Perturbating, Tuning, and Collaborating: Harnessing Vision Foundation Models for Single Domain Generalization on Medical Imaging

Chuang Liu[1,2*], **Yichao Cao**[3*], **YingYing Zhang**[1,2], **Xiu Su**[4†], **Haogang Zhu**[1,2 †]

[1]State Key Laboratory of Complex & Critical Software Environment, Beihang University, China
[2]Hangzhou International Innovation Institute, Beihang University, China
[3]School of Automation, Southeast University, China
[4]Big Data Institute, Central South University, China
{cliu_trans, zhangyingying, haogangzhu}@buaa.edu.cn, caoyichao@seu.edu.cn, xiusu@csu.edu.cn

Abstract

Single Domain Generalization (SDG) is critical in medical imaging applications. Recently, Vision Foundation Models (VFMs) have spearheaded a trend in AI development due to their robust generalizability and versatility. This work aims to fully explore the generalization capabilities of VFMs alongside the domain-specific expertise of specialized models, thoroughly investigating the boundaries of their respective capabilities, thereby collaboratively addressing SDG challenges within medical imaging. We propose a framework for **Colla**borative reasoning between **S**pecialized and **U**niversal models for **S**ingle **D**omain **G**eneralization (***CollaSU-SDG***) in medical imaging. Specifically, we first design a model-aware perturbation injection method from the perspective of single-source domain data, enabling differentiated and adaptive perturbation injection for two different scales of models. Then, a domain expansion adapter is designed for the VFM to adapt to the augmented single-source domain medical data. Lastly, we introduce an adaptive hierarchical transfer and dynamic dense prompting method that facilitate collaborative reasoning between the specialized and universal models, eliminating the need for explicit prompts. Through these designs, ***CollaSU-SDG*** fully leverages the strengths of both specialized and universal models, achieving robust out-of-distribution generalization capabilities on single-source domain data. Experimental results demonstrate that ***CollaSU-SDG*** significantly advances the state-of-the-art performance across a wide range of medical datasets. All the code will be publicly available.

Introduction

In the field of medical imaging, *Domain Adaptation* (DA) (Guan and Liu 2021) and *Domain Generalization* (DG) (Zhou et al. 2022) aim to tackle variations in data distributions that deviate from the traditional assumption of independent and identically distributed (IID) data (Carlucci et al. 2019b). Within this context, *Single-source Domain Generalization* (SDG) specifically focuses on the realistic challenge of developing methods that can generalize from a single source to multiple OOD target domains (Xu et al. 2023). This is particularly relevant in medical imaging, especially under conditions of limited data availability or stringent privacy concerns (Price and Cohen 2019). Significant contributions include Data Augmentation (Su et al. 2022a), Regularization (Huang et al. 2020a), and Self-supervised Learning (Carlucci et al. 2019a). However, the SDG problem remains far from being fully resolved.

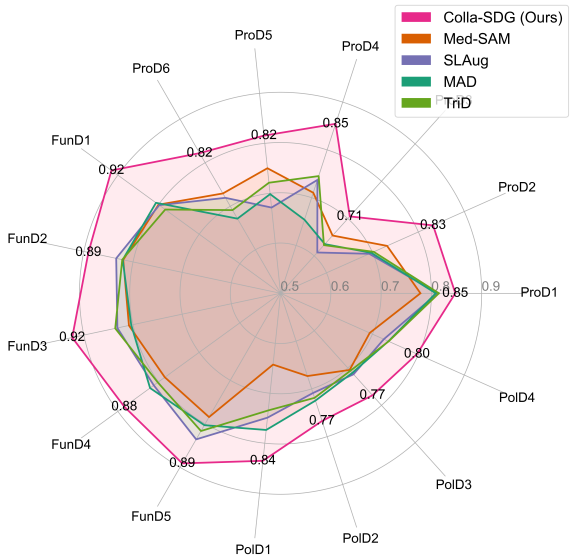

Figure 1: ***CollaSU-SDG*** outperforms state-of-the-art models on a broad range of datasets. "Pro", "Fun" and "Pol" denote three medical tasks. D_i denotes different domain of datasets.

Recent advancements in deep learning have been propelled by foundation models such as CLIP (Radford et al. 2021), LLaMA (Touvron et al. 2023), and GPT (Brown et al. 2020), impacting diverse fields including computer vision (CV) and natural language processing (NLP). Unlike earlier models, which were often specialized ($\mathcal{S}$-model), these are termed **Universal Model** ($\mathcal{U}$-model, or **Foundation Model**), emphasizing their robust generalization capabilities across a broad range of tasks (Zhou et al. 2023). Notably, the Segment Anything Model (SAM) (Kirillov et al. 2023b) emerges as the first foundation model tailored for image segmentation, showcasing remarkable performance across various domains

*These authors contributed equally.
†Corresponding author.

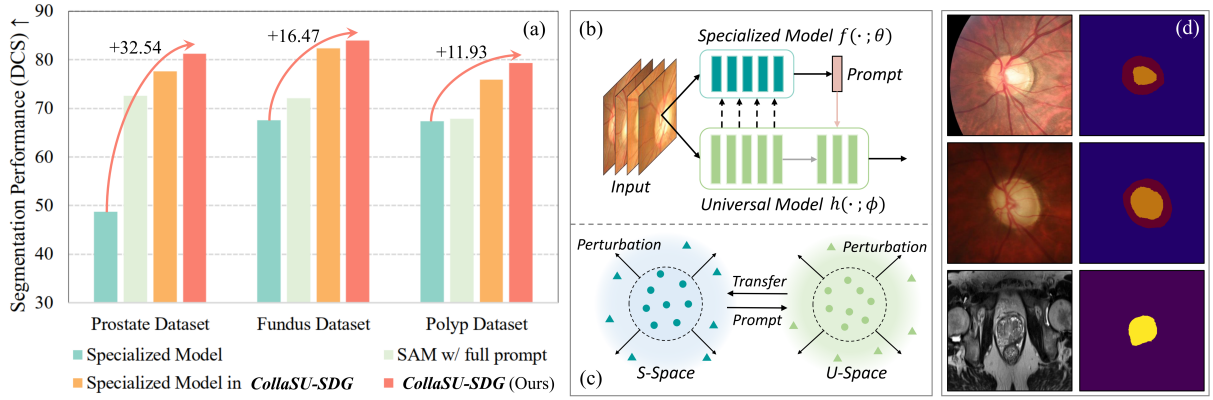

Figure 2: **(a)** shows the segmentation performance of the ***CollaSU-SDG*** method on three tasks. **(b)** The holistic framework of ***CollaSU-SDG***. **(c)** The proposed collaboration and perturbation strategy further activates and expands the feature space distribution, effectively countering inter-domain drift. **(d)** Visualizing clear and accurate segmentation results of ***CollaSU-SDG***.

such as natural scenes (Ji et al. 2023), medical imaging (Ma and Wang 2023), and remote sensing (Wang et al. 2023).

