Revisiting Masked Image Modeling with Standardized Color Space for Domain Generalized Fundus Photography Classification

Eojin Jang[1*], Myeongkyun Kang[2*], Soopil Kim[2,3], Min Sagong[4], and Sang Hyun Park[1,2†]

[1] Department of Interdisciplinary Studies, DGIST, Daegu, South Korea
{eojinjang,shpark13135}@dgist.ac.kr
[2] Robotics and Mechatronics Engineering, DGIST, Daegu, South Korea
[3] Division of Intelligent Robot, DGIST, Daegu, South Korea
[4] Department of Ophthalmology, Yeungnam University College of Medicine, Yeungnam Eye Center, Yeungnam University Hospital, Daegu, South Korea

Abstract. Diabetic retinopathy (DR) is a serious complication of diabetes, requiring rapid and accurate assessment through computer-aided grading of fundus photography. To enhance the practical applicability of DR grading, domain generalization (DG) and foundation models have been proposed to improve accuracy on data from unseen domains. Despite recent advancements, foundation models trained in a self-supervised manner still exhibit limited DG capabilities, as self-supervised learning does not account for domain variations. In this paper, we revisit masked image modeling (MIM) in foundation models to advance DR grading for domain generalization. We introduce a MIM-based approach that transforms images to achieve standardized color representation across domains. By transforming images from various domains into this color space, the model can learn consistent representation even for unseen images, promoting domain-invariant feature learning. Additionally, we employ joint representation learning of both the original and transformed images, using cross-attention to integrate their respective strengths for DR classification. We showed a performance improvement of up to nearly 4% across the three datasets, positioning our method as a promising solution for domain-generalized medical image classification.

Keywords: Domain Generalization · Masked Image Modeling · Fundus Photography.

1 Introduction

Diabetic retinopathy (DR) is one of the major complications of diabetes, which can lead to retinal damage, vision loss, and even blindness [11]. Therefore, early diagnosis is crucial for preventing DR, and color fundus photography (CFP) with

[*] Equal contribution.
[†] Corresponding author.

a computer-aided grading system allows ophthalmologists to perform rapid and accurate diagnoses. Though various methods have been proposed for DR grading [13], these methods exhibit low accuracy when applied to samples from different hospitals, indicating limited practical applicability due to poor generalization.

To address this, several domain generalization (DG) methods have been proposed for CFP [2, 9, 28, 31], aiming to achieve high accuracy on unseen domain datasets by training the model on multiple domain datasets. Specifically, existing methods address DG by (a) manipulating images [21, 30], (b) reducing the domain gap in feature space [9, 23, 26, 28], and (c) employing novel training strategies [2, 6, 31, 32]. Meanwhile, foundation models [5, 25] have recently emerged, exhibiting enhanced generalization accuracy, often surpassing existing DG approaches due to the scale and diversity of their training data [29]. Consequently, leveraging the robust representations of foundation models becomes promising to achieve high DG accuracy.

In line with this approach, foundation models such as RETFound [34] have been introduced for CFP. Despite the extensive training on 0.9 million CFP, RETFound's generalization capacity remains limited (as shown in our results). This limitation stems from the fact that the foundation model's self-supervised learning does not account for domain variations. For instance, masked image modeling (MIM) [14] in RETFound effectively captures contextual information within images by reconstructing randomly masked patches. However, large differences in color and contrast among CFP from different domains make it difficult to learn domain-invariant features. To address this challenge, we revisit MIM within foundation models for domain-generalized CFP classification.

Our proposed method adapts the MIM process to learn domain-invariant features by leveraging a standardized color space. To be specific, we first transform the image into a standardized color space based on the RGB statistics of the training dataset. Next, we randomly mask the original image and train the model to reconstruct it in the standardized space rather than in the original RGB space. This approach enables our model to standardize unseen images from any domain within a consistent color space. During this process, the encoder simultaneously learns domain-invariant features and contextual information. Notably, the transformed images exhibit a more consistent and enhanced contrast compared to the original images, yet they may lose certain details after transformation. Building on this MIM encoder training, we further propose the joint representation of the original and transformed images, as each image presents unique strengths for DR classification. We extract features from both images separately and apply cross-attention to align their features. By enhancing existing foundation models, which already demonstrate state-of-the-art accuracy across various fields [29], our approach represents a highly promising solution for domain-generalized medical image classification. In summary, the contributions are as follows: (i) We introduce a novel masked image modeling approach based on a standardized color space to train a domain-generalized feature encoder for CFP classification. (ii) Building on this feature encoder, we propose blending the original and transformed images for DR classification, employing cross-attention

