98.8	87.6	91.6
4	91.9	89.0	99.2	88.1	92.1
6	92.3	89.2	99.5	88.4	92.4
8	**92.6**	**89.4**	**99.7**	**88.8**	**92.6**
10	92.5	89.3	99.4	88.5	92.4
12	92.1	88.7	99.0	88.2	92.0

Table 7: Impact of the factorization step T on generalization performance. Experiments are conducted on PACS dataset.

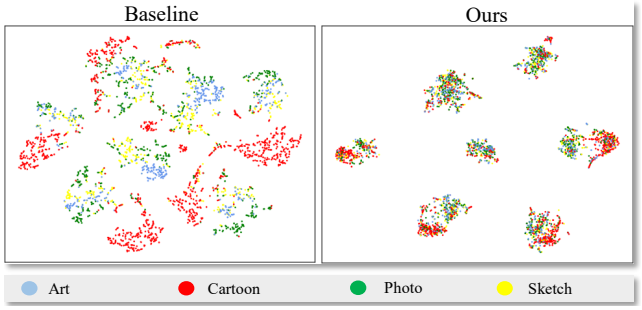

Figure 4: t-SNE feature space visualization between the Mamba baseline and the proposed DGFamba. Experiments conducted on the PACS dataset. A more generalized representation allows features from different domains (in different color) to be uniformly distributed.

tures usually rest more semantic and content information, which is less sensitive to the shift of cross-domain styles. Nevertheless, implementing on the fourth VMamba block ($\mathbf{F}_4$) still exhibits a clear accuracy improvement on unseen domains, namely, 0.7% on Art, 0.4% on Cartoon, 0.2% on Photo and 0.6% on Sketch, respectively.

On Factorization Steps T manipulates the factorization steps in the probability latent space. By default the factorization step T is set to 8. We further test the results when T shifts from 2 to 12, with an interval of 2. The results reported in Table 7 show that, the best performance on unseen domains is achieved when T is around 6 or 8. A factorization step that is too small or too large leads to a decline in performance, which may be explained that a small/large factorization step T can under-/over- fit the probability path, negatively impacting the overall generalization ability.

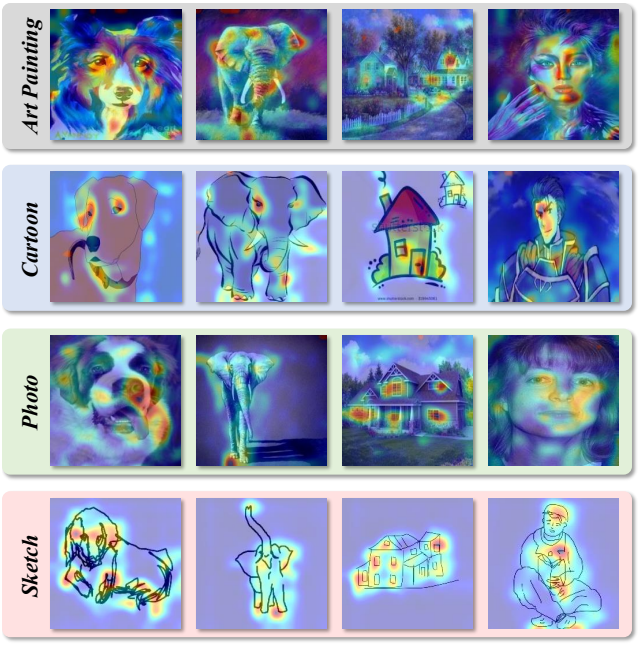

Figure 5: Attention map on the unseen target domain from four experiments on PACS.

Understanding Flow Factorized State Space

t-SNE visualization. We extract the feature embedding of the VMamba baseline and the proposed DGFamba, and inspect their distribution by t-SNE visualization. The results are displayed in Fig. 4a and b, respectively. The samples from different domains in the PACS dataset are labeled in different types of color. Ideally, a generalized representation allows the feature embeddings from different domains (in different color) to be more uniformly distributed, which corresponds to the observation in Fig. 4. This further indicates the generalization ability of DGFamba.

