Towards Generalizable Diabetic Retinopathy Grading in Unseen Domains

Haoxuan Che[1], Yuhan Cheng[1], Haibo Jin[1], and Hao Chen[1,2(✉)]

[1] Department of Computer Science and Engineering, The Hong Kong University of
Science and Technology, Kowloon, Hong Kong
`{hche,ychengbj,hjinag,jhc}@cse.ust.hk`
[2] Department of Chemical and Biological Engineering, The Hong Kong University of
Science and Technology, Kowloon, Hong Kong

Abstract. Diabetic Retinopathy (DR) is a common complication of diabetes and a leading cause of blindness worldwide. Early and accurate grading of its severity is crucial for disease management. Although deep learning has shown great potential for automated DR grading, its real-world deployment is still challenging due to distribution shifts among source and target domains, known as the domain generalization problem. Existing works have mainly attributed the performance degradation to limited domain shifts caused by simple visual discrepancies, which cannot handle complex real-world scenarios. Instead, we present preliminary evidence suggesting the existence of three-fold generalization issues: visual and degradation style shifts, diagnostic pattern diversity, and data imbalance. To tackle these issues, we propose a novel unified framework named Generalizable Diabetic Retinopathy Grading Network (GDRNet). GDRNet consists of three vital components: fundus visual-artifact augmentation (FundusAug), dynamic hybrid-supervised loss (DahLoss), and domain-class-aware re-balancing (DCR). FundusAug generates realistic augmented images via visual transformation and image degradation, while DahLoss jointly leverages pixel-level consistency and image-level semantics to capture the diverse diagnostic patterns and build generalizable feature representations. Moreover, DCR mitigates the data imbalance from a domain-class view and avoids undesired over-emphasis on rare domain-class pairs. Finally, we design a publicly available benchmark for fair evaluations. Extensive comparison experiments against advanced methods and exhaustive ablation studies demonstrate the effectiveness and generalization ability of GDRNet. The source code is released at https://github.com/chehx/DGDR.

1 Introduction

Diabetic Retinopathy (DR) is a leading cause of blindness, affecting millions of people worldwide, and early severity grading is vital for disease management [1].
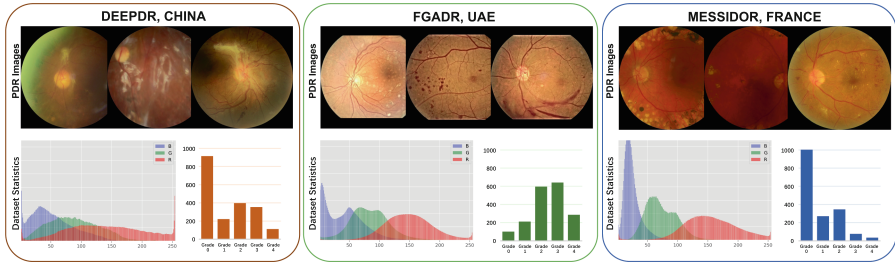

Fig. 1. The RGB statistics, category histograms and proliferative DR (PDR) samples from different datasets/domains. It can be observed that the existence of visual differences, image degradations and diverse diagnostic patterns from PDR samples and RGB statistics. Besides, the divergence among label histograms shows the data imbalance problem across domains and categories.

Although deep learning (DL) has shown promising results in automatic DR grading [2–4], its real-world deployment is still challenging. For instance, Google's DR grading system performed ideally in controlled lab settings [4], but failed to generalize well to complex scenarios which suffer from data shifts [5]. It is a common problem known as domain generalization (DG) [6], where the model performance significantly drops when applied to unseen domains different from the training data. Such an issue hinders the wide adoption and success of DL-based diagnostic tools in clinical practice [7].