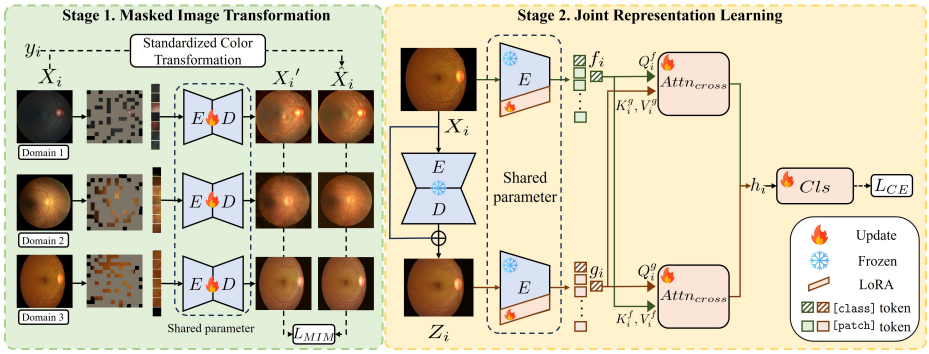

Fig. 1. Illustration of our proposed method, which comprises two stages: Stage 1. *Masked Image Transformation* and Stage 2. *Joint Representation Learning*. X_i represents the original image, $X_i{}'$ represents the transformed image, and $\hat{X}_i$ denotes the color-standardized image. In Stage 1, masked patches from X_i and the corresponding class label y_i are fed into the encoder E and decoder D, with training based on L_{MIM}. In Stage 2, using the blended image Z_i, features f_i and g_i extracted from X_i and Z_i via E are passed into the cross-attention module $Attn_{cross}$. The resulting features are averaged to produce h_i. For classification, h_i is then passed into the classification head Cls to predict the label, with cross-entropy loss L_{CE} used for training.

between them. This approach generates robust and informative features for unseen domain images. (iii) We achieve state-of-the-art accuracy in eye disease classification across three datasets, outperforming recent DG methods.

2 Method

Overview. The illustration of our proposed method is shown in Fig. 1. Our model consists of an encoder E, a decoder D, a multi-head cross-attention module $Attn_{cross}$, and a classification head Cls. The parameters of the encoder E and decoder D are initialized using a pre-trained foundation model (*i.e.*, RETFound [34]) to leverage its strong representational capacity, while the parameters in the other modules are initialized randomly. The proposed framework consists of two distinct stages: robust encoder training through *Masked Image Transformation* (Stage 1) and domain-generalized classification with *Joint Representation Learning* (Stage 2). In the Stage 1, we train E and D with a modified MIM approach that transforms images from any domain into a standardized color space. In the Stage 2, building on the trained encoder, we employ joint representation learning with cross-attention to leverage the complementary features of both original and transformed images.

Standardized Color Transformation. CFP from different domains exhibit distinct RGB color distributions. Therefore, aligning color distributions across domains is crucial to ensure that the same anatomical regions of the eye are mapped to comparable colors. Specifically, let C denote the number of channels,

H the image height, and W the image width. A given image $X_i \in \mathbb{R}^{C\times H\times W}$ is instance-normalized across the spatial dimensions [16] as: $(X_{i,c,h,w} - \mu_{i,c})/\sigma_{i,c}$, where $X_{i,c,h,w}$ denotes the intensity of the c-th channel at the h, w pixel of the i-th sample, $\mu_{i,c}$ denotes $\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} X_{i,c,h,w}$, and $\sigma^2_{i,c}$ denotes $\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}(X_{i,c,h,w} - \mu_{i,c})^2$. To transform any domain image into the standardized color, we calculate the mean, $\mu_c = \frac{1}{N^{tr}\times H\times W}\sum_{i=1}^{N^{tr}}\sum_{h=1}^{H}\sum_{w=1}^{W} X_{i,c,h,w}$, and the variance, $\sigma^2_c = \frac{1}{N^{tr}\times H\times W}\sum_{i=1}^{N^{tr}}\sum_{h=1}^{H}\sum_{w=1}^{W}(X_{i,c,h,w} - \mu_c)^2$, of all N^{tr} training images. The calculated σ_c and μ_c are then applied to transform the instance normalized image into the target color space, *i.e.*, the standardized color image $\hat{X}_i$. Formally,

