U-Net based Image Restoration for Diabetic Retinopathy Classification

Shivaansh Bhatia[1][0009−0004−5852−6109] and
Osmar R. Zaiane[1][0000−0002−0060−5988]

University of Alberta, Canada
{shivaans,zaiane}@ualberta.ca

Abstract. Variability in retinal fundus images across datasets, due to fluctuation in illumination, acquisition settings, or disparity in equipment, remains a major obstacle to developing robust early-stage diabetic retinopathy (DR) classification models. We propose an expert-guided retinal image restoration pipeline based on U-Net to harmonize retinal fundus images. Manually edited retinal images were used as expert-curated reference targets for supervised restoration training, and the dataset was subsequently processed using the trained restoration model for classification training and evaluation. Results show that the restored dataset achieves improved performance in accuracy, recall, precision, weighted F1 and macro F1-scores. These findings suggest that restoration-driven retinal image harmonization can improve the robustness of automatic DR classification under cross-dataset variability.

Keywords: fundus images · image restoration · diabetic retinopathy · classification

1 Introduction

Diabetic retinopathy (DR) is a major cause of blindness worldwide and is broadly categorized into non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) [1]. Automated deep learning methods have become increasingly important for large-scale DR screening and early disease detection [2] and hence robust classification models can assist in identifying early retinal abnormalities associated with DR.

However, variability across retinal fundus datasets remains a major challenge for automated DR classification. Differences in imaging equipment, illumination conditions, and acquisition settings introduce substantial inconsistencies between datasets. In addition, imaging artifacts and quality degradation can negatively affect feature extraction and reduce classification robustness and generalization performance [3].

Improving inter-dataset consistency and reducing image variability are essential for developing reliable, fast, and scalable DR screening systems while minimizing the need for dataset-specific model retraining. The proposed approach first trains a U-Net using high-quality fundus images and their degraded

pairs using custom augmentations, to learn a consistent image representation. The images from different datasets are inferred through the trained U-Net to reduced variations in image quality and appearance before they are passed to a CNN for DR classification.

1.1 Related Work

Numerous retinal image enhancement methods have been proposed to improve image quality. However, many rely on fixed global transformations that are not adaptable to variations in illumination and color distribution across datasets, hence limiting the classification stability.

Hayati et al. [4] proposed a CLAHE-based preprocessing framework that improved DR classification performance by applying computationally efficient formula-based contrast enhancement to grayscale fundus images. However, the approach was evaluated only on the APTOS dataset, which is relatively smaller and cleaner than datasets such as EyePACS [7].

Yi et al. [5] proposed the Medical Image Dataset Cleaning (MIDC) framework, which identifies low-credibility samples and aims to reduce the impact of noisy or mislabeled images through model disagreement analysis and separates them from reliable data. Its modular pipeline enables it to be applied across a wide range of medical image classification tasks. However, the framework removes samples identified as low-credibility, reducing the effective training dataset size and potentially discarding informative images.

To address these limitations, we propose a restoration-based retinal image harmonization framework using U-Net architecture [6]. The proposed pipeline reduces variability by correcting illumination inconsistencies and color variations while preserving the information from all fundus images, resulting in improved DR classification performance.

2 Models and Methodology

We propose a dual-stage framework consisting of: (1) retinal image restoration and (2) downstream DR classification, as shown in Fig. 1. In the first stage, a U-Net-based restoration model is used to harmonize retinal fundus images and reduce acquisition-related artifacts. In the second stage, classification experiments are performed using the restored dataset for model training and evaluation.

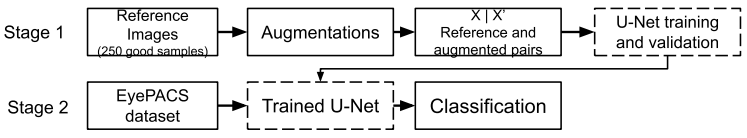


Fig. 1. Restoration and classification pipeline

2.1 Datasets

The proposed framework was evaluated using the publicly available EyePACS dataset [7], containing 35,126 high-resolution retinal fundus images categorized into five DR severity levels ranging from no DR to proliferative DR (labels 0–4).

A subset of 250 retinal images was selected and manually edited to adjust exposure, contrast, color balance, and image clarity, generating standardized reference targets for supervised restoration training. The remaining 34,876 images were retained for downstream classification experiments.

2.2 Augmentations

Custom augmentations were implemented to simulate common retinal imaging artifacts such as illumination imbalance, noise, blur, and contrast degradation. These augmentations were applied with randomized combinations and probabilities to increase data variability and improve restoration learning.