Recently, several studies have explored the DG problem and reported significant performance drops in the retinal vessel segmentation [8,9]. Similarly, in DR grading, previous works showed a significant decrease in performance when presented with unseen domains and attempted to solve this problem through the perspective of feature disentanglement [10] and domain-invariant feature learning [11]. Although these methods have improved performance towards unseen domains, they may not be effective in more complex real-world scenarios because they attribute the generalization issue only to limited domain shifts, such as simple visual discrepancies. However, the generalization issues across domains cannot be solely attributed to visual discrepancies [12].

In contrast to previous works, we argue that three factors contribute to poor generalization in DGDR: visual and degradation style shifts, diagnostic pattern diversity, and data imbalance. Specifically, as shown in Fig. 1, we first conduct a preliminary analysis of three public datasets/domains. First, style shifts arise due to various factors, not only limited to visual style discrepancies, but also factors such as variations in lighting conditions [13], image resolution [14], or the presence of artifacts or noise [15]. These factors have been neglected in previous works yet are essential for building a generalizable model. Second, domains may contain diverse diagnostic patterns, such as variations in lesion types, distribution, combination and severity in certain categories [16,17]. This diversity makes learning a generalizable model towards unseen domains challenging because they may contain partially-overlapped or even unknown diagnostic patterns. Finally, data
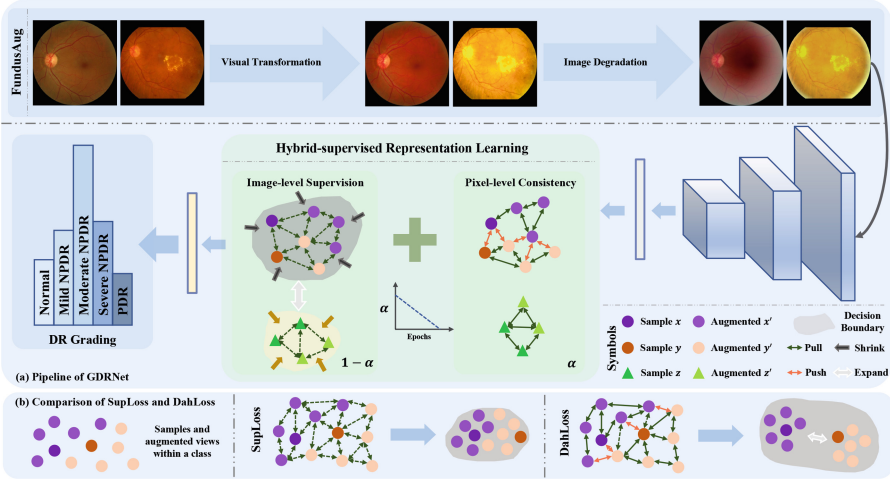
Fig. 2. The pipeline of GDRNet and a high-level visual understanding of DahLoss. FundusAug generates diverse, realistic augmented views, then leverages DahLoss to preserve pixel-level diagnostic patterns and learn generalizable features with sufficient intra-class variations. Moreover, DCR is applied to prevent minority classes from being underrepresented.

imbalance across categories and domains causes samples from specific datasets and minority classes to be underrepresented. Moreover, this imbalance can exacerbate the issue of omitting rare diagnostic patterns, leading to shortcuts in learning and poor generalization [18].

In this paper, we propose a novel framework, Generalizable Diabetic Retinopathy Grading Network (GDRNet) to address the DGDR problem. Our framework consists of three critical components: fundus visual-artifact augmentation (FundusAug), dynamic hybrid-supervised loss (DahLoss), and domain-class-aware re-balancing (DCR). By simulating visual transformations and image degradations, FundusAug enables the model to learn robust features that are less sensitive to style shifts caused by factors such as lighting conditions or artifacts and noise. DahLoss employs a hybrid-supervised learning paradigm to handle diagnostic pattern diversity and dynamically balances the influence of supervised and unsupervised learning. Jointly functioning with FundusAug, it enables the model to preserve pixel-level diagnostic information and learn generalizable features with sufficient intra-class variations. Furthermore, DCR assigns soft-balanced weights to each domain-class pair to prevent underrepresentation caused by data imbalance while avoiding undesired over-emphasis introduced by hard weighting. Finally, to evaluate generalization ability, we design a publicly available benchmark named Generalizable Diabetic Retinopathy Grading Benchmark (GDRBench), comprising eight popular datasets and two evaluation settings.