$$\hat{X}_{i,c,h,w} = \sigma_c\left(\frac{X_{i,c,h,w} - \mu_{i,c}}{\sigma_{i,c}}\right) + \mu_c. \tag{1}$$

An example of a standardized color image is shown in Fig. 2 (b). Although the original images from different domains are visually distinct, simply applying the scaling and shifting factors, *i.e.*, σ_c and μ_c, to the normalized images makes them appear similar. Note that σ_c and μ_c are calculated for each label, and different values are applied to different labeled images when transforming X_i to $\hat{X}_i$, since image features may vary depending on the class label. For simplicity, we omit the label notation in Eq. 1.

Masked Image Transformation. In contrast to the original MIM, which focuses solely on extracting self-reconstructive features, our approach aims to extract domain-invariant features. By mapping the same anatomical regions of the eye to comparable colors, the model learns to transform images into a consistent color representation, despite variations in color and contrast. Through this process, the encoder acquires semantically meaningful representations that generalize across domains. Specifically, we redesign the training task to transform the masked image into the standardized color image $\hat{X}_i$, rather than reconstructing the original image X_i. In MIM, this is formulated as:

$$L_{MIM} = \|(X_i^{p\prime} - \hat{X}_i^p)\odot(1 - M)\|_2^2, \tag{2}$$

where $X_i^{p\prime}$ and $\hat{X}_i^p$ indicate non-overlapping patches of $X_i{}'$ and $\hat{X}_i$, respectively. Also, M indicates a mask consisting of 0 and 1 for blocking random patches, and $\odot$ indicates element-wise multiplication. As a result, we can transform any images from unseen domains into the standardized color without using labels and obtain a robust feature encoder. Note that a robust feature encoder is required beyond standardizing the image, as $\hat{X}_i$ is often imperfect due to the absence of label information during testing. Therefore, revisiting MIM is a valid choice for training a domain-invariant feature encoder.

Joint Representation Learning. In this stage, the model is fine-tuned in a supervised manner for the downstream classification task using the trained encoder. With the encoder from Stage 1 kept frozen, we utilize LoRA (Low-Rank Adaptation [15]) for fine-tuning [35]. Although the transformed images provide more consistent and enhanced contrast, some details may be lost during the transformation, as shown in Fig. 2 (e). To address this, we blend the

RGB pixel values of X_i and $X_i{}'$, *i.e.*, $Z_i = (X_i + X_i{}') \times 0.5$. Then, by extracting features from Z_i, *i.e.*, $g_i = E(Z_i)$, where $g_i \in \mathbb{R}^{\mathtt{d}}$, the joint representation h_i is obtained by averaging the cross-attention features of f_i and g_i. Notably, since cross-attention has been utilized for learning domain-invariant feature representations in domain adaptation [27] and for aligning multi-modal features to achieve joint representation [19], these advantages encourage feature representations to be more domain-invariant and informative. For classification, the obtained joint representation h_i is passed through layer normalization and then fed into the classification head Cls to predict the label. The entire model is optimized with *cross-entropy loss* (denoted as L_{CE}) between the classification head's output and the ground truth y_i.

3 Experiments

Datasets. The proposed method is validated on three datasets. **4DR** [28] is a dataset containing CFP images for classifying five DR grades (*i.e.*, normal, mild, moderate, severe, and proliferative). It contains CFP images from four domains: IDRiD [24] (D1), DeepDRiD [22] (D2), Sustech-SYSU [20] (D3), and DR-V03 [4] (D4). **APTOS-Messidor** is similar to 4DR and contains CFP images from three domains: APTOS [17] (D1), Messidor [7] (D2), and Messidor-2 [1] (D3). **Glaucoma** is a dataset for classifying glaucoma (*i.e.*, normal and glaucoma). It contains CFP images from four domains: DRISHTI-GS [10] (D1), G1020 [3] (D2), LAG [18] (D3), and ORIGA [33] (D4).

