Supplementary Material — Divergent Domains, Convergent Grading: Enhancing Diabetic Retinopathy Grading Generalization

Sharon Chokuwa Muhammad H. Khan
Mohamed Bin Zayed University of Artificial Intelligence
United Arab Emirates
{sharon.chokuwa, muhammad.haris}@mbzuai.ac.ae

1. Quantifying the Domain Shift

To further quantify the domain shift empirically, we calculated the Kullback-Leibler (KL) divergences among the domains. Subsequently, a heatmap detailing pairwise KL divergences among all domains was constructed, as presented in Fig. 1. Lighter shades on the heatmap indicate higher KL divergence, while darker shades signify lower KL divergence. The smallest domain shift is observed between the DDR [1] and IDRID [3] datasets, characterized by a KL divergence of 1.34, whereas the largest disparity is recorded between RLDR [5] and DeepDR [2], with a divergence of 4.93. This holistic analysis provides insights into the extent of domain shift observed between domains.

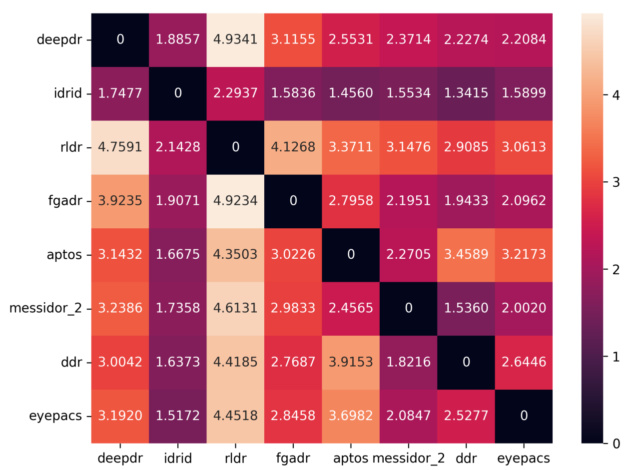

Figure 1. Heatmap visualization to show the KL divergences illustrating the extent of domain shift between domains. Lighter shades on the heatmap signify higher KL divergence, while darker shades indicate lower KL divergence.

2. Weighted Cross Entropy for DG

The weights for our weighted cross entropy are calculated as follows: For each domain S_d, the inverse class weights are calculated as the ratio of total samples (m_d) in S_d to the number of occurrences of each class y in that domain $\omega_{yd} = \frac{m_d}{n_{yd}}$. The inverse domain weights are calculated as the ratio of the total number of samples N to the number of samples in each domain $\theta_d = \frac{N}{N_d}$. The class weights are then normalized by dividing each weight by the maximum weight in that domain $\hat{\omega}_{yd} = \frac{\omega_{yd}}{\max(\omega_{yd})}$. The domain weights are normalized by dividing each weight by the maximum domain weight $\hat{\theta}_d = \frac{\theta_d}{\max(\theta_d)}$. Finally, the sample weights are calculated by multiplying the normalized class weights with the normalized domain weight for each class in each domain $w_{yd} = \hat{\omega}_{yd} \times \hat{\theta}_d$.

3. Dataset Proportions

The statistics of the datasets used for the SSL pretraining are presented in Tab. 1.

Dataset	Dataset Size
ORIGA	650
G1020	1020
ODIR-5K	8000
Drishti-GS	101
REFUGE	1200
RFMiD	1200
DIARETDB1	89
DRIONS-DB	110
DRIVE	40
JSIEC	997
CHASE-DB1	28
ROC	100
DR1 and DR2	2046
cataract_dataset	601
Fundus_Train_Val_Data	650
Total	16832

Table 1. Detailed breakdown of the compositions of the retinal datasets utilized during the SSL pretraining phase.