2   Methodology

An overview of GDRNet is shown in Fig. 2. It addresses the mentioned generalization issues, including style shifts, diagnostic pattern diversity, and domain-class data imbalance, by the proposed FundusAug, DahLoss, and DCR, respectively. Overall, GDRNet provides a unified solution to improve the generalization ability in unseen domains. This section will introduce each component in detail.

Fundus Visual-Artifact Augmentation. The external machine and internal retinal illumination conditions can cause differences in visual attributes [13], such as contrast and brightness, while image degradations, like spot artifacts and blurring, are also very common in fundus imaging [15,19], as depicted in Fig. 1. To bridge the style shifts caused by visual discrepancies and image degradations, we developed FundusAug, which is parameter-free and plug-and-use, designed to generate diverse and realistic augmented views. It employs five basic image transformations, including brightness, contrast, saturation, hue, and sharpness adjustments, to fill visual discrepancy gaps. Moreover, it uses extra four degradation-based image transformations, including halo simulation, hole generation, spot addition, and image blur, to address image degradation gaps. Specifically, given training data x with label y, the augmented view $\hat{x}$ of original data can be derived through the following equation:

$$\hat{x} := \text{FundusAug}(x) = \pi^n_{p_n,m_n}(\pi^{n-1}_{p_{n-1},m_{n-1}}(...\pi^1_{p_1,m_1}(x))), \tag{1}$$

where $\pi^n_{p_n,m_n}$ denotes transformation n with probability p_n and random intensity m_n. To reduce the parameter space of FundusAug while still ensuring image diversity, we implemented FundusAug by applying each operation with a parameter-free procedure that uniformly selects a process with a probability of 0.5. FundusAug can generate realistic augmented views while preserving their diagnostic semantics, as well as providing a robust foundation for subsequent generalizable feature learning by increasing image diversity. A detailed description and visualization of operations can be found in the appendix.

Dynamic Hybrid-Supervised Loss. While the supervised loss (SupLoss), e.g., the cross-entropy loss (CE), effectively guides the model to learn effective feature representations [20], it has two disadvantages in DGDR. First, SupLoss leads to dense features within categories while sufficient intra-class variations are crucial for effectively generalizing to unseen domains [21]. Second, the potential variety of diagnostic patterns in unseen domains requires the models to learn pixel-level lesion semantics as much as possible, while SupLoss lacks such functionality [22]. To tackle these issues, we proposed DahLoss to encourage models to learn features with sufficient intra-class variations and preserve diagnostic patterns, by introducing a hybrid-supervised paradigm to jointly leverage image-level severity supervision and pixel-level semantics consistency. A basic form of DahLoss is as

$$\mathcal{L}_{dhl} = (1 - \alpha)\mathcal{L}_{sup} + \alpha\mathcal{L}_{scon}, \tag{2}$$

where α decreasing within range $[0, 1]$ is a hyper-parameter to dynamically control the task focus, and $\mathcal{L}_{sup}$ and $\mathcal{L}_{scon}$ could be any supervised and self-supervised contrastive loss functions. In this paper, we adopt CE and instance discrimination loss [23] as $\mathcal{L}_{sup}$ and $\mathcal{L}_{scon}$. Specifically, within a multiviewed batch, let $i \in I \equiv \{1...N\}$ be the index of arbitrary augmented samples by FundusAug and $j(i) \in J \equiv \{1'...N'\}$ be the index of weakly-augmented samples originating from the same source sample, we denote $\mathcal{L}_{dhs}^i$ for sample i as

$$\mathcal{L}_{dhl}^i = -(1 - \alpha) \log p_t^i - \alpha \log \frac{\exp(f_i \cdot f_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(f_i \cdot f_a)/\tau)}, \tag{3}$$

where p_t^i denotes the predicted probability of the true class under one-hot encoding label, $f.$ denotes the l_2 normalized feature, the $\cdot$ symbol denotes the inner product, τ is the temperature parameter and $A(i) \equiv (J \cup I)/i$.