Experimental Details. We use an image size of 224×224. For validation set, a non-overlapping 20% subset of the training datasets from each domain is used to select the best model. We conduct each experiment three times with different seeds and report the final accuracy by averaging those results. In Stage 1, we set a batch size of 64 and train the model for 100 epochs under the leave-one-domain-out protocol. For the remaining settings, we use the default values in RETFound [34]. In Stage 2, we set a batch size of 32 and train the model for 50 epochs. LoRA sets a rank of 8 for all layers of the encoder except the first and last ones. For the remaining settings, we use the default values in DomainBed [12].

Experimental Scenarios. To demonstrate our method's superiority, we compare our method against various DG methods proposed from different perspectives. We compare our method with two data-level DG methods: Mixup [30], and FedDG [21]; four feature-level DG methods: Fishr [26], CauDR [28], ATFS-ViT [23], and SPSD-ViT [9]; and four training-level DG methods: CauIRL [6], DRGen [2], A2XP [32], and PLDG [31]. Additionally, we compare our method with the fine-tuned RETFound [34] using three approaches: full training (RET-FT), linear probing (RET-LP), and LoRA [15] (RET-LoRA). Notably, since CauDR [28], SPSD-ViT [9], DRGen [2], and PLDG [31] are proposed for CFP DG scenarios, outperforming these methods are particularly important. Also, to demonstrate our method's scalability, we conduct experiments on APTOS-Messidor and Glaucoma datasets. Achieving high accuracy in this experiment demonstrates strong reliability across various CFP datasets.

Table 1. Classification accuracy on the 4DR dataset. D1 represents the test accuracy of domain 1, with the model trained on images from D2, D3, and D4. **Bold** indicates the highest accuracy, and underline indicates the second-highest accuracy.

Method	D1	D2	D3	D4	Avg.
Mixup [30]	53.5	48.9	65.8	63.7	58.0
FedDG [21]	58.6	48.0	61.2	62.0	57.5
Fishr [26]	57.2	47.7	67.6	56.3	57.2
CauDR [28]	60.5	**54.2**	72.7	59.9	61.8
ATFS-ViT [23]	61.0	39.4	50.7	66.0	54.3
SPSD-ViT [9]	58.2	52.1	51.0	61.7	55.8
Ours	**62.2**	48.8	**80.7**	**68.3**	**65.0**

Method	D1	D2	D3	D4	Avg.
CauIRL [6]	51.9	51.4	67.8	57.0	57.0
DRGen [2]	56.0	50.0	59.4	67.9	58.3
A2XP [32]	51.7	34.3	36.3	56.0	44.6
PLDG [31]	58.5	51.3	75.8	64.0	62.4
RET-FT [34]	59.4	24.1	67.8	57.3	52.2
RET-LP [34]	42.3	41.9	61.3	49.3	48.7
RET-LoRA [34]	61.8	45.8	71.8	62.7	60.5

We conduct ablation experiments to assess the contribution of each stage to the final performance. Our training process comprises *Masked Image Transformation* (Stage 1, abbreviated as S.1) and *Joint Representation Learning* (Stage 2, abbreviated as S.2). Also, to evaluate the impact on transformed image, we conduct ablation studies by replacing the transformed image $X_i \oplus Im.$ ($X_i{}'$ in ours) with various images. We compare our method with $\hat{X}_i$, $X_i{}'$, and Z_i. To evaluate the impact on reconstructed image of RETFound, we exclude S.1 and S.2, and evaluate the accuracy with X_i, $\hat{X}_i$, and Z_i w/ R. (blend with a reconstructed image from RETFound). Additionally, we exclude S.2 and evaluate the accuracy with X_i, $\hat{X}_i$, $X_i{}'$, and Z_i, to perform ablation studies for more complex scenarios. Lastly, we exclude S.1 and evaluate the accuracy with Z_i w/ R..