Initially, a Laplacian of Gaussian (LoG) filter was applied to enhance retinal vessel structures in the reference images. The degraded image counterparts were then subjected to augmentations including contrast compression, Gaussian noise injection, random blurring, gamma correction, and color variations as shown in Fig. 2.

Standard preprocessing operations were additionally applied to all image pairs. Images were resized to 512×512 pixels, normalized to the range $[0, 1]$, and randomly flipped horizontally and vertically to improve training robustness and variability.

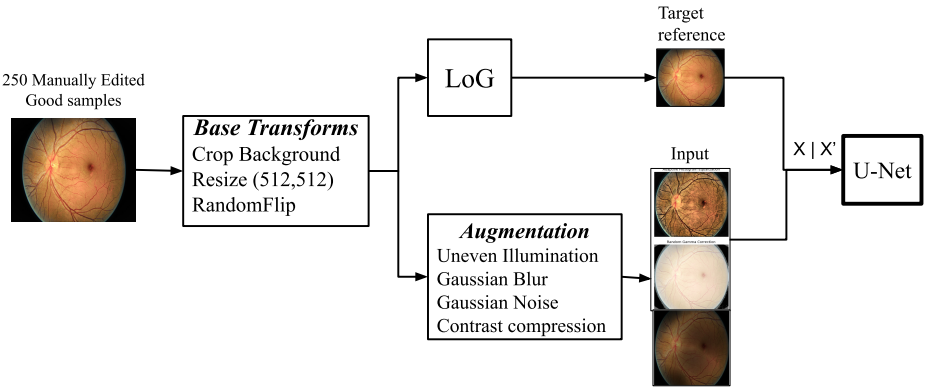


Fig. 2. Augmentation pipeline

2.3 Restoration Model

A U-Net architecture was used for retinal image restoration due to its effectiveness in medical image reconstruction and its ability to preserve spatial information through skip connections. The model was trained using supervised

reference–degraded image pairs to learn restoration mappings, while diverse augmentation combinations introduced sufficient variability to reduce overfitting.

The encoder consists of stacked double convolution blocks with convolution, batch normalization, and ReLU activation layers. The decoder reconstructs the image using transposed convolutions and corresponding skip connections from the encoder. A final (1×1) convolution layer generates the restored output image, which is optimized against the reference image during training.

2.4 Loss Function

A composite reconstruction loss function was used to preserve pixel-level accuracy, structural similarity, and retinal vessel details during restoration. The overall reconstruction loss ($\mathcal{L}_{\text{recon}}$) is defined as:

$$\mathcal{L}_{\text{recon}} = (0.6 \times \mathcal{L}_{\text{MSE}}) + (0.3 \times \mathcal{L}_{\text{SSIM}}) + (0.1 \times \mathcal{L}_{\text{grad}}) \tag{1}$$

where $\mathcal{L}_{\text{MSE}}$ denotes the mean squared error loss for pixel-wise reconstruction, $\mathcal{L}_{\text{SSIM}}$ represents the structural similarity loss [8], and $\mathcal{L}_{\text{grad}}$ denotes the gradient consistency loss based on Sobel filtering to preserve vessel boundaries and fine retinal structures. The weighting coefficients were empirically selected to prioritize reconstruction fidelity while conserving edge information.

3 Experiments

3.1 Restoration Experiments

The U-Net restoration model was trained using a fixed random seed of 42. The manually edited reference dataset was divided into 80:20 training-validation split, and custom augmentation permutations were applied to increase variability in the reference–degraded image pairs.

Training was performed for 300 epochs with a batch size of 16 on resized 512×512 images using an NVIDIA H100 GPU. The model was optimized with Adam [9] (learning rate 5×10^{-4}, weight decay 1×10^{-4}) and a cosine annealing scheduler [10]. Validation performance was evaluated using PSNR [11] and SSIM. The trained model was then used to generate harmonized EyePACS images (Fig. 3) for downstream classification.

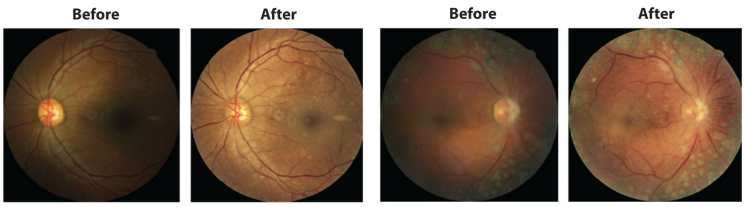


Fig. 3. Comparison of retinal images before and after restoration using the proposed U-Net.