As illustrated in Fig. 2 (b), the use of $\mathcal{L}_{sup}$ alone tends to force samples within the same class to cluster tightly together, resulting in a highly concentrated feature representation that may not be beneficial for generalization to unseen domains. However, by incorporating $\mathcal{L}_{scon}$, DahLoss can achieve a balance between the intra-class variation and inter-class distance in learned feature representations. Specifically, while $\mathcal{L}_{sup}$ encourages samples within the same class to cluster together, $\mathcal{L}_{scon}$ simultaneously pulls the augmented views of these samples closer to each other and pushes them far away from those of other samples. It results in a feature representation with both sufficient intra-class variation and clear inter-class separation. Moreover, DahLoss also leverages pixel-level consistency to preserve crucial diagnostic patterns. It is achieved by enforcing the model to maintain feature representations with semantics similarity between strong-weak augmented views originating from the same sample, which ensures that critical details and structures are learned. By jointly considering both image-level supervision and pixel-level consistency, DahLoss encourages the model to learn features with sufficient intra-class variations and preserve crucial diagnostic information, improving generalization performance in unseen domains. Finally, gradually decreasing the value of α during training would guide the model to focus more on the grading task, leading to a balance between representation learning and grading performance. In this paper, we simply set α decay from 1 to 0 linearly across training epochs to verify the effectiveness of DahLoss.

Domain-Class-Aware Re-balancing. The domain-class data imbalance can result in certain categories and diagnostic patterns in specific datasets being underrepresented, leading to biased and inaccurate model predictions [12]. To complicate the situation, adopting a hard balancing style, such as weighting domain-class pairs based on the reciprocal of occurrence frequency, could potentially cause performance degradation [24]. Such decay is owing to the fact that underrepresented domain-class pairs often have a small ratio, resulting in excessively high weights to dominate gradients. Inspired by [25], we designed the DCR method to address this issue. It assigns weights to each sample based on

Table 1. Comparison with state-of-the-art approaches under the DG test.

Target	APTOS			DeepDR			FGADR			IDRID			Messidor			RLDR			Average		
Metrics	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
ERM	75.0	44.4	38.9	77.0	39.5	34.3	66.2	32.0	27.1	82.3	50.0	44.1	79.1	60.7	43.4	75.9	36.5	35.7	75.9	43.8	37.3
DRGen [11]	79.9	58.1	40.2	83.0	38.7	34.1	69.4	41.7	24.7	84.7	44.6	37.4	79.0	60.1	40.5	79.5	43.1	37.0	79.3	47.7	37.3
Mixup [26]	75.3	62.6	43.2	75.3	29.0	25.2	66.7	42.3	32.3	78.8	39.0	27.6	76.7	54.7	32.6	76.9	43.6	37.7	75.0	45.2	33.1
MixStyle [27]	79.0	65.8	39.9	76.9	32.9	27.9	71.2	35.8	22.7	83.0	51.4	39.2	75.2	62.2	36.5	75.5	41.1	31.4	76.8	48.2	32.9
GREEN [3]	75.1	53.8	38.9	76.4	28.1	24.9	69.5	41.3	31.5	79.9	41.3	32.2	75.8	52.0	36.8	74.8	34.0	34.4	75.3	41.8	33.1
CABNet [2]	75.8	55.5	39.4	75.2	42.7	31.8	73.2	43.7	34.8	79.2	44.8	37.3	74.2	56.1	34.1	75.8	37.0	35.6	75.6	46.6	35.5
DDAIG [28]	78.0	67.1	41.0	75.6	37.6	32.2	73.6	42.0	33.8	82.1	37.4	27.0	76.6	58.4	35.3	75.6	36.1	27.7	76.9	46.4	32.8
ATS [29]	77.1	56.9	38.3	79.4	36.1	31.6	74.7	46.7	33.4	83.0	41.5	34.9	77.2	64.7	35.8	76.5	37.4	34.9	78.0	47.2	34.8
Fishr [30]	79.2	66.6	43.4	81.1	48.1	34.4	73.3	44.4	34.4	82.7	40.3	27.6	76.4	65.1	41.1	77.4	36.8	34.7	78.4	50.2	35.9
MDLT [12]	77.3	57.2	41.5	80.0	39.5	36.2	74.1	45.7	29.0	81.5	44.2	35.4	75.4	58.9	36.9	75.7	37.6	35.0	77.3	47.2	35.7
GDRNet	79.9	66.8	46.0	84.7	53.1	45.3	80.8	45.3	39.4	84.0	40.3	35.9	83.2	63.4	50.9	82.9	45.8	43.5	82.6	52.5	43.5