4 Results

Comparison against Recent DG Methods. Table 1 shows the accuracy on the 4DR dataset. The second-and third-place rankings of PLDG [31] and CauDR [28] suggest that DG methods for CFP achieve higher accuracy than those for natural images, demonstrating the effectiveness of tailored DG methods for CFP. Despite comparisons with various data-, feature-, and training-level DG methods, none of these approaches achieved high accuracy on the 4DR dataset. In contrast, our method achieves the highest average accuracy, demonstrating the superiority of the proposed approach. Notably, this superiority does not stem from the larger capacity of the foundation model [34], as RET-FT, RET-LP, and RET-LoRA (baseline) failed to outperform comparison methods (*e.g.*, CauDR). This suggests that employing S.1 and S.2 is a viable solution for CFP DG. Interestingly, RET-LoRA demonstrates higher accuracy than RET-FT. This aligns with the implicit results [35] that updating only a small number of parameters helps mitigate catastrophic forgetting and overfitting.

Scalability in Diverse Datasets. Table 2 shows the accuracy on the APTOS-Messidor and Glaucoma datasets. Overall, our method outperforms comparison methods, including the accuracy of the fine-tuned RETFound. Achieving the best accuracy on the APTOS-Messidor and Glaucoma datasets highlights that our

Table 2. Classification accuracy on the APTOS-Messidor and Glaucoma dataset.

Method	APTOS-Messidor				Glaucoma				
	D1	D2	D3	Avg.	D1	D2	D3	D4	Avg.
SPSD-ViT [9]	**60.5**	60.8	60.4	60.6	50.0	60.3	77.6	72.3	65.1
PLDG [31]	49.1	53.0	53.0	51.7	68.3	59.4	**78.9**	**76.8**	70.9
RET-FT [34]	45.7	54.7	56.7	52.4	66.3	42.2	61.2	59.7	57.4
RET-LP [34]	49.4	45.5	58.3	51.1	34.7	45.2	64.8	71.4	54.0
RET-LoRA [34]	56.1	56.0	62.8	58.3	50.5	42.0	64.8	73.1	57.6
Ours	58.3	**65.2**	**66.8**	**63.4**	**79.2**	**70.9**	78.3	70.9	**74.8**

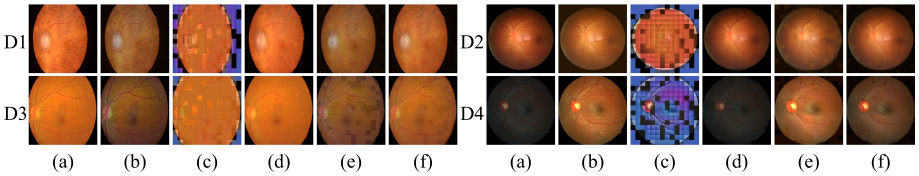

Fig. 2. The visualization of (a) original images X_i, (b) standardized color images $\hat{X}_i$, (c) reconstructed images from RETFound $X_i{}'$ w/ R., (d) $X_i{}'$ with a self-reconstructed image from the fine-tuned model, (e) $X_i{}'$ with a reconstructed image from ours, and (f) blend images with transformed images Z_i (ours). The disease labels for D1, D2, and D4 are Normal, whereas the label for D3 is Mild-DR.

method is robust across various DG scenarios for CFP. Though SPSD-ViT [9] and PLDG [31] were proposed for a dataset comprising APTOS-Messidor and EyePACS [8], excluding EyePACS, which contains 88,702 images, significantly degrades accuracy, indicating that previous approaches heavily rely on the EyePACS dataset. Even in a DG scenario with a limited number of samples, our approach successfully trains a robust classifier, demonstrating its consistent applicability.

Ablation Studies. Fig. 2 presents the visualization of reconstructed and transformed images. As shown in Fig. 2 (b) standardized color images $\hat{X}_i$, Eq. 1 reduces discrepancies between domains, as darker images become brighter (D2, D4) and brighter images become darker (D1, D3). In Fig. 2 (c) reconstructed images from RETFound $X_i{}'$ w/ R., the quality of the reconstructed image is very poor, indicating that the encountered domain shift significantly impacts the reconstruction performance. Comparing (d) the self-reconstructed image from the fine-tuned mode $X_i{}'$ and (e) $X_i{}'$ with a reconstructed image from ours in Fig. 2, (d) preserves the original texture, while (e) presents a more visually consistent result. However, (f) the blend image Z_i (ours) maintains the content of the input image while providing a consistent result. Furthermore, the stable reconstruction of D4 in Fig. 2 demonstrates that the proposed method performs robustly even under extreme conditions. This indicates that our approach is more effective in obtaining domain-invariant images, promoting better joint representation learning.