3.2 Classification Experiments

In order to justify the robustness and usefulness of fundus image restoration, we performed classification training and validation on both the original and restored EyePACS datasets using a 70:20:10 train-validation-test split over 100 epochs. Two widely used architectures, VGG16 [12] and MobileNetV3-Small [13], were selected to evaluate performance across models of different complexity.

ImageNet1K-pretrained weights [14] were used for initialization, while the final classification layers were reinitialized and retrained for DR classification. All images were resized to 224×224 pixels to match the input resolution requirements. Training was performed with a batch size of 128, an initial learning rate of 1×10^{-4}, and a weight decay of 5×10^{-4}.

4 Results

The restoration model achieved a best validation loss of 0.0684, with a PSNR of 23.00 dB and an SSIM of 86.90% as indicated in Fig. 4, indicating effective restoration of degraded retinal fundus images and preservation of important structural information.

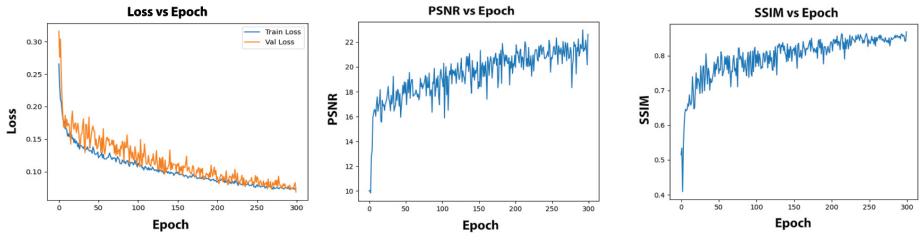


Fig. 4. Training and validation loss show gradual decrease in training and validation loss, PSNR score and SSIM score plot shows increasing scores.

Classification experiments were conducted on both the original and restored datasets using MobileNetV3-Small and VGG16 initialized with ImageNet pretrained weights and loss was calculated using weighted cross-entropy loss. The performance was evaluated using accuracy, macro F1-score, recall, and precision.

As shown in Table 1 below, restoration substantially improved classification performance across all evaluation metrics for both architectures. MobileNetV3-Small improved from a macro F1-score of 45.46% to 89.59%, while VGG16 improved from 53.21% to 96.04%. Notably, recall increased from 50.76% and 55.05% to 93.29% and 95.93% for MobileNetV3-Small and VGG16, respectively. Precision and accuracy also improved substantially, suggesting that the U-Net restoration framework generated more consistent retinal representations for downstream DR classification.

Table 1. Comparison of Validation Set Performance on Raw and Restored Retinal Images. Abbreviations: Acc. (Accuracy), F1-m (Macro F1-score), Rec. (Recall), and Prec. (Precision).

	Raw Images				Restored Images			
Model	**Acc.**	**F1-m**	**Rec.**	**Prec.**	**Acc.**	**F1-m**	**Rec.**	**Prec.**
MobileNetV3	56.80	45.46	50.76	86.27	**93.80**	**89.59**	**93.29**	**97.41**
VGG16	65.40	53.21	55.05	88.84	**96.13**	**96.04**	**95.93**	**98.12**

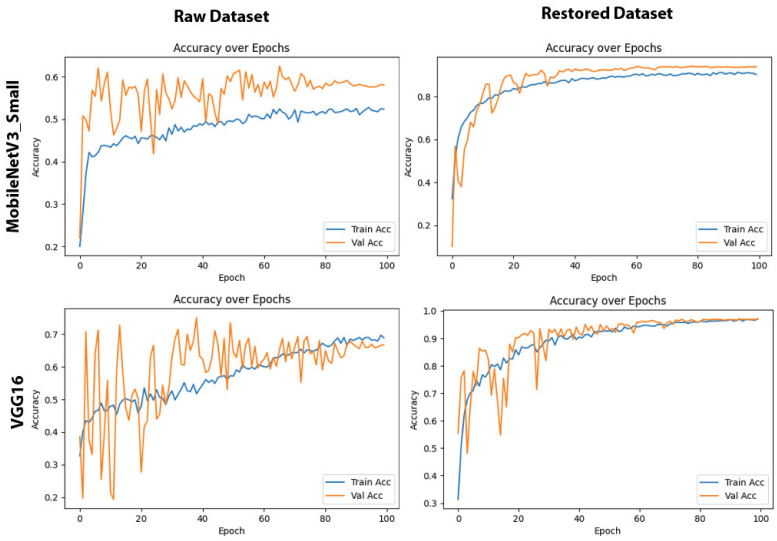


Fig. 5. Raw vs Processed Dataset Classification Accuracy Plots for MobileNetV3_Small and VGG16.