the occurrence probability q_c^d of its category c and domain d. Specifically, DCR calculates the weight w_c^d for category c in domain d as

$$w_c^d = \frac{\sum_{d' \in D} \sum_{j=1}^{N} (q_j^{d'})^\beta}{(q_c^d)^\beta},$$ (4)

where D denotes the set of domains, N is the amount of classes, and β with a range of $[0, 1]$ is a hyperparameter that adjusts the balancing intensity. When β approaches to 0, DCR assigns weight more equally, and when β closes to 1, it acts more like the naive hard balancing method. By introducing β, DCR enables more nuanced weighting of samples based on their domain and class, reducing the risk of over-emphasizing underrepresented samples in the loss function. By considering the occurrence probabilities of all categories across all domains, DCR mitigates the data imbalance problem of underrepresented class-domain pairs. These two make DCR an effective solution for handling domain-class data imbalance and improving the generalizability of DR grading models.

3 Experiments

Experimental Settings, Implementation Details and Evaluation Metrics. To comprehensively analyze and evaluate our framework, we designed the GDRBench involving two generalization ability evaluation settings and eight popular public datasets. First, GDRBench preserves the classic leave-one-domain-out protocol (DG test), which requires leaving one domain for evaluation and training models on the rest. It involves six datasets, including DeepDR [16], Messidor [17], IDRID [31], APTOS [32], FGADR [33], and RLDR [34]. Further, we designed an extreme single-domain generalization setting (ESDG test), which follows the train-on-single-domain protocol with datasets mentioned above but adds two extra large-scale datasets, DDR [35] and EyePACS [36] for evaluation. It simulates real-world generalization issues, in which models are trained only on thousands of samples but are required to generalize well on one hundred thousand images from multiple hospitals and areas. We used ResNet50 pre-trained on ImageNet as the backbone and a fully connected layer as the linear classifier.

Table 2. Ablation studies on proposed components under the DG test.

Method	VT	ID	DCR	$\mathcal{L}_{dhl}$	APTOS	DeepDR	FGADR	IDRiD	Messidor	RLDR	Average
ERM	–	–	–	–	75.04	77.02	66.19	82.31	79.10	75.86	75.92
Model A	✓	–	–	–	77.16	80.22	74.87	82.83	80.48	80.23	79.30
Model B	–	✓	–	–	75.38	79.13	68.35	82.72	77.94	78.93	77.08
Model C	–	–	✓	–	75.58	78.76	67.34	82.59	78.86	78.39	76.92
Model D	–	–	–	✓	77.67	81.22	75.34	81.98	79.09	80.57	79.31
Model E	✓	✓	–	–	77.28	83.41	77.44	83.63	80.22	80.82	80.46
Model F	✓	✓	✓	–	77.37	83.90	77.91	83.44	82.43	83.33	81.39
Model G	✓	✓	–	✓	80.40	85.07	78.65	83.68	84.91	81.97	82.45
GDRNet	✓	✓	✓	✓	77.67	84.69	80.78	84.01	83.16	82.91	82.57

For evaluation, we report three critical metrics, namely accuracy (ACC), the area under the ROC curve (AUC), and macro F1-score (F1). We used **bold** and underline to indicate the first- and second-highest scores. A detailed illustration of the datasets and implementation settings can be found in the appendix.