Table 3. Classification accuracy on the 4DR dataset with various transformed images for joint representation, *i.e.*, $X_i \oplus Im.$. X_i indicates an original image, and $\hat{X}_i$ indicates a standardized color image. $X_i{}'$ indicates a transformed image, Z_i indicates a blend image, and Z_i w/ R. indicates a blend image with a reconstructed image from RETFound.

$X_i \oplus Im.$	S.1	S.2	D1	D2	D3	D4	Avg.
$\hat{X}_i$	✓	✓	57.4	47.3	64.7	65.0	58.6
$X_i{}'$	✓	✓	60.5	45.5	77.2	<u>68.2</u>	<u>62.8</u>
Z_i	✓	✓	62.2	**48.8**	**80.7**	**68.3**	**65.0**
X_i			61.8	45.9	71.7	62.7	60.5
$\hat{X}_i$			58.5	42.3	76.8	59.7	59.3
Z_i w/ R.			59.9	39.4	74.7	62.0	59.0

$X_i \oplus Im.$	S.1	S.2	D1	D2	D3	D4	Avg.
X_i	✓		**62.8**	43.2	75.9	66.5	62.1
$\hat{X}_i$	✓		58.5	45.1	74.7	62.5	60.2
$X_i{}'$	✓		55.6	<u>48.1</u>	73.0	62.6	59.8
Z_i	✓		<u>62.4</u>	42.5	<u>80.1</u>	63.2	62.1
Z_i w/ R.		✓	60.1	47.4	76.0	61.6	61.3

Table 3 presents the ablation results for transformed image on the 4DR dataset. The ablation results on $\hat{X}_i$, $X_i{}'$, and Z_i with both S.1 & S.2 demonstrate that the transformed image is beneficial than applying Eq. 1, as the accuracies of $X_i{}'$ and Z_i are higher than $\hat{X}_i$. This is attributed to the transformed images providing more consistent results compared to applying Eq. 1, as shown in Fig. 2. Additionally, the blend image Z_i demonstrates higher accuracy than using the transformed images $X_i{}'$ alone, supporting the effectiveness of the blending approach. To investigate whether using transformed images without our approach can still achieve high accuracy. The results, excluding S.1 and S.2, indicate that utilizing only the original images X_i achieves the highest accuracy, suggesting that employing transformed images without our approach does not enhance accuracy. Also, similar results are shown with S.1, as neither $X_i{}'$ nor Z_i exceed the accuracy of X_i. Lastly, the accuracy of with S.2 (Z_i w/ R.) is higher than with nothing (Z_i w/ R.), indicating that using joint representation with cross-attention is beneficial even without employing the fine-tuned encoder from S.1. Despite RETFound's poor reconstruction capability, Z_i w/ R. achieves accuracy comparable to X_i, as the original image is blended into Z_i. In summary, we confirmed that the blended transformed image, *i.e.*, Z_i, is the most appropriate approach for joint representation learning, *i.e.*, with S.2. Additionally, fine-tuning to obtain domain-invariant images, *i.e.*, with S.1, significantly improves the robustness of the fine-tuned encoder. These results demonstrate that both stages are crucial and synergistically improve accuracy when combined.

5 Conclusion

We introduced a novel approach for improving DG in DR classification by employing MIM with a standardized space. This approach standardizes color representation across domains, allowing our model to learn robust and domain-invariant features. Additionally, we leveraged cross-attention between the original and transformed images to enhance DG capability by integrating complementary information from both. Our method demonstrated superior DG accuracy on three CFP datasets, significantly outperforming existing state-of-the-art DG

