

A The Details of our Datasets

A.1 OD/OC (Fundus) segmentation dataset

The OD/OC (Fundus) segmentation dataset is primarily utilized for the segmentation of the optic disc (OD) and optic cup (OC), comprising five public datasets collected from different medical centers. These are denoted as domain A (BinRushed [1]), domain B (Magrabia [1]), domain C (REFUGE [39]), domain D (ORIGA [72]), and domain E (Drishti-GS [49]). Following previous methods [8], each image is center-cropped and resized to 256×256 and normalized by min-max normalization. For the evaluation, we employ the Dice Similarity Coefficients (DSC) [%] (the higher the better) to quantitatively assess the segmentation results.

A.2 The Prostate segmentation dataset

The Prostate segmentation dataset comprises 116 MRI instances from six different clinical centers, aggregated from three public datasets, including NCI-ISBI13 [2], I2CVB [25], and PROMISE12 [30] datasets. Following the methodologies described in [32, 35], the dataset is preprocessed to standardize the field of view for the prostate region and resized to 384×384 . The assessment of prostate segmentation performance also employs the Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD).

A.3 The Natural image classification dataset PACS

The image classification dataset PACS is a specialized dataset for studying domain generalization in image classification [26]. The PACS dataset contains 9,991 images across seven categories, collected from four distinct domains: *photo*, *sketch*, *cartoon*, and *art painting*. PACS poses a challenging scenario for single-source domain generalization (SDG) due to the significant shift between domains. Following the methodology in [47, 57], we use a random single domain as the source to train the model and evaluate its performance on the remaining three domains.

We select the data from one organization as the source domain for training and hold out the rest as the target domains for testing. Taking Fundus dataset as an example, after training and validating on domain 'A', testing is conducted on the remaining four domains (B,C,D,E) which denotes as 'A to Rest'.

B Implementation Details

we employ the AdamW optimizer [36] on Optical Disc (OD) / Optic Cup (OC) segmentation tasks and Prostate segmentation tasks, with $\beta = [0.9, 0.999]$ and utilize the SGD optimizer on natural image classification tasks. The initial learning rates are set as follows: for prostate segmentation, $l_0 = 0.01$, for OD/OC segmentation, $l_0 = 0.001$ and for natural image classification, $l_0 = 0.01$. These rates decay according to the polynomial rule $l_t = l_0 \times (1 - \frac{t}{T})^{0.9}$, where l_t denotes the learning rate at epoch t , and T represents the total number of epochs, which are set to 200 for prostate segmentation and 100 for the joint segmentation of OD and OC, with the batch size set as 8. For natural image classification tasks, we follow [46] to train our model 40 epochs.

Submission ID: 4543. 2024-08-01 05:25. Page 11 of 1-13.

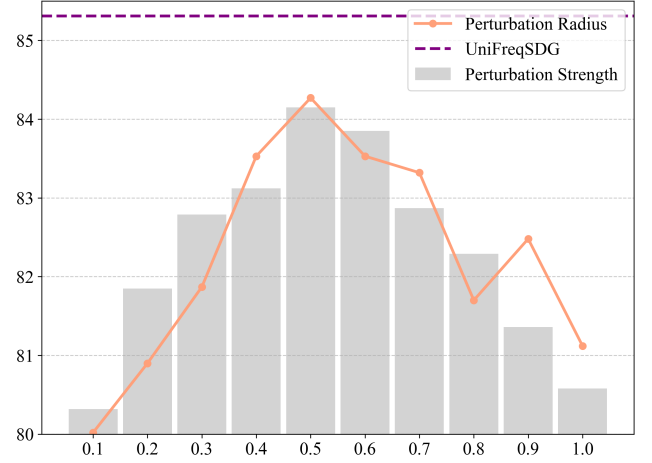


Figure 8: Comparison between fixed-parameter low-frequency perturbation schemes and our LSP method.

Table 8: Experiments on the Fundus dataset w.r.t different loss terms.