Comparison with Other Methods. We conducted a comprehensive experiment to evaluate our framework, comparing it with a vanilla baseline (ERM) and other state-of-the-art methods from various categories, including ophthalmic disease diagnosis (OSD) [2,3], domain generalization techniques (DGT) [11,26–29], and feature representation learning (FRL) [12,30]. These chosen methods can be adopted in DGDR with minimal or no modifications, and their brief descriptions are in the appendix. We employed a standard DG augmentation pipeline [12] for all methods except DRGen with a default augmentation strategy [11]. Table 1 shows the quantitative results of the DG test, where the row of "target" indicates the test domain. GDRNet performs better than other methods and significantly improves at least two of AUC, ACC, and F1 on all sub-tests except the test on IDRiD, whose small scale leads to unobvious diagnostic pattern diversity. Typically, ODS methods do not consider domain shifts, and they thus fail to generalize as well as other methods. As expected, DGT and FRL methods consistently improve the performance compared with ERM due to specific designs towards limited domain shifts. However, GDRNet still outperforms these methods markedly due to it handles three-fold generalization issues via increasing training diversity via realistic augmentation, learning generalizable features with sufficient intra-class variations, preserving diagnostic patterns, and rebalancing the minority of samples. Overall, the result shows the effectiveness of GDRNet.

Ablation Study of Proposed Components. To evaluate the effectiveness of our proposed components, we conducted an extensive ablation study under the DG test and presented the AUC score achieved by different models in Table 2. We first examined the individual effects of FundusAug with only visual transformation (VT), FundusAug with only image degradation (ID), DCR and DahLoss

Table 3. Comparison with state-of-the-art approaches under the ESDG test.

Source	APTOS			DeepDR			FGADR			IDRID			Messidor			RLDR			Average		
Metrics	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
ERM	66.4	53.2	31.6	70.7	47.3	31.2	55.3	5.6	7.1	69.6	56.5	33.9	70.6	51.3	33.7	70.1	27.3	26.4	67.1	40.2	27.3
DRGen [11]	69.4	60.7	**35.7**	**78.5**	39.4	31.6	59.8	6.8	8.4	70.8	67.7	30.6	77.0	64.5	37.4	78.9	19.0	21.2	72.4	43.0	27.5
Mixup [26]	65.5	49.4	30.2	70.7	49.7	33.3	58.8	5.8	7.4	70.2	64.0	32.6	71.5	63.0	32.6	72.9	27.7	27.0	68.3	43.3	27.2
MixStyle [27]	62.0	48.8	25.0	53.3	32.0	14.6	51.0	7.0	7.9	53.0	53.5	19.4	51.4	57.6	16.8	53.5	18.3	6.4	54.0	36.2	15.0
GREEN [3]	67.5	52.6	33.3	71.2	44.6	31.1	58.1	5.7	6.9	68.5	60.7	33.0	71.3	54.5	33.1	71.0	31.9	27.8	67.9	41.7	27.5
CABNet [2]	67.3	52.2	30.8	70.0	55.4	32.0	57.1	6.1	7.5	67.4	62.7	31.7	72.3	63.8	35.3	75.2	23.0	25.4	68.2	43.8	27.2
DDAIG [28]	67.4	48.7	31.6	73.2	38.5	29.7	59.9	5.0	5.5	70.2	60.2	33.4	73.5	69.1	35.6	74.4	25.4	23.5	69.8	41.2	26.7
ATS [29]	68.8	51.7	32.4	72.7	52.4	33.5	60.3	5.3	5.7	69.1	66.6	30.6	73.4	64.8	32.4	75.0	24.2	23.9	69.9	44.2	26.4
Fishr [30]	64.5	61.7	31.0	72.1	61.0	30.1	56.3	6.0	7.2	71.8	48.0	30.6	74.3	52.0	33.8	78.6	19.3	21.3	69.6	41.3	25.7
MDLT [12]	67.6	53.3	32.4	73.1	50.2	33.7	57.1	7.1	7.8	71.9	61.7	32.4	73.4	58.9	34.1	76.6	29.0	30.0	70.0	43.4	28.4
GDRNet	69.8	52.8	35.2	76.1	40.0	**35.0**	**63.7**	7.5	**9.2**	72.9	70.0	35.1	78.1	65.7	40.5	79.7	44.3	37.9	73.4	46.7	32.2