From Fig. 5, the restored datasets demonstrate faster and more stable convergence than the raw datasets, with reduced divergence between the training and validation losses throughout training. This behavior suggests that the proposed harmonization reduces inter-image variability across mini-batches, resulting in more consistent feature learning. Furthermore, substantial performance gains were observed across all architectures, with MobileNet improving from 56.40% to 93.80% accuracy and VGG16 from 65.40% to 96.13%, indicating that retinal image harmonization through the proposed U-Net restoration framework improves downstream classification performance.

To further assess generalization, 10% of the dataset was reserved for testing. The results in Table 2 show that both models maintained strong performance on the unseen subset. Similar improvements were observed after restoration, with VGG16 achieving a macro F1-score of 90.96% and MobileNetV3-Small achieving 86.94%.

Table 2. Comparison of Test Set Performance on Raw and Restored Retinal Images. Abbreviations: Acc. (Accuracy), F1-m (Macro F1-score), Rec. (Recall), and Prec. (Precision).

	Raw Images				**Restored Images**			
Model	**Acc.**	**F1-m**	**Rec.**	**Prec.**	**Acc.**	**F1-m**	**Rec.**	**Prec.**
MobileNetV3	50.79	44.86	48.32	83.59	**89.12**	**86.94**	**92.30**	**95.67**
VGG16	59.89	51.80	53.82	86.94	**94.34**	**90.96**	**94.42**	**96.62**

5 Conclusion

Variability in retinal fundus datasets, including uneven illumination and differences in color grading, remains a challenge for automated DR classification. In this work, we proposed a U-Net-based restoration framework to harmonize retinal images and reduce acquisition-related variability.

In addition to visual assessment, classification experiments using MobileNetV3-Small and VGG16 demonstrated consistent improvements in classification accuracy, macro F1-score, precision, and recall on both validation and test sets after restoration. These results suggest that restoration-driven image harmonization can improve downstream DR classification performance and hence can be considered a reliable way to reduce inter-dataset variability to tackle acquisition artifacts, resulting in robust, scalable and reliable automated DR classification system. Through extensive experimentation, we found that the quality of image restoration depends strongly on how closely the simulated combinations of degradations resemble real imaging artifacts. However, this work could be further strengthened through comparisons with existing methods and by conducting cross-dataset evaluations on previously unseen datasets to better assess generalization and mitigate dataset-specific biases. Future work will investigate class-specific conditional restoration strategies and analyze regions of interest (ROIs) using class activation mapping (CAM) to improve the interpretability and diagnostic relevance of the proposed framework.

Acknowledgments. The authors would like to acknowledge the computational resources provided by Digital Alliance of Canada, which supported the training and evaluation of the proposed models.

Disclosure of Interests. The Authors declare no conflict of interests.

References

1. Aiello, L.P., Gardner, T.W., King, G.L., Blankenship, G., Cavallerano, J.D., Ferris III, F.L., Klein, R.: Diabetic retinopathy. Diabetes Care **21**(1), 143–156 (1998)
2. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016)
3. Gong, Y., Liu, G., Xue, Y., Li, R., Meng, L.: A survey on dataset quality in machine learning. Information and Software Technology **162**, 107268 (2023). https://doi.org/10.1016/j.infsof.2023.107268
4. Hayati, M., Muchtar, K., Roslidar, Maulina, N., Syamsuddin, I., Elwirehardja, G.N., Pardamean, B.: Impact of CLAHE-based image enhancement for diabetic retinopathy classification through deep learning. Procedia Computer Science **216**, 57–66 (2023) https://doi.org/10.1016/j.procs.2022.12.111
5. Yi, S., Chen, Z.: MIDC: Medical image dataset cleaning framework based on deep learning. Heliyon **10**(19), e38910 (2024). https://doi.org/10.1016/j.heliyon.2024.e38910
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
7. EyePACS: EyePACS diabetic retinopathy detection dataset. https://www.kaggle.com/c/diabetic-retinopathy-detection
8. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015). https://doi.org/10.48550/arXiv.1412.6980
10. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017). https://doi.org/10.48550/arXiv.1608.03983
11. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 2366–2369 (2010). https://doi.org/10.1109/ICPR.2010.579
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015). https://doi.org/10.48550/arXiv.1409.1556
13. Howard, A., Sandler, M., Chu, G., et al.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324 (2019). https://doi.org/10.1109/ICCV.2019.00140
14. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848