These research advancements inspire us to explore the use of VFMs to address challenging SDG tasks in medical images. This raises a critical question: ***given the significant gap between medical and natural images, how can VFMs be effectively applied to specific medical domains and adapted to the substantial stylistic differences across medical image domains***? In addition, the performance of SAM relies on the precise locations of explicit prompts (Yue et al. 2024). That means a precise manual guidance or well-performing specialist detector is required for accurate prompts, this complexity further limits SAM's direct application in the SDG tasks.

To fully leverage the generalization capabilities of foundational models and the specialized expertise of domain-specific models, our work aims to achieve collaborative computation between these two types of models to address the SDG task in the medical field. Our approach, termed "***Perturbating***, ***Tuning***, and ***Collaborating***," designs an architecture that promotes collaborative inference between $\mathcal{S}$- and $\mathcal{U}$- models (see Fig. 2-b). The "***Perturbating***" component involves model-aware perturbation injection (see Fig. 2-c), which dynamically adapts to the capacities of different models to achieve differentiated perturbations on single-source data, thereby enhancing the capabilities of $\mathcal{S}$- and $\mathcal{U}$-models on single-domain data. The "***Tuning***" module is designed as an domain expansion adapter for the $\mathcal{U}$-Model, facilitating the adaptation of the foundational model to single-domain medical data. Finally, the "***Collaborating***" strategy comprises two aspects: adaptive hierarchical transfer for the $\mathcal{S}$-Model and dynamic dense prompting for the $\mathcal{U}$-Model. The experimental results indicate that this collaborative paradigm has achieved significant improvements in medical SDG tasks (see Fig. 2). The contributions are summarized as follows:

- We propose a method, named ***CollaSU-SDG***, for SDG in the medical image segmentation. This approach synergizes the domain expertise of specialized ($\mathcal{S}$-) models with the generalization capabilities of universal ($\mathcal{U}$-) models, alleviating SDG challenges effiectively.
- We introduce a collaborative strategy that employs dynamic dense prompts from specialized models to guide foundation models. Additionally, the integration of adaptive hierarchical transfer at multiple scales enhances the dynamic prompting capabilities.
- We design model-aware perturbation injection and domain expansion adapter from both data and structural perspectives to enhance SDG capabilities. This approach involves adaptive frequency domain perturbations within both ($\mathcal{S}$-) and ($\mathcal{U}$-) models.
- As shown in Fig. 1, extensive experiments conducted on sixteen datasets across three medical tasks with ***CollaSU-SDG*** demonstrate significant improvements across all metrics, with DSC increasing by 7.31% to 84.00% on the fundus dataset and by 8.39% to 81.29% on the prostate dataset. Our CollaSU-SDG method significantly establishes new state-of-the-art results.

Related Work

Single-source Domain Generalization (SDG). SDG methods (Liu et al. 2024; Chen et al. 2022) aim to derive robust and invariant features from source data alone. Traditional SDG techniques include Image Transformations (Volpi and Murino 2019), Adversarial Learning (Su et al. 2021), Model-based Augmentation (Yue et al. 2019) and Feature-based Augmentation (Zhou et al. 2022). To counteract the risk of overfitting due to domain shifts, diverse data augmentation methods have been developed (Xu et al. 2020; Huang et al. 2020b; Zhou et al. 2021). Adversarial methods (Zhong et al. 2022) employ an adversarial domain synthesizer to create new domains through interpolation, ensuring semantic consistency with mutual information regularization. These methods have inspired our approach to inject perturbations.

Vision Foundation Model (VFM). Foundation models such as SAM (Kirillov et al. 2023b), DINOV2 (Oquab et al. 2023) and CLIP (Radford et al. 2021) represent significant advancements. These large-scale, universal models, have gained substantial attention due to their comprehensive pre-training

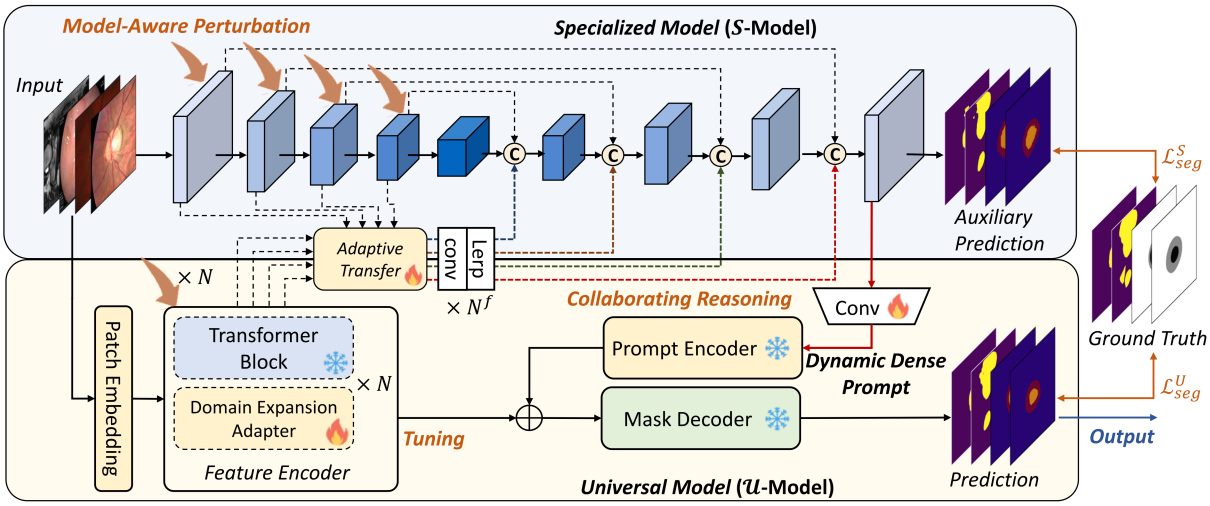

Figure 3: The overall architecture of proposed ***CollaSU-SDG*** approach, where the model-aware perturbation injection (with the "*Perturbating*" process), domain expansion adapter for $\mathcal{U}$-Model (with the "*Tuning*" process), adaptive hierarchical transfer and dynamic dense prompting (with the "*Collaborating*" process) are shown.

across diverse datasets, enabling them to adapt flexibly to various tasks (Brown et al. 2020). The exceptional generalization capabilities of these models have prompted further exploration of their application to specialized tasks, showing their transformative impact across the technological landscape (Ji et al. 2023; Ma and Wang 2023).