λ_{seg}	λ_{adi}	λ_{con}	λ_{sty}	λ_{dec}	Fundus Dataset
1.0	1.0	1.0	1.0	1.0	84.39
1.0	1.0	1.0	1.0	0.5	84.96
1.0	1.0	1.0	0.5	1.0	84.40
1.0	1.0	0.5	1.0	1.0	83.29
1.0	1.0	1.0	1.0	2.0	83.96
1.0	1.0	1.0	2.0	1.0	83.30
1.0	1.0	2.0	1.0	1.0	83.37
1.0	0.5	1.0	0.5	0.5	84.87
0.5	1.0	1.0	0.5	0.5	84.32
1.0	2.0	1.0	0.5	0.5	84.78
2.0	1.0	1.0	0.5	0.5	85.12
1.0	1.0	1.0	0.5	0.5	85.31

C Extended experiments about the UniFreqSDG framework

C.1 The impact of hyperparameters for loss terms

We also present the results for detailed weight settings for loss terms to Tab. 8. Based on these results, our UniFreqSDG method is not particularly sensitive to the hyperparameters for loss terms. In this work, we select a set of hyperparameters that exhibit the best performance as the default settings.

C.2 Fixed v.s. Learnable Parameters in LSP Module

In this section, we analyze the low-frequency perturbation scheme with fixed parameters in LSP and demonstrate the advantages of our adaptive spectral perturbation scheme. Specifically, we conduct detailed ablation experiments on two adaptive parameters in LSP: the low-frequency (LF) radius (r) and the spectral perturbation strength (α). When determining the optimal value of a parameter, such as the

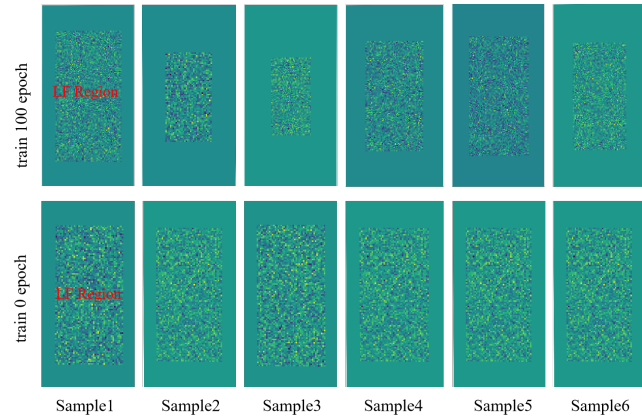


Figure 9: Visualization of the learned low-frequency (LF) masks. We visualized the process when FreqUniFreqSDG was trained on the Fundus dataset. Specifically, we visualized the features after the first block of the model’s encoder during the training process. The first column represents the early stage of training, where the learned LF masks showed little difference; the second column shows that after sufficient training, the model began to capture the LF masks of different samples in a personalized manner.

Table 9: Comparisons with existing augmentation methods on Fundus with as the backbone. The baseline is the vanilla Unet directly trained on the single source domain.

Augmentation	Optical Disc / Cup Segmentation (DSC \uparrow)					Avg. DSC \uparrow
	A	B	C	D	E	
Mixup [69]	78.86	82.59	81.39	81.75	80.36	80.99
CutMix [68]	80.97	83.11	82.54	82.38	81.50	82.10
MixStyle [78]	68.26	80.06	76.23	75.65	84.26	75.29
DSU [28]	69.07	79.14	79.64	73.36	73.37	74.91
UniFreqSDG_m	83.82	86.27	84.84	85.36	85.31	85.31

low-frequency radius, we maintain another hyperparameter, the adaptive perturbation intensity, constant (for example, a random value of 0.5). As shown in Fig. 7, the overall performance of the fixed-parameter perturbation method is lower than our adaptive perturbation approach, as it does not take into account the dynamic distribution of the spectrum in the input images. Furthermore, it is noteworthy that when the perturbation intensity is set to one, and the perturbation radius is one (covering the entire spectral range), there is a significant decline in segmentation performance.

C.3 Visualization of Learnable LF Masks

Our adaptive LSP is capable of learning the range of low frequencies for each sample and each channel based on input features. To verify that our LSP module can indeed identify different spectral sizes, we visualize the low frequencies learned by the model during the training process. As shown in Fig. 8, at the beginning of training, there is no significant change in the low-frequency range. This is because the model has not yet accumulated enough knowledge of source domain samples, and thus the learned low-frequency masks appear to be randomly generated. After 100 epochs of training, the

Table 10: Average Surface Distance (ASD) Metric Comparison on Prostate segmentation dataset. \mathcal{D}_i denotes the single training domain setting. We mark the top results in red and the second in blue.