Activation Map Visualization. Fig. 5 visualizes the attention map of the proposed DGFamba on the unseen target domain across all four experimental settings. We use the class activation map (CAM) (Zhou et al. 2016) to compute the heat map and layout on the original image. It can be seen that DGMamba highlights the key local regions of each category, despite the style shift on unseen target domain.

Conclusion

In this paper, we proposed a flow factorized state space learning scheme to harness the emerging selective state space modeling (SSM) for visual domain generalization. Its general idea is to learn a style-invariant state space embedding by first randomizing the styles and then aligning the pre- and post- hallucinated state embedding in the latent flow space. The proposed DGFamba consists of three key components, namely, state style randomization (SSR), state flow encoding (SFE) and state flow constraint (SFC). Extensive experiments demonstrated that DGFamba significantly outperforms existing CNN and ViT based methods, and the concurrent DGMamba.

References

Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in Terra Incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 456–473.

Bi, Q.; Yi, J.; Zheng, H.; Ji, W.; Zhan, H.; Huang, Y.; Li, Y.; and Zheng, Y. 2024a. Samba: Severity-aware Recurrent Modeling for Cross-domain Medical Image Grading. In *Annual Conference on Neural Information Processing Systems*.

Bi, Q.; Yi, J.; Zheng, H.; Zhan, H.; Huang, Y.; Ji, W.; Li, Y.; and Zheng, Y. 2024b. Learning Frequency-Adapted Vision Foundation Model for Domain Generalized Semantic Segmentation. In *Annual Conference on Neural Information Processing Systems*.

Bi, Q.; You, S.; and Gevers, T. 2024a. Generalized Foggy-Scene Semantic Segmentation by Frequency Decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1389–1399.

Bi, Q.; You, S.; and Gevers, T. 2024b. Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 801–809.

Blanchard, G.; Deshmukh, A. A.; Dogan, Ü.; Lee, G.; and Scott, C. 2021. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1): 46–100.

Cha, J.; Chun, S.; Lee, K.; Cho, H.-C.; Park, S.; Lee, Y.; and Park, S. 2021. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, 22405–22418.

Chen, C.; Tang, L.; Liu, F.; Zhao, G.; Huang, Y.; and Yu, Y. 2022. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. In *Advances in Neural Information Processing Systems*, volume 35, 33302–33315.

Chu, X.; Jin, Y.; Zhu, W.; Wang, Y.; Wang, X.; Zhang, S.; and Mei, H. 2022. DNA: Domain Generalization with Diversified Neural Averaging. In *Proceedings of the 39th International Conference on Machine Learning*, 4010–4034. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Dou, Q.; de Castro, D. C.; Kamnitsas, K.; and Glocker, B. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in neural information processing systems*, 32.

Du, Y.; Xu, J.; Xiong, H.; Qiu, Q.; Zhen, X.; Snoek, C. G.; and Shao, L. 2020. Learning to learn with variational information bottleneck for domain generalization. In *European conference on computer vision*, 200–216.

Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. 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*, 1657–1664.

Fang, T.; Lu, N.; Niu, G.; and Sugiyama, M. 2020. Rethinking importance weighting for deep learning under distribution shift. In *Advances in Neural Information Processing Systems*, volume 33, 11996–12007.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2021. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*, volume 34, 23885–23899.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. In *International Conference on Learning Representations*.

He, H.; Bai, Y.; Zhang, J.; He, Q.; Chen, H.; Gan, Z.; Wang, C.; Li, X.; Tian, G.; and Xie, L. 2024. Mambaad: Exploring State Space Models for Multi-class Unsupervised Anomaly Detection. *arXiv preprint arXiv:submit/4242424*. Section 4.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-Challenging Improves Cross-Domain Generalization. In *European Conference on Computer Vision*, 124–140.

Huang, Z.; Wang, H.; Zhao, J.; and Zheng, N. 2023. idag: Invariant dag searching for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19169–19179.