Methodology

The objective of SDG is to train a model $f(\cdot;\theta) : \mathcal{X} \rightarrow \mathcal{Y}$ using data solely from one source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ to generalize effectively across multiple unseen target domains $\mathcal{D}_t = \{x_i^t\}_{i=1}^{N_t}$. As the target domain data are not available during training, the optimization objective for $f(\cdot;\theta)$ can be mathematically formulated as follows:

$$\min_{\theta} \frac{1}{N_s} \sum_{(x,y)\in\mathcal{D}_s} \mathbb{E}_{\widetilde{x}=g(x;\psi)} \mathcal{L}(\mathcal{S}(\widetilde{x};\theta), y) \tag{1}$$

where $\mathcal{S}(\cdot;\theta)$ denotes the specialized model to be optimized with θ, $g(\cdot;\psi)$ denotes the data augmentation process (sometimes a style transfer model with ψ as parameters), $\mathcal{L}$ denotes the overall loss function, and $\widetilde{x}$ represents the augmented single-source domain data. In this work, as shown in Fig. 3, we aim to leverage the generalization capabilities of a universal model (denoted as $\mathcal{U}(\cdot;\phi)$) to assist a specialized model in achieving SDG in specific domains, and consequently revise the optimization objective as follows:

$$\min_{\theta,\phi} \frac{1}{N_s} \sum_{(x,y)\in\mathcal{D}_s} \mathbb{E}_{\widetilde{x}=g(x;\psi)} \mathcal{L}(\mathcal{S}(\widetilde{x};\theta), \mathcal{U}(\widetilde{x};\phi), y) \tag{2}$$

where ϕ denotes the learnable weights corresponding to the universal model and θ represents the specialized ones.

Model-Aware Perturbation Injection

To fully activate and enlarge the feature space of Specialiezed ($\mathcal{S}$-) and Universal ($\mathcal{U}$-) models, we propose a model-aware perturbation injector (MPI), as shown in Fig. 4-(a). This approach is primarily based on two objectives: *i)* Sufficient perturbation injection is employed to maximize the diversity of styles in single-source domain data. *ii)* The intensity and range of injected perturbations are dynamically adjusted based on the model's tolerance to perturbations.

Adaptive Low-Frequency Separation. For a given intermediate feature $\mathcal{X}_i \in \mathbb{R}^{H\times W\times C}$, with H, W, and C denoting the height, width, and number of channels respectively, we firstly perform a 2D FFT (Chi, Jiang, and Mu 2020) for each channel independently to obtain the corresponding frequency representations $\mathcal{F}(\mathcal{X}_i) \in \mathbb{R}^{H'\times W'\times C'}$. The Fourier transformation for each channel is computed independently to get the corresponding amplitude $\mathcal{A}_i$ and phase information $\mathcal{P}_i$. To inject perturbations within the low-frequency range that contains the style information, we introduce a learnable circular low-frequency mask $\mathcal{M}_r$ to obtain the LF components as $\mathcal{F}^l(\mathcal{X}_i) = \mathcal{M}_i^r \cdot \mathcal{A}_i$, where $\mathcal{M}_i^r$ is a mask with values of 0 except in the central $\pi \cdot r^2$ region. And r is the radius factor, determining the size of the low-frequency (LF) mask.

Model-Aware Perturbation Intensity. It's well-known that specialized and universal models possess differing model capacities, prompting a critical question: *What intensity of spectral perturbations should be injected for different models and different feature layers?* To this end, we design a model-aware perturbation injector that jointly perceives the encoder features of both $\mathcal{S}$- and $\mathcal{U}$-models, and adaptively adjust the perturbation intensities, enabling the optimal perturbation injection for two types of models.

Specifically, we denote the features of different layers in the $\mathcal{U}$- and $\mathcal{S}$-model's encoder as $\{x_i^u\}_{i=1}^{N_u}$ and $\{x_i^s\}_{i=1}^{N_s}$. Then, a global pooling layer GAP and convolution layer $\phi_{1\times1}$ are adopted to aggregate and concatenate these features:

$$\mathcal{X}_c = \phi_{1\times1}([\{GAP(x_i^u)\}_{i=1}^{N_u}, \{GAP(x_i^s)\}_{i=1}^{N_s}]). \tag{3}$$

where the representation $\mathcal{X}_c$ is further mapped into a channel-

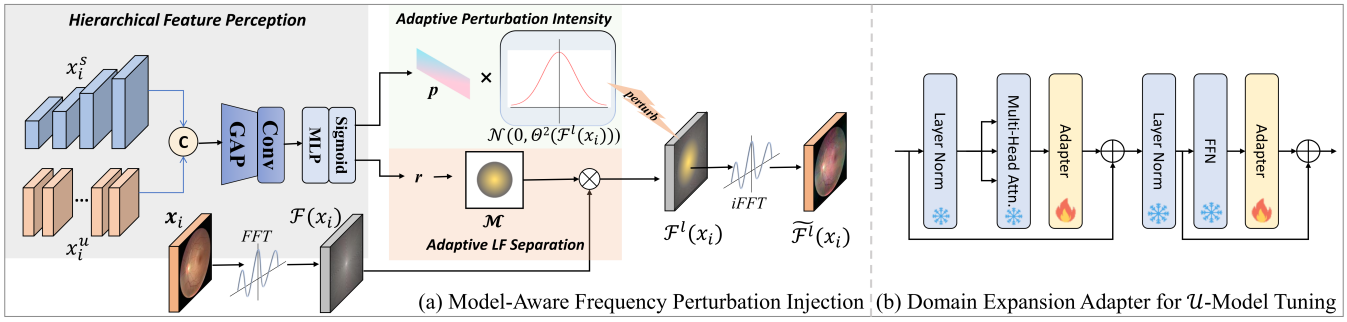

Figure 4: **(a)** The workflow of the model-aware frequency perturbation injection; **(b)** The enhanced Transformer block with two Domain Expansion Adapters in our $\mathcal{U}$-Model tuning.

wise value $\mathcal{E} = [e_1, e_2, \ldots, e_{2c}]$ indicative of the varied perturbation strengths and radius between the network layers in the $\mathcal{U}$- and $\mathcal{S}$-model:

$$\mathcal{E} = (Sigmoid\,(MLP\,(\mathcal{X}_c))) \tag{4}$$

where the $\mathcal{E}$ can be divided into perturbation strength $\boldsymbol{p} = \mathcal{E}[c:2c]$ and radius $r = \mathcal{E}[0:c]$. These dynamically adjusted parameters ensure that the MPI can flexibly modulate the perturbation intensity and range for both $\mathcal{S}$- and $\mathcal{U}$-models based on the multi-scale features from them.