(Models A-D, respectively). Our results demonstrate that all proposed components effectively improve model generalization performance and address the three-fold generalization issues discussed in this paper. We then combined these components to analyze their joint effects (Models E-G). Notably, DahLoss and FundusAug contribute significantly to the improvement, demonstrating their capability to increase the training diversity, handle diagnostic pattern diversity, and dynamically balance the influence of supervised and unsupervised learning. Finally, we achieved the best performance by combining all the components in GDRNet, showing that our components play complementary roles. It is worth mentioning that we can observe different performance trends across the various datasets in the ablation study, which indicates the complexity of generalization issues in DGDR. It also suggests that different domains may have varying grades of severity in the three-fold generalization issues identified in this paper.

Generalization from a Single Source Domain. Further, to comprehensively investigate the generalization performance, we introduced the ESDG test. It is a more demanding evaluation setting than the DG test because it requires models trained on a single source dataset to generalize to new domains with significantly larger scales and different data distributions. In contrast, the DG test involves training on multiple domains to fill domain gaps naturally. The quantitative results are presented in Table 3, where the row of "source" indicates the training dataset. As expected, the average performance of all methods decreases significantly, indicating the difficulty of ESDG. Although some methods outperform others in the DG test, such as MixStyle, due to their specific design to leverage the discrepancy of domains, they fail to generalize in the strict ESDG test. Despite more strict requirements, GDRNet still outperforms other methods in average performance and at least one metric in all sub-tests, owing to its effective designs towards three-fold generalization issues.

4 Conclusion

In this paper, we tackled the three-fold generalization issues that hinder the generalizability of DR grading, including style shifts, diagnostic pattern diversity,

and data imbalance. To overcome these challenges, we proposed a novel and unified framework called GDRNet, incorporating three effective components: FundusAug, DahLoss, and DCR. FundusAug enables the generation of realistic augmented views, while DahLoss leverages supervised and unsupervised learning to preserve diagnostic patterns and increase intra-class variations of features. Finally, DCR softly handles the data imbalance across categories and domains to avoid potential performance decay. Together, these three components work synergistically to improve the generalization performance of the model. GDRNet achieved superior performance on both DG and ESDG tests of the proposed publicly available GDRBench, demonstrating its effectiveness and robustness in addressing the three-fold generalization issues in DR grading. Overall, our work provides valuable insights and practical solutions for improving the generalization capability of deep learning in medical image analysis, and has the potential to benefit real-world clinical applications.

References

1. Cho, N.H., Shaw, J., Karuranga, S., Huang, Y., da Rocha Fernandes, J., et al.: IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pract. **138**, 271–281 (2018)
2. He, A., Li, T., Li, N., Wang, K., Fu, H.: CabNet: category attention block for imbalanced diabetic retinopathy grading. IEEE Trans. Med. Imaging **40**(1), 143–153 (2020)
3. Liu, S., Gong, L., Ma, K., Zheng, Y.: GREEN: a graph REsidual rE-ranking network for grading diabetic retinopathy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 585–594. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_56
4. Beede, E., et al.: A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2020)
5. Heaven, W.D.: Google's medical AI was super accurate in a lab. Real life was a different story. MIT Technol. Rev. **4**, 27 (2020)
6. Wang, J., et al.: Generalizing to unseen domains: a survey on domain generalization. IEEE Trans. Knowl. Data Eng. (2022)
7. Li, T., et al.: Applications of deep learning in fundus images: a review. Med. Image Anal. **69**, 101971 (2021)
8. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: DoFE: domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Trans. Med. Imaging **39**(12), 4237–4248 (2020)
9. Lyu, J., Zhang, Y., Huang, Y., Lin, L., Cheng, P., Tang, X.: AADG: automatic augmentation for domain generalization on retinal image segmentation. IEEE Trans. Med. Imaging **41**(12), 3699–3711 (2022)