Kim, D.; Yoo, Y.; Park, S.; Kim, J.; and Lee, J. 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9619–9628.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International conference on machine learning*, 2649–2658. PMLR.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826. PMLR.

Li, B.; Shen, Y.; Yang, J.; Wang, Y.; Ren, J.; Che, T.; Zhang, J.; and Liu, Z. 2023. Sparse mixture-of-experts are domain generalizable learners. In *The Eleventh International Conference on Learning Representations*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, Broader and Artier Domain Generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5543–5551.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2312.00752*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

Long, S.; Zhou, Q.; Li, X.; Lu, X.; Ying, C.; Luo, Y.; Ma, L.; and Yan, S. 2024. DGMamba: Domain Generalization via Generalized State Space Model. *arXiv preprint arXiv:2404.07794*.

Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8690–8699.

Noori, M.; Cheraghalikhani, M.; Bahri, A.; Vargas Hakim, G. A.; Osowiechi, D.; Ayed, I. B.; and Desrosiers, C. 2024. Tfs-vit: Token-level feature stylization for domain generalization. *Pattern Recognition*, 149: 110213.

Perry, R.; Von Kügelgen, J.; and Schölkopf, B. 2022. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*, volume 35, 10904–10917.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378: 686–707.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Song, Y.; Keller, A.; Sebe, N.; and Welling, M. 2023a. Flow Factorized Representation Learning. *Advances in Neural Information Processing Systems*, 36.

Song, Y.; Keller, T. A.; Sebe, N.; and Welling, M. 2023b. Latent Traversals in Generative Models as Potential Flows. In *International Conference on Machine Learning*, 32288–32303.

Sultana, M.; Naseer, M.; Khan, M. H.; Khan, S.; and Khan, F. S. 2022. Self-distilled vision transformer for domain generalization. In *Proceedings of the Asian Conference on Computer Vision*, 3068–3085.

Tan, Z.; Yang, X.; and Huang, K. 2024. Rethinking multi-domain generalization with a general learning objective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5385–5394.

Wang, Y.; Chen, J.; Wang, M.; Li, H.; Wang, W.; Su, H.; Lai, Z.; Wang, W.; and Chen, Z. 2023. A Closer Look at Classifier in Adversarial Domain Generalization. In *Proceedings of the 31st ACM International Conference on Multimedia*, 280–289.

Wang, Y.; Li, H.; Cheng, H.; Wen, B.; Chau, L.-P.; and Kot, A. 2022. Variational Disentanglement for Domain Generalization. *Transactions on Machine Learning Research*.

Wang, Z.; Zheng, J.-Q.; Zhang, Y.; Cui, G.; and Li, L. 2024. Mamba-unet: Unet-like Pure Visual Mamba for Medical Image Segmentation. *arXiv preprint arXiv:2402.05079*.

Xiao, Y.; Tang, Z.; Wei, P.; Liu, C.; and Lin, L. 2023. Masked Images are Counterfactual Samples for Robust Fine-Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20301–20310.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A Fourier-Based Framework for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Yao, X.; Bai, Y.; Zhang, X.; Zhang, Y.; Sun, Q.; Chen, R.; Li, R.; and Yu, B. 2022. PCL: Proxy-based Contrastive Learning for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7097–7107.

Yi, J.; Bi, Q.; Zheng, H.; Zhan, H.; Ji, W.; Huang, Y.; Li, Y.; and Zheng, Y. 2024. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8159–8168.

Zhang, C.; Zhang, M.; Zhang, S.; Jin, D.; Zhou, Q.; Cai, Z.; Zhao, H.; Liu, X.; and Liu, Z. 2022. Delving Deep into the Generalization of Vision Transformers under Distribution Shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7277–7286.

Zhao, Y.; Zhong, Z.; Yang, F.; Luo, Z.; Lin, Y.; Li, S.; and Sebe, N. 2021. 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*, 6277–6286.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, 561–578.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain generalization with mixstyle. In *International Conference on Learning Representations*.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2024. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3): 822–836.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.