Frequency Perturbation Injector. We firstly model the LF spectrum as a multivariate Gaussian distribution based on a statistical analysis of the frequency distribution. The variance, derived from samples, is calculated as follows:

$$\Theta^2(\mathcal{F}^l(\mathcal{X}_i)) = \frac{1}{\pi r^2}\sum_u\sum_v \left[\mathcal{F}^l(\mathcal{X}_i) - \mathbb{E}[\mathcal{F}^l(X_i)]\right]^2 \tag{5}$$

where πr^2 denotes the LF region, u, v indicate the position of spectral maps, and Θ^2 reflects element variation intensity due to domain shifts. Higher Θ^2 indicates greater variability. Then, we inject noise into the LF components that representing the style information (Guo et al. 2023):

$$\widetilde{\mathcal{F}}^l(\mathcal{X}_i) = \mathcal{F}^l(\mathcal{X}_i) + \boldsymbol{p}\cdot\zeta, \quad \zeta \sim \mathcal{N}(0, \Theta^2(\mathcal{F}^l(\mathcal{X}_i))) \tag{6}$$

where ζ is a sampled Gaussian noise, and $\boldsymbol{p}$ is a matrix of channel scores from Eq. 4. Finally, using the inverse Fast Fourier Transform ($iFFT$), we combine the perturbed LF components $\widetilde{\mathcal{F}}^l(\mathcal{X}_i)$ with the high-frequency components $\mathcal{F}^h(\mathcal{X}_i)$:

$$\widetilde{\mathcal{X}}_i = \mathcal{F}^{-1}([\widetilde{\mathcal{F}}^l(\mathcal{X}_i), \mathcal{F}^h(\mathcal{X}_i)]) \tag{7}$$

This approach effectively expands the single-source domain training data with a learnable spectral perturbation module.

Domain Expansion Adapter for $\mathcal{U}$-Model

To adapt the $\mathcal{U}$-Model to the distinct characteristics of the medical domain, we employ an Adapter-based Tuning Method inspired by the LoRA method (Hu et al. 2021). As illustrated in Fig. 4-(b), Adapter blocks are integrated into each Vision Transformer (ViT) block, specifically after the Multi-Head Attention (MHA) layer and the Feed Forward Network (FFN), to facilitate tuning. The structure of the Adapter block includes three components: a dimension-reducing FFN (FFN^{down}), a ReLU activation function, and a dimension-increasing FFN (FFN^{up}). This design transforms dimensions through the FFN layers, enabling the learning of medical knowledge within a low-rank space. Enhancements to the standard Transformer Layer of the ViT involve the integration of an Adapter structure, with the modifications as follows:

$$\mathcal{X}' = \mathcal{X} + \text{Adapter}\,(\text{MHA}\,(\text{LN}\,(\mathcal{X}))) \tag{8}$$

$$\mathcal{X}^o = \text{LN}\,(\mathcal{X}') + \text{Adapter}\,(\text{FFN}\,(\text{LN}\,(\mathcal{X}'))) \tag{9}$$

where $\mathcal{X}'$ represents the intermediate feature representations after the application of the MultiHead Attention and Adapter modules. $\mathcal{X}^o$ denotes the output of the Transformer Layer following the FFN and a subsequent Adapter application. The Adapter structure is defined as follows:

$$\text{Adapter}(\mathcal{X}) = \mathcal{X} + \text{FFN}^{up}(\text{ReLU}(\text{FFN}^{down}(\mathcal{X}))) \tag{10}$$

Here, each Adapter module consists of two FFN layers separated by an activation function, forming a residual connection that enhances the Transformer's ability to adapt to nuances in the medical data.

Collaborating Reasoning with $\mathcal{S}$- & $\mathcal{U}$-Models

The motivations behind collaborating reasoning are twofold: *i)* to leverage the powerful generalizability of $\mathcal{U}$-models to enhance the cross-domain generalization capabilities of $\mathcal{S}$-models; *ii)* to utilize $\mathcal{S}$-models as prompt generators, thereby enhancing the expertise of $\mathcal{U}$-models in target domains.

Adaptive Hierarchical Transfer (AHT) for $\mathcal{S}$-Model. In CollaSU-SDG framework as shown in Fig. 3, the decoder of the $\mathcal{S}$-model integrates three types of features: specialized features $\{x_i^s\}_{i=1}^{N_s}$, decoder features $\{x_i^d\}_{i=1}^{N_d}$, and universal features $\{x_i^u\}_{i=1}^{N_u}$. The first two types are inherent to the $\mathcal{S}$-model, while the universal features are introduced from the $\mathcal{U}$-model. We select representative hierarchical embeddings $\{x_i^u\}_{i=1}^{N_u}$ from $\mathcal{U}$'s backbone, where each layer's embeddings encompass image representations at varying scales and semantic levels. Given the specialized features $\mathcal{X}_i^s$ as a query, AHT reasons about the relationships of the $\mathcal{S}$- and $\mathcal{U}$-features to directly output the semantic enhanced universal features $\hat{\mathcal{X}}^u$. Firstly, we process the specialized features x_i^s using a

convolutional layer $\phi_{1\times1}$ followed by a global average pooling layer GAP to obtain the aggregated features $\mathcal{Q}_i$:

$$\mathcal{Q}_i = \tau(\phi_{1\times1}(\text{GAP}(\mathcal{X}_i^s))), \quad \mathcal{Q} \in \mathbb{R}^{1\times b\times c} \tag{11}$$

where the τ denotes the reshape operation, b and c denote the batch size and dimension of features, respectively. Then we concatenate the universal features $\{x_i^u\}_{i=1}^{N_u}$:

$$\mathcal{K} = cat(\tau(\text{GAP}(\mathcal{X}_j^u)), \quad \mathcal{K} \in \mathbb{R}^{N_u\times b\times c} \tag{12}$$