10. Che, H., Jin, H., Chen, H.: Learning robust representation for joint grading of ophthalmic diseases via adaptive curriculum and feature disentanglement. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 523–533. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_50

11. Atwany, M., Yaqub, M.: DRGen: domain generalization in diabetic retinopathy classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 635–644. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_61

12. Yang, Y., Wang, H., Katabi, D.: On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision. pp. 57–75. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20044-1_4

13. Shen, Z., Fu, H., Shen, J., Shao, L.: Modeling and enhancing low-quality retinal fundus images. IEEE Trans. Med. Imaging **40**(3), 996–1006 (2020)

14. Wang, X., Xu, M., Zhang, J., Jiang, L., Li, L.: Deep multi-task learning for diabetic retinopathy grading in fundus images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2826–2834 (2021)

15. Liu, H., Li, H., Wang, X., Li, H., et al.: Understanding how fundus image quality degradation affects CNN-based diagnosis. In: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, pp. 438–442. IEEE (2022)

16. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., et al.: DeepDRiD: diabetic retinopathy-grading and image quality estimation challenge. Patterns **3**(6), 100512 (2022)

17. Abràmoff, M.D., Lou, Y., Erginay, A., et al.: Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest. Ophthalmol. Vis. Sci. **57**(13), 5200–5206 (2016)

18. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., et al.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (2020)

19. Che, H., Chen, S., Chen, H.: Image quality-aware diagnosis via meta-knowledge co-embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19819–19829 (2023)

20. Zhang, Y., Hooi, B., Hu, D., Liang, J., Feng, J.: Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. Adv. Neural. Inf. Process. Syst. **34**, 29848–29860 (2021)

21. Duboudin, T., Dellandréa, E., et al.: Encouraging intra-class diversity through a reverse contrastive loss for single-source domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 51–60 (2021)

22. Islam, A., Chen, C.F.R., Panda, R., et al.: A broad study on the transferability of visual representations with contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8845–8855 (2021)

23. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)

24. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning, pp. 4334–4343. PMLR (2018)

25. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

26. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: International Conference on Learning Representations (2017)

27. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with MixStyle. In: International Conference on Learning Representations

28. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13025–13032 (2020)

29. Yang, F.E., Cheng, Y.C., Shiau, Z.Y., Wang, Y.C.F.: Adversarial teacher-student representation learning for domain generalization. Adv. Neural. Inf. Process. Syst. **34**, 19448–19460 (2021)

30. Rame, A., Dancette, C., Cord, M.: Fishr: invariant gradient variances for out-of-distribution generalization. In: International Conference on Machine Learning, pp. 18347–18377. PMLR (2022)

31. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., et al.: Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)

32. APTOS: Aptos 2019 blindness detection website. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed 20 Feb 2022

33. Zhou, Y., Wang, B., Huang, L., Cui, S., Shao, L.: A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. IEEE Trans. Med. Imaging **40**(3), 818–828 (2020)

34. Wei, Q., et al.: Learn to segment retinal lesions and beyond. In: International Conference on Pattern Recognition, pp. 7403–7410. IEEE (2021)

35. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf. Sci. **501**, 511–522 (2019)

36. EYEPACS: Kaggle eyepacs dataset. https://paperswithcode.com/dataset/kaggle-eyepacs. Accessed 20 Feb. 2023