For our DG model the datasets utilised are presented in

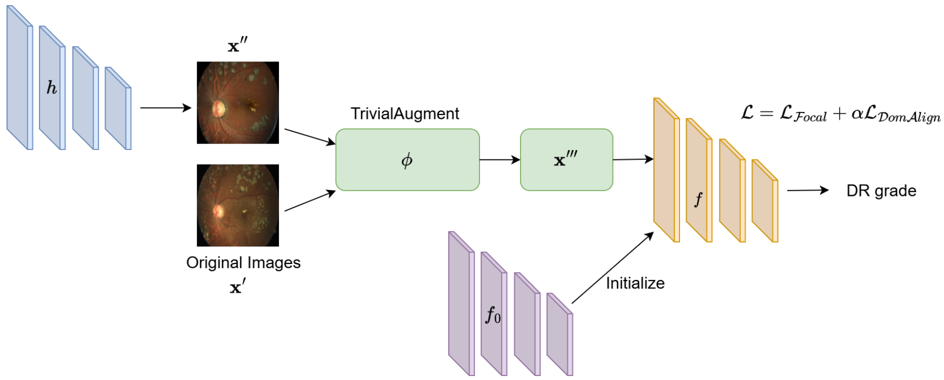

Figure 2. Overview of our method

Tab. 2, which illustrates the dataset sizes as well as geographic origins.

Dataset	Dataset size	Dataset origin
DeepDR	1600	Different hospitals in China
Messidor-2	1744	France
IDRID	516	India
APTOS	3656	Rural India
FGADR	1842	UAE hospitals
RLDR	1593	USA
DDR	12497	23 provinces in China
EyePACS	88698	USA

Table 2. OOD datasets sizes and origins used for our DG model.

4. Additional Implementation Details Set-Up

SSL pretraining: We employ a batch size of 128. The model is trained for 200 epochs, utilizing a learning rate of 0.0003 and a weight decay of 1e-4.

DR Augmentations: Our Dreambooth utilized the hyperparameters outlined in Tab. 3. When Textual Inversion finetuning we utilize a training batch size of 1, a learning rate of 0.0005, and a maximum of 1000 training steps. Consistently, we employ the same instance and class prompts as employed in Dreambooth fine-tuning, maintaining a real guidance strength of 0.5, a Stable Diffusion guidance scale of 7.5, conducting 1000 denoising steps and 200 inference steps. Our code builds on [4] for generating the images.

DG training and evaluation: Tunable parameters ε, C and α were experimentally chosen to be 0.1, 5 and 10, respectively. The batch size is configured as 128, with training conducted over 200 epochs (for our main results), employing a learning rate of 0.001 and SGD optimizer.

5. Generative model finetuning data

What implications arise from incorporating target data in finetuning the latent diffusion model? To ensure that the datasets used in the OOD evaluations are distinct from the ones used in finetuning the latent diffusion model, only the EyePACS dataset was used for the latter purpose. This dataset was chosen due to its expansive size and rich variety of image representations across all grades, particularly grade 4 compared to the other domains, ensuring an ample size of samples for the fine-tuning process. We required approximately 200 images for training and 500 images as class images in each grade's Dreambooth model. Our rationale for this approach was to ensure that there is no data leakage between the diffusion model and the DG model, since in a DG setting, we assume that the target data is completely unseen throughout the training process.

Hyperparameter	Value
Pretrained model	CompVis/stable-diffusion-v1-4
With prior preservation	True
Prior loss weight	1.0
Instance prompt	*'a photo of fundus with $\langle$grade name$\rangle$'*
Class prompt	*'a photo of fundus'*
Resolution	512
Train batch size	2
Learning rate	5e-6
Learning rate scheduler	constant
Max train steps	30000
Number class images	500

Table 3. Finetuning Dreambooth hyperparameters and their values.

References

[1] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511 – 522, 2019. 1

[2] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, Adrian Galdran, J.M. Poorneshwaran, Hao Liu, Jie Wang, Yerui

Chen, Prasanna Porwal, Gavin Siew Wei Tan, Xiaokang Yang, Chao Dai, Haitao Song, Mingang Chen, Huating Li, Weiping Jia, Dinggang Shen, Bin Sheng, and Ping Zhang. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, page 100512, 2022. 1

[3] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. 1

[4] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2

[5] Qijie Wei, Xirong Li, Weihong Yu, Xiao Zhang, Yongpeng Zhang, Bojie Hu, Bin Mo, Di Gong, Ning Chen, Dayong Ding, et al. Learn to segment retinal lesions and beyond. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7403–7410. IEEE, 2021. 1