Finally, a dot-product operation is adopted to generate an association map $\mathcal{S}_i$, which captures the associations between the specialized and universal feature vectors in$\mathcal{Q}_i$ and $\mathcal{K}$:

$$S_i = \text{Softmax}\left(\frac{\mathcal{Q}_i \times \mathcal{K}^\top}{\sqrt{c}}\right), \quad S_i \in \mathcal{R}^{b\times N_s} \tag{13}$$

At this point, the combined universal features can be calculated as $\hat{\mathcal{X}}^u = sum(S_i \times \mathcal{X}_j^u)$. To fuse the combined features $\hat{\mathcal{X}}^u$ into the decoder of specialized model, convolutional layers and linear interpolation (abbreviated as Lerp in Fig. 3) operations are used to adjust the channel and spatial dimensions, respectively:

$$\mathcal{X}_{i+1}^d = \text{Decoder}(\mathcal{X}_i^d, cat(\mathcal{X}_i^s, \text{conv}(\text{Lerp}(\hat{\mathcal{X}}^u)))) \tag{14}$$

Dynamic Dense Prompting for $\mathcal{U}$-Model. For segmentation tasks in medical SDG, two output modalities are considered: the first utilizes a specialized model, enhanced with generic features for direct output, and the second employs outputs from a universal model, such as the SAM. While the specialized model directly outputs predictions, its cross-domain generalization capability is generally weaker compared to SAM. However, SAM typically requires interactive inputs, such as point or box prompts, for specific target segmentation. In this work, we have developed a strategy employing **D**ynamic **D**ense **P**rompting **(DDP)** by the specialized model to guide SAM, aiming to enhance SAM's segmentation precision without requiring manual intervention for each image.

In the decoder of specialized model, the feature maps of the final layer are strongly correlated with the categories: the model applies a convolution layer followed by a sigmoid activation function directly to the feature maps to generate per-pixel predictions. Therefore, we propose using the final layer's feature maps as prompt information to substitute for the learnable dense embeddings in SAM, thereby guiding SAM's output. The specific operations are as follows:

$$\mathcal{Y}^u = \mathcal{U}(\mathcal{X}, \text{PromptEncoder}(Conv(\mathcal{X}^d)); \phi) \tag{15}$$

where $\mathcal{Y}^u$ and ϕ denotes the segmentation results and weights of the universal model $\mathcal{U}(\cdot)$, $\mathcal{X}$ denotes the perturbated features and $\mathcal{X}^d$ represents the dynamic dense prompt from the specialized model.

Training Objective

In this work, two distinct loss functions are employed: segmentation loss $\mathcal{L}_{seg}^U$ for the $\mathcal{U}$-model and segmentation loss $\mathcal{L}_{seg}^S$ for the $\mathcal{S}$-model. The segmentation loss $\mathcal{L}_{seg}$ comprises a combination of Dice loss $\mathcal{L}_{dice}$ and Cross-Entropy loss $\mathcal{L}_{ce}$.

Task	Prostate Segmentation						Avg.
Single Seen Site	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	
	DSC↑						
CSDG (Ouyang et al. 2022a)	80.72	68.00	59.78	72.40	68.67	70.78	70.06
MaxStyle (Chen et al. 2022)	81.25	70.27	62.09	58.18	70.04	67.77	68.27
EFDM (Zhang et al. 2022)	80.87	69.78	63.16	65.39	69.84	67.15	69.37
SLAug (Su et al. 2022b)	81.20	69.32	60.92	73.72	67.15	71.93	70.71
TriD (Chen et al. 2023)	81.50	70.28	62.89	74.52	72.12	69.11	71.74
MAD (Qu et al. 2023)	80.87	69.78	63.16	65.39	69.84	67.15	69.37
Med-SAM-full (Ma et al. 2024)	77.82	73.19	65.46	71.04	75.02	72.94	72.58
SAM-full (Kirillov et al. 2023a)	72.32	73.31	61.53	64.46	68.89	61.39	66.98
CollaSU-SDG (Ours)	**84.76**	**83.21**	**70.58**	**85.46**	**81.55**	**82.20**	**81.29**
	+3.26	**+9.90**	**+5.12**	**+10.94**	**+6.53**	**+9.26**	**+8.39**

Table 1: Performance Comparison of our ***CollaSU-SDG*** with SOTA methods on Prostate segmentation task.

Thus, our overall optimization objective can be expressed by the following equation:

$$\mathcal{L}_{seg} = \mathcal{L}_{dice} + \mathcal{L}_{ce}, \mathcal{L}_{total} = \lambda_{seg}^U \mathcal{L}_{seg}^U + \lambda_{seg}^S \mathcal{L}_{seg}^S \tag{16}$$

where λ_{seg}^U and λ_{seg}^S are hyperparameters to balance the weights of the loss between the $\mathcal{S}$- and $\mathcal{U}$-Models. A detailed discussion of the hyperparameters for the loss terms is provided in Ablation Studies Section. Based on the experiments, we set λ_{seg}^U and λ_{seg}^S to 1 and 0.1, respectively. During inference, all the perturbation operations are removed.

Experiments

Experimental Setup

We introduce three representative medical image segmentation tasks: Prostate, Fundus and Polyp segmentation task, following (Liu, Dou, and Heng 2020; Chen et al. 2024), to validate the effectiveness of our method. We use a universal model based on SAM-b and a specialized model based on ResNet34 Unet (Ronneberger, Fischer, and Brox 2015). To further validate the generality of our proposed Collaborative framework, we provide experimental results of various $\mathcal{S}$ models (such as transformer) in the appendix. The Dice score metric (DSC) and Average Surface Distance (ASD) are utilized for evaluation on fundus and prostate dataset. The DSC, enhanced-alignment metric (E_θ^{max}) (Fan et al. 2018), and structural similarity metric (S_α) (Cheng and Fan 2021) are utilized for evaluation on polyp dataset.

Main Results

We extensively select domain generalization methods from both the medical field and the natural image domain as comparisons to demonstrate the superiority of our approach, such as ERM setting: learning a model on a single source domain without any generalization techniques, some feature perturbation methods: MixStyle (Zhou et al. 2021), MAD (Qu et al. 2023), DSU (Li et al. 2022) and TriD (Chen et al. 2023); some medical SDG methods: CSDG (Ouyang et al. 2022a), MaxStyle (Chen et al. 2022) and SLAug (Su et al. 2022b). Additionally, we select several segmentation works based on SAM such as Vallina SAM (Kirillov et al. 2023a) and MedSAM (Ma et al. 2024) for further fairness comparison.