methods. This work highlights a promising pathway for advancing DG in medical image classification, particularly in applications requiring consistent performance across diverse clinical settings.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00516124) and the National IT Industry Promotion Agency(NIPA), an agency under the MSIT and with the support of the Daegu Digital Innovation Promotion Agency (DIP), the organization under the Daegu Metropolitan Government and Smart HealthCare Program funded by the Korean National Police Agency(KNPA) (No. 220222M01) and the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.RS-2025-02219277).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abràmoff, et al.: Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA ophthalmology (2013)
2. Atwany, M., Yaqub, M.: Drgen: domain generalization in diabetic retinopathy classification. In: MICCAI. Springer (2022)
3. Bajwa, M.N., Singh, G.A.P., Neumeier, W., Malik, M.I., Dengel, A., Ahmed, S.: G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In: IJCNN. IEEE (2020)
4. Benítez, V.E.C., Matto, I.C., Román, J.C.M., Noguera, J.L.V., García-Torres, M., Ayala, J., Pinto-Roa, D.P., Gardel-Sotomayor, P.E., Facon, J., Grillo, S.A.: Dataset from fundus images for the study of diabetic retinopathy. Data in brief (2021)
5. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2021)
6. Chevalley, M., Bunne, C., Krause, A., Bauer, S.: Invariant causal mechanisms through distribution matching. arXiv:2206.11646 (2022)
7. Decencière, et al.: Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology (2014)
8. Dugas, E., Jared, Jorge, Cukierski, W.: Diabetic retinopathy detection. kaggle.com/competitions/diabetic-retinopathy-detection (2015), kaggle
9. Galappaththige, C.J., Kuruppu, G., Khan, M.H.: Generalizing to unseen domains in diabetic retinopathy classification. In: WACV (2024)
10. Gómez-Valverde, et al.: Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. Biomedical optics express (2019)
11. Group, E.T.D.R.S.R., et al.: Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification: Etdrs report number 10. Ophthalmology (1991)
12. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv:2007.01434 (2020)
13. Gulshan, et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama (2016)

14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. ICLR (2021)
16. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
17. Karthik, Maggie, Dane, S.: Aptos 2019 blindness detection. kaggle.com/competitions/aptos2019-blindness-detection (2019), kaggle
18. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: CVPR (2019)
19. Li, X., Hou, Y., Wang, P., Gao, Z., Xu, M., Li, W.: Trear: Transformer-based rgb-d egocentric action recognition. IEEE Transactions on Cognitive and Developmental Systems (2021)
20. Lin, L., Li, M., Huang, Y., Cheng, P., Xia, H., Wang, K., Yuan, J., Tang, X.: The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading. Scientific Data (2020)
21. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: CVPR (2021)
22. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns (2022)
23. Noori, M., Cheraghalikhani, M., Bahri, A., Hakim, G.A.V., Osowiechi, D., Ayed, I.B., Desrosiers, C.: Tfs-vit: Token-level feature stylization for domain generalization. Pattern Recognition (2024)
24. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. Medical image analysis (2020)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. PMLR (2021)
26. Rame, A., Dancette, C., Cord, M.: Fishr: Invariant gradient variances for out-of-distribution generalization. In: ICML. PMLR (2022)
27. Wang, X., Guo, P., Zhang, Y.: Unsupervised domain adaptation via bidirectional cross-attention transformer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer (2023)
28. Wei, H., Shi, P., Miao, J., Zhang, M., Bai, G., Qiu, J., Liu, F., Yuan, W.: Caudr: A causality-inspired domain generalization framework for fundus-based diabetic retinopathy grading. Computers in Biology and Medicine (2024)
29. Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Ling, P., Wang, B., Chen, H., Zheng, J.: Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In: CVPR (2024)
30. Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., Zhang, W.: Adversarial domain adaptation with domain mixup. In: AAAI (2020)
31. Yan, S., Yu, Z., Liu, C., Ju, L., Mahapatra, D., Betz-Stablein, B., Mar, V., Janda, M., Soyer, P., Ge, Z.: Prompt-driven latent domain generalization for medical image classification. IEEE Transactions on Medical Imaging (2024)
32. Yu, G., Hwang, H.: A2xp: Towards private domain generalization. In: CVPR (2024)
33. Zhang, et al.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual international conference of the IEEE engineering in medicine and biology. IEEE (2010)

34. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature (2023)
35. Zhu, Y., Shen, Z., Zhao, Z., Wang, S., Wang, X., Zhao, X., Shen, D., Wang, Q.: Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis. In: ISBI. IEEE (2024)