Task	Seen Site	Prostate Segmentation (ASD \downarrow)						Avg.
		\mathcal{D}_1	\mathcal{D}_2	\mathcal{D}_3	\mathcal{D}_4	\mathcal{D}_5	\mathcal{D}_6	
ERM [48]	ResNet-34	7.54	8.87	13.30	11.97	9.98	7.65	9.89
MixStyle [78]	ResNet-34	4.98	5.77	6.30	5.21	5.98	6.26	5.75
CSDG [41]	EfficientNet-B2	3.51	4.08	4.56	3.58	4.46	4.17	4.06
MaxStyle [5]	ResNet-34	3.40	3.80	4.32	3.23	3.67	4.12	3.77
EFDM [71]	ResNet-34	3.45	3.82	4.35	3.37	3.89	4.03	3.82
SLAug [50]	EfficientNet-B2	3.31	3.74	4.23	3.22	3.79	3.91	3.67
TriD [8]	ResNet-34	3.28	3.69	4.15	3.14	3.67	3.81	3.70
UniFreqSDG_m	ResNet-34	0.89	1.89	2.79	0.83	2.00	2.19	1.77

Table 11: Comparison of Multi-source domain generalization results on Prostate dataset (%). We mark the top results in red and the second in blue. \mathcal{D}_i denotes the single testing domain and the rest domains used for training.

Task	Prostate Segmentation (DSC \uparrow)						Avg.
	\mathcal{D}_1	\mathcal{D}_2	\mathcal{D}_3	\mathcal{D}_4	\mathcal{D}_5	\mathcal{D}_6	
JiGen [4]	85.45	89.26	85.92	87.45	86.18	83.08	86.22
BigAug [70]	85.73	89.34	84.49	88.02	81.95	87.63	86.19
FEDG [33]	86.43	89.59	85.30	88.95	85.93	87.39	87.27
DOFE [55]	89.79	87.42	84.90	88.56	86.47	87.72	87.48
RAM-DISR [80]	87.56	90.20	86.92	88.72	87.17	87.93	88.08
DCAC [d]	91.76	90.51	86.30	89.13	83.39	90.56	88.61
TriD [8]	91.63	90.71	86.91	89.42	88.67	90.11	89.57
UniFreqSDG_m	92.14	91.84	89.59	90.89	90.30	91.32	91.01

model is now able to learn the diverse low-frequency distributions present in the samples.

C.4 Comparisons with other augmentation methods

In our experiments, to demonstrate the effectiveness of UniFreqSDG, we also compared it with several classic image data augmentation methods: i.e., Mixup [69] and CutMix [68], as well as two state-of-the-art feature-level augmentation methods, i.e., MixStyle [78] and DSU [28] (These two methods have already been compared in Tables 1 and 2.). As shown in Tab. 9, both image-level augmentations (Mixup and Cutmix) and feature-level augmentation methods have led to performance improvements, indicating that these data augmentation techniques can introduce a certain degree of diversity to the input samples, thereby enhancing the model’s generalization performance. Moreover, compared to image-level augmentations, our approach achieved better generalization performance, which suggests that the UniFreqSDG method can significantly enhance the diversity of input samples during the training process. Additionally, compared to feature-level data augmentation methods, our model also has significant advantages, which may be due to our method’s ability to generate diverse samples while also effectively preserving domain-invariant content features.

C.5 The results of the ASD evaluation metrics

We conducted comparative experiments on the prostate dataset using the ASD metrics. As shown in Tab. 10, our method still demonstrates a significant advantage, proving its effectiveness.

C.6 Multi-source domain generalization

We conducted multi-source domain generalization experiments on the prostate dataset. Specifically, we trained on multiple source domains and then tested on a single remaining target domain. We conducted comparisons with SOTA methods, including FedDG [33],

Table 12: Performance comparisons of different *Sim* function on Fundus dataset [8].

<i>Sim</i> Function	Optical Disc / Cup Segmentation (DSC \uparrow)					Avg. DSC \uparrow
	A	B	C	D	E	
L1 loss	83.16	86.59	85.39	83.75	85.12	84.80
KL-div	83.57	86.50	83.28	83.58	85.24	84.43
JS-div	82.17	85.40	82.49	83.99	84.26	83.66
Cosine	83.82	86.27	84.84	85.36	85.31	85.31

DOFE[55], RAM-DISR [80], and DCAC. Results in Tab. 11 show that our *UniFreqSDG_m* outperforms all the previous methods.

C.6.1 The ablation study of different Sim function. In Tab. 12, we also demonstrate the impact of different similarity measurement functions (Eq. 17) on the model’s performance. Besides the Cosine Similarity metric, there are also KL-Divergence (KL-Div), JS-Divergence (JS-Div), and L1 loss (L1) considered. These functions to some extent all contribute to enhancing the proposed UniFreqSDG’s learning of distinct features. However, overall, the performance improvement brought by the Cosine Similarity metric is the most significant.