Methods	Optical Disc / Cup Segmentation (DSC↑)					Avg. ↑	Optical Disc / Cup Segmentation (DSC↑)					Avg. ↑
	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$		$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	
	Optical Disc DSC (↑)						Optical Cup DSC (↑)					
MixStyle (Zhou et al. 2021)	75.67	83.35	82.86	68.86	79.54	78.06	60.84	62.60	73.77	61.44	66.79	66.73
CSDG (Ouyang et al. 2022b)	78.40	82.02	81.46	75.51	81.09	79.70	65.11	70.79	76.19	65.26	65.28	68.53
MaxStyle (Chen et al. 2022)	77.40	80.95	79.59	76.69	81.95	79.32	65.44	67.62	74.52	66.05	64.84	67.10
SAN-SAW (Peng et al. 2022)	76.42	80.79	81.17	78.83	78.00	79.04	59.01	65.51	73.23	62.36	64.42	64.31
EFDM (Zhang et al. 2022)	78.83	84.83	82.25	82.13	81.45	81.90	62.75	65.94	72.20	61.62	63.02	64.10
DSU (Li et al. 2022)	76.88	82.17	81.12	82.36	83.09	81.12	61.26	70.16	74.10	63.19	59.65	65.67
SLAug (Su et al. 2022b)	79.83	83.42	83.18	81.17	83.57	82.23	64.53	71.30	75.94	64.52	67.12	68.28
TriD (Chen et al. 2023)	78.35	82.19	83.62	80.18	81.65	81.12	66.67	70.85	74.13	67.53	66.96	69.23
MAD (Qu et al. 2023)	80.63	82.10	80.37	82.08	80.31	81.10	67.14	66.57	72.40	68.38	69.37	68.77
Med-SAM-full (Ma et al. 2024)	80.00	81.99	80.83	78.45	78.05	79.86	62.55	64.99	65.28	65.46	63.20	64.31
SAM-full (Kirillov et al. 2023a)	75.74	73.49	76.39	77.63	74.42	75.73	61.75	64.42	64.56	64.78	62.98	63.70
CollaSU-SDG (Ours)	**91.58** **+10.95**	**89.07** **+4.24**	**92.37** **+8.75**	**88.41** **+6.05**	**89.09** **+6.00**	**90.10** **+7.20**	**79.87** **+12.73**	**75.25** **+3.95**	**82.54** **+6.35**	**75.35** **+6.97**	**76.45** **+7.08**	**77.89** **+7.42**

Table 2: Performance Comparison of our ***CollaSU-SDG*** with SOTA methods on Fundus segmentation task.

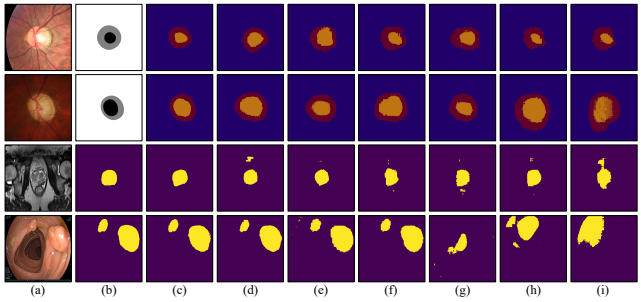

Figure 5: Comparisons across different SDG Methods: fundus (top rows), prostate (third rows) and Polyp (last rows) with ground truth (b) and predictions (c-i). The subfigures (a) to (g) correspond to: (a) input image, (b) ground truth, (c) ***CollaSU-SDG***, (d) MedSAM (Ma et al. 2024), (e) MAD (Qu et al. 2023), (f) TriD (Zhou et al. 2021), (g) SLAug (Su et al. 2022b), (h) CSDG (Xu et al. 2020) and (i) baseline.

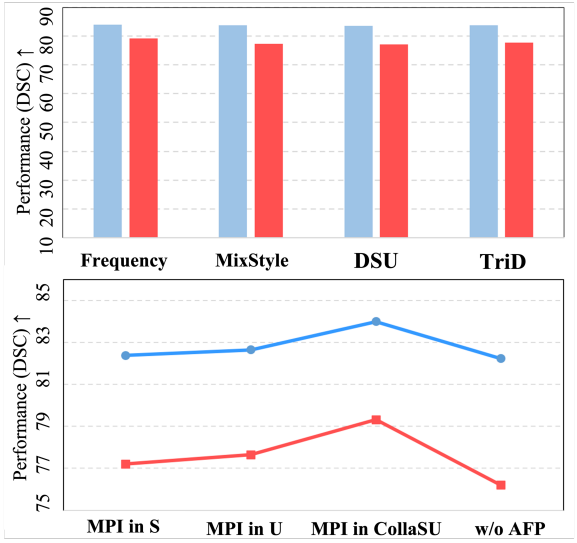

Figure 6: **(a)** Ablation study of different perturbation methods in the feature space of the specialized and universal model on two datasets. **(b)** Ablation study of our MPI in input space on two datasets. (Blue/red denote fundus/prostate datasets.)

As mentioned in the Introduction section, we cannot obtain the ground-truth bounding box as a prompt during the testing phase in practical medical applications. Therefore, the SAM and Med-SAM were implemented in the whole box mode termed as "SAM-full" and "MedSAM-full".

Tab. 1, Tab. 2 and Tab. 3 illustrate the quantitative results of compared methods on the prostate, fundus and polyp task, respectively. The results demonstrate that our method outperforms other SDG-based methods on all segmentation metrics (DSC, E_θ^{max} and S_α) across three tasks as shown in the Fig. 2 (a). Fig. 5 shows a visual comparison of segmentation results with other methods. It is clearly observed that our method generates few misclassified predictions in the unseen target domain with varying morphologies and styles.

Ablation Studies

Our collaborative framework achieves robust generalization in the medical SDG segmentation task, demonstrating significant potential. To thoroughly investigate the effectiveness of different components, we designed extensive ablation experiments. Additionally, in the appendix, we analyze more detail settings of the proposed method.

Impact of Adapter Tuning for VFM. In Variant 2 of Tab. 4, we demonstrate the impact of Adapters in the CollaSU-SDG. By incorporating two effective Adapters, our method shows significant improvements across all evaluation metrics. These results indicate that injecting the medical scene information and knowledge into the visual fundamental model is both necessary and highly effective.

Collabrating *v.s.* Fintuning. To validate the advantages of our collabrating-based approach, we compared it with state-of-the-art fine-tuning methods for visual foundation models (VFMs), focusing on the "Collabrating *vs.* fine-tuning". For example, we considered methods like LoRA (Hu et al. 2022) Adapter (Wu et al. 2023) and Rein (Wei et al. 2024).

Methods	$\mathcal{D}_1$			$\mathcal{D}_2$			$\mathcal{D}_3$			$\mathcal{D}_4$			Average		
	DSC	E_ϕ^{max}	S_α	DSC	E_ϕ^{max}	S_ϕ	DSC	E_ϕ^{max}	S_ϕ	DSC	E_ϕ^{max}	S_α	DSC	E_ϕ^{max}	S_α
MixStyle (Zhou et al. 2021)	74.05	83.93	81.02	69.59	82.77	79.69	67.37	80.58	77.38	72.74	83.68	82.58	70.94	82.74	80.17
CSDG (Ouyang et al. 2022b)	73.86	81.59	80.36	68.54	80.10	81.34	67.32	81.28	77.14	70.38	80.36	80.78	70.03	80.83	79.91
MaxStyle (Chen et al. 2022)	74.73	82.21	80.41	69.07	80.56	80.70	68.75	80.33	79.21	71.38	80.49	80.03	70.98	80.90	80.09
EFDM (Zhang et al. 2022)	72.52	82.60	80.37	69.39	79.08	79.59	69.98	80.86	78.67	73.27	82.01	82.88	71.29	80.86	78.77
SLAug (Su et al. 2022b)	74.89	82.14	78.29	70.87	79.12	75.35	71.69	80.87	80.26	72.48	81.67	81.13	72.48	80.95	78.89
TriD (Chen et al. 2023)	73.44	81.32	79.63	71.88	78.56	77.35	70.75	80.73	76.21	73.53	81.02	80.53	72.40	80.41	78.43
MAD (Qu et al. 2023)	77.35	82.75	82.49	72.46	80.03	78.83	71.32	81.73	80.37	73.48	81.94	80.62	73.65	81.61	80.58
Med-SAM-full (Ma et al. 2024)	64.27	77.31	72.96	67.32	79.38	76.21	70.53	82.73	81.54	69.41	81.65	80.76	67.88	80.27	77.87
SAM-full (Kirillov et al. 2023a)	60.73	74.86	73.51	62.31	75.98	74.47	65.79	78.73	76.57	68.78	80.36	79.27	64.40	77.48	75.96
CollaSU-SDG	**83.52**	**90.65**	**86.47**	**76.72**	**87.19**	**83.83**	**77.32**	**87.20**	**83.24**	**79.71**	**90.07**	**86.40**	**79.32**	**88.55**	**83.92**
	+6.17	**+6.72**	**+3.98**	**+4.26**	**+4.42**	**+2.49**	**+6.00**	**+4.47**	**+1.70**	**+6.23**	**+6.06**	**+3.52**	**+5.67**	**+6.19**	**+3.34**

Table 3: Performance Comparison of our ***CollaSU-SDG*** (Ours) with SOTA methods on Polyp segmentation task.

	EDA	MPI_S	MPI_U	AHT	DDP	Fundus	Prostate
Baseline	-	-	-	-	-	67.53	48.75
Variant 1	-	-	-	-	✓	82.23	79.13
Variant 2	✓	-	-	-	✓	82.57	79.63
Variant 3	-	✓	-	-	✓	82.38	79.38
Variant 4	-	-	✓	-	✓	82.65	79.59
Variant 5	-	-	-	✓	✓	82.78	79.67
CollaSU-SDG	✓	✓	✓	✓	✓	**84.00**	**81.29**

Table 4: Ablation experiments in ***CollaSU-SDG*** for the Fundus and Prostate segmentation tasks.

Method	VFM-arch	publications	Fundus	Polyp
LoRA (Hu et al. 2022)	SAM-b	ICLR22	77.89	72.78
Adapter (Wu et al. 2023)	SAM-b	Arxiv23	78.53	73.43
Rein (Wei et al. 2024)	SAM-b	CVPR24	80.36	75.35
Collabrating-only	SAM-b	-	**82.78**	**77.69**
CollaSU-SDG	SAM-b	-	**84.00**	**79.32**

Table 5: Ablation experiments comparing Collabrating-only and fine-tuning-only approaches for the Fundus segmentation and Polyp segmentation tasks.

Specifically, we removed the perturbation and tuning components and used only the *Collabrating* aspect for comparison with other methods. As shown in Tab. 5, our method outperforms other advanced fine-tuning approaches on both datasets, showing the superiority of the *Collabrating* concept.

Effect of Perturbation Injection. Firstly, we tested different perturbation schemes in Fig. 6 (a), such as MixStyle, TriD, and DSU (Li et al. 2022). We found that all perturbation schemes improved the models' segmentation performance, indicating that feature-level perturbations can help visual foundation models combat inter-domain drift in single-source domain generalization. In addition, we experimented with injecting perturbations into the feature space of both specialized and universal models. As shown in Fig. 6 (b), the highest segmentation performance is achieved when the spectral perturbation is injected simultaneously into both the specialized and universal models. This demonstrates that the perturbation enhances the resilience of both models to inter-domain shifts, thereby facilitating better collaboration.

Effect of Dense Prompt Strategy. From variant 1 in Tab. 4, it is observable that the inclusion of the Dynamic Dense Prompting (DDP) strategy significantly enhances the model segmentation performance. This improvement is likely due to the fact that DDP injects dense semantic prompts from the specialized model into the decoder of the foundation model, aiding the foundation model in better understanding the semantic information distribution of the feature embeddings, thereby producing more reliable segmentation masks.

Effect of Adaptive Hierarchical Transfer (AHT) for a Specialized Model. From the ablation variant 4 and variant 5 in the tab. 4, it is evident that omitting the multi-scale fusion strategy results in a significant decrease in segmentation performance. This may be attributed to our collaborative framework's reliance on domain knowledge exchange between the specialized and universal models. Initially, the universal model transmits multi-scale semantic features to the specialized model, aiding it in combating domain shifts.

Conclusion

In this work, we introduce a collaborative reasoning method between Specialized and Universal models for Single Domain Generalization (***CollaSU-SDG***) in bridging the modality gap between universal and medical imaging domains. By effectively tuning VFMs with domain-specific medical images, we adapt these models to function robustly across diverse medical imaging contexts. Additionally, the introduction of frequency-domain perturbations enhances the resilience of the models against out-of-distribution data, further bolstering their generalization capabilities. The implementation of a dynamic dense prompt method facilitates a seamless collaboration between the specialized and universal models, eliminating the need for explicit points or bounding boxes from manual guidance. We hope the ***CollaSU-SDG*** framework can provide new perspectives for future research in the single domain generalization fields.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant No.2021ZD0140407, the Beijing Natural Science Foundation L222152 and the National Natural Science Foundation of China under Grant No. U21A20523.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019a. Domain Generalization by Solving Jigsaw Puzzles. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2224–2233.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019b. Domain Generalization by Solving Jigsaw Puzzles. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, C.; Li, Z.; Ouyang, C.; Sinclair, M.; Bai, W.; and Rueckert, D. 2022. Maxstyle: Adversarial style composition for robust medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 151–161. Springer.

Chen, Z.; Pan, Y.; Ye, Y.; Cui, H.; and Xia, Y. 2023. Treasure in Distribution: A Domain Randomization Based Multi-source Domain Generalization for 2D Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*.

Chen, Z.; Pan, Y.; Ye, Y.; Lu, M.; and Xia, Y. 2024. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11184–11193.

Cheng, M.-M.; and Fan, D.-P. 2021. Structure-Measure: A New Way to Evaluate Foreground Maps. *International Journal of Computer Vision*, 2622–2638.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast Fourier Convolution. *Neural Information Processing Systems,Neural Information Processing Systems*.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Guan, H.; and Liu, M. 2021. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3): 1173–1185.

Guo, J.; Wang, N.; Qi, L.; and Shi, Y. 2023. ALOFT: A Lightweight MLP-like Architecture with Dynamic Low-frequency Transform for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24132–24141.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, J. E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, abs/2106.09685.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020a. Self-Challenging Improves Cross-Domain Generalization. *ArXiv*, abs/2007.02454.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020b. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 124–140. Springer.

Ji, G.-P.; Fan, D.-P.; Xu, P.; Cheng, M.-M.; Zhou, B.; and Gool, L. V. 2023. SAM struggles in concealed scenes — empirical study on "Segment Anything". *Science China Information Sciences*, 66: 1–3.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023a. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023b. Segment anything. *arXiv preprint arXiv:2304.02643*.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and Duan, L. 2022. Uncertainty Modeling for Out-of-Distribution Generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Liu, C.; Cao, Y.; Su, X.; and Zhu, H. 2024. Universal Frequency Domain Perturbation for Single-Source Domain Generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6250–6259.

Liu, Q.; Dou, Q.; and Heng, P.-A. 2020. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, 475–485. Springer.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Ma, J.; and Wang, B. 2023. Segment Anything in Medical Images. *ArXiv*, abs/2304.12306.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Ouyang, C.; Chen, C.; Li, S.; Li, Z.; Qin, C.; Bai, W.; and Rueckert, D. 2022a. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4): 1095–1106.

Ouyang, C.; Chen, C.; Li, S.; Li, Z.; Qin, C.; Bai, W.; and Rueckert, D. 2022b. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4): 1095–1106.

Peng, D.; Lei, Y.; Hayat, M.; Guo, Y.; and Li, W. 2022. Semantic-Aware Domain Generalized Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2594–2605.

Price, W. N.; and Cohen, I. G. 2019. Privacy in the age of medical big data. *Nature Medicine*, 37–43.

Qu, S.; Pan, Y.; Chen, G.; Yao, T.; Jiang, C.; and Mei, T. 2023. Modality-Agnostic Debiasing for Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24142–24151.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science*.

Su, X.; You, S.; Xie, J.; Zheng, M.; Wang, F.; Qian, C.; Zhang, C.; Wang, X.; and Xu, C. 2022a. ViTAS: Vision transformer architecture search. In *European Conference on Computer Vision*, 139–157. Springer.

Su, X.; You, S.; Zheng, M.; Wang, F.; Qian, C.; Zhang, C.; and Xu, C. 2021. K-shot nas: Learnable weight-sharing for nas with k-shot supernets. In *International Conference on Machine Learning*, 9880–9890. PMLR.

Su, Z.; Yao, K.; Yang, X.; Wang, Q.; Sun, J.; and Huang, K. 2022b. Rethinking Data Augmentation for Single-source Domain Generalization in Medical Image Segmentation.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Volpi, R.; and Murino, V. 2019. Addressing Model Vulnerability to Distributional Shifts Over Image Transformation Sets. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7979–7988.

Wang, D.; Zhang, J.; Du, B.; Tao, D.; and Zhang, L. 2023. Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. *ArXiv*, abs/2305.02034.

Wei, Z.; Chen, L.; Jin, Y.; Ma, X.; Liu, T.; Ling, P.; Wang, B.; Chen, H.; and Zheng, J. 2024. Stronger Fewer & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28619–28630.

Wu, J.; Ji, W.; Liu, Y.; Fu, H.; Xu, M.; Xu, Y.; and Jin, Y. 2023. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. arXiv:2304.12620.

Xu, Q.; Zhang, R.; Wu, Y.-Y.; Zhang, Y.; Liu, N.; and Wang, Y. 2023. SimDE: A Simple Domain Expansion Approach for Single-Source Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 4798–4808.

Xu, Z.; Liu, D.; Yang, J.; Raffel, C.; and Niethammer, M. 2020. Robust and Generalizable Visual Representation Learning via Random Convolutions. In *International Conference on Learning Representations*.

Yue, W.; Zhang, J.; Hu, K.; Xia, Y.; Luo, J.; and Wang, Z. 2024. Surgicalsam: Efficient class promptable surgical instrument segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6890–6898.

Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A. L.; Keutzer, K.; and Gong, B. 2019. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2100–2110.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8035–8045.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems*, 35: 338–350.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; Peng, H.; Li, J.; Wu, J.; Liu, Z.; Xie, P.; Xiong, C.; Pei, J.; Yu, P. S.; University, L. S. M. S.; University, B.; University, L.; University, M.; University, N. T.; of California at San Diego, U.; University, D.; of Chicago, U.; and Research, S. 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *ArXiv*, abs/2302.09419.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *ICLR*.